Constraining the Higgs and heavy neutral lepton sectors with the CMS detector

Doctoral dissertation presented by

Angela Taliercio

in fulfilment of the requirements for the degree of Doctor in Sciences.

Thesis support committee

Prof. Giacomo Bruno (Supervisor) UCLouvain, Belgium

-

Thesis Jury

Prof. Vincent Lemaitre	President	Université Catholique de Louvain
Prof. Giacomo Bruno	Promoteur	Université Catholique de Louvain
Prof. Celine Degrande		Université Catholique de Louvain
Prof. Didar Dobur		Ghent University
Prof. Maria Cepeda		CIEMAT

Contents

Introduction

"Dimidium facti qui coepit habet"
Quintus Horatius Flaccus (65 - 8 B.C.)

The ambitious goal of every physicist is to understand and explain the laws that rules the Universe.
The Newton's gravitational law and the Maxwell's equations describe all the macroscopic phenomena in our planet thus the problem that puzzled the scientists from the last century until now is to understand the microscopic behaviour of particles.
We know that there are four fundamental forces which rules the Universe: the strong interaction, which allows the stability of nuclei, the electromagnetic interaction, which is responsible of the inter-molecular forces, the weak interaction, which deals with the radiative decay of unstable atoms and the gravitational force, known and studied from the beginning of our history.
The mathematical framework in which three of these forces are accommodated is the Standard Model of particle physics; it is a very powerful theory which is strongly confirmed by experimental observations and describes all the known interactions except the gravity.
The Standard Model mathematical formulation is derived starting from the physics strongholds, as the conservation of energy and the invariance under Lorentz transformation. It classifies the matter into two families of particles: the quarks, which have non integer spin and obey at the Fermi-Dirac statistics, and the bosons, which have integer spin and obey at the Bose-Einstein statistics. Depending on their physical properties, those particles interact together, in particular bosons mediate the interaction between quarks.
For symmetry requirements, the Standard Model firstly accommodates massless particles, in contrast with the experimental observations, then, with the introduction of the Brout Englert and Higgs mechanism, this problem is solved and particles get mass by interacting with the Higgs boson field. The discovery of the Higgs boson in 2012 shed light on this mechanism.

Even if the Standard Model is proven by experiments, it is not the *theory of everything* for two main reasons: the gravity is not accommodated and the asymmetry between the production of particles and antiparticles (in favor of particles) is not explained, thus there should be a more general theory which incorporates Standard Model prediction and heals its problems.
The effort of the scientific community of the last decades was, then, to find new physical phenomena beyond the Standard Model, in order to prove new theories which can accommodate all the physical phenomena known in our Universe.
The most important research center in the particle physics field is the European Organization for Nuclear Research (CERN) which hosts the biggest and most powerful accelerator of the world, the Large Hadron Collider (LHC). It collides two protons beams, at 13 TeV center of mass energy, in four interaction points, where four experiments are located: the Compact Muon Solenoid (CMS), A Toroidal LHC ApparatuS (ATLAS), A Large Ion Collider Experiment (ALICE) and LHC beauty (LHCb).
My thesis deals with Beyond the Standard Model (BSM) scenarios with both displaced and prompt signature. For the displaced analysis a Heavy Neutral Lepton analysis in the semileptonic final state is performed. I am looking for right-handed neutrinos with mass between 1 and 15 GeV. Those particles are accommodated in Majoran and Dirac hypothesis, exclusion limits on the mass and on the lepton-neutrino coupling are presented for both models.
On the other hand, also BSM scenarios with prompt signature are investigated within the Higgs framework. A ttH multileptonic differential analysis is presented targeting both EFT results and SM unfolding results. Deep Neural Network is largely use for regression and to discriminate SM for EFT signature.
Following the stream of investigating the Higgs properties a HH analysis in $b\bar{b}4l$ final state is presented. This analysis takes the advantage of a very clear signature and a small amount of background that maximise the sensitivity search, limits on the anomalous values of the Higgs trilinear coupling and on the SM hypothesis are shown. All these analyses target the full Run2 dataset of the CMS experiment.

Chapter 1

The Standard Model of Particle Physics

"Nil posse creare de nihilo"
Lucrezio, *De Rerum Natura*

1.1 The fundamental building blocks of Nature

The Standard Model of Particle Physics (SM) is the most powerful theory we have now to describe three of the four fundamental interactions: the strong, the weak and the electromagnetic; gravity is not included.
The model describes the matter as *fermions* fields of half-integer spin 1/2 and *bosons* fields of integer spin 0 or 1.
In particular it predicts the presence of twelve fermions (and their antiparticles) that are classified according to how they interact: six *quarks*, that can interact via strong and the electroweak interaction (see Sec 1.2 - 1.3) and other six *leptons* which, instead, interact only via the electroweak interaction (see Sec 1.3).
The SM particles are divided into 3 families, depending on their mass (1.1); the number of families is not predicted by the theory but came from experimental results.

The quantitative description of the interactions is found by studying the density of lagrangian of the SM.
This mathematical function has several properties, one of the most important is that symmetries of the lagrangian reflect particular conservation laws (Noether theorem) observed in nature; for example the invariance under Lorentz transformation of a physics system involves the conservation of the four-momentum and thus of the total energy.
Another important symmetry transformation is the *gauge* one which means

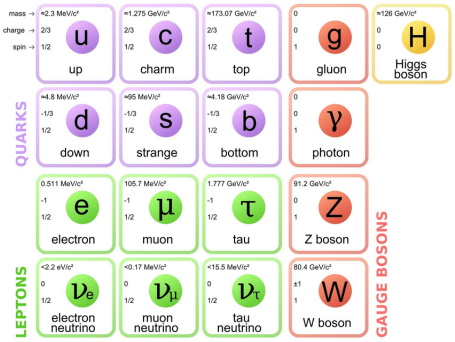


Figure 1.1: Quarks and leptons families

that starting from a field A_μ and a density of lagrangian $\mathcal{L}$, the infinitesimal transformation:

$$A'_\mu \rightarrow A_\mu + \delta\theta \tag{1.1}$$

where θ is a generic function, leaves the lagrangian unchanged:

$$\mathcal{L}_{A_\mu} = \mathcal{L}'_{A'_\mu} \tag{1.2}$$

If θ depends on coordinate, the transformation is said to be *local*, otherwise is *global*.

The invariance of the lagrangian under global gauge transformation is correlated with the *global* conservation of energy in the system, which means that the final energy must be equal to the initial one; instead invariance of the lagrangian under *local* gauge transformation is correlated with the *local* conservation of energy, which ensure the energy conservation in the evolution of the system between the initial and final states.

We will see that the introduction of new fields is necessary if we want the lagrangian to be invariant under local gauge transformation.

In order to have a renomalizable theory, no divergent terms should appear in the calculation.

The lagrangian of the SM satisfied all of these requirements and is described by the $SU(3) \times SU(2) \times U(1)$ gauge group, where $SU(3)$ describes the strong interaction, $SU(2)$ the weak interaction and $U(1)$ the electromagnetic interaction.

The invariance under local gauge transformation of the $SU(3)$ group predicts

the presence of eight spin 1 boson fields, the gluons, which mediate the strong interaction; the same is true for the invariance under $SU(2) \times U(1)$ group which predicts the presence of the 4 spin 1 bosons: the Z and $W^{\pm}$ for the weak interaction and the γ for the electromagnetic interaction. The lagrangian of the SM does not contain mass terms, hence the introduction of a new scalar sector, namely the Higgs sector, is needed.
This new term of the SM lagrangian predicts the presence of the Higgs boson, whose coupling with particles gives them the mass.
An overview of the fundamental interactions follows below.

1.2 The Strong interaction

Experiments done in the last century have shown that protons and neutrons (named also nucleons) are not elementary particles but they are formed and held together by a *see* of quarks and gluons; the interaction which allows the stability of the nucleons and then of the nuclei is the strong interaction.
The equivalent of the electric charge in the electromagnetic interaction, for the strong interaction is the color charge; experiments proof the existence of three different color (with their anticolor) charges.
The formalism and the theoretical justification of this behaviour of the matter comes from the *quantum cromodinamics* which is described by the $SU(3)$ color group; the lagrangian density [1] is:

$$\mathcal{L} = \bar{\psi}(x)(i\partial^{\mu}\gamma_{\mu})\psi(x) \tag{1.3}$$

where $\psi(x)$ is the quark field spinor and γ are the Dirac matrices [1].
The lagrangian have to be invariant under $SU(3)$ gauge transformation of the form:

$$\psi'(x) \rightarrow U\psi(x) \tag{1.4}$$

where U is a rotation under $SU(3)$ color group.
If the matrix U does not depend on x the equivalence between $\mathcal{L}$ and $\mathcal{L}'$ is immediate so the lagrangian is invariant under global gauge transformation; if

$U = U(x)$ the situation is different and the lagrangian becomes:

$$\mathcal{L}' = \bar{\psi}'(x)(i\partial^\mu\gamma_\mu)\psi'(x) =$$
$$\bar{\psi(x)}U(x)^+(i\partial^\mu\gamma_\mu)U(x)\psi(x) =$$
$$\bar{\psi(x)}(i\partial^\mu\gamma_\mu)\psi(x) + \bar{\psi(x)}U(x)^+(i\partial^\mu\gamma_\mu)U(x)\psi(x).$$

To make the lagrangian invariant it is necessary to redefine the transformation of the derivative:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + igG_\mu \tag{1.5}$$

where G_μ are vectorial fields called the gluon field.
With this re-definition of the derivative the lagrangian is:

$$\mathcal{L}' = \bar{\psi}'(x)(iD^\mu\gamma_\mu)\psi'(x) \tag{1.6}$$

Imposing the invariance under $SU(3)$ transformation, $\mathcal{L}' = \mathcal{L}$, the relation for the gluon field are obtained:

$$G'_\mu = U(x)G_\mu U^+(x) - \frac{i}{g}U(x)\partial U^+(x) \tag{1.7}$$

The self interaction of the gluon are described by the gluon strength tensor:

$$G_{\mu\nu} = \partial_\mu G_\nu(x) - \partial_\nu G_\mu(x) + ig_s[G_\mu(x), G_\nu(x)] \tag{1.8}$$

where g_s is the strength of the interaction and it depends on the energy. The commutator $[G_\mu(x), G_\nu(x)]$ is not zero so $SU(3)$ is not abelian: this means that the vertices with more than one gluon are allowed by the theory (for the electromagnetic interaction this is not true).
Now, including the gluons field, the lagrangian of the QCD is:

$$\mathcal{L} = \bar{\psi}(x)(iD^\mu\gamma_\mu)\psi(x) - \frac{1}{4}Tr[G_{\mu\nu}, G^{\mu\nu}] \tag{1.9}$$

where with respect to 1.6, the gluon strength tensor was added.
The mass term ($G_\mu G^\mu$) for the gluon does not exist, so they are massless.
Beside the quark mass, the only free parameter in the Lagrangian is $\alpha_s = \frac{g_s^2}{4\pi}$.

The function $\alpha_s = \frac{g_s^2}{4\pi}$ varies with the energy and characterize the strength of two or more colored particles interacting strongly; its behaviour is shown in Fig 1.2:

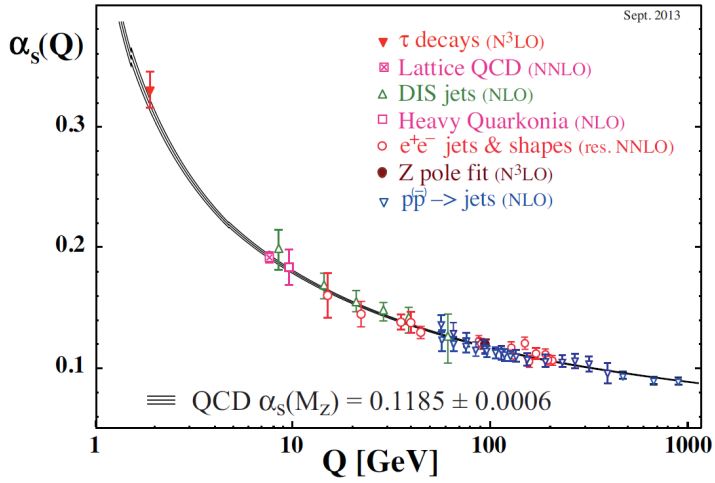


Figure 1.2: Strong coupling constant

The shape of the curve shows two important properties of the strong interaction: the *confinement of the color* and the *asymptotic freedom*.
The confinement of the colour means that at low energy α_s is so high that it's impossible to divide two quarks; in this sense colored particle should not exist in nature.
However at very high energy α_s become smaller and the quarks could exist in free state; that represents the asymptotic freedom.

1.3 The Electroweak Interaction

The first step to study the electroweak interaction, is to consider just the weak interaction.
This theory starts from electron and his neutrino with their Dirac's fields: $e(x)$, $\nu(x)$ and considers the left handed and the right handed parts of those field: $e_L(x), \nu_L(x), e_R(x)$. For the neutrino and in particular from the β-decay, only the left-handed part ν_{eL} couples [1]; right handed neutrino are not predicted in the SM but some experimental evidence, such as neutrino oscillation, incorpo-

rates the right handed neutrino in a more complex framework (see Sec).
The group which describes this interaction is the weak isospin group, $SU(2)$; it treats differently the left and right handed part of electrons and neutrinos: the left handed part of those particle are held together in a left spinor:

$$\psi(x)_L = \begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \end{pmatrix} \tag{1.10}$$

while the right handed part of the electron, $e_R(x)$, is treated as a singlet.
The left and the right handed part of electron can be written as:

$$e_L = \frac{1}{2}(1 - \gamma_5)e(x) \tag{1.11}$$

$$e_R = \frac{1}{2}(1 + \gamma_5)e(x) \tag{1.12}$$

Due to the electron mass term, those fields are clearly not a solution of the Dirac equation. However, the electron mass is very small thus, for the electroweak theory, the electron is assumed to be massless.
After this assumption, the density of lagrangian for the Dirac's fields is [1]

$$\mathcal{L}_0(x) = (\bar{\nu}_{eL}(x), \bar{e}_L(x))(i\gamma^\mu \partial_\mu) \begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \end{pmatrix} + e_R i\gamma^\mu \partial_\mu \nu_R(x) \tag{1.13}$$

The theory must be invariant under local $SU(2)$ of weak isospin tranformations. This leads to the introduction of three (as the number of the vector basis in the adjoint rappresentation) vectorial fields that will be related to the $W^\pm$ and Z bosons existing in nature. For the fundamental representation of the SU(2) group we have as generators the Pauli spin matrices τ_1, τ_2 and τ_3. The corresponding vector fields will be denoted as W^1_λ, W^2_λ and W^3_λ and we combine them into a Hermitian 2 x 2 matrix with zero trace:

$$W_\lambda(x) = W^a_\lambda(x)\frac{\tau_a}{2} \tag{1.14}$$

where τ_a for a = 1, 2, 3, are the generators of SU(2) and correspond to the Pauli spin matrices.
The strength tensor is:

$$W_{\lambda\rho}(x) = \partial_\lambda W_\rho - \partial_\rho W_\lambda + ig[W_\lambda(x), W_\rho(x)] = W^a_{\lambda\rho}\frac{\tau_a}{2} \tag{1.15}$$

where:

$$W^a_{\lambda\rho} = \partial_\lambda W^a_\rho - \partial_\rho W^a_\lambda + g\epsilon_{abc}W^b_\lambda W^c_\rho \tag{1.16}$$

thus the density of lagrangian, taking into account the new fields is,

$$\mathcal{L}^{weak}_{int}(x) = \frac{1}{2}Tr[W_{\lambda\rho}(x), W^{\lambda\rho}(x)] + (\bar{\nu}_{eL}(x), \bar{e}_L(x))i\gamma^\mu(\partial_\mu + igW_\lambda)\begin{pmatrix}\nu_{eL}(x)\\ e_L(x)\end{pmatrix}$$
$$+e_R i\gamma^\mu\partial_\mu\nu_R(x)$$

The effect of the boson's fields can be clarified by the following replacement:

$$W^\pm_\lambda = \frac{1}{\sqrt{2}}(W^1_\lambda \mp W^2_\lambda) \tag{1.17}$$

so the density of lagrangian is:

$$\mathcal{L}(x) = -g(\bar{\nu}_{eL}(x), \bar{e}_L(x))\gamma^\lambda W^a_\lambda\frac{\tau_a}{2}\begin{pmatrix}\nu_{eL}(x)\\ e_L(x)\end{pmatrix} \tag{1.18}$$

The coupling between the electron and the neutrino can be understood as the absorbtion (or emission) of a W boson.
The third component, wihch couples to the left neutrino and not to the right one, can't be interpreted as the photon field because the electromagnetic interaction treats equally particles with the same charge and so cannot distinguish between two different helicity states.
The interaction is described by the Feynman diagram in Fig 1.3.

To understand the atypical coupling involving the left neutrinos, the electromagnetic interaction is needed; thus it is necessary to include the electromagnetic group in the previous lagrangian, imposing the invariance both for $SU(2)$ and for $U(1)$ gauge transformation.

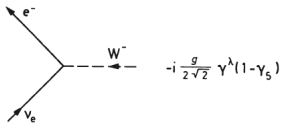


Figure 1.3: Feynman dyagram of the elementary coupling between electron and his neutrino

The invariance for $U(1)$ transformation means that the lagrangian must be unchanged if the phase of ψ_L and ψ_R are changed:

$$\begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \end{pmatrix} \rightarrow e^{i\varphi'} \begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \end{pmatrix} \tag{1.19}$$

$$e_R(x) \rightarrow e^{i\varphi} e_R(x) \tag{1.20}$$

The $U(1)$ group is called the *weak hypercharge* Y group, where Y will be defined later. The hypercharge of the left handed electrons is y_L and hypercharge of the right handed electrons is y_R, those are two constant that will be set below.
If fermions are combined into a spinor $\psi(x)$, then the transformation of the U(1) hypercharge group can be written as:

$$\psi(x) = \begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \\ e_R(x) \end{pmatrix} \rightarrow e^{i\chi Y} \begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \\ e_R(x) \end{pmatrix} = e^{i\chi Y} \psi(x) \tag{1.21}$$

where:

$$Y = \begin{pmatrix} y_L & 0 & 0 \\ 0 & y_L & 0 \\ 0 & 0 & y_R \end{pmatrix} \tag{1.22}$$

The invariance under the $U(1)$ group is guaranteed with the introduction of the vectorial field:

$$B_{\lambda\rho} = \partial_\lambda B_\rho - B_\rho \partial_\lambda \tag{1.23}$$

and thus, with the redefinition of the covariant derivate:

$$D_\lambda = \partial_\lambda + igW_\lambda^a(x)T_a + ig^{'}B_\lambda(x)Y \tag{1.24}$$

the lagrangian:

$$\mathcal{L} = -\frac{1}{2}Tr[W_{\lambda\rho}(x), W^{\lambda\rho}(x)] - \frac{1}{4}B_{\lambda\rho}(x)B^{\lambda\rho}(x) + \bar{\psi}(x)i\gamma_\mu D^\mu\psi(x) \tag{1.25}$$

The coupling term is:

$$\begin{gathered}\mathcal{L}^{'} = -\bar{\psi}(x)\gamma^\lambda(igW_\lambda^a T_a + ig^{'}B_\lambda Y)\psi(x) = \\ -\frac{g}{\sqrt{2}}(W_\lambda^+\bar{\nu}_{eL}\gamma^\lambda e_L + W_\lambda^-\bar{e}_L\gamma^\lambda\nu_{eL}) \\ -\frac{1}{2}(gW_\lambda^3 + 2y_L g^{'}B_\lambda)\bar{\nu}_{eL}\gamma^\lambda\nu_{eL} \\ +\frac{1}{2}(gW_\lambda^3 - 2y_L g^{'}B_\lambda)\bar{e}_L\gamma^\lambda e_L - y_R g^{'}B_\lambda\bar{e}_R\gamma^\lambda e_R\end{gathered}$$

Noting that $g^{'}$ is a free parameter, y_L is set equal to $-\frac{1}{2}$ without loss of generality; those are a linear combination of W_3 and B which have physics reality:

$$Z_\lambda = \frac{1}{\sqrt{g^2 + g^{'2}}}(gW_\lambda^3 - g^{'}B_\lambda) \tag{1.26}$$

$$A_\lambda = \frac{1}{\sqrt{g^2 + g^{'2}}}(g^{'}W_\lambda^3 + gB_\lambda) \tag{1.27}$$

Defining:

$$sin\theta_W = \frac{g^{'}}{\sqrt{g^2 + g^{'2}}} \tag{1.28}$$

$$cos\theta_W = \frac{g}{\sqrt{g^2 + g^{'2}}} \tag{1.29}$$

it results:

$$Z_\lambda = cos\theta_W W_\lambda^3 - sin\theta_W B_\lambda \tag{1.30}$$

$$A_\lambda = sin\theta_W W_\lambda^3 + cos\theta_W B_\lambda \tag{1.31}$$

and the density of the lagrangian for the $SU(2) \times U(1)$ theory is:

$$\mathcal{L} = -\frac{g}{\sqrt{2}}(W_\lambda^+ \bar{\nu}_{eL}\gamma^\lambda e_L + W_\lambda^- \bar{e}_L\gamma^\lambda \nu_{eL})$$
$$-\sqrt{g^2 + g^{'2}} Z_\lambda \{\frac{1}{2}\bar{\nu}_{eL}\gamma^\lambda \nu_{eL} - \frac{1}{2}\bar{e}_L\gamma^\lambda e_L - sin^2\theta_W(-\bar{e}_L\gamma^\lambda e_L + y_R\bar{e}_R\gamma^\lambda e_R)\}$$
$$-\frac{gg^{'}}{\sqrt{g^2 + g^{'2}}} A_\lambda(-\bar{e}_L\gamma^\lambda e_L + y_R\bar{e}_R\gamma^\lambda e_R)$$

By taking into account that the electromagnetic interaction does not distinguish between left and right handed particles and that the strength of the interaction is proportional to the electric charge, both y_R and e can be set to:

$$y_R = -1, e = \frac{gg^{'}}{\sqrt{g^2 + g^{'2}}} \tag{1.32}$$

The density of lagrangian can be written in a more compact form:

$$\mathcal{L} = -e\left\{A_\lambda \mathcal{J}_{em}^\lambda + \frac{1}{\sqrt{2}sin\theta_W}(W_\lambda^+ \bar{\nu}_{eL}\gamma^\lambda e_L + W_\lambda^- \bar{e}_L\gamma^\lambda \nu_{eL}) + \frac{1}{sin\theta_W cos\theta_W} Z_\lambda \mathcal{J}_{NC}^\lambda\right\} \tag{1.33}$$

where:

$$\mathcal{J}_{em}^\lambda = -\bar{e}_L\gamma^\lambda e_L - \bar{e}_R\gamma^\lambda e_R = -\bar{e}\gamma^\lambda e \tag{1.34}$$

$$\mathcal{J}_{NC}^\lambda = \frac{1}{2}\bar{\nu}_{eL}\gamma\lambda\nu_{eL} - \frac{1}{2}e_L\gamma^\lambda e_L - sin\theta_W \mathcal{J}_{em}^\lambda \tag{1.35}$$

The existence of the *neutral current* and of the neutral Z boson are in agreement with the theory.
The density of lagrangian does not contain any mass term, and this is in contrast with experimental results which measured, for example, the mass of the electrons and of the gauge bosons Z and $W^{\pm}$.
It is necessary to introduce a mechanism to both give mass to the particles and to be compatible with the electroweak theory (invariant under SU(2) and U(1) transformation).
This particular *piece* of the electroweak lagrangian useful for that purpose was derived by the physicists R. Brout, F. Englert and P. Higgs and deal with the *symmetry breaking* of the electroweak theory, as explained in the following section.

1.4 Higgs mechanism

The spontaneous symmetry breaking corresponds to the case when a the system shows a symmetry in the lagrangian that is no longer valid at the ground state, so the symmetry is *broken*.
The easiest way to show this particular behaviour is to consider a potential of a scalar field ρ:

$$V(\rho) = -\frac{1}{2}\mu^2\rho^2 + \frac{1}{4}\lambda\rho^4 \tag{1.36}$$

where $\mu^2 > 0$ and $\lambda > 0$, as shown in Fig 1.4

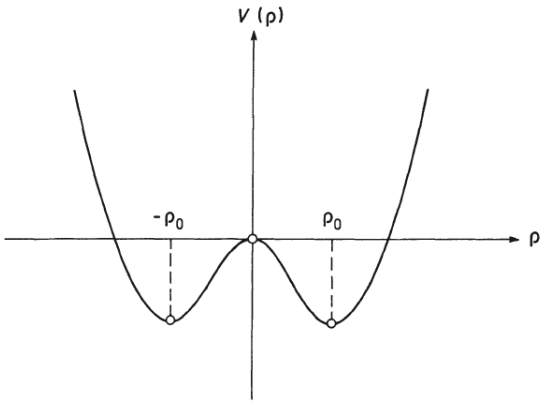


Figure 1.4: Shape of the V_ρ potential in function of ρ

The ground state of the potential can be found by:

$$\frac{\partial V(x)}{\partial x} = 0 \tag{1.37}$$

and this gives the solution $\rho = \pm\rho_0$ where $\rho_0 = \sqrt{\frac{\mu^2}{\lambda}}$.
In the ground state the particle is either at $\rho = \rho_0$ or $\rho = -\rho_0$ and neither of these position, taken alone, shows the symmetry of the potential for transformation as $\rho \to -\rho$; the symmetry is spontaneous broken and the system can decay in one of the two ground states with equal probability.
It is quite easy to see that spontaneous symmetry breaking is associated with a degeneration of the ground state.
An analogue mechanism could be used to give mass at the particle in the SM framework.
Weinberg and Salam incorporates the idea of the spontaneous symmetry breaking in the SM by proposing the presence of a scalar field named the *Higgs field* to give mass at the particles [2]. In the easiest version, it is sufficient to introduce two complex fields: φ_1 and φ_2 arranged in a doublet [1]

$$\varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \end{pmatrix} \tag{1.38}$$

To be coherent with the theory the Higgs potential must be invariant under $SU(2)$ and $U(1)$ transformation. The easiest choice for the new lagrangian density for the field φ is:

$$\mathcal{L}_\varphi = (\partial_\mu \varphi^+)(\partial^\mu \varphi) - V(\varphi) \tag{1.39}$$

where:

$$V(\varphi) = k\varphi^+\varphi + \lambda(\varphi^+\varphi)^2 \tag{1.40}$$

As the previous discussion, the constant $\lambda > 0$ for the theory to be stable, while for k we have no boundary condition, however to have the spontaneous symmetry breaking we must considered $k = -\mu^2 < 0$.
With the substitution:

$$\frac{\rho}{\sqrt{2}} = \sqrt{\varphi^+\varphi} \tag{1.41}$$

the potential is now:

$$V(\varphi) = -\frac{1}{2}\mu^2\rho^2 + \frac{1}{4}\lambda\rho^4 \tag{1.42}$$

The minimum of the potential, as in the previous example is:

$$\rho_0 = \sqrt{\frac{\mu^2}{\lambda}} \tag{1.43}$$

Coming back to the doublet, the ground state of the potential is:

$$\varphi = e^{i(\frac{\tau}{2})\psi} \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}\rho_0 \end{pmatrix} \tag{1.44}$$

thus the ground state is infinitely degenerate.

A possible expectation value for the ground state is:

$$< 0|\varphi|0 >= \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}\rho_0 \end{pmatrix} \tag{1.45}$$

To construct the interaction between the Higgs field and the other particles, it is necessary to construct a density of lagrangian which satisfies the invariance under $SU(2) \times U(1)$ theory; the easiest choice is the Yukawa density of lagrangian:

$$\mathcal{L}_{Yuk} = -c_e \bar{e}_R \varphi^+ \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} + h.c \tag{1.46}$$

It is obvious that the coupling is invariant under local isospin transformation.
We now further demand that the coupling is invariant under hypercharge transformations. For this we need to assign a suitable hypercharge y_H to the Higgs field φ. The elementary processes described by:

$$e_L \rightarrow e_R + \varphi_2 \tag{1.47}$$

should conserve hypercharge. This yields the condition:

$$y_H = y_L - y_R = \frac{1}{2} \tag{1.48}$$

In this way the Yukawa density of lagrangian is invariant under U(1) transformation.
The covariant derivative with the addition of the Higgs hypercharge becomes:

$$D_\lambda = (\partial_\lambda + igW_\lambda^a \frac{\tau_a}{2} + ig^{'} B_\lambda y_H) \tag{1.49}$$

and thus the final lagrangian could be written as:

$$\mathcal{L} = -\frac{1}{2}Tr[W_{\lambda\rho}, W^{\lambda\rho}] - \frac{1}{4}B_{\lambda\rho}B^{\lambda\rho}$$
$$+(\nu_{eL}e_L)i\gamma^\lambda D_\lambda \begin{pmatrix}\nu_{eL}\\ e_L\end{pmatrix} + \bar{e}_R i\gamma^\lambda D_\lambda e_R$$
$$-c_e\bar{e}_R\varphi^+ \begin{pmatrix}\nu_L\\ e_L\end{pmatrix} - c_e^*(\bar{\nu}_{eL}, \bar{e}_L)\varphi e_R$$
$$+(D_\lambda\varphi^+)(D^\lambda\varphi) - V(\varphi)$$

and without loss of generality c_e, the *Yukawa constant*, could be set $c_e >= 0$.
The total number of unknown parameters is five: $g, g^{'}, c_e, \lambda, \mu^2$.
The final step to give mass to the particles is to break the symmetry of the lagrangian by moving to the ground state, so calculating:

$$< 0|\varphi^+|0 > \left(-igW_\lambda^a\frac{\tau_a}{2} - ig^{'}B_\lambda y_H\right)\left(igW^{\lambda a}\frac{\tau_a}{2} - ig^{'}B^\lambda y_H\right) < 0|\varphi|0 > \tag{1.50}$$

That gives for the gauge bosons:

$$m_W^2 = \frac{e^2\rho_0^2}{4sin^2\theta_W} \tag{1.51}$$

$$m_Z^2 = \frac{e^2\rho_0^2}{4sin^2\theta_W cos^2\theta_W} \tag{1.52}$$

and for the electron:

$$m_e = c_e\frac{\rho_0}{\sqrt{2}} \tag{1.53}$$

The Higgs interaction is proportional to the vacuum expectation value, ρ_0, and to the Yukawa constant which is unknown.
With this mechanism the neutrino remains massless, however the neutrino mass could be accommodated in the SM framework with the see-saw mechanism, see Sec. 1.6.
Expanding the potential around the vacuum state we end up with another particle, the Higgs boson with mass: $m_H = 2\lambda\rho_0^2$; the theory does not predict the

value of this mass: it has to be derived experimentally.
The unknown values of the SM are: $e, sin\theta_W, m_e, m_W^2, m_H^2$. This model could be expanded to include the other quarks by replacing the electron spinors with the quark one:

$$\psi = \begin{pmatrix} \nu_L(x) \\ e_L(x) \\ e_R(x) \\ . \\ . \\ . \\ t_R(x) \end{pmatrix} \tag{1.54}$$

and thus we obtain for the SM:

$$\mathcal{L} = -\frac{1}{2}Tr[W_{\lambda\rho}W^{\lambda\rho}] - \frac{1}{4}B_{\lambda\rho}B^{\lambda\rho} + \bar{\psi}iD^{\lambda}\gamma_{\lambda}\psi + \mathcal{L}_{Yuk} + (D\varphi^{\lambda})^{+}(D\varphi_{\lambda}) - V(\varphi) \tag{1.55}$$

where:

$$\mathcal{L}_{Yuk} = \bar{\psi}(x)C_i\varphi_i(x)\psi(x) + h.c \tag{1.56}$$

The summary of the known particle with their properties is given in Fig 1.5.

Type	1st gen.	2nd gen.	3st gen.	I_3	Y	Q	$SU(3)_C$
Quarks	$\begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$\begin{pmatrix} c_L \\ s_L \end{pmatrix}$	$\begin{pmatrix} t_L \\ b_L \end{pmatrix}$	$\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$	$1/3$	$\begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$	triplet
	u_R	c_R	t_R	0	$4/3$	$2/3$	
	d_R	s_R	b_R	0	$-2/3$	$-1/3$	
Leptons	$\begin{pmatrix} \nu_{e,L} \\ e_L \end{pmatrix}$	$\begin{pmatrix} \nu_{\mu,L} \\ \mu_L \end{pmatrix}$	$\begin{pmatrix} \nu_{\tau,L} \\ \tau_L \end{pmatrix}$	$\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$	-1	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	singlet
	e_R	μ_R	τ_R	0	-2	1	
	$\nu_{e,R}$	$\nu_{\mu,R}$	$\nu_{\tau,R}$	0	0	0	

Figure 1.5: Summary of all known elementary particles with their properties

1.5 Higgs boson phenomenology

Experimental confirmation of the reality of the Higgs boson comes from the experiment; indeed in july 2012 the first evidence of a resonance, decaying into two γ or two Z bosons, with the mass of 125 GeV [3].
Before the discovery of the Higgs boson and depending on its mass, several decay modes were studied and the decay rate calculated, as shown in Fig 1.6.

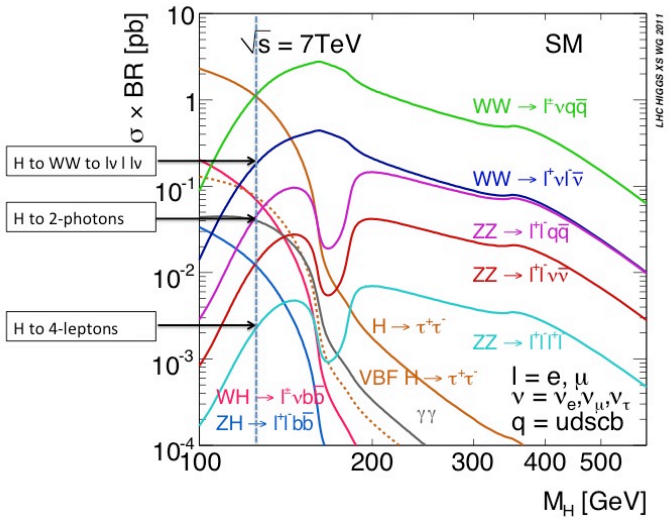


Figure 1.6: Possible decay mode of the Higgs boson in function of his mass

Even if the $\gamma\gamma$ decay mode, as shown in Fig 1.6, is not the favourite one, it was used to search the Higgs boson, together with $H \rightarrow ZZ \rightarrow 4\,leptons$ because of the good signal to background discrimination and the optimal mass resolution. The diphoton and 4 leptons invariant mass are reported in Fig 1.7.

In Fig 1.7 (a) a narrow peak at 125 GeV over a smoothly falling background in the invariant mass distribution of two photons emerges; the same peak is found in the 4 leptons invariant mass (Fig 1.7 (b)), dominated by ZZ non resonant background process; there was the evidence of a new particle with 125 GeV of mass with a 5σ significance.
It is worth to clarify that the peak at 91 GeV arises from the Z→4l decay, the Feynman diagram of this process is shown in Fig 1.8.

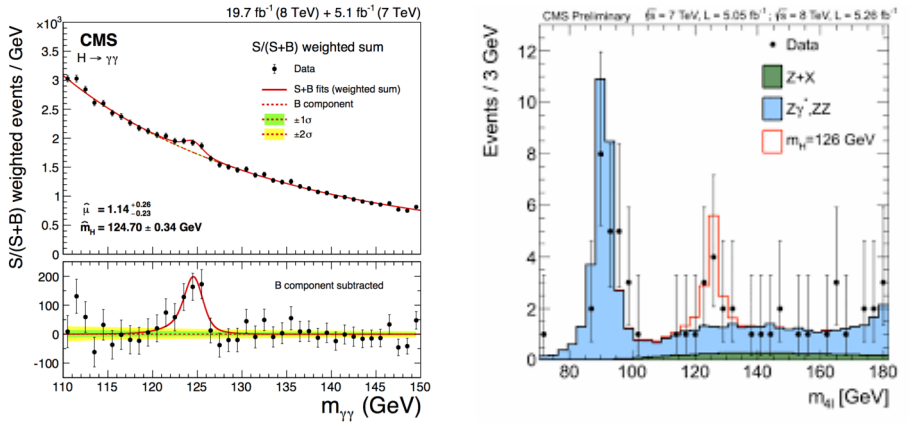


Figure 1.7: Invariant 4 leptons mass (a) and diphoton invariant mass (b) in events collected at 7 TeV

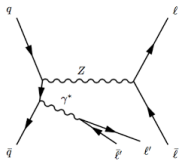


Figure 1.8: Faynman dyagram of the Z→4l decay.

The first evidence of the Higgs boson was obtained combined the data of CMS and ATLAS experiment from the first LHC data-taking (RunI), corresponding roughly at the luminosity of $20 fb^{-1}$.
After RunI, from 2016 to 2018 (RunII), the luminosity collected by CMS was one order of magnitude higher achieving $137 fb^{-1}$. During RunII around 8 million Higgses were produced, corresponding to an incredible amount of data to be analysed in Fig 1.9 [4].

The first measurement that was done with the combination of RunI and 2016 (36 fb^{-1}) only was the Higgs boson mass, a milestone for the Higgs mechanism. $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ \rightarrow 4l$ were combined together. The former was chosen for the small branching fraction but very clean final state topology, that means diphoton invariant reconstructed with high precision, and the latter was chosen for the large signal-to-background ratio and precise reconstruction of the final-state decay products. The results are shown in Fig

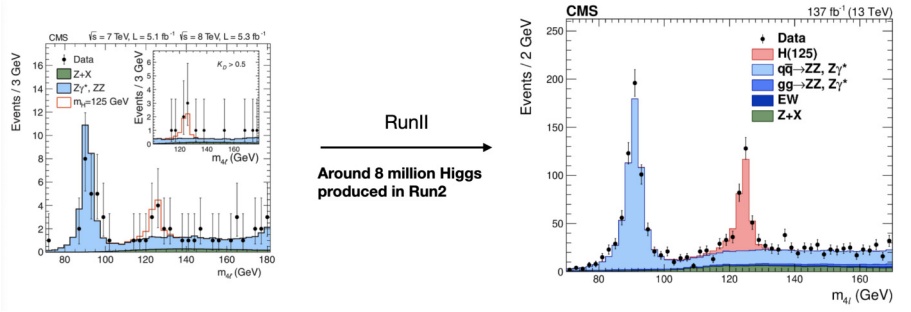


Figure 1.9: Comparison between RunI and RunII [4]

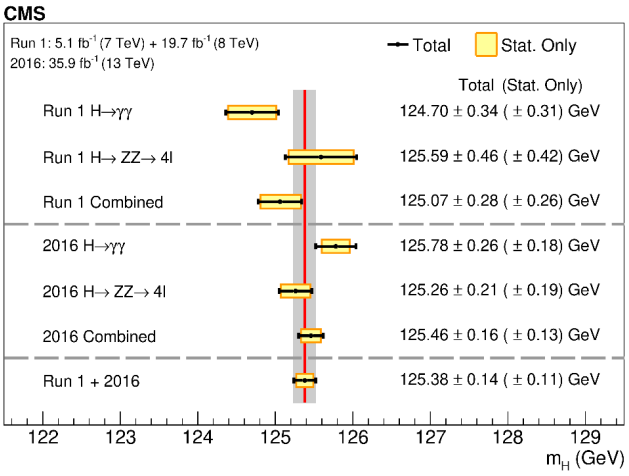


Figure 1.10: Higgs mass measurement with the CMS experiment combining $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ [5]

1.10 [5].

The effort of the scientist's community was then focused to fully determine the Higgs boson properties starting from his couplings to fermions and vector bosons: a wide range of final state were studied, and lots of exiting results came out. Among them it's worth to cite the first observation of Higgs boson coupling to the second generation fermion ($H \to b\bar{b}$ with 5.6σ [6]) and first evidence of Higgs boson coupling to the third generation fermion ($H \to \mu^+\mu^-$ with 3σ [7]). A summary of the latest CMS results is in Fig 1.11.

From the plot in Fig 1.11 no evidence of physics beyond the SM is found. In parallel with the determination of the Higgs coupling to fermions, it was

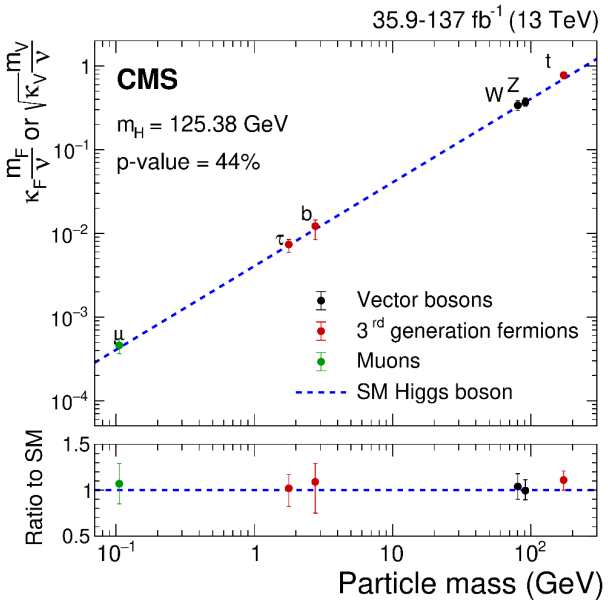


Figure 1.11: Higgs couplings with fermions as a function of the parton mass [7]

possible to investigate scenarios as the Higgs self coupling and the so called differential cross section measurement of the Higgs observables, which means to seek a deviation from the SM measuring the cross section in bins of a certain observable. During this thesis I will show results that I've obtained for both of these phase-space, I'll leave a more precise introduction and explanation in the next sections.

To conclude the overview of the Higgs results, it's worth to mention the production modes that we can access at LHC. Fig 1.12 gives a nice picture of the situation we are facing this days, with the 13 TeV center of mass energy.

Going in decreasing order, the most important Higgs production mechanism is the gluon gluon fusion with cross section of about 48.6 pb. The second most important mechanism is the vector boson fusion (VBF) where the Higgs boson is produced in association with a jet pair of large invariant mass, the cross section is one order of magnitude lower with respect to the gluon-gluon fusion and it's 3.8 pb.

The Higgs boson can also be product in association with vector bosons with a clear signature of one vector boson and a Higgs candidate, this process has a cross section of 2.3 pb. And the last but no the least Higgs production mechanism is the association with the top quark in the ttH final state; this mechanism is of particular interest because of the direct determination of the Yukawa cou-

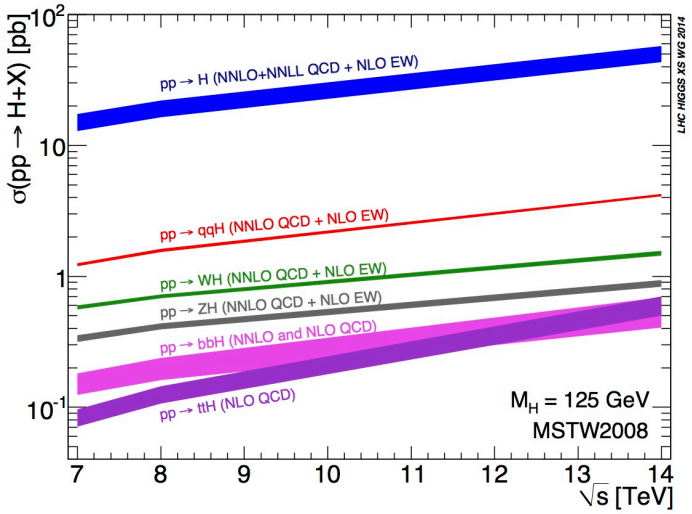


Figure 1.12: Higgs boson production mechanism at LHC as a function of the center of mass energy

pling for the top quark and has a cross section of 0.5 pb.
Tab 1.1 gives the branching ratio of the most important Higgs boson decay mode.

Topology	Branching ratio
$H \to \bar{b}b$	58.4%
$H \to W^+W^-$	21.4%
$H \to \tau^+\tau^-$	6.27%
$H \to ZZ$	2.62%
$H \to \gamma\gamma$	0.23%
$H \to Z\gamma$	0.15%
$H \to \mu^+\mu^-$	0.002%

Table 1.1: Branching ratio of the most important Higgs boson decay mode.

In the next subsections we will focus on Higgs boson pair production and Higgs associated production with top quarks.

1.5.1 Higgs boson pair production

In the SM, the Higgs self-couplings are uniquely determined by the structure of the scalar potential [8]:

$$V = \frac{m_h^2}{2}h^2 + \lambda_3 v h^3 + \lambda_4 v h^4 \tag{1.57}$$

where $\lambda_3 = \lambda_4 = \frac{m_h^2}{2v^2}$.
It is evident the prediction of a vertex with three and four Higgs bosons; in particular the vertex with three Higgs bosons are proportional to the λ_3 constant, and the determination of this parameter, which is predicted and calculated in the SM framework, could be a strong proof of the stability of the SM theory.
The di-Higgs phenomenology is dominated by the very tiny cross section of 31 fb, due to the destructive interference of the box and triangle diagrams [9] (Fig 1.15 - 1.13).

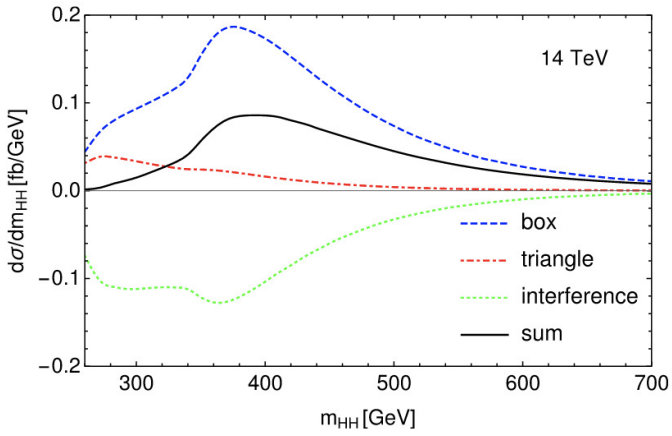


Figure 1.13: Box and triangle interference in the di-Higgs production. It is possible to see the contribution of the box (blue line) and triangle (red line) with their interference (green line) and the overall sum (black line) [9]

The HH phenomenology is well summarized in Fig 1.13 where it is possible to see how the box and triangle diagrams contribute to the final cross section. The interference term is always negative, this leads to a substantial reduction of the overall sum (black line).
With a very low cross section and an overall contamination of backgrounds mainly from QCD and $t\bar{t}$ production, the di-Higgs searches are very peculiar

and challenging. Fig 1.14 shows the cross section of the different HH production modes as a function of the center of mass energy.

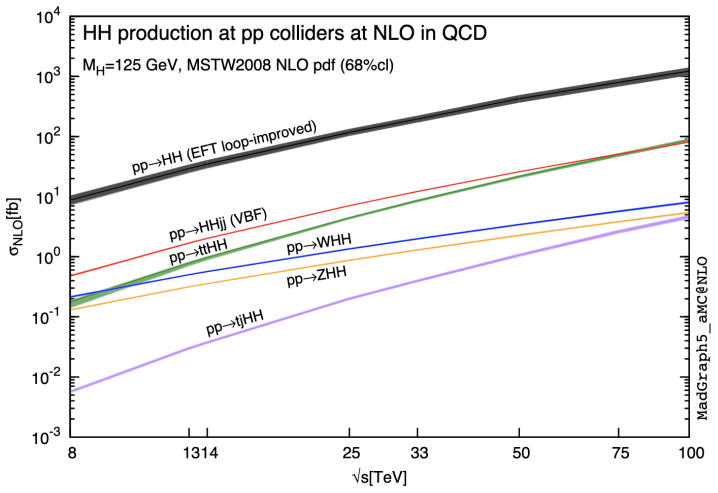


Figure 1.14: Di-Higgs production mechanism at LHC as a function of the center of mass energy [9]

Analogously to single-Higgs production, all the channels lead to a final state involving two Higgs bosons. Going in the decreasing order of cross section, we can see the peculiarity of the most important production mechanism of the HH:

- via gluon-gluon fusion where the two Higgs are produced on mass shell by a radiation from a heavy quark loop and so it depends on λ_3 and on the y_t Yukawa coupling for the top quark;

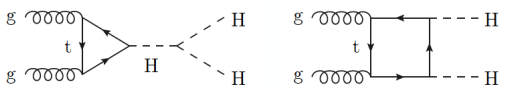
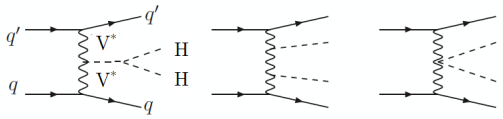
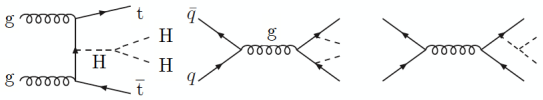


Figure 1.15: Gluon gluon fusion di Higgs production

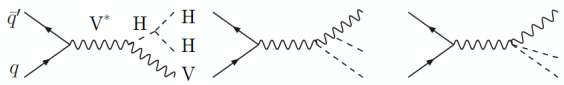
- via vector boson fusion which has a cross section of one order of magnitude lower than the gluon fusion;



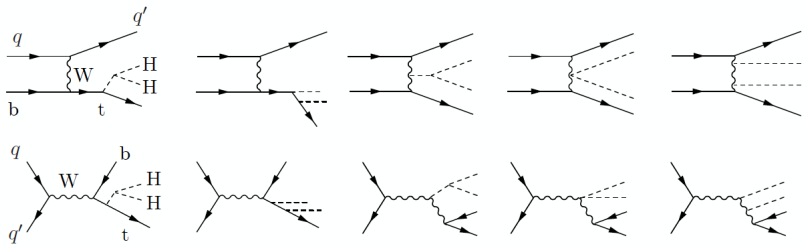
- via top quark associated production;



- via vector boson associated production: the HH are produced in association with a on shell vector boson;



- via single top associated production: the HH production comes in association with the top quark production and that is the only process that is sensitive at the same time to the HH couplings to vector bosons and to top quarks and to their relative phase.



The cross section of these mechanism with the relative center of mass energy are listed in Fig 1.16.

Because of their extremely tiny cross section, the HH channels were hidden during the full RunI: the phenomenology and actual investigation of this production started only with RunII. Even now, with the full luminosity of RunII, we have not enough data to measure the trilinear coupling or the HH cross section.
To access experimentally the di-Higgs phase space one should find a trade of between keeping the branching ratio high enough and enhance the signal purity, selecting and combining different Higgs decay as is shown in Fig 1.17.

The branching ratio of each combination of di-Higgs channel is specified in Fig 1.17. To keep the branching ratio as high as possible, the majority of di-Higgs searches are forced to have one Higgs decaying into two b quarks.

Production mode	σ[fb]
	$\sqrt{s} = 8\,\text{TeV}$
Gluon fusion	$10.15^{+4.1\%}_{-5.7\%}$ (scale) $\pm 3.1\%$(PDF) $\pm 2.6\%(\alpha_s) \pm 5.0\%$(top)
VBF	$0.459^{+3.2\%}_{-3.6\%}$ (scale) $\pm 2.6\%$(PDF + α_s)
$t\bar{t}$HH	$0.174^{+2.8\%}_{-10.6\%}$(scale) $\pm 3.9\%$(PDF + α_s)
W^+HH	$0.145^{+0.43\%}_{-0.52\%}$(scale) $\pm 2.8\%$(PDF + α_s)
W^-HH	$0.0677^{+1.0\%}_{-1.2\%}$ (scale) $\pm 3.7\%$(PDF + α_s)
ZHH	$0.143^{+2.7\%}_{-2.3\%}$ (scale) $\pm 2.6\%$(PDF + α_s)
tjHH	$0.00540^{+5.4\%}_{-3.1\%}$ (scale) $\pm 5.6\%$(PDF + α_s)
	$\sqrt{s} = 13\,\text{TeV}$
Gluon fusion	$33.49^{+4.3\%}_{-6.0\%}$ (scale) $\pm 2.1\%$(PDF) $\pm 2.3\%(\alpha_s) \pm 5.0\%$(top)
VBF	$1.62^{+2.3\%}_{-2.7\%}$ (scale) $\pm 2.3\%$(PDF + α_s)
$t\bar{t}$HH	$0.772^{+1.7\%}_{-4.5\%}$ (scale) $\pm 3.2\%$(PDF + α_s)
W^+HH	$0.329^{+0.32\%}_{-0.41\%}$(scale) $\pm 2.2\%$(PDF + α_s)
W^-HH	$0.173^{+1.2\%}_{-1.3\%}$ (scale) $\pm 2.8\%$(PDF + α_s)
ZHH	$0.362^{+3.4\%}_{-2.6\%}$ (scale) $\pm 1.9\%$(PDF + α_s)
tjHH	$0.0281^{+5.2\%}_{-3.2\%}$ (scale) $\pm 4.5\%$(PDF + α_s)
	$\sqrt{s} = 14\,\text{TeV}$
Gluon fusion	$39.59^{+4.4\%}_{-6.0\%}$ (scale) $\pm 2.1\%$(PDF) $\pm 2.2\%(\alpha_s) \pm 5.0\%$(top)
VBF	$1.95^{+1.8\%}_{-2.3\%}$ (scale) $\pm 2.4\%$(PDF + α_s)
$t\bar{t}$HH	$0.949^{+1.8\%}_{-4.8\%}$ (scale) $\pm 3.2\%$(PDF + α_s)
W^+HH	$0.368^{+0.33\%}_{-0.39\%}$(scale) $\pm 2.1\%$(PDF + α_s)
W^-HH	$0.197^{+1.2\%}_{-1.3\%}$ (scale) $\pm 2.7\%$(PDF + α_s)
ZHH	$0.414^{+3.5\%}_{-2.7\%}$ (scale) $\pm 1.8\%$(PDF + α_s)
tjHH	$0.0364^{+3.7\%}_{-1.3\%}$ (scale) $\pm 4.7\%$(PDF + α_s)

Figure 1.16: Cross section of the principal production channel for HH event

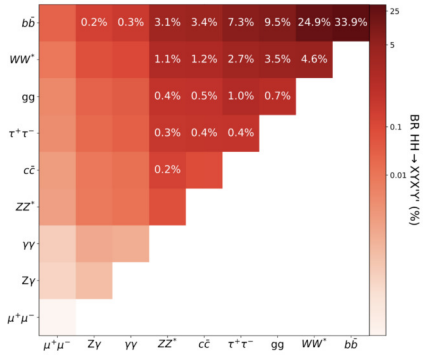


Figure 1.17: Branching ration of di-Higgs production in different topologies

Using the 2016 data, the CMS community published $HH \to b\bar{b}\gamma\gamma/b\bar{b}\tau^+\tau^-/b\bar{b}b\bar{b}$ and their combination. The choice of the final state was given by the considerations done above, all these analysis were aiming to have an upper limit on the di-Higgs cross section and a constraint on the λ_3 parameter. In Fig 1.18 the results for each analysis and the result of the combination is shown.

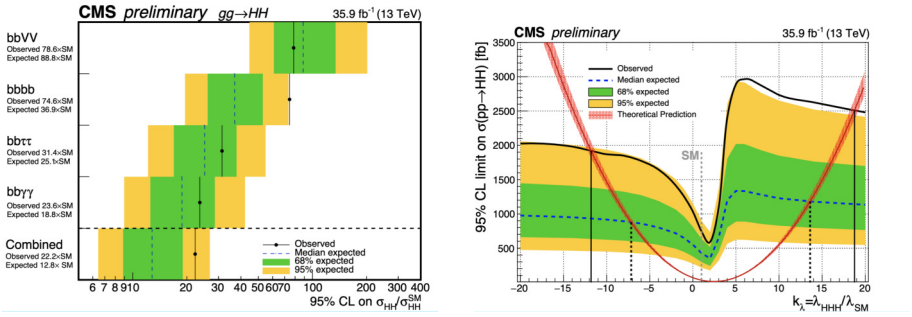


Figure 1.18: On the left: upper limit on the signal strength for the di-Higgs channels and their combination. On the right constrained on k_λ for the di-Higgs combination. Both plots are done with CMS 2016 data

The upper limit on the signal strength, which is defined as $\sigma_{mesured}/\sigma_{SM}$, is 12.8 at 95% CL and the k_λ, defined as $\lambda_{3measured}/\lambda_{3SM}$, is constrained to be within -7.1 and 13.6.
For the purpose of this thesis, we will focus on $HH \to \bar{b}bZZ \to \bar{b}b4l$ analysis was the first HH analysis done with the full RunII dataset.

1.5.2 Higgs associated production with top quarks

With the purpose to fully determine the Higgs properties, the Yukawa coupling should be investigated. The strength of the coupling of the Higgs boson with the fermions goes with the mass of the particle, in particular, given the mass of the fermion, m_f the Yukawa coupling (y_f) is $y_f = \sqrt{2}m_f/\nu$, where ν is the vacuum expectation value of the Higgs field.
Given this relation, it is extremely important to measure this quantity in the phase-space where it is mostly emphasised: the Higgs associated production with top quark. Looking at numbers, given: $\nu = 246GeV$ and $m_t = 173.4$ the predicted value of y_f is of the order of one. A deviation of y_t from the SM prediction would unambiguously indicate the presence of new physics beyond the SM.

The Yukawa coupling of the Higgs to fermions has been studied and determined with a large overall uncertainties during the LHC RunI. Increasing the luminosity of a factor one, with Run II, it was possible to constrain the value of $k_t = y_t/y_t^{SM}$ to be within - 0.9 and -0.7 or 0.7 and 1.1 at 95% CLs probing the associated production of the Higgs in association with a top quark pair and subsequent decay to leptonic final states. The measured has been improved by adding all the other final states and now it's quoted to $k_t = 1.01^{+0.11}_{-0.11}$, the best result so far obtained with the RunII data.
A summary of the results obtained for the Yukawa coupling in both RunI and RunII can be found in Fig 1.19 [10].

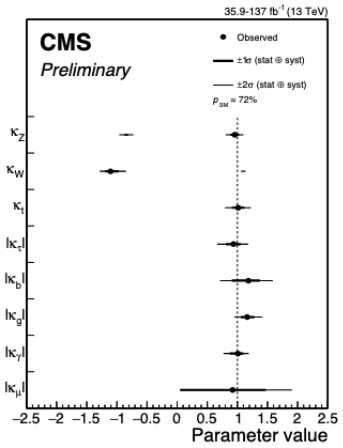


Figure 1.19: Summary of the couplings Yukawa coupling for each particle. The thick (thin) black lines report the 1σ (2σ) confidence intervals. Results were obtained during RunI and RunII with a luminosity up to $137 fb^{-1}$ [10]

The coupling modifiers are consistent with unity, except for k_W. The preferred negative value of k_W results from the interference between diagrams contributing to tH production.
No evidence of deviation with respect to the SM expectation has been found, this doesn't however mean that the BSM physics is excluded. Thanks to the impressive amount of data that we have available, it is possible to perform an analysis that looks in all the possible deviation of the theory from data in bins some predefined observables, this procedure goes with the name of differential cross section measurement and will be fully explained in the next chapters. For the ttH system, the observables that are most sensitive to BSM effect are the Higgs transverse momentum and the ttH mass.

A deviation in these vertices (Fig 1.20) of the ttH production can be accommodated in the contest of Effective Field Theory (EFT) [11].

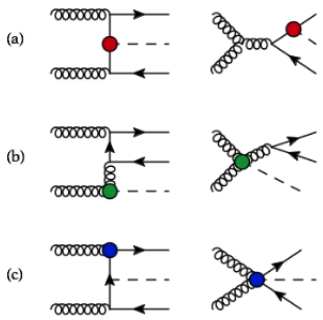


Figure 1.20: Example of vertices that can be modified with an EFT approach for ttH production

The anomalous interactions between SM particles may lead to energy growth of the scattering amplitudes through the introduction of new Lorentz structures or via spoiling delicate unitarity cancellation in the SM amplitudes. This energy growth can impact the kinematic distributions significantly, a feature that emphasizes the importance of the EFT interpretations of differential measurements and its capability to outperform inclusive ones. In the context of this thesis, a ttH differential and EFT analysis will presented in following sections.

1.6 Heavy Neutral Leptons

Together and in parallel with the Higgs physics, evidence of BSM interaction may come also from neutrino phenomenology.
Because of absence of right-handed neutrino fields in the SM, the couplings between neutrinos and scalar fields is not allowed; then it is not possible to accommodate in any sense the neutrino mass. Nonetheless, the observation of flavor neutriono oscillations, which means a mixing between several neutrino flavours in flight, are possible only if neutrinos have mass. This represents a very strong evidence of physics beyond the SM.
From the theoretical point of view, the main way that we have to give mass to neutrinos is to introduce a heavy states N with right-handed chirality, also called Heavy Neutral Leptons (HNL). In this minimal SM extension, neutrinos can couple with the Higgs field and acquire mass (Dirac mass) as: $m_\nu \sim y_\nu^2 * v^2/m_N$ where y_ν is the Yukawa coupling and v is the vacuum expectation

value of the Higgs field. The smallness of neutrino mass with respect to all the particles that acquire mass in the same way is puzzling itself. It's possible to describe neutrinos also in the Majorana framework, where each particle is the charge conjugate of itself. In the latter option, the mass term can be generated by some unknown physics mechanism way above the electroweak scale, thus not accessible at the energy that we have now.
Since neutrinos can be described in both Dirac and Majorana framework, the interchange of this description makes the neutrino "split" into a heavy and a light component mass terms, this mechanism is well-known as the see-saw mechanism [12].
At this point it is clear why the determination of the neutrino mass is as important as the determination of the Higgs mass in the SM.
Lots of effort has been made to constraint the neutrino mass: from cosmology experiments [13] - [14] to direct measurements in the tritium decay [15]. The common result indicate that the neutrino masses are orders of magnitude smaller compared to the other SM particles.
Beyond the gauge invariant mass term, HNLs can also help understand several other problems in cosmology and high energy physics. For example a stable HNL may be a possible candidate for dark matter, while any heavy partner might help explain the matter-antimatter asymmetry of the early universe as a second and third generation of heavy neutrinos would increase the amount of CP-violation.
The mass range of HNL has been scanned from keV to GeV scale, in Fig 1.21 [16] is shown the current limits on the mixing parameters with three lepton families (denoted as $|V_{Nl}|$) and the mass (M_N), prospects with future colliders are also included.

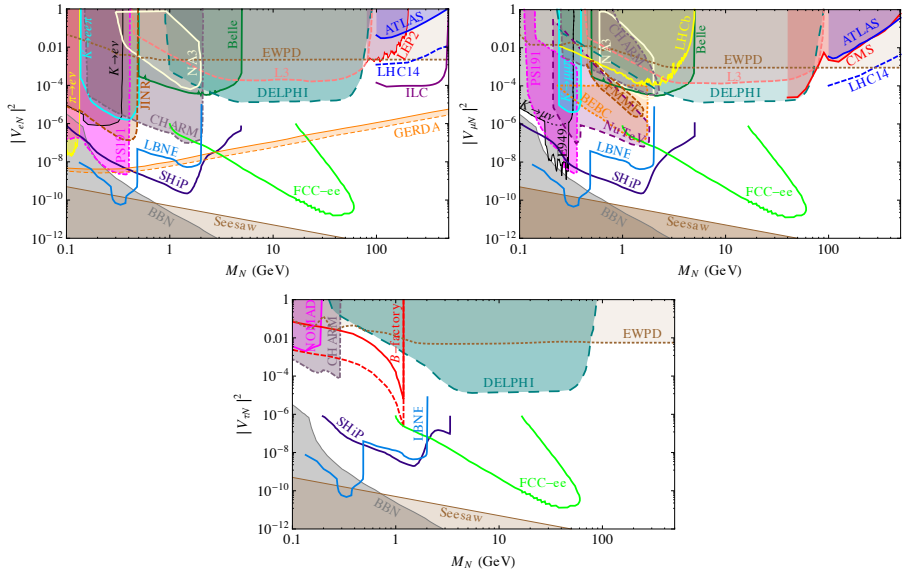


Figure 1.21: Current limits on the HNL masses and their couplings to the three lepton families, and projected sensitivity achievable in planned future facilities [16].

In the context of high energy physics, since HNL is a singlet in the SM, it cannot interact with any of the SM particle through the electroweak nor the strong interaction, it can however mix with the SM neutrino and the products of the mixing can be explored.
It holds a relation between the mass of the HNL, the coupling to the leptons and the life-time: $\tau_N = M_N^{-5}|V_{Nl}^{-2}|$; this implies that the decay product of the HNL, with long lifetime, are non-prompt and emerge from a secondary vertex, spatially displaced with respect to the interaction point.
There have been several searches for HNLs in CMS, ATLAS and LHCb. The CMS experiment reported on a search for HNLs using events with two same-sign leptons and at least one jet is searched for using data collected during 2016 in proton-proton collisions at a center-of-mass energy of 13 TeV and corresponding to an integrated luminosity of 35.9 fb^{-1}.
The limits shown in Fig 1.22 are written as function of the mass of the HNL and of the coupling with leptons and are the most restrictive direct limits up to now [17].

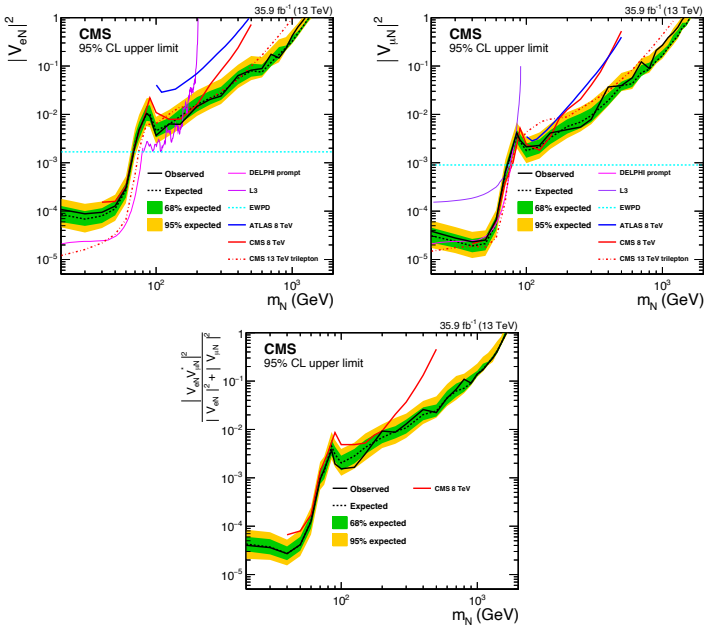


Figure 1.22: Exclusion region at 95% CL in the $|\mathrm{V_N}|^2$(top-left), $|\mathrm{V_N}|^2$(top-right) and, $|V_{\mathrm{N}e}V_{\mathrm{N}\mu}^*|^2/(|V_{\mathrm{N}}|^2 + |V_{\mathrm{N}}|^2)$ (bottom) vs. m_{N} plane [17].

In this thesis we will focus on the direct production of an HNL in a W decay where the SM neutrino oscillates into a HNL, and the HNL afterwards decay into a W boson and a charged lepton (see Fig 1.23), spanning a not yet explored phase space were the HNL mass is constrained to be within 1 and 15 GeV (Sec 5).

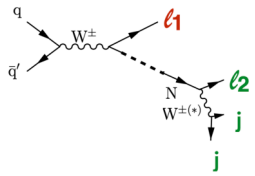


Figure 1.23: Typical diagrams for the production of a HNL at the LHC (N) through its mixing with a SM neutrino, leading to a final state with two charged leptons and two jets.

Chapter 2

LHC accelerator and *CMS* detector

The European Organization for Nuclear Research (CERN) was born as a European research organization focused on particle physics and technology development. Now the activities of CERN catalyses the scientific activity in the high energy physics domain, with about 10000 people using the facilities, from more than 800 institutes and universities of 76 different countries in the world. The CERN laboratories host the Large Hadron Collider (LHC), designed to accelerate protons inside a 26.7 km long tunnel to a centre-of-mass energy of 14 TeV. The LHC is the largest and most powerful particle accelerator ever built and represents today the frontier of the research machine for the high energy physics.
The particle beams, accelerated by a complex accelerator chain, collide in four interaction points at LHC, instrumented with detectors. In one of these four points is installed the Compact Muon Solenoid (CMS) experiment, designed to explore the physics at the TeV scale.
This chapter introduces the properties and operations of the LHC and the structure of the CMS detector used to collect the data analysed for this thesis.

2.1 LHC accelerator

The *Large Hadron Collider* is a circular accelerator of 27 km of diameter.
It is designed to collide protons at a centre-of-mass energy of 14 TeV with an instantaneous luminosity of $\mathcal{L} = 10^{34} cm^{-2} s^{-1}$ as well as lead ions at a centre-of-mass energy of $2.76 GeV$ per nucleon and $\mathcal{L} = 10^{27} cm^{-2} s^{-1}$ [18].
In the LHC, two separate, counter-rotating particle beam lines are kept in orbit in two magnetic channels thanks to the field generated by superconducting niobium-titanium coils. The particles are steered and divided into bunches by a magnetic field of 8.3T generated by a current of about 11 kA in the 1232

dipole magnets, each measuring 15 metres of length and 35 tonnes of weight. In one bunch there are at least 10^{11} protons. The stability of the beam dynamics is ensured by 392 quadrupoles magnets measuring 5 to 7 metres of length, that focus the particles and keep them in a narrow beam. Special quadrupoles are installed in front of the collision points to squeeze the beams and increase the proton density in the collisions. Superconducting magnets are cooled with superfluid helium-4 and kept to a working temperature of 1.9K.
As shown in Fig 2.1, the LHC is just the last element of an injection chain made of several smaller particle accelerators. The hydrogen atoms are stripped of their electrons in a duo-plasmatron source and are accelerated to an energy of 50 MeV in the Linear Accelerator (LINAC2), which feeds the Proton Synchrotron Booster (PSB) where protons are accelerated to 1.4 GeV. The beam is then injected into the Proton Synchrotron (PS) for a further acceleration to 25 GeV, and subsequently into the Super Proton Synchrotron (SPS) where protons reach an energy of 450 GeV. The proton beams are finally transferred to the two LHC beam pipes, where the beams are accelerated and shaped into proton bunches thanks to radio-frequency cavities operated at 400MHz. Once the proton reaches the nominal energy and the beam dynamics is stabilized, protons are brought to collide in four points along the LHC ring.

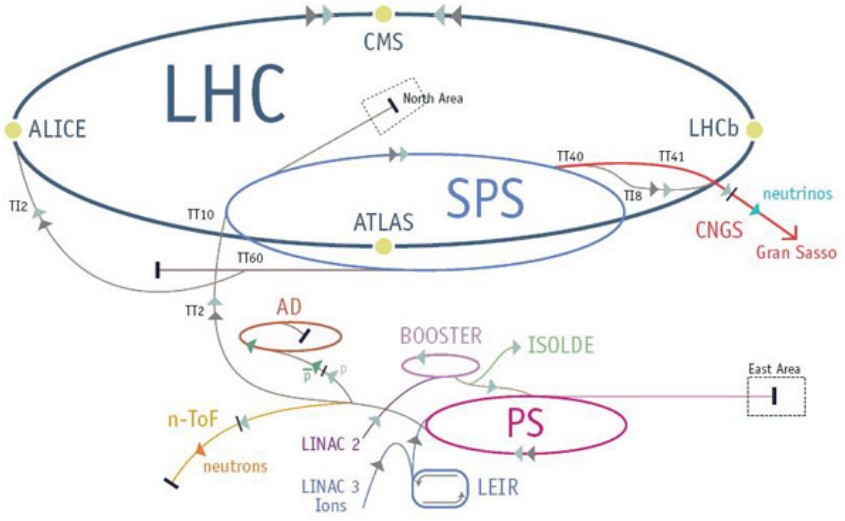


Figure 2.1: Accelerator chain at LHC

An important parameter of the LHC machine is the instantaneous luminosity $\mathcal{L}$ of the collisions, that depends on the beam properties as [18]:

$$\mathcal{L} = \frac{N_b^2 n_b f_{rev} \gamma}{4\pi\sigma_x\sigma_y} F$$

where N_b is the number of particle in each of the n_b bunches per beam, f_{rev} is the revolution frequency of the particles in the beam, γ is the Lorentz factor to take into account the relativistic contraction of the electromagnetic field, σ_x and σ_y describes respectively the average of the bunches in the x and y direction. The factor F accounts for the geometric reduction of the instantaneous luminosity due to not perfect geometric allineation of the bunches at the interaction point.

The nominal values of the LHC machine parameters are summarized in Table 2.1.

$\sqrt{s}$	center of mass energy	14 TeV
Δt_b	bunch spacing	25 ns
N_b	particle per bunch	1.15×10^{11}
n_b	bunches per beam	2808
f_{rev}	revolution frequency	11.2 kHz
σ_{xy}	transverse bunch size	16.7 μm
σ_z	longitudinal bunch size	7.55 cm

Table 2.1: Nominal parameter at LHC machine.

The integrated luminosity $\int Ldt$ is a measure of the number of collisions produced in a certain temporal interval dt; in order to obtain the integrated luminosity per year the recovery time of the machine must not be taken into account.
The luminosity is the coefficient of proportionality between the number of events N, produced for a specific process, and its cross section σ:

$$N = L \times \sigma$$

In the accelerator there are four collision points instrumented with the experiments:

- "A Toroidal LHC ApparatuS" (ATLAS) and the "Compact Muon Solenoid" (CMS) experiments are installed in the diametrically opposite of the LHC, where the highest instantaneous luminosity of collision is produced. They are designed as hermetic, multi-purpose detectors designed to study the *high pt* physics and the Higgs sector;
- "A Large Ion Collider Experiment" (ALICE) is built to study heavy ion collisions and quark-gluon plasmas, to shed light on *strong interaction* phenomena as the formation of quarks, baryons and mesons;
- "LHC beauty" (LHCb) one-arm spectrometer devoted to the study of CP-violation in B hadrons.

In the following part of the chapter the CMS detector is described .

2.2 The CMS experiment

The CMS detector has been built to explore the physics at the TeV scale with many different signatures and final states. It was consequently designed as a multi-purpose detector, that hermetically surrounds the interaction point in the underground cavern of Point 5 in Cessy, and is instrumented with several subsystems developed for the identification and measurement of different types of particles. The detector has a cylindrical structure with a diameter of 15 m and a length of 21.5 m, and an overall weight of about 12500 t [19].
Collisions take place in the centre of the CMS experiment every 25 ns, implying that new waves of particles leave the interaction point before those produced in the previous bunch crossing have even escaped the external surface of the detector.
In addition, multiple proton interactions can take place within each bunch crossing. These two effects are globally denoted as out-of-time and in-time "pileup" and overlap to the signal of interest represented by the hard-scatter interaction. These challenging conditions requires the design of a detector highly granular, fast in its response, and resistant to the radiation. At the same time, it must be capable of precisely measuring the energy and the momentum of the final state particles and to identify them.

2.2.1 Coordinate system

A right-handed Cartesian coordinate system is used to describe the detector and the collision products. It is defined with its centre in the nominal interac-

tion point, the x axis pointing to the centre of the LHC ring, the y axis pointing upwards, and the z axis pointing in the anticlockwise proton beam direction. Given the cylindrical structure of the detector, a polar system is also used. The azimuthal angle φ is defined in the (x, y) plane, called *transverse* plane, as the angle formed with respect to the positive x axis, and the radial coordinate in this plane is denoted as r. The polar angle θ is defined in the (r, z) plane as the angle formed with the z axis, as shown in Fig 2.2 it is usually converted into the pseudorapidity $\eta = ln(tan(\frac{\theta}{2}))$[1].

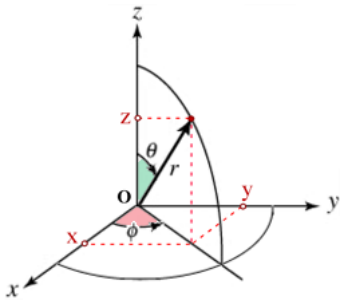


Figure 2.2: Polar system

The spatial separation of two particles can be expressed in terms of their angular distance as $(\Delta R)^2 = (\Delta\eta)^2 + (\Delta\varphi)^2$.
The projection of the momentum of a particle onto the transverse plane is referred to as the "*transverse momentum*" or p_T.

2.2.2 Detector structure

The CMS detector is made of a central section, or "barrel", and two forward regions, or "endcaps", as it can be observed in the schematic representation of Fig 2.3.

CMS is instrumented with multiple, concentric layers of detectors to identify and measure the particles produced in the collisions. The interaction point is surrounded by pixel tracker and strip tracker detectors to precisely measure the positions of the interaction points (or "collision vertices") and the trajec-

[1]The pseudorapidity is a generalization of the rapidity which is defined as $y = \frac{1}{2} ln \frac{E+p_z c}{E-p_z c}$. Pseudorapidity is preferred in high energy physics because is invariant under Lorentz transformation, which is not true for the rapidity

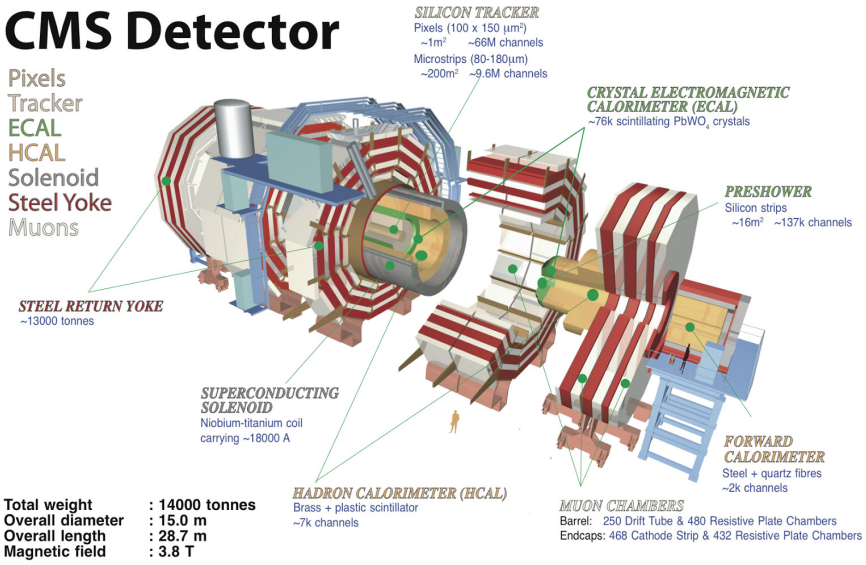


Figure 2.3: A schema of the Compact Muon Solenoid

tory and momentum of the charged particles.

The electromagnetic and the hadron calorimeters are located around the tracking systems and designed to absorb electrons, photons, and hadrons within their volume to measure the energy deposited. Muons can traverse the calorimeters and are measured in muon tracking systems located in the outermost part of the detector.

The core of the experiment is a niobium-titanium superconducting solenoid of 6m of diameter. It is operated at a temperature of 4.5K and generates a 3.8T magnetic field along the z axis [19]. This strong magnetic field is used to bend the charged particles and measure their transverse momentum with the tracking subdetectors. The tracker and the calorimeters systems are located inside the solenoid, which imposes tight constraints to their size and, in the case of the calorimeters, requires high density materials to contain the incoming particles and their secondary interaction products. The return field of the magnet has an intensity of about 2T and is used to measure the transverse momentum with the muon detectors located inside the iron structure that surrounds the solenoid. This causes the muons trajectories to be bent in opposite directions in the inner tracker and muon systems, a characteristic feature to which the CMS experiment owes its logo.

The information from the individual subdetectors are often redundant and can be combined to improve the reconstruction of final state objects.

2.2.3 Inner tracking systems

The inner tracking system is designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from the LHC collisions; it covers a region of pseudorapidity equal to $|\eta| < 2.5$ [19]. It consists of a volume of 5.6 m of length and 2.4 m of diameter instrumented with silicon sensors, sensitive to the passage of charged particles.
Thanks to the uniform magnetic field (3.8 T) within the tracking detector volume, a charged particle follows an elliptical orbit with radius proportional to the momentum and the mass of the particle, thus the information on the position of charged particles within each silicon detector, or *hits*, are combined to measure the radius of the orbit and consequently the momentum and charge of these particles. The spatial measurement provided by the tracking system also allows for the determination of the hard scattering interaction point (*primary vertex*) and its discrimination against additional interactions from the pileup in the event. It also allows for the reconstruction of in-flight decays such as those of B hadrons (*secondary vertex*).
To fulfill the requirement of precise spatial measurement while being exposed to a large flux of particles, the tracking detector is finely segmented and equipped with fast readout on-board electronics. Moreover, to minimize the impact of the tracking measurement on the passage of charged particles, its design is optimized to use a minimum amount of material. The silicon detector technology deployed in the CMS tracking system addresses these needs by providing a large surface of thin, finely segmented, active detectors.
The detector occupancy rapidly decreases with the radial distance r as the particle flux with a r^{-2} dependence: thus, depending on the flux of particle to which the detector must be exposed, the inner tracking system is divided in 3 zones [19]:

- the innermost region ($r < 20$ cm), which faces with high radiation and particle flux, and consists of pixel detectors disposed in three cylindrical layers in the barrel and two disks in the endcap. The pixel element has size of 100 μm x 150 μm;
- the medium region (20 cm $< r <$ 55 cm) which is instrumented by microstrip silicon modules of size 10 cm x 80 μm;

- the outermost region (55 cm $< r <$ 110 cm) which uses silicon microstrip modules with the size of 25 cm x 180 μm.

A schematic view of the inner tracking system is reported in Fig 2.4.

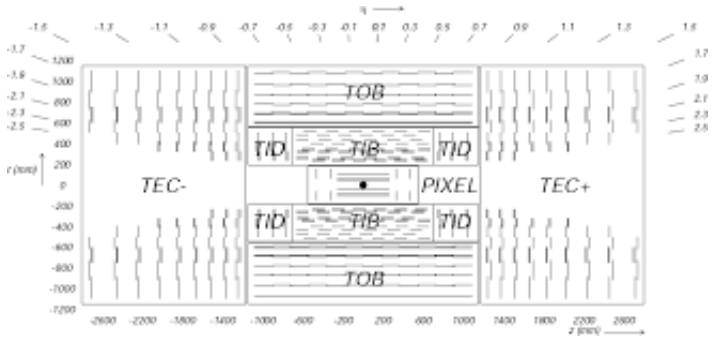


Figure 2.4: Longitudinal view of the CMS inner tracking system layout. The pixel detector is located directly in front of the interaction point. The strip tracker detector is composed of the tracker inner barrel (TIB) and tracker inner disks (TID), surrounded by the tracker outer barrel (TOB) and the tracker endcaps (TEC).

The transverse momentum resolution of the tracker is a function of the particle momentum, is shown in Fig 2.5. For muons, the efficiency is about 99% over most of the acceptance range.

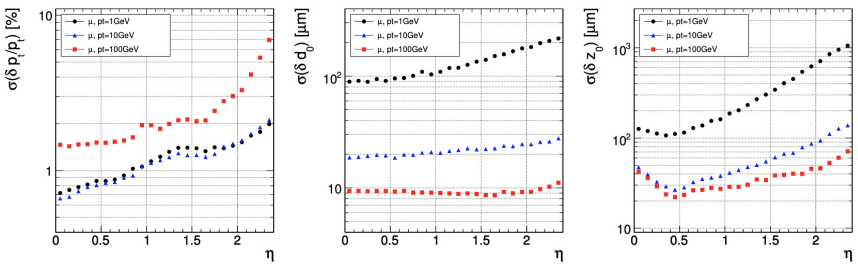


Figure 2.5: Resolution of several track parameters for single muons with transverse momenta of 1, 10 and 100 GeV.

2.2.4 Pixel detector

The most inner part of the tracking is instrumented with silicon pixel detector. It contributes to very precise measurement of the tracks the $r-\varphi$ and z planes,

therefor is responsible of small impact parameter resolution and of secondary vertex reconstruction.
This high granular detector is essential for the precise reconstruction of secondary vertices from b and tau decays; it forms seed tracks for the outer track reconstruction and high level triggering.
With the scope of having the same track resolution on the $r - \varphi$ and z, the different cells of size of 100 μm x 150 μm are rearranged in different geometries depending on the pseudorapidity region that they cover. Fig 2.6 shows a longitudinal overview of the pixel detector.

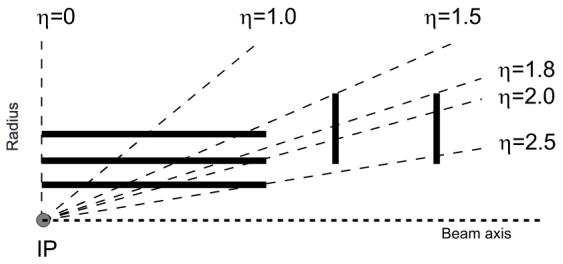


Figure 2.6: Pixel detector geometry

It is possible to distinguish three barrel layers (BPix) with two endcap disks (FPix). BPix are 53 cm long layers located at 4.4, 7.3 and 10.2 cm while the FPix are disks from 6 to 15 cm radius located at $z = \pm 34.5$ and $z = \pm 46.5$ cm. BPix (FPix) contain 48 million (18 million) pixels covering a total area of 0.78 (0.28) m^2. The chosen arrangement for layers and disks ensure the presence of 3 traking points over almost the full η range. For this reason in the highest η region, the 2 disk points are combined with the lowest possible radius of the 4.4 cm layer [19].
The proximity to the beam pipe requires a high-radiation and high-flux tolerant design. The chosen sensors are n+ pixel on n- substrate detector that allows partial depleted operation even at very high particle flux. The drift of the electrons in $\simeq$4T electromagnetic field is perpendicular to the barrel pixels, this leads to a charge spreading of the collected signal on more then one pixel. This leads to a 15-20 μm spatial resolution. To achieve a position resolution of 15 μm over the full range, a charge-sharing is induced between the neighbouring pixel. This condition is due to the 20° tilted in the forward detectors and is also enhanced by the ExB drift. The radiation damage will reduce de-

pletion depth or the increase in bias voltage, thus the spacial resolution will be degraded every year. To cope with this problem, the innermost layers the mechanics and the cabling of the pixel system has been designed to allow a yearly access if needed.
The lifetime of the innermost layer with the full LHC luminosity is around 2 year.
To summarise the pixel detector performances, Fig 2.7 shows the efficiency over the full pseudorapidity range.

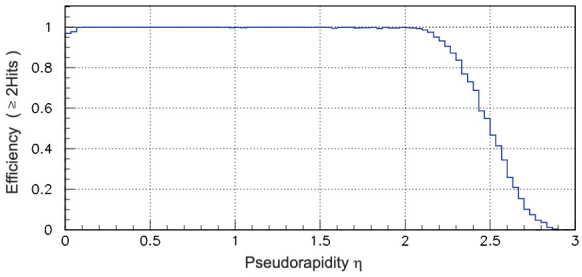


Figure 2.7: Pixel detector geometry [19]

2.2.5 Silicon Strip Tracker

The furthermost tracker region, between 20 and 116 cm, is instrumented with silicon strip layers and disks and is composed by three different subsystem: tracker inner barrel and disks (TIB and TID), tracker outer barrel (TOB), tracker EndCaps (TEC):

- TIB and TID extend in radius towards 55 cm and are composed of 4 barrel layers, supplemented by 3 disks at each end. Thanks to TIB and TID it is possible to have at least 4 hits for each trajectory in the $r - \varphi$ plane, the geometry is such as the TIB strips are parallel to the beam pipe and the TID disks are radial.
  Depending on the distance to the beam pipe, the strip pitch of the TIB varies from 80 μm to 120 μm leading to a single point resolution from 23 to 35 μm. In the TID the mean pitch varies between 100 m and 141 m.

- TOB it consists of 6 barrel layers 500 m thick micro-strip sensors with strip pitches that varies from 183 to 122 m. TOB surrounded TIB/TID

detectors and it extends till 118 cm, it provides a single point resolution of 53 to 35 m.

- TEC is composed of 9 disks, carrying up to 7 rings of silicon micro-strip detectors and covers all the z range beyond the TOB. The thickness of the rings goes from 320 to 500 μm with radial strips of 97 m to 184 m average pitch.

The sensors in the strip tracker are single sided p-on-n type silicon micro-strip sensors done on a 6 inch wafers in a standard planar process which leaded to a significant cost reduction with respect to the standard 4 inch wafers.
Silicon Strip Tracker geometry is shown in Fig 2.8

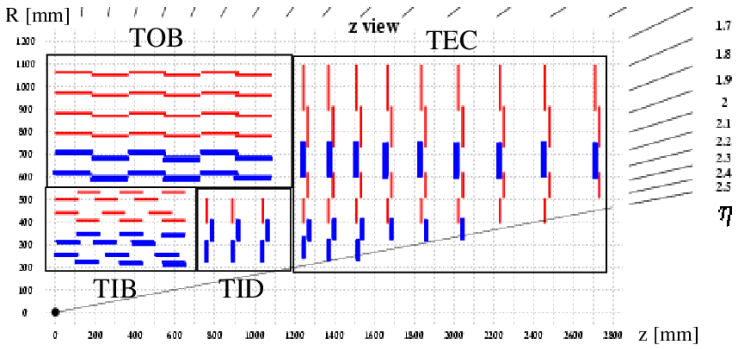


Figure 2.8: Silicon Strip Tracker layout in a longitudinal section. The thin red and thick blue lines represent single-sided and double-sided silicon modules, respectively

2.2.6 Strip Tracker Electronics

The physical requirement for the CMS experiment and, in general, for all the LHC experiment, are pushing further the technology in terms of speed and amount of data to be transferred. All the front end electronics have to be in sync with the overall LHC clock and since also time of flights particle have to be taken into account, local adjustment should be made locally. Thus, the overall timing synchronization is a very challenging task. The CMS electronics consist on two main parts: one way is the electronic that transmits the data from the detector to the control room, the other one is a bi-directional control chain that has to deliver clock and trigger signals to each detector and exchanges control information between control room and front-end electronics. All signal are transmitted through 100m of optical fiber between front-end and control room.

On either end, the light information is converted to electrical signals and vice-versa.
In Fig 2.9 there is a sketch of the CMS Silicon Strip Tracker electronics.

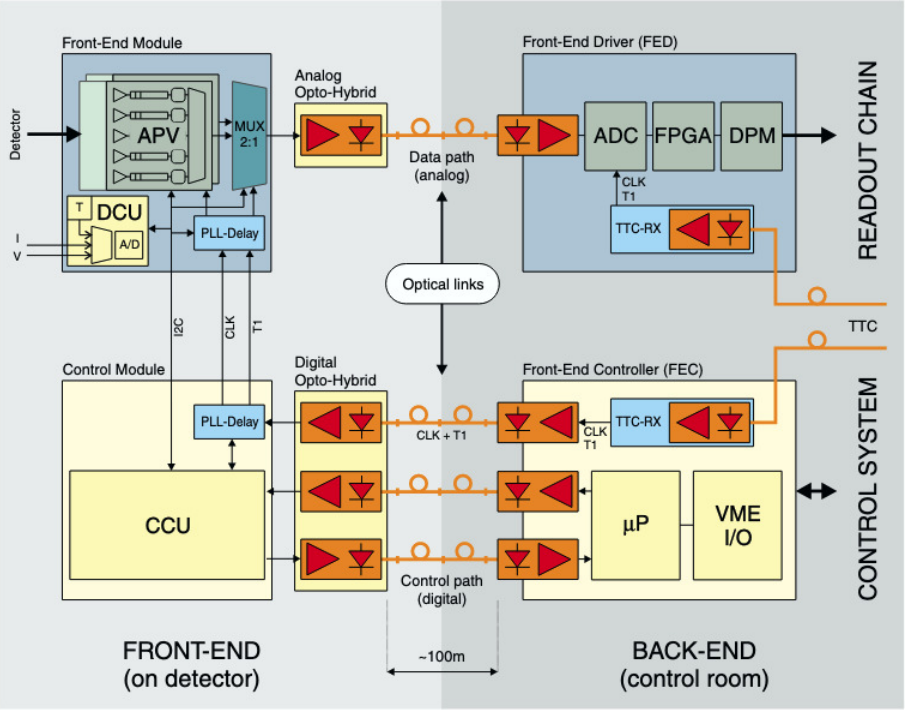


Figure 2.9: Schematic view of the CMS Silicon Strip Tracker back-end and front-end electronic.

Several components can be distinguished on the board: the timing, trigger and control (TTC) takes care of the centralized clock and of the trigger of the optical fibers, then there are local receiver boards TTCrx that gives these signals to the electronic. The Front-end controller (FEC) that adds to these signals a control and sends them to the front-end control module using a digital optical link. The information are given to the Communication and Control Unit (CCU) that interprets the data and passes them to the front-end module. The parameter of the detector, as temperature and voltage are controlled by the Detector Control Unit (DCU) that talks back with the CCU that transmit the data to the FED again. In this thesis we will focus mainly on the APV25 front-end amplifiers, that takes analog signals and multiplexed them to the

APVMUX chip and sent over the analog optical link. The data are digitalized and preprocessed by the FED that takes also the clock from the TTC [19].

APV chips

The APV chip is a front-end amplifier for CMS silicon strip tracker Fig 2.10 .

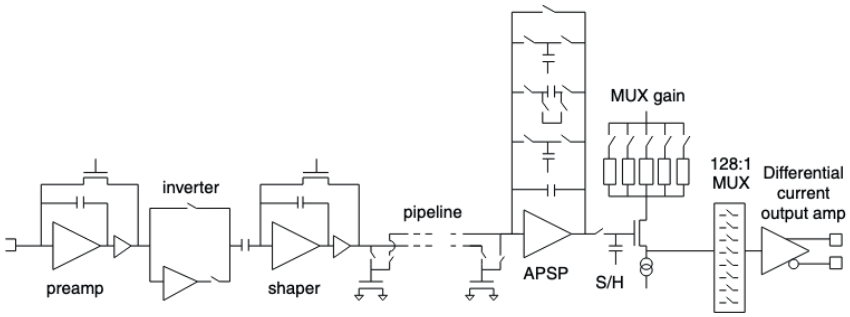


Figure 2.10: Block dyagram of the APV chip.

There are three main components: the pre-amplifier, the shaper and the deconvolution filter for each of its 128 channels. After the pre-amplifier it is possible to select the polarity of the signal with the inverter. The CR-RC shaper is a 50 ns peaking time. The output of the shaper is sampled each 25ns that corresponds to 40MHz and then is fed to a pipeline of adjustable size. The signal is extracted at the end of the pipeline under a certain trigger request. It is possible to set the chip for deconvolution algorithm to combine the samples of three consecutive bunches in a voltage pulse which has rise and fall time fairly equal to the 25 ns bunch crossing interval. At the end all the signal of the 128 channels are multiplexed onto a single line with a differential current amplifer output. The current version of the APV chip series is called APV25S1, which is manufactured in the 0.25 μm submicron process as its predecessor APV25S0.
Inelastic interactions between hadrons and the nuclei of silicon sensors can heavily ionise the bulk through the production of highly ionising particles (HIPs), such as recoiling nuclei and nuclear fragments. Those HIP events caused a mismeasurement of the charge collected by the APV sensors and a huge recovery time in the detector. In Sec ... it will be describe how to deal with this effect.

2.2.7 Tracker upgrade for Phase I

The pixel detector was changed during the long shut down I. This choice was motivated mainly by two factors: the buffer size and read out speed of the Read Out Chip (ROC) and the radiation damage of the components.
For the former the main issue was that the ROC, as it was developed, was too slow and would have caused an important read out inefficiency of around 16% (50%) with 25 ns (50 ns) bunch spacing; the inefficiency increase exponentially with the luminosity. This would lead to a substance degradation of the tracking performances. In addition due to radiation damage, the pixel detector was not meant to last for all the Phase I data taking. Furthermore the detector has lots of passive material that degrades the measurements due to multiple scattering and the detector designed was not fully hermetic.
The goal of the upgrade was to have a fully efficient detector with 4 hits coverage up to $|\eta| < 2.5$ and with less material that can sustain the full luminosity of RunII $2x10^{34}cm^2s^1$.
The pixel detector was then replaced with a new pixel detector with 1 layer and 1 disk more (to reach the 4 hits plateau) plus a new and more efficient ROC [20]. The passive material was significantly reduced by moving the readout electronics and connectors further out.
The redesigned pixel tracker improved the performances of 30/40% with respect to the previous version: the pattern recognition, track parameter resolution, vertexing, and b-tagging performance of the upgraded detector are significantly better than in the previous detector. The improvement in tracking efficiency, fake rate, impact parameter resolution, and vertexing all contribute to significantly increase the b-tagging performance of the new detector.
The improvement in the b tagging resolution vertex resolution was of a crucial importance for many analysis as the Higgs and the Neutrino analyses that will be presented in this thesis. A feeling of the improvement among the two versions of the tracker can be read in Fig 2.11.

The plots shown the b tagging efficiency of the Combined Secondary Vertex algorithm versus the light quark efficiency; the (a) is with a luminosity low enough that no multiple collisions are allow while the (b) is with the RunII luminosity. The red line is the performance of the new tracker while the blue one is the previous version of the tracker. The results obtained with the new tracker shows a significantly better performance even at zero pileup events where the performance of the previous tracker was not degraded by dynamic data loss.

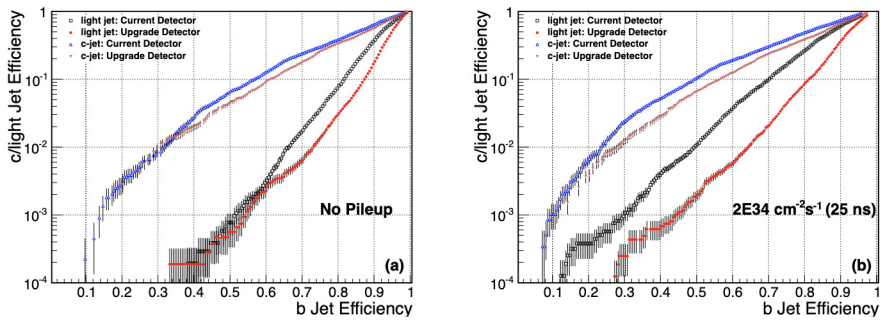


Figure 2.11: b tag efficiency of the Combined Secondary Vertex algorithm versus the light quark efficiency. (a) is with a luminosity low enough that no multiple collisions are allow while the (b) is with the RunII luminosity. The red line is the performance of the new tracker while the blue one is the previous version of the tracker [20]

2.2.8 Electromagnetic calorimeter

The electromagnetic calorimeter, or ECAL, is designed to measure the energy of incident electrons and photons. The CMS ECAL is a homogeneous and highly granular calorimeter made of lead tungstate crystals ($PbWO_4$) [19]. The energy measurement is based on the electromagnetic showering of the electron or photon, detected through the scintillation light produced in the crystal, and on the *containment* of the shower in the detector. The crystals are at the same time the dense interacting material and the active scintillating medium, resulting in an excellent energy resolution.

The choice of $PbWO_4$ is motivated by its high density (8.28 g/cm^3), small radiation length ($X_0 = 0.89cm$) and short Molière radius (R = 2.2 cm). These parameters ensure an excellent containment of the electromagnetic shower within the crystals, which have a length of approximately 25 X_0. The lead tungstate is radiation hard and about 80% of its scintillation light is produced within 25 ns, making it ideal for the high instantaneous luminosity collisions of the LHC and the proton bunch spacing. The disadvantage of this material is the relatively low light yield, corresponding to about 30 photons per MeV of deposited energy, which requires the usage of photodetectors with internal amplification.

The barrel part of the ECAL is made of 61200 crystals with a frontal transverse section of 22 × 22 mm^2 and a length of 23 cm, and ensures the coverage of the region $\eta < 1.479$. The two endcaps are each made of 7324 crystals of a frontal transverse section of 28.62×28.62 mm^2 and a length of 22 cm, and extend the coverage up to $\eta < 3.0$.

The tracking and calorimeter detectors in CMS provide complementary measurements.
The former can identify only charged particles, and the precision of its momentum measurement increases as p_T decreases, because of the larger curvature of the trajectory. Inversely, the latter can measure both charged and neutral particles with a resolution that increases with the particle energy itself.
The resolution of ECAL depends on 3 terms: the first one is a stochastic term that depends on the number n of scintillation photons (or elementary information carriers in general) produced in the interaction and follows a poissonian distribution, the second term is related to detector inhomogeneities, resulting in an error that amounts to a constant fraction of E, the third term accounts for the noise in the detector and does not depend on E. The combined effect of these three factors results in a dependence of the calorimetric energy resolution σ on the particle of energy E as:

$$(\frac{\sigma}{E})^2 = (\frac{S}{\sqrt{E}})^2 + (\frac{N}{\sqrt{E}})^2 + C^2$$

where S, N and C denote the stochastic, noise, and constant terms, respectively.
In ECAL test beam studies performed with incident electrons, the values S = 2.8%, N = 12%, and C = 0.3% were obtained.

2.2.9 Hadronic calorimeter

Hadrons typically traverse the ECAL volume without being stopped. The hadronic calorimeter, or HCAL, is designed to absorb them within its volume and measure their energy. While electrons and photons interact only electromagnetically with the calorimeter material, the hadrons interact strongly with nuclei of HCAL's material and thus the flux of produced particle is higher than the one produced in ECAL. For this reason the transverse and longitudinal lenght of the hadronic shower is higher than electromagnetic one, thus the hadronic shower is more challenging to contain in the detector. Taking into account the production of neutral particles due to hadron-nucleon interaction, the scenario becomes even more complicated.
These effect limit the intrinsic resolution on hadron energies, which can be improved offline with the usage of the particle flow reconstruction techniques.
The detector is complemented by a outer hadron calorimeter (HO) located outside the solenoid, which extends the total interaction depth to about 11 λ_0,

where the constant indicates the average interaction length of hadrons in the calorimeter. The energy measurement in the forward region is complemented by the forward hadronic calorimeter (HF), that is located 11.2m away from the interaction point and measures hadron interactions up to $|\eta| = 5.2$.
Because of the higher radiation levels in the forward region, the HF is composed of steel absorbers and quartz fibres that produce light by Cherenkov effect, measured by photomultiplier tubes (PMTs). Fibres of two different lengths are installed to estimate the electromagnetic and hadronic components of the shower. The global layout of the HCAL is illustrated in Fig 2.12.

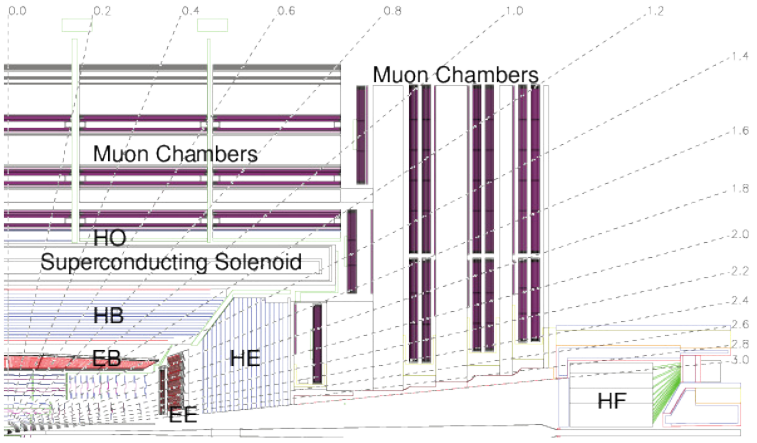


Figure 2.12: Longitudinal view of the HCAL layout, where are visible the barrel (HB) and endcaps (HE) detectors located inside the solenoid, the outer detector (HO) outside the solenoid, and the forward calorimeter (HF).

The overall HCAL performance is dominated by the imperfect containment of the hadronic shower, which results in a resolution sampling term of about $\frac{110\%}{\sqrt{E}}$ and a constant term of 9% [19], as measured in test beams with pions.

2.2.10 Muon system

Because of their typical energy, muons produced in collisions at the LHC have minimal energy loss rates. As a consequence, they traverse the ECAL, the HCAL, and the solenoid volumes without being stopped and are identified and measured in the muon detectors located in the outermost part of CMS. The muon momentum is measured using the return field of the solenoid inside the iron structure in which the muon detectors are embedded, and complements the measurement from the inner tracker previously discussed. CMS is instru-

mented with three types of muon detectors, chosen accordingly to the expected background rates and uniformity of the magnetic field, as it is illustrated in Fig 2.13 [19].

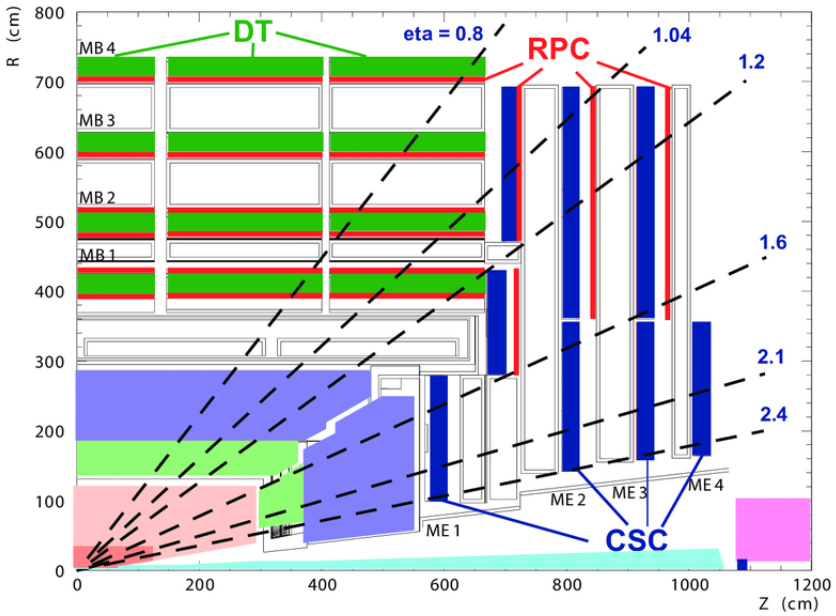


Figure 2.13: CMS muon system

Drift tube chamber

In the barrel region, CMS deploys 250 drift tubes (Fig 2.14), called DT, uniformly distributed in the five barrel sections, or "wheels". Each wheel hosts four concentric rings of DT stations, organized in twelve contiguous sectors. The basic element of the DT detector is a rectangular cell of transverse size 4.2 × 1.3 cm^2, containing an anode wire and filled with a Ar/CO_2 gas mixture. Electrodes on the top and bottom of the cell ensure a constant field and a uniform drift velocity of about 55 $\mu m/s$, while cathodes are placed on the sides of the cell. DT cells are organized in three groups of four elements (three "super-layers", SLs) that together make a DT chamber. Muons traversing each group ionize the gas, and their position and angle are measured from the time needed by the electrons to drift toward the anode wires.
Each DT cell has a spatial resolution of about 200 μm, resulting in a resolution

of 80-120 μm for the global chamber measurement [19].

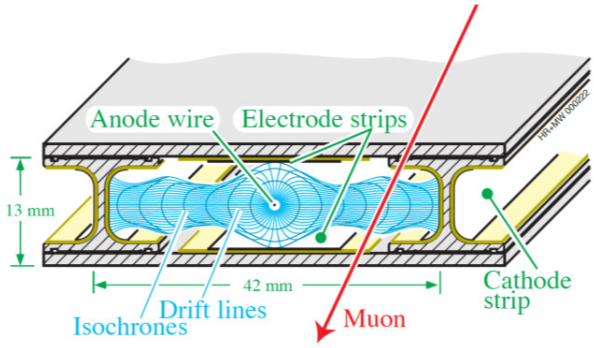


Figure 2.14: Longitudinal section of drift tubes

Because of the relative long delay time of DT, those detectors cannot operate at high particle flux and magnetic field, thus in the endcap they are replaced with cathode strip chambers.

Cathode strip chambers detectors

Cathode strip chambers detectors, namely CSCs, (Fig 2.15) are used to instrument the endcap regions of CMS ($0.9 < |\eta| < 2.1$). As explained before, the different choice of detector technology is imposed by the higher background rates and the stronger magnetic field. CSCs are detectors designed in a trapezoidal shape and made of six layers of anode wires interposed between seven segmented cathode plates disposed in the perpendicular direction. CSCs contain a $Ar/CO_2/CF_4$ gas mixture, which is ionized upon the passage of a muon. The resulting signals induced on the wires and on the strips are interpolated and provide a position measurement in the (r, φ) plane (anode wires) and along the z direction (strips). The CSC is a fast detector, capable of identifying the bunch crossing of a pp collision, and achieves a spatial resolution of 40-150 μm [19].

Resistive plate chambers

Resistive plate chambers, namely RPCs, (Fig 2.16) are installed in both the barrel and endcaps and cover the region $|\eta| < 1.6$. RPCs are formed by two

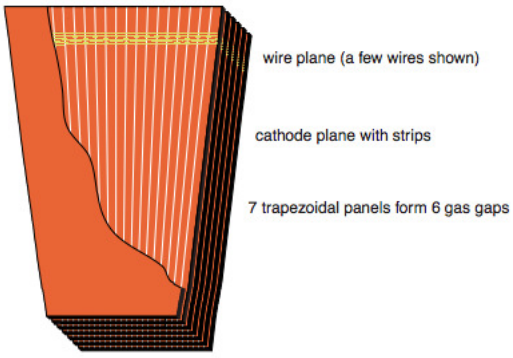


Figure 2.15: Cathode strip tubes

gaps consisting of two resistive bakelite layers of 2 mm thickness separated by a 2 mm volume filled with a $C_2H_2F_4/i$ - C_4H_{10}/SF_6 gas mixture [19]. The detector is operated in avalanche mode and, when traversed by a muon, an avalanche is generated by the high electric field inside the gas volume and is read out by strips located on the outer surface of the gap. Although RPCs have a modest spatial resolution of 0.8 -1.2 cm, they have excellent timing properties with a resolution of the order of the ns, allowing for the determination of the pp bunch crossing.

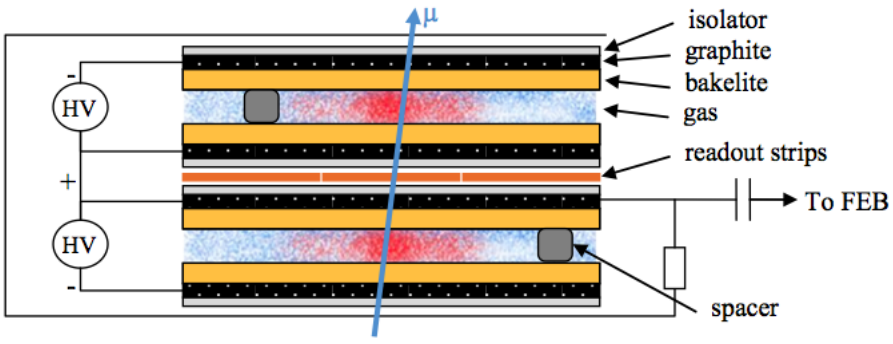


Figure 2.16: Longitudinal view of resistive plate chamber

2.3 Luminosity measurement in CMS

The precise determination of the luminosity is a key parameter in collider experiments. Together with the center of mass energy, it is one of the two main

figures of merit that drives the construction of new experiment and that can quantify the potential of producing novel massive particle. In principle the luminosity is a process-independent quantity and, known the cross section of a given process can be fully determine. The fundamental limitation on precise determination of this ideal cross section motivate the choice of other technique to measure the luminosity.
The different detectors that measurement this quantity are called luminometer and there are two main algorithms that CMS follows: the pixel cluster counting (PCC) and the zero counting method. The luminosity can be written as:

$$\mathcal{L} = \frac{R_{vis}}{\sigma_{vis}}$$

where R_{vis} is the measured rate of a given luminometer and σ_{vis} is the visible cross section in the detector.
For each bunch interaction a number of different observables (hits, tracks or clusters) can be measured in the detector:

$$< N_{observable} >=< N_{observable/interaction} > \mu$$

where μ is the average number of interaction. In a minimum bias events, which are events with very low or no trigger requirement, μ can be expressed in terms of an underling cross section σ_0:

$$\mu = \frac{\sigma_0}{f} L$$

where f = 11 245.6 Hz is the revolution frequency at LHC. The visible cross section can be written as:

$$\sigma_{vis} =< N_{observable/interaction} > \sigma_0$$

using the definition of μ:

$$\mu = \frac{\sigma_{vis}}{< N_{observable/interaction} >} \frac{L}{f} \Rightarrow$$
$$\mathcal{L} = \frac{< N_{observable/interaction} > f}{\sigma_{vis}}$$

The unknown parameter of this formula are $< N_{observable/interaction} >$ and σ_{vis}: σ_{vis} can be determined looking at the spacial dispersion in the x-y axis of the beam using the so called Van Der Mer scan [21] while $< N_{observable/interaction} >$ (usually is $< N_{cluster/interaction} >$) and can be determined counting the number of clusters per each bunch slot. This procedure is called pixel cluster counting (PCC) [22].
To determine the uncertainties on the luminosity two type of errors are taken into account: the tail of the pixel hits that goes in the subsequent bunch slot and the exponential decay activation of the material that cover the detector, these two effects are taken into account in the overall procedure.
The zero counting method [22] exploits the underling possonian probability of having a null interaction in a pp collision. The number of null collisions can be written as:

$$P_{\mu}(0) = e^{-\mu}$$

where $P_{\mu}(0)$ is the poissonian probability of having no interaction. It is possible to determine μ and then the luminosity as:

$$\frac{d\mathcal{L}}{dt} = \mu \frac{f}{\sigma_{vis}}$$

In Fig 2.17 the main CMS luminometers are highlighted.

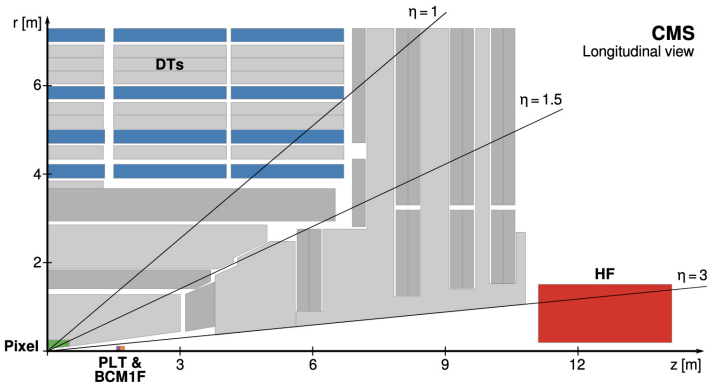


Figure 2.17: Longitudinal section of CMS where the main luminometers are highlighted. It is possible to see the PLT, BCM1F, DTs, and HF whose characteristics will be discussed in the text [22]

Pixel detector and DTs are used for the offline luminosity measurement and use the PCC; the results of the two subdetector are then compared to estimate the systematic uncertainties on the luminosity.
The online monitoring of the luminosity is left to the Pixel Luminosity Telescope (PLT), the Fast Beam Conditions Monitor (BCM1F) and the Hadron Forward calorimenter. PLT are a total of 48 silicon sensors arranged into 16 "telescopes", eight at either end of CMS outside the pixel endcap. It measures the luminosity using the zero counting algorithm where, the overall mean rate is estimated using the fraction of events with no triple coincidence.
BCM1F consists of a total of 24 sensors mounted on the same carriage as the PLT and uses single-crystal diamond sensors. The peculiarity of this detector is the very fast read out response of 6.25 ns, and is used for the online measurement of the luminosity with the zero counting algorithm. Last but not least there is the Hadron Forward (HF) that are placed in the most forward η and, thanks a dedicate readout system, it can substain the full 40MHz flux of LHC. As the PLT and the BCM1F also the HF uses an algorithm based on the zero counting concept.

2.4 Trigger system

One of the most difficult challenge of the detectors of this time is the capacitance of distinguish and store events useful for the analysis from the others. This is not simple. At LHC every 25 ns the beams collides and in a framework as CMS experiment, about 20 M events per second were produced but just around 100 per second can be saved, thus a fast mechanism capable to select good events is of vital importance for every kind of analysis. This task is assigned to complex trigger systems.
The CMS experiment has a trigger system able to select and store the information of 100 events per second. Two trigger levels are needed:

- level-1 trigger; it is a hardware trigger with the task of performing a first selection of the events that it has to be efficient and fast as much as possible. It takes the information from the fastest detectors (ECAL and RPCs) and uses algorithms with no iteration.
  This trigger reduces the events rate from 20 MHz to 100 kHz.

- High Level Trigger (HLT); it is a software trigger, slower than the first one, which works offline. It uses all the possible information of the event and reduces the event rate at 300 Hz.

To optimize the trigger efficiency, the Global Muon Trigger (GMT) is used. For each bunch crossing it takes all the principal information of the event and chooses at least four muon candidate caming from DT and barrel RPCs and an equal number from CSCs and endcup RPCs. From the selected tracks other requirements were applied and, at the end, only four muons are selected.

2.5 CMS upgrade PhaseII

The next milestone of the next years will be the launch of the High Luminosity LHC, HL-LHC [23]. The goal of the project is to increase the instantaneous luminosity by a factor of 10 beyond the LHC design values. From the end of 2027, the HL-LHC will collect 10 times more data than the LHC would expect to collect by the end of 2024. This helps detect very rare processes and improve SM accuracy measurements. In addition, CMS and other LHC experiments have begun ambitious detector upgrades to take full advantage of the increased amount of data. Displaced vertices will benefit from the programmed upgrades of the ATLAS and CMS detectors for HL-LHC. They will increase the coverage of the forward area, help to have better temporal and spatial resolution, and they will add new features as track triggers [24]. The following brief description can be found in reference [25–29]. At the HL-LHC, the instantaneous luminosity is about 5 times higher than at the LHC, causing 140-200 pp collisions at each beam crossing. In this harsh environment, object reconstruction and particle identification is more complicated due to tracks coming from nearby vertices, the detector upgrade is essential. Therefore, the coverage of $|\eta|$ and timing and spatial resolution are crucial for separating different events from each other. The inner CMS tracker has four additional cylindrical layers that enclose the area with $|z|$. Cover $|z| < 200$mm, first layer placed at 28 mm, up to 12 end cap disks including $|\eta|$ improve coverage from current 2.4 to almost 4 (Fig 2.18). Additional modules will be installed on the external CMS tracker.

A module called p_T will be able to correlate the signals from the sensors and identify pairs of hits (called "stubs") that match particles above p_T = 2 GeV. In addition, these stubs are provided as input to the L1 trigger, which enables the L1 trigger to make use of track-finding. An additional muon chamber is installed on the endcap. They are included in the L1 muon trigger. An additional hit of the algorithm-improved endcap allows for higher trigger efficiency on displaced muon tracks, regardless of the HL-LHC's high occupancy envi-

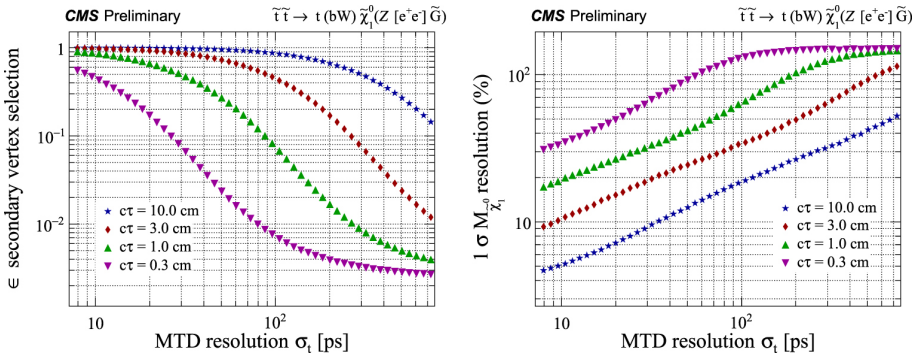


Figure 2.18: Efficiency (left) and mass resolution (right) as a function of the timing resolution of the MTD for reconstruction of the $\tilde{\chi}_0^1$ mass in the SUSY GMSB example of $\tilde{\chi}_0^1 \to \tilde{G}e^+e^-$, with mass of $\tilde{\chi}_0^1 = 700\,\text{GeV}$, considering events with a separation of primary and secondary vertices by more than 3 in both space and time [24].

ronment. The CMS MIP Timing Detector (MTD) consists of a cylinder and endcap parts consisting of a single layer module placed between the tracker and the calorimeter, covering up to $|\eta| \sim 3$. MTD improves reconstruction by collecting time information on charged particles and combining tracking and time measurement. This design provides a time resolution of approximately 30-40 ps for a 0.7 GeV p_T threshold at the start of the HL-LHC, and the time resolution in the barrel drops to 50-60 ps at the end. The introduction of a timing detector helps reduce the pile-up effect. The CMS detector improvements for the HL-LHC will globally enhance the lepton reconstruction acceptance and efficiency thanks to the wider coverage in the forward regions and the new timing and trigger features. Preliminary results on the offline tracking performance over the full acceptance of the CMS tracker are excellent, with further improvements expected as the detector design and simulation algorithms are optimized. In Fig 2.19, the resolution of the transverse momentum and the transverse impact parameter for single muons with $p_\text{T} = 10\,\text{GeV}$ as a function of the pseudorapidity, both with the current detector and after the implementation of the HL-LHC upgrades, is shown.

In Fig 2.20 the resolution of the vertex position in the x, y, and z coordinates is shown as a function of the number of tracks associated to the vertex. The vertex position resolution is almost independent of the amount of PU in the event and the longitudinal resolution is only 50% worse than the transverse one, as expected given the pixel dimensions of the inner tracker modules.

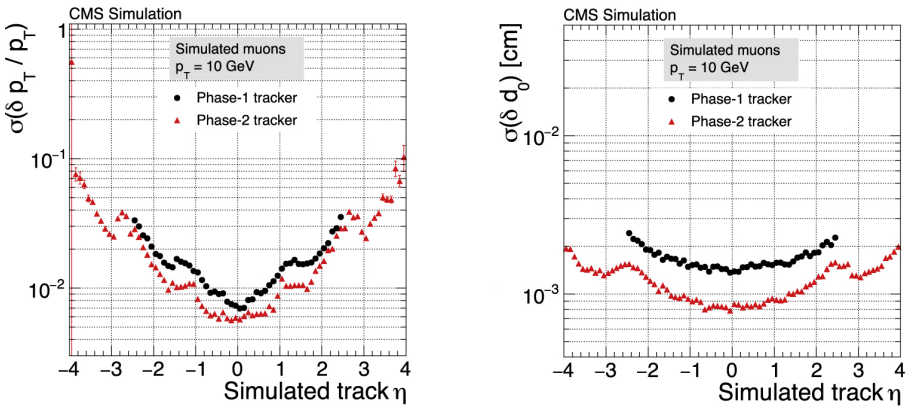


Figure 2.19: Relative resolution of the transverse momentum (left) and transverse impact parameter (right) as a function of the pseudorapidity for the current (black dots) and the upgraded (red triangles) CMS tracker, using single isolated muons with a transverse momentum of 10 GeV [24].

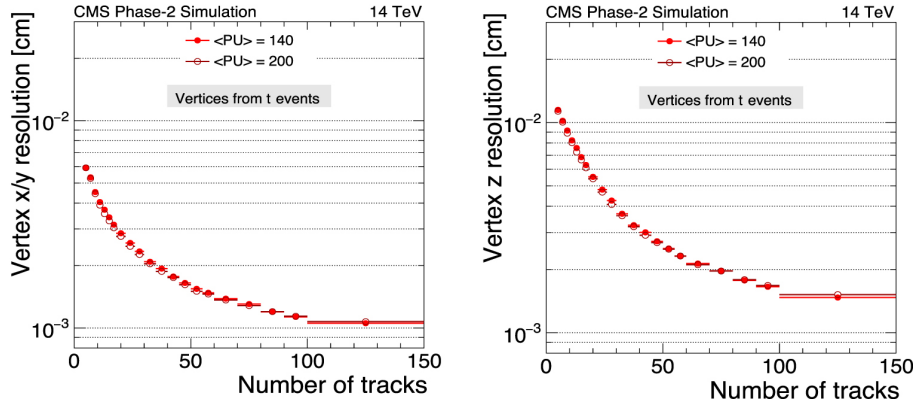


Figure 2.20: Relative resolution of the transverse momentum (left) and transverse impact parameter (right) as a function of the pseudorapidity for the current (black dots) and the upgraded (red triangles) CMS tracker, using single isolated muons with a transverse momentum of 10 GeV [24].

The long-lived HNL search will significantly benefit from most of those upgrades. Moreover, the timing records may be used as a further discriminating variable among the HNL signal and backgrounds. Considering the projected time precision of approximately 40 ps, the MTD detector could in principle be capable of locate the delayed leptons originated from long-lived HNL decays.

Chapter 3

Physics object reconstruction

The raw detector information is combined and used to reconstruct "physics objects", which are the input of all the data analyses. A global event reconstruction is performed to identify a few elementary objects: charged and neutral hadrons, electrons, photons, and muons. These are subsequently combined to reconstruct more complex objects such as semileptonic τ decays, jets, and missing transverse momentum.

3.1 Global event reconstruction

The high granularity of the tracker and ECAL detectors, the strong magnetic field to separate neutral and charged particles, the hermeticity of the HCAL and HF, and the excellent muon system resolution make the CMS detector ideally suited to identify and measure the individual final state particles. The particle flow (PF) algorithm [30], which uses the information of all detectors, is designed to exploit the redundant measurements from the CMS subsystems and reconstruct physics objects from raw detector data, performing a global event reconstruction.
A schema of the interaction of the particles in CMS is illustrated in Fig 3.1.

The trajectory of charged particles, or tracks, are reconstructed from the hits in the tracker systems, and matched to deposits in the ECAL only (electrons) or in the HCAL as well (charged hadrons). The absence of a track in front of a calorimetric deposit reveals the passage of a photon or a neutral hadrons. Finally, the presence of a track in the muon systems identifies the interaction of a muon. The combination of these redundant measurements has a positive impact on all physics objects, and the largest effects are observed in jets, tau leptons and missing transverse momentum.

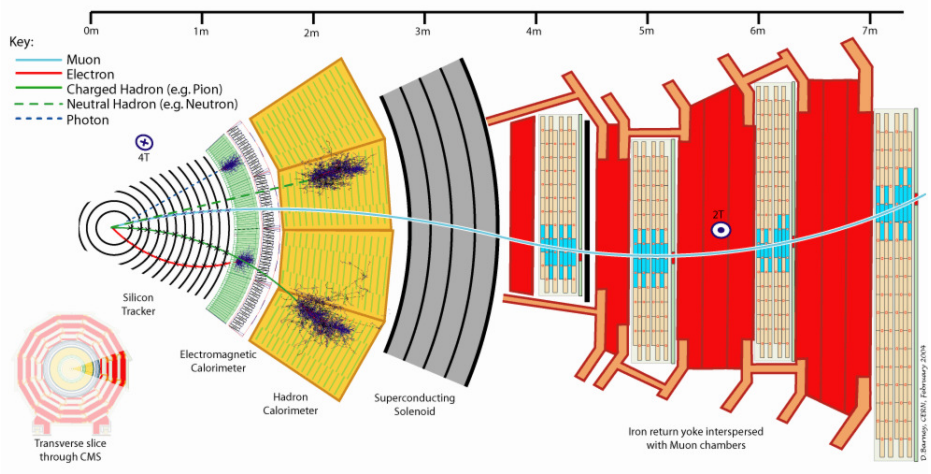


Figure 3.1: Schematic representation of different particle's signature in CMS detector

3.2 Muon reconstruction

Muons leave a very clean signature in the CMS detector thanks to their interactions in the muon detectors. As a consequence, muon tracks are reconstructed with dedicated algorithms that are independent from the iterative PF tracking discussed above, and are based on a Kalman filter method that accounts for the muon energy loss in the detector materials. In the standard CMS reconstruction, muons are reconstructed independently both in the tracker detector (*tracker track muon*) and in the muon system (*stand-alone muon track*); starting from those objects two kind of reconstruction are possible [31]

- Tracker muons reconstruction (*inside-out*): all tracker tracks with $p_T > 0.5$ GeV and a total momentum larger that 2.5 GeV, are considered tracker muons if the trajectory extrapolated to the muon system fits at least one muon segment.

- Global muons reconstruction (*outside-in*): for each standalone-muon track, a matching tracker track is found by comparing parameters of the two tracks propagated onto a common surface. At large momentum ($p_T > 200$ GeV), this technique improves the momentum resolution compared to the tracker only fit.

Thanks to the high reconstruction efficiency in both tracker and muon systems, about 99% of muons are reconstructed either as tracker or global muons, and those candidates that share the same inner tracks are merged into a single

object.

The global and tracker muon algorithms can be optimized depending on the analysis requirements. Three more muon objects can be distinguished:

- Soft muon: the selection requires the candidate to be a tracker muon with the additional requirement that a muon segment is matched in both x and y coordinates. This selection is optimized for muons with low p_T (< 10 GeV) and used for the B physics.
- Tight muon: the candidate must be a global muon with $\chi^2/d.o.f < 10$ and must have at least one muon chamber hit included in the global-muon track fit. In addition, his corresponding tracker track must match at least two muon station. With this selection, the rate of muons from decays in flight is significantly reduced at the price of a few percent loss in efficiency reconstruction.
  This alghoritm is used for W, Z and H physics.
- Particle flow muon: the muon has to be reconstructed both as a tracker muon and as global muon, and depending on the isolation[1] of the muons, the information of other detectors are taken into account. This algorithm is useful for the identification of muons within a jet.

The default algorithm for muon momentum assignment in CMS is called the "sigma switch" [31]. This algorithm chooses the best value between the momentum estimates given by the tracker-only fit and by the global fit. The global fit is chosen when both the fits give muon p_T above 200 GeV and the charge-to-momentum ratios agrees within two standard deviations from the tracker only measurement. In all the other case the tracker only fit is taken.
The muon transverse momentum resolution ranges between 1 to 6%, depending on the η coordinate, for muons with p_T < 100 GeV, and is better than 10% for central muons of p_T = 1TeV.

3.3 Electron reconstruction

Electrons are reconstructed by associating a track reconstructed in the silicon detector with a cluster of energy in the ECAL.

[1] This variable will be treated in the next section, it estimates how much a particle is far from the others

Tracker algorithms have to take into account the non-Gaussian energy loss, and clustering algorithms have to collect the bremsstrahlung photon energy deposits that can be located away from the electron interaction point in the ECAL. The electron reconstruction algorithm addresses these two effects with a dedicated tracking and an advanced energy clustering [32].
The latter are built by regrouping PF ECAL clusters in "superclusters". This procedure identifies a seed cluster and gathers together the energy deposits associated to bremsstrahlung photons. Preshower energy clusters in the endcaps are also taken into account in the procedure. The supercluster aggregation depends on the cluster E_T and exploits the correlation between their η and φ positions, preferring clusters that are spread along the direction because of the magnetic field.
Tracks are reconstructed with a Gaussian sum filter (GSF) method [32]. In contrast to the Kalman filter, the GSF method accounts for the large bremsstrahlung energy emission by approximating the radiated energy loss probability with a sum of Gaussian distributions. The GSF tracking is initiated, or "seeded", by to complementary algorithms. An ECAL-seeding procedure makes an estimate for the track position starting from the PF superclusters, while a tracker-seeding relies on the general charged particle iterative tracks and looks for a correspondence with a PF supercluster.
GSF tracks and PF superclusters are associated into an electron candidate if they satisfy some loose requirements on their qualities and matching. They are subsequently used to estimate the electron charge and its momentum, the latter being computed from a combination of GSF track curvature and supercluster total energy. To improve the resolution, electrons are classified in five categories depending on their quality and bremsstrahlung properties. The momentum resolution for electrons produced in Z boson decays ranges between 1.7 and 4.5% depending on the electron category and position in the detector.

3.4 Jet reconstruction

As quarks and gluons undergo an hadronization process, the estimation of their initial momentum requires the recollection and measurement of the hadronization products. Jets are thus reconstructed by clustering the PF candidates with the anti-k_T algorithm [33] [34].
The algorithm iteratively combines PF candidates that are close to each other according to a metric, that is defined to produce jets of an approximate conic shape clustered around the hardest particles in the event. The size of the jet

cone is determined by the distance parameter R at which the algorithm is operated.
The jet four momentum is computed as the vector sum of the clustered PF candidates four momenta, and a set of corrections are applied to calibrate the jet response using the information of generated particles in a simulation [34]. These corrections of the jet energy scale take into account the contribution from pileup in the event, nonlinearities in the detector response to hadrons, and residual differences between the data and the simulation used for the method.

3.5 Missing transverse momentum and energy reconstruction

The existence of undetected final state particles such as neutrinos can be indirectly inferred from the imbalance of the total transverse momentum vector sum. The negative projection of this vector onto the transverse plane is denoted as missing transverse momentum (p_T^{miss}).
The p_T^{miss} vector is reconstructed with the PF algorithm [35] as the negative of the vectorial sum of the transverse momenta of the PF candidates reconstructed in the event. Since inefficiencies of the tracking algorithm, minimal thresholds in the calorimeter energy estimation, and non linearities of the energy response of the calorimeters for hadronic particles, can introduce a bias in the p_T^{miss} determination, a correction is applied by propagating to the p_T^{miss} sum the jet energy corrections.
In particular, the corrected p_T^{miss} vector is estimated as:

$$p_T^{miss,corr} = p_T^{miss} - \sum_{jet}(p_T^{corr} - p_T)$$

3.6 Dataset definition

All the data recorded from CMS are funnelled into streams, integrated in the High Level Trigger (HLT) menu, and further organised in a hierarchical structure of primary datasets and secondary datasets/dedicated skims. A dataset is defined based on the final state particles reconstructed from the CMS detector. According to the results of the HLT selection, datesets are divided into primary dataset of non-exclusive events. With the goal of easing the data access and and facilitating the event processing, skims can be applied to primary dataset

to obtain secondary dataset. In the analyses chapters will be specified the sets of primary dataset used and how the exclusive set of events was selected.

3.7 Multivariate Variate analysis techniques

For the purpose of this thesis several machine learning techniques will be used. In this section a complete overview of these methods is given. Before starting, is important to clarify and understand why the use of machine learning is becoming more and more crucial in all the various fields: the goal of machine learning is to understand the hidden structure of a given dataset and create a model that can best describe it. Based on the user needs, machine learning can be used:

- to classify events (classification) $\rightarrow$ starting from a set of input features, the network divides the event in categories as signal-like and background-like
- to regress an observable (regression) $\rightarrow$ the network predicts a numeric value from a given input

Two types of machine learning approaches can be distinguished:

- supervised learning: in which the desire input and output data are labeled from the user. The network is trained with inputs that are labeled with their desired outputs. The purpose of this method is that the network will find a function that adjusts its own output to the expected output by minimising a given figure of merit (details will be given in the next sections). A clear example is a classification problem where we want to distinguish signal and background events: we can assign to the signal events (known from the simulation) a label equal to 1 while in the background a label equal to 0, give to the network some input features (as the kinematic of the event) and then ask the network to understand the function that can disentangle signal from background
- unsupervised learning: in this method the data are not labeled from humans and the network is left free to discover hidden patterns within a dataset. This kind of approach usually requires lots of data to have a meaningful result

These two different approaches have in common the way in which the machine learning works: we need a subset of our entire dataset to train the network, in more simple words to ask the network to understand some behavior from the inputs; and another subset of the dataset that is used for testing the performances of the training.
The performances of the network are evaluate with a figure of merit called the Area Under the ROC (Reciver Operating Characteristic) curve.
The ROC is a probability curve of the True Positive Rate (TPR) versus False Positive Rate (FPR). If we take a classification problem, we can define TPR the rate of positive response when the correct answer is positive, in easier words the probability of associating 1 to the class-1 events; the FPR is instead the probability of associating 1 to a class-0 event. An illustration of the ROC is shown in Fig 3.2.

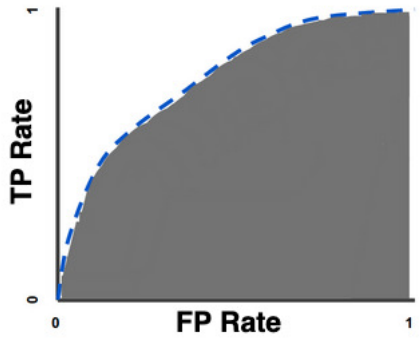


Figure 3.2: Illustration of the ROC curve as a probability curve of the True Positive Rate (TPR) versus False Positive Rate (FPR)

The area under the ROC curve tells how the model is capable of distinguishing between classes, the higher is the Area Under the Curve (AUC) the better is the model at predicting classes-0 events at 0 and classes-1 events at 1. If AUC = 1 it means that the classifier can always associate 0 to class-0 events (and viceversa), AUC = 0 means that the classifier always misidentify events and classes while AUC = 0.5 means that the classifier is right in the 50% of cases (random responce).
A good classifer thus must then have a AUC » 50% to trust its performances. The most important problem that may occur in the process of learning is the so called 'overtraining' problem. The overtraining happens when the network learns some peculiar features from the training sample and thus is not able to

make good predictions on an unseen dataset. To check if the issue is present in our network one can simply compare the performances of the training and testing sample (looking for instance at the Area Under the ROC), if they are the same this means that no overtrainig has occurred in the process.
The hyperparameters are all the adjustable parameters of the network, the tuning of the hyperparameters is extremely important to have good performances of the network. In this thesis two types of machine learning will be used: Boosted Decision Tree (BDT) and Deep Neural Network (DNN), in the coming subsections the characteristics of the two methods will be analysed.

3.7.1 Boosted Decision Tree Learning

Let's start our introduction talking about Decision Tree. A decision tree is a structure that is built by iteratively by asking questions to a partition of data and reaching a solution. To clarify the ideas, let's say that we want to distinguish apples from pears based on the width and the height of each fruit, using these inputs one can ask several questions to guess the fruit. A visual representation of this problem is given in Fig 3.3.

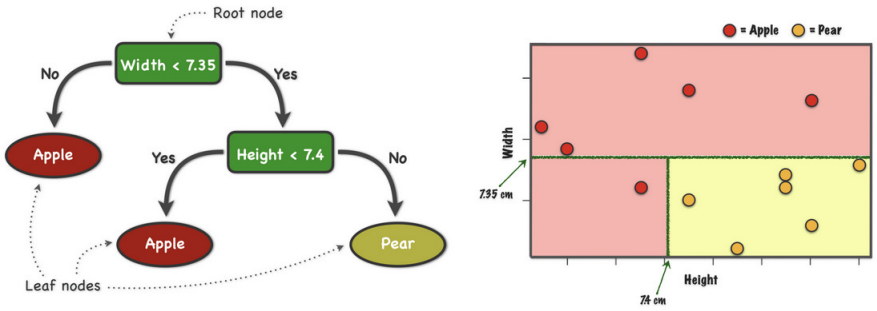


Figure 3.3: Schematic representation of a decision tree

Each question that we ask to the network is called root node, the answer of a root node is a leaf node. This division can also be visualize into a 2D plane that is divided in a number of boxes with each one corresponding to a classification (apple/pear) for an observation.
Switching back to physics, the idea is to create a decision tree that is able to make good predictions on both the training data an any unseen observation. To reach this goal is important to understand how many splits we need and how we can define a good split.

The definition of a good split (in the apple/pear example a good split is a split that takes all the apple in a box and all the pears into another one) is related to the concept that we want that the leaf nodes are more homogeneous than the parent node; if $P(x_i)$ is the relative frequency of observation i, the entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

and can be used to measure the purity of the leaf node. In principle, we could minimise only this function to find the best splitting, but we could end up in situation with 2 main boxes one very huge with almost all the dataset and another one very small but also very pure. To solve this situation we could instead maximise the difference of the entropy between the parent and the leaf node, such a feature goes under the name of 'information gain'. The splitting procedure is stopped when there is no more improvement in the information gain.
To make this procedure stronger we could use an ensemble of decision trees - boosted decision tree. All the trees are connected in series: the output of each tree is weighted and goes as input of another tree, then final goal is to minimise a function that takes the information from the tree ensamble. Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate.
Among all the possible hyperparameters for a boosted decision tree we could cite the most important here:

- Max depth: the maximum depth of the tree. The higher is this the max depth the more complex the model is
- Ntrees: the number of decision tree to be used in the gradient-boosted decision tree
- Shrinkage: the contribution of each tree to the final algorithm can be weighted to slow down the learning by the algorithm. This weighting is called a shrinkage or a learning rate

3.7.2 Deep learning

Generally speaking, deep learning is an algorithm that given a X as input tries to find a non random Y output. To find the relation X $\rightarrow$ Y, deep learning uses the so called neural network that are represented in Fig 3.4

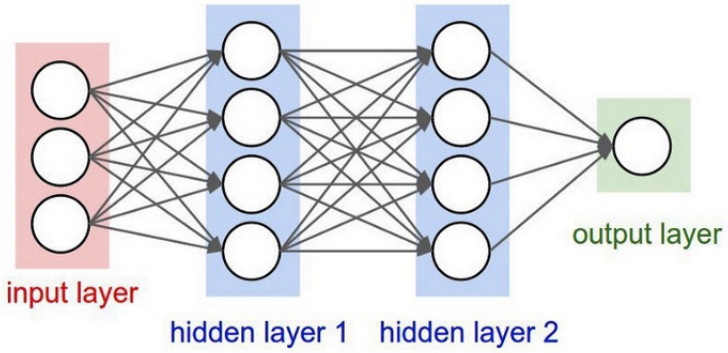


Figure 3.4: Schematic representation of a deep neural network

The structure of a neural network is composed by input, inner and output layers and each layer is composed by several nodes. Input layers take the input observables from the dataset and pass this information to the inner layers following a math relation shown in Fig 3.5

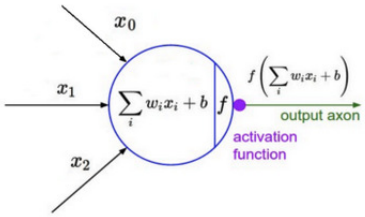


Figure 3.5

where w_i are the weights and b is the bias parameter, that function is the function that actually do the learning and is called activation function.
There could be several types of activation functions given the problem that we

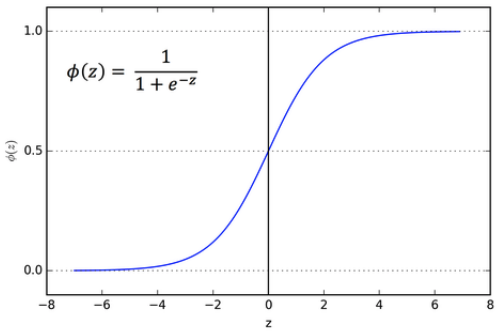


Figure 3.6: Sigmoid function

want to solve with machine learning. The most common and used activation function for a classification problem is the sigmoid function (Fig 3.6)

It has a S slope and gives output in 0 to 1 range. After the network passes its inputs all the way to the output, a function called loss function evaluates how good the prediction is. Given the needs of the human, several loss functions can be defined, the most common loss function for a classification problem is the mean square error:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

where $\hat{Y}_i$ represent the prediction and Y_i is the expected output. The goal of the network is then to minimise this loss function by adjusting per each iteration the weights and the bias parameter.
The 'deep' part of the network consist in adding several inner layer to make the model more complex.
The most common hyperparameters are:

- Batch size: the number of samples processed before the model is updated
- Epochs: the number of complete passes through the training dataset
- Learning rate: how much to change the model in response to the estimated error each time the model weights are updated (same as the shrinkage for BDTs)

In the context of this thesis, the DNN will be used for classification and for prediction, with a customised loss function, more detailed will be in Sec 6.8.2.

Chapter 4

HH analysis

4.1 Introduction

The discovery of the Higgs boson by the ATLAS and CMS collaborations [36, 37] was the first experimental proof of the predicted mechanism of electroweak symmetry breaking [38–43]. From 2012 on the effort of the scientific community focused on the determination of the Higgs properties.
The measurement of the Higgs boson self-coupling would provide an independent test of the SM and shed light on the Higgs mechanism by verifying the shape and intensity of the Higgs scalar field potential [44]. A measurement of the Higgs boson trilinear coupling can be obtained by measuring the production of pairs of Higgs bosons (HH) at the LHC.
In the context of HH searches, due to the very low cross section of the process one should find a trade-off between having a reasonably high branching ratio (this can be translated into have enough events to set reasonable limits) and a very high signal purity (this can be translated into suppress enough the background and enhance the signal).
Studies of pair production of Higgs bosons allow us to probe different regions of the anomalous couplings space and of the resonant invariant mass spectrum. A combination of different channels is therefore needed to obtain the best possible sensitivity for the HH production.
Depending on the HH decay channels, a wide variety of such channels will be obtained. The four-lepton (4ℓ) decay channel is the rarest observed so far at the LHC but it has the largest signal-to-background ratio. The analysis presented in this thesis consists of the search for HH pairs where one Higgs decays to 4ℓ and the other to a pair of b jets and is the first analysis in the CMS collaboration that investigates this particular phase space. The choice of this final state takes advantage of the high branching fraction of the $b\bar{b}$ decay channel, partially compensating for the large yield reduction from the tiny branching fraction of the 4ℓ decay channel.

The SM production cross section has been computed at NNLO in quantum chromodynamics (QCD), including NNLL corrections and finite top quark mass effects at NLO. Its value is $\sigma_{HH} = 31.05^{+2.2\%}_{-5.0\%}(\text{QCD scale}) \pm 2.1\%(\text{PDF}) \pm 2.1\%(\alpha_S) \pm 2.7\%(\text{top})\, fb$ in proton-proton (pp) collisions at 13 TeVfor a Higgs boson mass of 125 GeV [45–48].
The exploration of the Higgs boson pair production in $\bar{b}b4l$ decay channel requires to identify and reconstruct several different types of final state objects and to use them for the selection of signal-like events.
This in turn requires the reconstruction of the $H \rightarrow \bar{b}b$ and $H \rightarrow ZZ \rightarrow 4l$ decays and the usage of their properties to identify the specific signature of signal events and to reject background ones.
The data used are those collected by the CMS experiment in the full Run 2 data-taking and corresponding to an integrated luminosity of 137 fb^{-1} .
The theoretical $\sigma \times BR$ of the process studied is about 10 ab; this means that at the Run 2 LHC's energy and luminosity a direct measurement of the experimental cross section is not feasible, thus the goal of the analysis is to set an upper limit on this observable and to constrain the Higgs boson trilinear coupling, by implementing a cut-based selection and then using a Boosted Decision Tree to improve the results.

4.1.1 Publication plans and my contribution

I led this analysis from the beginning and I touched all the aspects: from the simulation of the signal process to the evaluation of the upper limit on different observables. In the following sections I'll describe in full detail the flow of the analysis.
This analysis is currently approved and public and is waiting for submission to the JHEP journal; it will be also described in details in the H-HH combination paper that has been submitted to Nature and in the HH combination standalone paper.

4.2 Dataset

This analysis uses a data sample recorded by the CMS experiment during 2016, 2017 and 2018, corresponding to a total integrated luminosity of 137 fb^{-1}.

Run-range	Dataset	Integrated luminosity
273150-275376	/DoubleMuon/Run2016B-17Jul2018-ver2-v1/MINIAOD /DoubleEG/Run2016B-17Jul2018-ver2-v1/MINIAOD /MuonEG/Run2016B-17Jul2018-ver2-v1/MINIAOD /SingleElectron/Run2016B-17Jul2018-ver2-v1/MINIAOD /SingleMuon/Run2016B-17Jul2018-ver2-v1/MINIAOD	5.892 fb^{-1}
275656-276283	/DoubleMuon/Run2016C-17Jul2018-v1/MINIAOD /DoubleEG/Run2016C-17Jul2018-v1/MINIAOD /MuonEG/Run2016C-17Jul2018-v1/MINIAOD /SingleElectron/Run2016C-17Jul2018-v1/MINIAOD /SingleMuon/Run2016C-17Jul2018-v1/MINIAOD	2.646 fb^{-1}
276315-276811	/DoubleMuon/Run2016D-17Jul2018-v1/MINIAOD /DoubleEG/Run2016D-17Jul2018-v1/MINIAOD /MuonEG/Run2016D-17Jul2018-v1/MINIAOD /SingleElectron/Run2016D-17Jul2018-v1/MINIAOD /SingleMuon/Run2016D-17Jul2018-v1/MINIAOD	4.353 fb^{-1}
276831-277420	/DoubleMuon/Run2016E-17Jul2018-v1/MINIAOD /DoubleEG/Run2016E-17Jul2018-v1/MINIAOD /MuonEG/Run2016E-17Jul2018-v1/MINIAOD /SingleElectron/Run2016E-17Jul2018-v1/MINIAOD /SingleMuon/Run2016E-17Jul2018-v1/MINIAOD	4.117 fb^{-1}
277932-278808	/DoubleMuon/Run2016F-17Jul2018-v1/MINIAOD /DoubleEG/Run2016F-17Jul2018-v1/MINIAOD /MuonEG/Run2016F-17Jul2018-v1/MINIAOD /SingleElectron/Run2016F-17Jul2018-v1/MINIAOD /SingleMuon/Run2016F-17Jul2018-v1/MINIAOD	3.186 fb^{-1}
278820-280385	/DoubleMuon/Run2016G-17Jul2018-v1/MINIAOD /DoubleEG/Run2016G-17Jul2018-v1/MINIAOD /MuonEG/Run2016G-17Jul2018-v1/MINIAOD /SingleElectron/Run2016G-17Jul2018-v1/MINIAOD /SingleMuon/Run2016G-17Jul2018-v1/MINIAOD	7.721 fb^{-1}
281207-284068	/DoubleMuon/Run2016H-17Jul2018-v1/MINIAOD /DoubleEG/Run2016H-17Jul2018-v1/MINIAOD /MuonEG/Run2016H-17Jul2018-v1/MINIAOD /SingleElectron/Run2016H-17Jul2018-v1/MINIAOD /SingleMuon/Run2016H-17Jul2018-v1/MINIAOD	8.857 fb^{-1}

Table 4.1: 2016 dataset used.

The 2016 analysis relies on five different primary datasets (PDs), *DoubleEG*, *DoubleMuon*, *MuEG*, *SingleElectron*, and *SingleMuon*, each of which combines a certain collection o f high-level trigger (HLT) paths. In order to avoid duplicate events from different PDs, events are taken as follow:

- from DoubleEG, if they pass the diEle or triEle triggers,
- from DoubleMuon, if they pass the diMuon or triMuon triggers and fail the diEle and triEle triggers,
- from MuEG, if they pass the MuEle or MuDiEle or DiMuEle triggers and fail the diEle, triEle, diMuon, and triMuon triggers,

- from SingleElectron, if they pass the singleElectron trigger and fail all the above triggers,
- from SingleMuon, if they pass the singleMuon trigger and fail all the above triggers.

The HLT paths used for 2016 collisions data are listed with the associated primary datasets in Table 4.2.

HLT path	primary dataset
HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	DoubleEG
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL	DoubleMuon
HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL	DoubleMuon
HLT_TripleMu_12_10_5	DoubleMuon
HLT_Mu8_TrkIsoVVL_Ele17_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu17_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL	MuonEG
HLT_Ele25_eta2p1_WPTight	SingleElectron
HLT_Ele27_WPTight	SingleElectron
HLT_Ele27_eta2p1_WPLoose_Gsf	SingleElectron
HLT_IsoMu20 OR HLT_IsoTkMu20	SingleMuon
HLT_IsoMu22 OR HLT_IsoTkMu22	SingleMuon

Table 4.2: Trigger paths used in 2016 collision data.

The 2017 analysis relies on five different primary datasets (PDs), *DoubleEG*, *DoubleMuon*, *MuEG*, *SingleElectron*, and *SingleMuon*, each of which combines a certain collection o f high-level trigger (HLT) paths. In order to avoid duplicate events from different PDs, events are taken as follow:

- from DoubleEG, if they pass the diEle or triEle triggers,
- from DoubleMuon, if they pass the diMuon or triMuon triggers and fail the diEle and triEle triggers,

Run-range	Dataset	Integrated luminosity
297046-299329	/DoubleMuon/Run2017B-31Mar2018-v1/MINIAOD /DoubleEG/Run2017B-31Mar2018-v1/MINIAOD /MuonEG/Run2017B-31Mar2018-v1/MINIAOD /SingleElectron/Run2017B-31Mar2018-v1/MINIAOD /SingleMuon/Run2017B-31Mar2018-v1/MINIAOD	4.792 fb^{-1}
299368-300676	/DoubleMuon/Run2017C-31Mar2018-v1/MINIAOD /DoubleEG/Run2017C-31Mar2018-v1/MINIAOD /MuonEG/Run2017C-31Mar2018-v1/MINIAOD /SingleElectron/Run2017C-31Mar2018-v1/MINIAOD /SingleMuon/Run2017C-31Mar2018-v1/MINIAOD	9.755 fb^{-1}
302030-303434	/DoubleMuon/Run2017D-31Mar2018-v1/MINIAOD /DoubleEG/Run2017D-31Mar2018-v1/MINIAOD /MuonEG/Run2017D-31Mar2018-v1/MINIAOD /SingleElectron/Run2017D-31Mar2018-v1/MINIAOD /SingleMuon/Run2017D-31Mar2018-v1/MINIAOD	4.319 fb^{-1}
303824-304797	/DoubleMuon/Run2017E-31Mar2018-v1/MINIAOD /DoubleEG/Run2017E-31Mar2018-v1/MINIAOD /MuonEG/Run2017E-31Mar2018-v1/MINIAOD /SingleElectron/Run2017E-31Mar2018-v1/MINIAOD /SingleMuon/Run2017E-31Mar2018-v1/MINIAOD	9.424 fb^{-1}
305040-306462	/DoubleMuon/Run2017F-31Mar2018-v1/MINIAOD /DoubleEG/Run2017F-31Mar2018-v1/MINIAOD /MuonEG/Run2017F-31Mar2018-v1/MINIAOD /SingleElectron/Run2017F-31Mar2018-v1/MINIAOD /SingleMuon/Run2017F-31Mar2018-v1/MINIAOD	13.50 fb^{-1}

Table 4.3: 2017 dataset used.

- from MuEG, if they pass the MuEle or MuDiEle or DiMuEle triggers and fail the diEle, triEle, diMuon, and triMuon triggers,
- from SingleElectron, if they pass the singleElectron trigger and fail all the above triggers,
- from SingleMuon, if they pass the singleMuon trigger and fail all the above triggers.

The HLT paths used for 2017 collisions data are listed with the associated primary datasets in Table 4.4.

The data samples used in 2018 are listed in Table 4.5.

The 2018 analysis relies on four different primary datasets (PDs), *DoubleMuon*, *MuEG*, *EGamma*, and *SingleMuon*, each of which combines a certain collection of high-level trigger (HLT) paths. In order to avoid duplicate events from different PDs, events are taken as follow:

HLT path	primary dataset
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	DoubleMuon
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8	DoubleMuon
HLT_TripleMu_12_10_5	DoubleMuon
HLT_TripleMu_10_5_5_D2	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL_DZ	MuonEG
HLT_Ele35_WPTight_Gsf_v	SingleElectron
HLT_Ele38_WPTight_Gsf_v	SingleElectron
HLT_Ele40_WPTight_Gsf_v	SingleElectron
HLT_IsoMu27	SingleMuon

Table 4.4: Trigger paths used in 2017 collision data.

Run-range	Dataset	Integrated luminosity
315252-316995	/DoubleMuon/Run2018A-17Sep2018-v2/MINIAOD /MuonEG/Run2018A-17Sep2018-v1/MINIAOD /SingleMuon/Run2018A-17Sep2018-v2/MINIAOD /EGamma/Run2018A-17Sep2018-v2/MINIAOD	14.00 fb^{-1}
317080-319310	/DoubleMuon/Run2018B-17Sep2018-v1/MINIAOD /MuonEG/Run2018B-17Sep2018-v1/MINIAOD /SingleMuon/Run2018B-17Sep2018-v1/MINIAOD /EGamma/Run2018B-17Sep2018-v1/MINIAOD	7.10 fb^{-1}
319337-320065	/DoubleMuon/Run2018C-17Sep2018-v1/MINIAOD /MuonEG/Run2018C-17Sep2018-v1/MINIAOD /SingleMuon/Run2018C-17Sep2018-v1/MINIAOD /EGamma/Run2018C-17Sep2018-v1/MINIAOD	6.94 fb^{-1}
320673-325175	/DoubleMuon/Run2018D-PromptReco-v2/MINIAOD /MuonEG/Run2018D-PromptReco-v2/MINIAOD /SingleMuon/Run2018D-PromptReco-v2/MINIAOD /EGamma/Run2018D-PromptReco-v2/MINIAOD	31.93 fb^{-1}

Table 4.5: 2018 dataset used.

- from EGamma, if they pass the diEle or triEle or singleElectron triggers,
- from DoubleMuon, if they pass the diMuon or triMuon triggers and fail the diEle and triEle triggers,

- from MuEG, if they pass the MuEle or MuDiEle or DiMuEle triggers and fail the diEle, triEle, singleElectron, diMuon, and triMuon triggers,
- from SingleMuon, if they pass the singleMuon trigger and fail all the above triggers.

The HLT paths used for 2018 collisions data are listed in Table 4.6.

HLT path	primary dataset
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_v*	EGamma
HLT_DoubleEle25_CaloIdL_MW_v*	EGamma
HLT_Ele32_WPTight_Gsf_v*	EGamma
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8_v*	DoubleMuon
HLT_TripleMu_10_5_5_DZ_v*	DoubleMuon
HLT_TripleMu_12_10_5_v*	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*	MuEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	MuEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	MuEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v*	MuEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ_v*	MuEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL_DZ_v*	MuEG
HLT_IsoMu24_v*	SingleMuon

Table 4.6: Trigger paths used in 2018 collision data.

4.2.1 Simulation

Signal Samples

Signal samples of the SM gg $\rightarrow$HH $\rightarrow$ 4ℓprocess are generated at the next-to-leading-order (NLO) in QCD using Powheg for all the three years with an effective field theory description of the fermion loops. Showering of parton-level events is done using PYTHIA8.209 for all the three years, with no specific matching requirement. Samples are generated with the NNPDF 3.1 NNLO parton distribution functions (PDFs) [49]. In Table 4.7 are reported the SM signal samples used for all the three years.
Signal samples for Beyond Standard Model (BSM) hypotheses with different value of the trilinear Higgs boson couplings are reported in Table 4.8. These samples are generated privately with different values of k_λ.

year	σ	$\sigma \times BR$	MC sample
2016	31.05 fb	0.00448 fb	/SM-HH-NLO_POWHEG_GEN-SIM/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
2017	31.05 fb	0.00448 fb	/HH_bb4l_2017_NLO_152kPart2_gen-sim/ilmargje-MiniAOD-7c2f65f61e72ee224088fc41982d0d62/USER
2018	31.05 fb	0.00448 fb	/HH_bb4l_2018_NLO_152kPart2_gen-sim/ilmargje-processed_gen-sim-premix_DR_MiniAODstep3-0bd58594e6ade05f64e0c3a8301c3139/USER

Table 4.7: MC samples used for the SM HH signal process.

year	k_λ	σ	MC sample
2016	-1	131.9 fb	/HH_bb4l_2016_NLO_kLambda_-1_GEN-SIM/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
	0	70.38 fb	/HH_bb4l_2016_NLO_k_lambda_0/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
	2	13.81 fb	/HH_bb4l_2016_NLO_kLambda_2_GEN-SIM/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
	3	18.67 fb	/HH_bb4l_2016_NLO_kLambda_3_GEN-SIM/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
	5	94.82 fb	/HH_bb4l_2016_NLO_kLambda_5_GEN-SIM/ilmargje-MiniAOD-53f8667ba4b240d5eafd36e71bf34742/USER
2017	-1	131.9 fb	/HH_bb4l_2017_NLO_k_lambda_-1_GEN-SIM/ilmargje-MiniAOD-320a10544d366c5c2cbf9f153fc62ac6/USER
	0	70.38 fb	/HH_bb4l_2017_NLO_k_lambda_0_GEN-SIM/ilmargje-MiniAOD-320a10544d366c5c2cbf9f153fc62ac6/USER
	2	13.81 fb	/HH_bb4l_2017_NLO_k_lambda_2_GEN-SIM/ilmargje-MiniAOD-320a10544d366c5c2cbf9f153fc62ac6/USER
	3	18.67 fb	/HH_bb4l_2017_NLO_k_lambda_3_GEN-SIM/ilmargje-MiniAOD-320a10544d366c5c2cbf9f153fc62ac6/USER
	5	94.82 fb	/HH_bb4l_2017_NLO_k_lambda_5_GEN-SIM/ilmargje-MiniAOD-320a10544d366c5c2cbf9f153fc62ac6/USER
2018	-1	131.9 fb	/HH_bb4l_2018_NLO_k_lambda_-1_GEN-SIM/ilmargje-MiniAOD-0bd58594e6ade05f64e0c3a8301c3139/USER
	0	70.38 fb	/HH_bb4l_2018_NLO_k_lambda_0_GEN-SIM/ilmargje-MiniAOD-0bd58594e6ade05f64e0c3a8301c3139/USER
	2	13.81 fb	/HH_bb4l_2018_NLO_k_lambda_2_GEN-SIM/ilmargje-MiniAOD-0bd58594e6ade05f64e0c3a8301c3139/USER
	3	18.67 fb	/HH_bb4l_2018_NLO_k_lambda_3_GEN-SIM/ilmargje-MiniAOD-0bd58594e6ade05f64e0c3a8301c3139/USER
	5	94.82 fb	/HH_bb4l_2018_NLO_k_lambda_5_GEN-SIM/ilmargje-MiniAOD-0bd58594e6ade05f64e0c3a8301c3139/USER

Table 4.8: BSM HH samples used for BSM HH signal Hypotheses.

Background Samples

SM Higgs processes constitute a background for this analysis. Descriptions of the SM Higgs boson to ZZ production are obtained using the POWHEG V2 [50–52] generator for the main production modes: gluon fusion (H) including quark mass effects [53], vector boson fusion (VBF) [54], and associated production (WH, H, bbH and H [55]). In the case of WH and H the MINLO HVJ extension of POWHEG is used [56]. The description of the decay of the Higgs boson to four leptons is obtained using the JHUGEN generator [57]. In the case of WH, H and H, the Higgs boson is allowed to decay to H$\rightarrow\rightarrow 2\ell$2X such that 4-lepton events where two leptons originate from the decay of associated , W bosons or top quarks are also taken into account in the simulation. Showering of parton-level events is done using PYTHIA8.209, and in all cases matching is performed by allowing QCD emissions at all energies in the shower and vetoing them afterwards according to the POWHEG internal scale. All samples are generated with the NNPDF 3.1 NLO parton distribution functions (PDFs) [49].

Production of ZZ via quark-antiquark annihilation is generated at next-to-leading order (NLO) using POWHEG V2 [58] and PYTHIA8, with the same settings as for the Higgs signal.

The$\rightarrow ZZ$ process is simulated at leading order (LO) with MCFM [59,60]. In order to match the$\rightarrow$ H $\rightarrow ZZ$ transverse momentum spectra predicted by POWHEG at NLO, the showering for MCFM samples is performed with dif-

ferent PYTHIA8 settings, allowing only emissions up to the parton-level scale ("wimpy" shower).

Table 4.9 summarizes these MC simulation datasets used for this analysis.

Process	Dataset Name	$\sigma \times BR(\times\epsilon_{\text{filter}})$
$gg \to H \to\to 4\ell$	/GluGluHToZZTo4L_M125_13TeV_powheg2_JHUgenV709_pythia8/[1]	12.18 fb
$qq \to Hqq \to qq \to 4\ell qq$	/VBF_HToZZTo4L_M125_13TeV_powheg2_JHUgenV709_pythia8/[1]	1.044 fb
$q\bar{q} \to W^{+}H \to W^{+} \to 4\ell + X$	/WplusH_HToZZTo4L_M125_13TeV_powheg2-minlo-HWJ_JHUgenV709_pythia8/[1]	0.232 fb
$q\bar{q} \to W^{-}H \to W^{-} \to 4\ell + X$	/WminusH_HToZZTo4L_M125_13TeV_powheg2-minlo-HWJ_JHUgenV709_pythia8/[1]	0.147 fb
$q\bar{q} \to H \to\to 4\ell + X$	/ZH_HToZZ_4LFilter_M125_13TeV_powheg2-minlo-HZJ_JHUgenV709_pythia8/[1]	0.668 fb
$gg \to ttH \to tt \to 4\ell + X$	/ttH_HToZZ_4LFilter_M125_13TeV_powheg_JHUgenV709_pythia8/[1]	0.393 fb
$gg \to bbH \to bb \to 4\ell + X$	/bbH_ToZZTo4L_M125_13TeV_JHUGenV7011_pythia8/[1]	0.135 fb
$qq \to\to 4\ell$	/ZZTo4L_13TeV_powheg_pythia8/[1]	1.256 pb
$gg \to\to 4e$	/GluGluToContinToZZTo4e_13TeV_MCFM701/[1]	0.00159 pb
$gg \to\to 4\mu$	/GluGluToContinToZZTo4mu_13TeV_MCFM701/[1]	0.00159 pb
$gg \to\to 4\tau$	/GluGluToContinToZZTo4tau_13TeV_MCFM701/[1]	0.00159 pb
$gg \to\to 2e2\mu$	/GluGluToContinToZZTo2e2mu_13TeV_MCFM701/[1]	0.00319 pb
$gg \to\to 2e2\tau$	/GluGluToContinToZZTo2e2tau_13TeV_MCFM701/[1]	0.00319 pb
$gg \to\to 2\mu 2\tau$	/GluGluToContinToZZTo2mu2tau_13TeV_MCFM701/[1]	0.00319 pb
$Z \to \ell\ell$ + jets	/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/[1]-v1	5765.4 pb
WWZ	/WWZ_TuneCP5_13TeV-amcatnlo-pythia8/[1]_ext1-v2	0.1651 pb
WZZ	/WZZ_TuneCP5_13TeV-amcatnlo-pythia8/[1]_ext1-v2	0.05565 pb
ZZZ	/ZZZ_TuneCP5_13TeV-amcatnlo-pythia8/[1]_ext1-v2	0.01398 pb
$TTZ(\ell\ell)$	/TTZToLL_M-1to10_TuneCP5_13TeV-amcatnlo-pythia8/[1]-v1	0.04695 pb
$TTZ(jets)$	/ttZJets_TuneCP5_13TeV_madgraphMLM_pythia8/[1]_ext1-v2	0.259 pb
$TTW(\ell\nu)$	/TTWJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8/[1]_ext1-v2	0.2149 pb

[1] RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic for 2016
[1] RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14 for 2017
[1] RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15 for 2018

Table 4.9: Background Monte Carlo samples and cross sections.

Pileup Reweighting

For each year, corresponding simulation samples are reweighted to match the pileup (PU) distribution in data. An example of reweighting procedure for 2018 data is shown in Fig. 4.1.

4.3 Object selection

The analysis targets the final state with $HH \to b\bar{b}4l$ where $l = \mu/e$; in this section I'll explain the selection used to identify leptons and jets. Selection criteria are inspired by the suppression of the QCD background (mainly hadronic jets that leads to non-isoleted leptons in the detector volume) and by the physics of the process; in this case, giving the mean life time of the SM Higgs: $1.5610^{22}s$, we are dealing with two Higgses that decay in the primary vertex and thus the product of the decay are expected to be isolated with respect to the other tracks. This considerations are reflected in the selection criteria below.

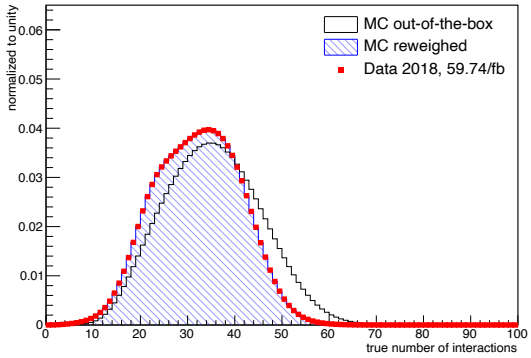


Figure 4.1: Distribution of pileup in 2018 Data and MC, before and after the application of pileup weights.

4.3.1 Electrons

To reject the majority of QCD background while preserving the highest possible efficiency, electrons are preselected using loose cuts on the track-cluster matching observables.
As explained in Chapter 3, the electrons are reconstructed both in the tracker volume and in the ECAL calorimeter. The track-cluster matching is the procedure that uses observable as the fraction of energy in the calorimeter (p_{in}) and the momentum extrapolated to the surface of the ECAL from the track at the exit of the tracker (p_{out}) and the η / φ positions extrapolated from both subdetectors to associate the electrons clusters in the calorimeter with the most probable tracks in the tracker [61].Requirements on the transverse momentum (p_T) p_T > 7 GeV, on $|\eta| < 2.5$ and $|d_{xy}| < 0.5cm$, $|d_z| < 1cm$, which are the spacial distance in the xy (z) plane between the primary vertex and the track of the electron, are further applied. Electrons selected with the criteria above are called loose electrons.
Loose electrons are isolated using a Gradient Boosted Decision Tree (GBDT) that takes as inputs observables fromthe electromagnetic cluster, the matching between the cluster and the electron track, observables based exclusively on tracking measurements as well as particle-flow (PF) isolation sums. The full list is in Table 4.10.

The classifier was trained on Drell-Yan plus jets MC sample in the context of the H $\rightarrow ZZ \rightarrow 4\ell$ analysis [62].

observable type	observable name
6*cluster shape	RMS of the energy-crystal number spectrum along η and φ; $\sigma_{i\eta i\eta}$, $\sigma_{i\varphi i\varphi}$ super cluster width along η and φ 'ratio of the hadronic energy behind the electron supercluster to the supercluster energy, H/E circularity $(E_{5\times5} - E_{5\times1})/E_{5\times5}$ sum of the seed and adjacent crystal over the super cluster energy R_9 for endcap traing bins: energy fraction in pre-shower E_{PS}/E_{raw}
2*track-cluster matching	energy-momentum agreement E_{tot}/p_{in}, E_{ele}/p_{out}, $1/E_{tot} - 1/p_{in}$ position matching $\Delta\eta_{in}$, $\Delta\varphi_{in}$, $\Delta\eta_{seed}$
5*tracking	fractional momentum loss $f_{brem} = 1 - p_{out}/p_{in}$ number of hits of the KF and GSF track N_{KF}, N_{GSF} reduced χ^2 of the KF and GSF track χ^2_{KF}, χ^2_{GSF} number of expected but missing inner hits probability transform of conversion vertex fit χ^2
3*isolation	particle-flow photon isolation sum particle-flow charged hadrons isolation sum particle-flow neutral hadrons isolation sum
1*For PU-resilience	mean energy density in the event: ρ

Table 4.10: Overview of input variables to the identification classifier.

To ensure that leptons are consistent with a common primary vertex (PV) we used the significance of the impact parameter to the event vertex, $|\text{SIP}_{\text{3D}} = \frac{\text{IP}}{\sigma_{\text{IP}}}|$, where IP is the lepton impact parameter in three dimensions at the point of closest approach with respect to the primary interaction vertex, and σ_{IP} the associated uncertainty. Therefore, a "primary lepton" is a lepton satisfying $|\text{SIP}_{\text{3D}}| < 4$.

4.3.2 Muons

For the analysis we defined loose muons, muons that satisfy: $p_T > 5$, $|\eta| < 2.4$, $|d_{xy}| < 0.5$ cm, $|d_z| < 1$ cm. Further more, muons have to be reconstructed by either the Global Muon or Tracker Muon algorithm. Standalone Muon tracks that are only reconstructed in the muon system are rejected.
A disclaimer is done also on the muon p_T: a loose muon that has p_T < 200 GeV is considered in the analysis if it also passes the PF muon ID (see chapter 3), a loose muon that has p_T > 200 GeV is considered in the analysis if it passes the PF muon ID or the Tracker High- ID, the definition of which is shown in Table 4.11.

To deal with situations when a single muon can be incorrectly reconstructed as two or more muons an additional “ghost-cleaning” step is performed:

- Tracker Muons that are not Global Muons are required to be arbitrated;
- if two muons are sharing 50% or more of their segments then the muon with lower quality is removed.

Table 4.11: The requirements for a muon to pass the Tracker High- ID. Note that these are equivalent to the Muon POG High- ID with the global track requirements removed.

Plain-text description	Technical description
Muon station matching	Muon is matched to segments in at least two muon stations **NB: this implies the muon is an arbitrated tracker muon.**
Good measurement	$\frac{}{\sigma} < 0.3$
Vertex compatibility $(x - y)$	$d_{xy} < 2$ mm
Vertex compatibility (z)	$d_z < 5$ mm
Pixel hits	At least one pixel hit
Tracker hits	Hits in at least six tracker layers

To subtract contributions of pileup events, a $\Delta\beta$ correction is applied, where $\Delta\beta = \frac{1}{2}\sum_{\text{PU}}^{\text{charged had.}}$ gives an estimate of the energy deposit of neutral particles (hadrons and photons) from pileup vertices. The relative isolation of muons is defined:

$$\text{RelPFiso} = \frac{\sum^{\text{charged had.}} + \max(\sum^{\text{neutral had.}} + \sum^{\text{photon}} - \Delta\beta, 0)}{\text{lepton}} \tag{4.1}$$

The isolation working point for muons was optimized in Ref. [4] and the working point was chosen to be RelPFiso$(\Delta R = 0.3) < 0.35$.
The cut on the impact paramenter is the same used for electrons: $|\text{SIP}_{\text{3D}}| < 4$

4.3.3 Photons for Final State Radiation recovery

Final State Radiation (FSR) is the process of the associated emission of a photon with a lepton. If this effect is neglected in the analysis, the reconstruction of the lepton momentum is biased since the energy loss for the photon emission is not considered. To account for these events, a strict selection of PF photons is done. The cuts of the FSR selection explained below had been optimised for the Higgs searches, more details can be found in Ref. [4]:

1. The preselection of PF photons is done by requiring $p_{T,\gamma} > 2$ GeV, $|\eta^{\gamma}| < 2.4$, and a relative particle-flow isolation smaller than 1.8. The

latter variable is computed using a cone of radius $R = 0.3$ constructed around the photon, and summing all the energy contributions of the hadrons/photons that falls in the cone and gives an indication on whether the photon is isolated or not (smaller the value highest the isolation).

2. Supercluster veto: PF photons that match with any electron passing both the loose ID and SIP cuts are removed. The matching is performed by directly associating the two PF candidates.

3. Photons are associated to the closest lepton in the event among all those pass both the loose ID and SIP cuts.

4. photons that do not satisfy the cuts $\Delta R(\gamma, l)/E^2_{T,\gamma} < 0.012$, and $\Delta R(\gamma, l) < 0.5$ are discarded.

5. If more than one photon is associated to the same lepton, the lowest-$\Delta R(\gamma, l)/E^2_{T,\gamma}$ is selected.

6. For each FSR photon that is selected, that photon is excluded from the isolation sum of all the leptons in the event that pass both the loose ID and SIP cuts. This concerns the photons that are in the isolation cone and outside the isolation veto of said leptons ($\Delta R < 0.4$ AND $\Delta R > 0.01$ for muons and $\Delta R < 0.4$ AND ($\eta^{\text{SC}} < 1.479$ OR $\Delta R > 0.08$) for electrons).

4.3.4 Jets

Jets are reconstructed by using the anti-k_T clustering algorithm out of particle-flow candidates, with a distance parameter $R = 0.4$, after rejecting the charged hadrons that are associated to a pileup primary vertex.
To reduce instrumental background, CMS community recommendation is to use a set of cuts on the number of neutral/charged hadron fraction that depends on the η of the jets, in particular for this analysis only the so called tight ID jets are used . The efficiency of the tight jets is around 98-99% in all the η regions with a background rejection of around 98% [63]. In addition, jets that come from pile up events are reconstructed and discarded using a multivariate classifier that is fed with input variables as the trajectories of tracks associated to the jets, the topology of the jet shape and the object multiplicity [63].
To target the b-jtes coming from the Higgs boson, jets are required to have $|\eta| < 2.4$, $p_T > 20$ GeV and to be isolated from identified leptons by a distance

of ΔR(jet,lepton/photon) > 0.3.
The energy resolution of the jets is degraded by the non linear response of the calorimeter; to face this problem a set of Jet Energy Correciton (JEC) is needed and applied[1].
The algorithm used to select jets coming from b quark is the Deep Combined Secondary Vertex [65].
The deep combined b-tagging algorithm is based on the reconstruction of the secondary decay vertex of the weakly decaying b hadrons.
The discriminating variable able to distinguish b quark jet from non-b quark jets is built using a deep neural network that is fed with different topological and kinematics variables; in particular:

- the distance between the primary vertex and the secondary vertex in the transverse plane must be greater than 100 μm and lower than 2.5 cm;
- the distance between the primary vertex and the secondary vertex in the transverse plane divided by its error has to be greater than 3;
- the invariant mass of charged particle associated to the vertex must be lower tha 6.5 GeV;
- the vertex must not be compatible with a K^0_S decay vertex.

The output of the algorithm is a continuous distribution from 0 to 1 whose shape depends on the quark that generates the jet (Fig 4.3).

To equalize the performances in data and MC of the algorithm, a set of weights, as know as Scale Factors, are computed event per event using the following formula:

$$SF_{weight} = \Pi_i^{N_{jets}} SF(D_i, p_{Ti}, \eta_i, h_{flavour}) \tag{4.2}$$

where D_i is the b tagging discriminant and $h_{flavour}$ is the hadron flavour.
The scale factors are applied and computed separately for all three years. As an example in Figure 4.3 is shown the distribution of the b tagger discriminator of the jet with the highest b tagger score before (in blue) and after (red) applying

[1]Corrections consist in L1 Pileup, L2 Relative Jet Correction, L3 Absolute Jet Correction for both Monte Carlo samples and data. Residual calibration for data are also applied [64]

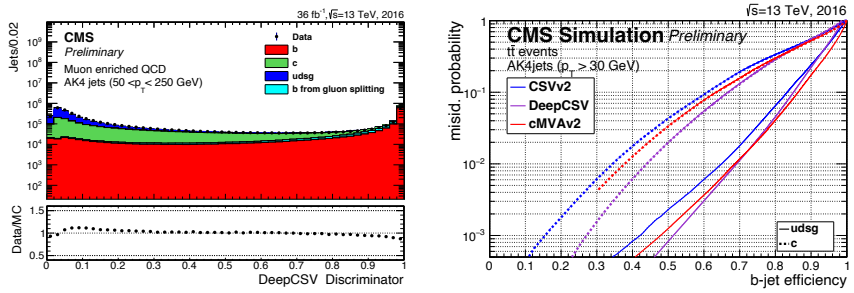


Figure 4.2: Left: distribution of the DeepCSV discriminators for ak4 jets in a muon enriched jet sample. The markers correspond to the data. The stacked, coloured histograms indicate the contributions of the different jet flavours in the simulation. Right: performance of the b jet identification efficiency algorithms demonstrating the probability for non-b jets to be misidentified as b jet as a function of the efficiency to correctly identify b jets. The lines shown are for CSVv2, DeepCSV, and cMVAv2. CSVv2 is the older version of DeepCSV where a non-deep neural network is used to discriminate among jets flavor. cMVAv2 uses also the information from the soft leptons inside jets, while CSVv2, DeepCSV do not [65]

the scale factor event weights. The plot is done using the HH 2018 signal sample.

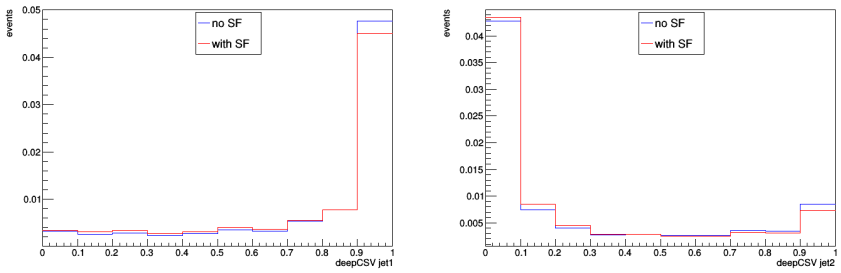


Figure 4.3: This plot shows the distribution of the b tagger discriminator before (in blue) and after (red) applying the scale factor event weights. Plots are is done using the HH 2018 signal sample. Left: b tagger discriminator of the jets with the highest b tagger score; right: b tagger discriminator of the jets with the second highest b tagger score

4.3.5 Pre-firing

For the years 2016 and 2017 the ECAL gradual timing shift was not propagated to the L1 trigger primitives properly. This resulted in a wrong association of a

large fraction of high η trigger primitives to the previous bunch crossing. Since L1 trigger does not fire over two consecutive bunch crossing, this resulted in events that self vetoed if a large deposit of ECAL energy was found in the region $2 < \eta < 3$. This effect is not described in the Monte Carlo simulations, therefore a weight is computed for each event and it is applied to the MC simulation for the years 2016 and 2017.

4.3.6 Summary of the object selection

The requirements on all objects used for the analysis are summarized in the Table 4.12. In addition, a "ghost-cleaning" procedure is applied to the muons, as described in Sec. 4.3.2. A lepton is declared **loose** if it passes the reconstruction, kinematics and d_{xy}/d_z cuts and declared **tight** if it passes in addition the identification, isolation and SIP3D cut.

Table 4.12: Summary of physics object selection criteria.

Electrons
$^{e} > 7\text{GeV} \qquad \lvert\eta^{e}\rvert < 2.5$
$\mathrm{d_{xy}} < 0.5\,\mathrm{cm} \qquad \mathrm{d_{z}} < 1\,\mathrm{cm}$
$\lvert\mathrm{SIP_{3D}}\rvert < 4$
Muons
Global or Tracker Muon
Discard Standalone Muon tracks if reconstructed in muon system only
Discard muons with muonBestTrackType==2 even if they are globa or tracker muons
$\mathrm{p_{T}^{\mu}} > 5\ \mathrm{GeV} \qquad \lvert\eta^{\mu}\rvert < 2.4$
$\mathrm{d_{xy}} < 0.5\,\mathrm{cm} \qquad \mathrm{d_{z}} < 1\,\mathrm{cm}$
$\lvert\mathrm{SIP_{3D}}\rvert < 4$
PF muon ID if $p_T < 200$ GeV, PF muon ID or High-p_T muon ID (Table 4.11) if $p_T > 200$ GeV
$\mathcal{I}^{\mu}_{\mathrm{PF}} < 0.35$
FSR photons
$\mathrm{p_{T}^{\gamma}} > 2\ \mathrm{GeV} \qquad \lvert\eta^{\gamma}\rvert < 2.4$
$\mathcal{I}^{\gamma}_{\mathrm{PF}} < 1.8$
$\Delta R(\ell,\gamma) < 0.5 \qquad \frac{\Delta R(\ell,\gamma)}{(\gamma)^2} < 0.012\ \mathrm{GeV}^{-2}$
Jets
$\mathrm{p_{T}^{jet}} > 20\ \mathrm{GeV} \qquad \lvert\eta^{\mathrm{jet}}\rvert < 2.4$
$\Delta R(\ell/\gamma, \mathrm{jet}) > 0.3$
Cut-based jet ID (tight WP)
Jet pileup ID (tight WP)

4.4 Event selection

The objects defined in the section above are combined to properly reconstruct the $HH \to b\bar{b}4l$ final state.
Events are required to have at least one good primary vertex PV) fullfilling the following criteria: high number of degree of freedom ($N_{PV} > 4$), collisions restricted along the $z-$axis ($z_{PV} < 24$ cm) and small radius of the PV ($r_{PV} < 2$ cm).
The first step of the analysis is the reconstruction of the Higgs boson decaying into 4 leptons. The 4l candidate is built from events that contain exactly 4 tight leptons (Sec 4.3.6) where the FSR photons are subtracted as described in Sec 4.3.3. To ensure that the lepton are well isolated, electrons which are within $\Delta R < 0.05$ of selected muons are discarded. The construction and selection of four-lepton candidates proceeds according to the following sequence:

1. **Z candidates** are built as pairs of selected leptons of opposite charge and matching flavour (e^+e^-, $\mu^+\mu^-$) that satisfy $12GeV/c^2 < m_{\ell\ell(\gamma)} <$ 120 GeV/c^2, where the Z candidate mass includes the selected FSR photons if any.

2. **ZZ candidates** are built as pairs of non-overlapping Z candidates. The Z candidate with reconstructed mass $m_{\ell\ell}$ closest to the nominal Z boson mass is denoted as Z_1, and the second one is denoted as Z_2. ZZ candidates are required to satisfy the following list of requirements:

   - **Ghost removal** : $\Delta R(\eta, \varphi) > 0.02$ between each of the four leptons.
   - **lepton** p_T: Two of the four selected leptons should pass $p_{T,i} >$ 20 GeV/c and $p_{T,j} > 10$ GeV/c.
   - **QCD suppression**: all four opposite-sign pairs that can be built with the four leptons (regardless of lepton flavor) must satisfy $m_{\ell\ell} >$ 4 GeV/c. Here, selected FSR photons are not used in computing $m_{\ell\ell}$, since a QCD-induced low mass dilepton (eg. J/Ψ) may have photons nearby (e.g. from π_0).
   - Z_1 **mass**: $m_{Z1} > 40$ GeV/c^2
   - **'smart cut'**: defining Z_a and Z_b as the mass-sorted alternative pairing Z candidates (Z_a being the one closest to the nominal Z boson

mass), require NOT($|m_{\mathrm{Za}} - m_{\mathrm{Z}}| < |m_{\mathrm{Z1}} - m_{\mathrm{Z}}|$ AND $m_{\mathrm{Zb}} < 12$). Selected FSR photons are included in m_{Z}'s computations. This cut discards 4μ and $4e$ candidates where the alternative pairing looks like an on-shell Z + low-mass $\ell^+\ell^-$.

- **four-lepton invariant mass**: $115 < m_{4l} < 135$ GeV/c^2

3. Events containing at least one selected ZZ candidate + 2 jets form the **signal region**.

If in an event there is more than one ZZ candidate, since the Higgs boson usually decays into high p_T objects, the ZZ candidate with the highest values of the scalar sum of the 4-lepton p_T is chose.
Fig 4.4 shows the invariant 4-leptons mass for the three years at this step of the analysis.

The ZZ non-resonant production represents the majority of the background and is divided into $qq \to ZZ$ and $gg \to ZZ$. The latter doesn't seem to behave as the former under the Z peak. It is worth to clarify that the peak at 91 GeV arises from the Z→4l decay. The contribution gg induced shows a similar peak but the cross section of this process is much lower (since it is loop induced) with respect to the qq induced process, thus the gg induced process is less visible in the plot. Higgs processes (violet background) are the only processes, together with the signal (red line) that peak at 125 GeV. The dark green background is the so called ZX contribution that come from events with one or more non-prompt letpons that contaminate the signal region, a precise treatment and estimation of this process is described in Sec 4.5. A good data-mc agreement appeared for all the years.
To take into account the $H \to b\bar{b}$ contribution, the 4-lepton selection is refined by adding a further requirement of at least two jets in the event. This addition will mainly reduce the background contribution from SM Higgs. If there is more then two jets in the event, the Higgs candidate is built from the two jets with the highest b tagging score. An important thing to notice is that in this analysis the b tagger requirement is not used to further discard events; b tagging and a further discrimination of the events are done using machine learning techniques (Sec 4.6).

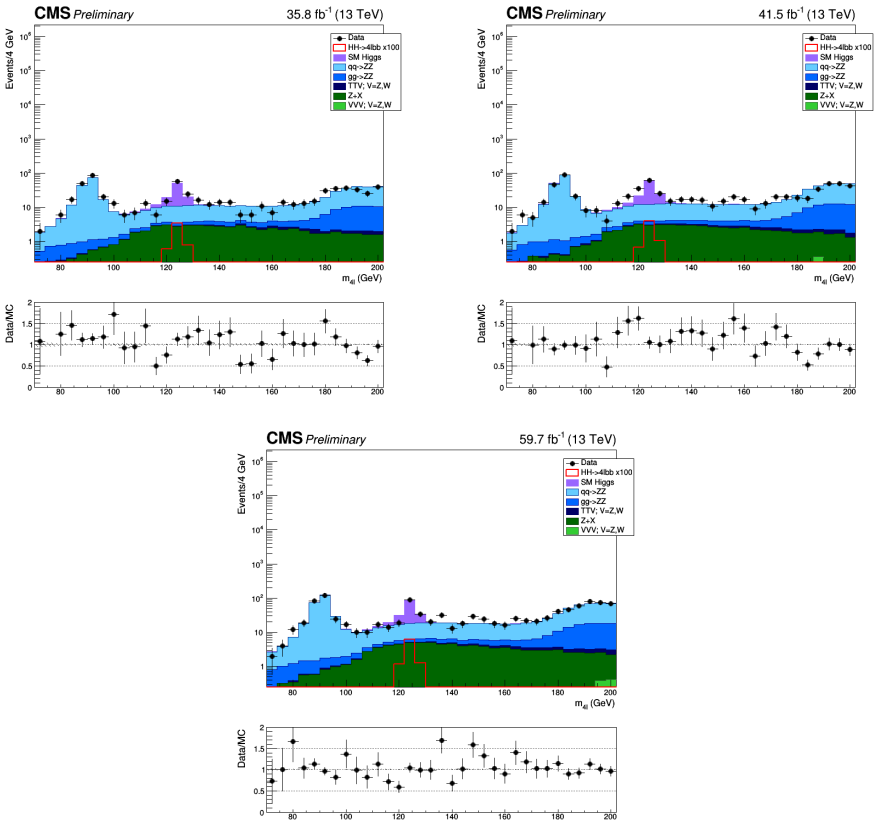


Figure 4.4: Four-lepton invariant mass after the full four leptons selection for 2016 (top left), 2017 (top right) and 2018 (bottom) datasets.

bb candidate selection study

A study was performed on the SM HH signal Monte Carlo sample in order to verify that the way chosen to select the two jets is the most efficient. A cone of radius $\Delta R = 0.4$ is built around each RECO jet selected; then we search for a GEN b jet coming from the decay of the Higgs boson and we check if the considered RECO jet is matching with the GEN jet, inside the cone previously defined. For this study three different way for selecting the two jets for building the di-jet candidate are considered:

- select the two highest p_T jets in the event
- select the two highest b tagger score jets in the event
- select the highest b tagger score jet and the highest p_T jet in the event (verifying that they are not the same jet).

Figure 4.5 reports the result of this study. The method that selects the two highest p_T jets in the event (blue line) is not very efficient since for most of the times only one of the RECO jet selected matches with a GEN jet. With this method there are some cases where none of the two jets selected match with a GEN jet coming from the Higgs boson. For the method that selects the highest b tagger score jet and the highest p_T jet in the event (green line), it's more frequent that both of the RECO jets selected match correctly with the signal GEN jets, but also with this method there are some cases where none of the two jets selected match with a GEN jet coming from the Higgs boson. The method that selects two highest b tagger score jets in the event (red line), instead, is the most efficient one since in most of the cases (even more cases than the previous method) both of the selected RECO jets are matched with GEN jets coming from the decay of the Higgs boson, even if there are few cases where both of the selected RECO jets are not coming from signal GEN jets.

The third method (2 highest b tagger jets) for selecting the two jets for building the di-jet candidate is chosen, since it is the most efficient.

Figure 4.5 right, shows the invariant mass built from the two jets selected with the three methods. It is possible to notice that the invariant mass built from the RECO jets selected with the 2-highest-b-tagger-jets method (red line) has less events in the right tail of the distribution with respect to the other methods (green and blue lines), since with this method less background jets (meaning, jets not coming from the Higgs boson decay in two b jets) are selected.

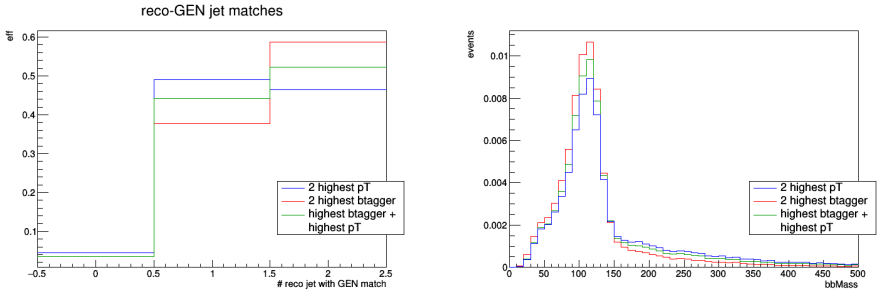


Figure 4.5: Right: RECO-GEN jet matching study performed on the signal Monte Carlo sample. RECO jets selected with one of the three method considered are matched with GEN jets coming from the decay of the Higgs boson. For the method that selects the two highest p_T jets (blue line) most of the times only one of the RECO jet selected matches with a GEN jet; for the method that selects the highest b tagger score jet and the highest p_T jet (green line) most of the times both of the RECO jets selected matches with a GEN jet but there are cases where none of the two RECO jets selected match with a signal GEN jet; for the method that selects the two highest b tagger score jets in the event (red line) most of the times both the selected RECO jets are matched with GEN jets coming from the decay of the Higgs boson and less background jets are selected.
Left: Di-jet invariant mass built from RECO jets selected with the three different methods considered: method that selects the two highest p_T jets (blue line), method that selects the highest b tagger score jet and the highest p_T jet (green line), method that selects the two highest b tagger score jets in the event (red line).

4.5 Background estimation

Background processes can be divided into reducible and irreducible sources: an irreducible background is a background that can have the same signature of the signal and thus survives the cuts that are used to define a phase space region in which the signal is enhance (so called Signal Region) while the reducible background is a background source that enter in the SR even if is originated from processes that should not pass the SR cuts. The reducible background ($Z + X$) originates from processes that contain one or more non-prompt leptons. The main sources of non-prompt leptons are non-isolated electrons and muons coming from decays of heavy-flavour mesons, mis-reconstructed jets (usually originating from light-flavour quarks) and electrons from γ conversions. In the following section we will understand how to deal with this processes.

$q\bar{q} \rightarrow ZZ$ Background

The $q\bar{q} \rightarrow ZZ$ background is generated at NLO, while the fully differential cross section has been computed at NNLO [66], but is not yet available in a partonic level event generator. Therefore NNLO/NLO correction for the $q\bar{q} \rightarrow ZZ$ background process are applied to the Monte Carlo sample deferentially as a function of $m(ZZ)$.

Additional NLO electroweak corrections which depend on the initial state quark flavor and kinematics are also applied to the $q\bar{q} \rightarrow ZZ$ background process in the region $m(ZZ) > 2m(Z)$ where the corrections have been computed.

$gg \rightarrow ZZ$ Background

The $gg \rightarrow ZZ$ background is simulated at LO with the generator MCFM 7.0 [59, 60, 67]. An exact calculation beyond the LO does not exists for the $gg \rightarrow ZZ$ background, but it has been shown [68] that the soft collinear approximation is able to describe the background cross section and the interference term at NNLO. The NNLO k-factor for the signal is obtained as a function of $m_{4\ell}$ using the HNNLO v2 Monte Carlo program [69–71] by calculating the NNLO and LO $gg \rightarrow H \rightarrow 2\ell 2\ell'$ cross sections at the small H boson decay width of 4.07 MeV and taking their ratios.

Standard Model Higgs-boson Background

The SM Higgs boson processes represent one of the main backgound for this analysis, especially the associated production with a pair of top quarks. Descriptions of the SM Higgs boson to ZZ production are obtained using the POWHEG V2 generator for the main production modes as described in section 4.2.1.

Reducible Background

To estimate the amount of non-prompt objects that contaminate the signal region the first step is to measure the probability for electrons and muons (f_e and f_μ) that do pass the **loose** selection criteria (defined in Section 4.3.1 and 4.3.2) to also pass the final selection criteria (defined in Section 4.4); this probability, referred to as *fake rate*, will be applied in dedicated control regions to extract the expected background contamination in the SR (4l + 2 jets).

To estimate the Z+X contribution in the SR we select samples of $Z(\ell\ell) + e + at\ least\ 2\ isolated\ jets$ and $Z(\ell\ell) + \mu + at\ least\ 2\ isolated\ jets$ events that are expected to be completely dominated by final states which include a Z boson, a fake lepton and at least 2 jets isolated with respect to the leptons (see Sec 4.3.4), so that the phase space is as close as possible to our signal region. These events are required to have two same flavour, opposite charge leptons with $p_T > 20$ GeV for the first lepton and $p_T > 10$ GeV for the second lepton, passing the tight selection criteria, thus forming the Z candidate. In addition, there is exactly one lepton passing the loose selection criteria as defined above and at least two isolated jets. The loose lepton is used as the probe lepton for the fake rate measurement. The invariant mass of this lepton and the opposite sign lepton from the reconstructed Z candidate should satisfy $m_{2\ell} > 4$ GeV.
The fake ratios are evaluated using the tight requirement $|M_{\text{inv}}(\ell_1\ell_2) - M_{\text{Z}}| < 7$ GeV, to reduce the contribution from photon (asymmetric) conversions populating low masses. The fake rates are measured in bins of the transverse momentum of the loose lepton and also they are computed separately for the barrel and the endcap region. To further check the procedure, fake rate curves, obtained with 2016 dataset, are first evaluated in $Z(\ell\ell) + e$ and $Z(\ell\ell) + \mu$ control region and in $Z(\ell\ell) + e + at\ least\ 2\ not\ isolated\ jets$ and $Z(\ell\ell) + \mu + at\ least\ 2\ not\ isolated\ jets$ and then compared to $Z(\ell\ell)+e+at\ least\ 2\ isolated\ jets$ and $Z(\ell\ell) + \mu + at\ least\ 2\ isolated\ jets$ control region (Fig 4.6). What can be seen from Fig 4.6 is that the blu marker ($Z(\ell\ell) + e$ and $Z(\ell\ell) + \mu$ control region) and the red marker ($Z(\ell\ell) + e + at\ least\ 2\ not\ isolated\ jets$ and $Z(\ell\ell) + \mu + at\ least\ 2\ not\ isolated\ jets$) are in agreement, that means that adding 2 not isolated jets in $Z(\ell\ell) + e$ and $Z(\ell\ell) + \mu$ control region does not change the phase space in which the probe lepton is extracted, in particular from both cases we observed the same probability of the probe lepton to be a fakeble object as a function of p_T.
With the requirement of 2 isolated jets (dark green marker) we force the probe lepton to be far away from jets, we discharge all the events with a fakeble object coming from the surroundings of the jets (mostly bjets misreconstructed as leptons); this lowering the fake rate probability expecially at low p_T, where most of the bjets are misreconstructed as leptons.

Since in this analysis we asked for at least 2 isolated jets, the fake rate is estimated in $Z(\ell\ell) + \ell + at\ least\ 2\ isolated\ jets$.
In Fig 4.7 - Fig 4.8 the fake rate distribution for 2017 and 2018 dataset are

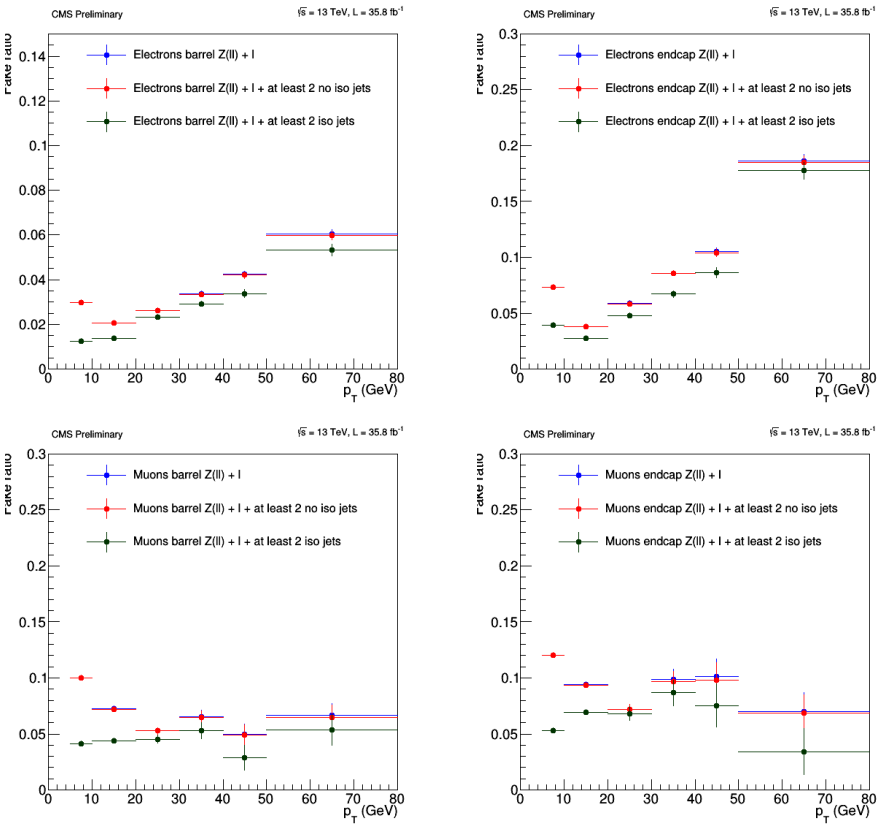


Figure 4.6: Fake rate as a function of the probe p_T for electrons and muons measured in $Z(\ell\ell) + \ell$, $Z(\ell\ell) + \ell + at\ least\ 2\ not\ isolated\ jets$ and $Z(\ell\ell) + \ell + at\ least\ 2\ isolated\ jets$ sample in 13 TeV. The fake rates are shown after the removal of WZ contribution from MC for 2016 dataset. The blu marker ($Z(\ell\ell) + e$ and $Z(\ell\ell) + \mu$ control region) and the red marker ($Z(\ell\ell) + e + at\ least\ 2\ not\ isolated\ jets$ and $Z(\ell\ell) + \mu + at\ least\ 2\ not\ isolated\ jets$) are in agreement, that means that adding 2 not isolated jets in $Z(\ell\ell) + e$ and $Z(\ell\ell) + \mu$ control region does not change the phase space in which the probe lepton is extracted, in particular from both cases we observed the same probability of a the probe lepton to be a fakeble object in function of p_T. With the requirement of 2 isolated jets (dark green marker) we force the probe lepton to be far away from jets, we discharge all the events with a fakeble object coming from the surroundings of the jets (mostly bjets misreconstructed as leptons); this lowering the fake rate probability expecialy at low p_T, where most of the bjets are misreconstructed as leptons.

shown: the blue marker represents the fake rate computed in $Z(\ell\ell) + \ell$ while the dark green marker represents the fake rate computed in $Z(\ell\ell)+\ell+at\ least\ 2\ isolated\ jets$.

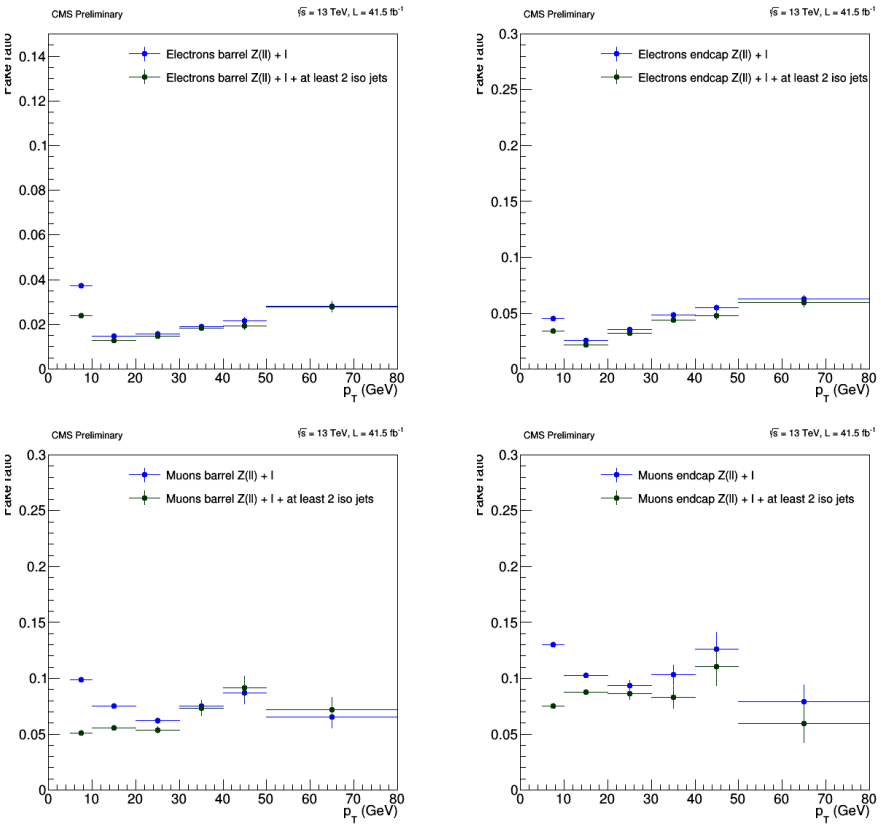


Figure 4.7: Fake rate as a function of the probe p_T for electrons and muons measured in a $Z(\ell\ell) + \ell$ sample in 13 TeV data. The fake rates are shown after the removal of WZ contribution from MC for 2017 dataset.

Fake Rate Application

Two control samples are obtained as subsets of four lepton events which pass the first step of the selection (First Z step, see Section 4.4), requiring an additional pair of loose leptons of same flavour and opposite charge, that pass the SIP_{3D} cut. The events must satisfy all kinematic cuts applied for the Higgs phase space selection (see 4.4).

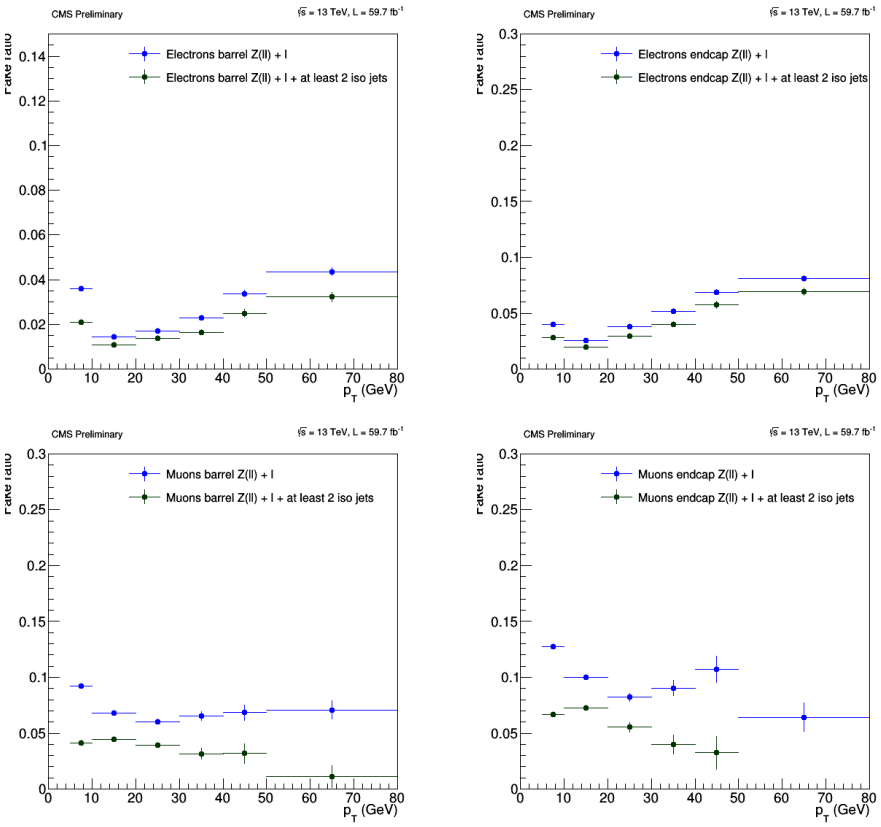


Figure 4.8: Fake rate as a function of the probe p_T for electrons and muons measured in a $Z(\ell\ell) + \ell$ sample in 13 TeV data. The fake rates are shown after the removal of WZ contribution from MC for 2018 dataset.

The first control sample is obtained by requiring that the two loose leptons which do not form the Z_1 candidate, do not pass the final identification and isolation criteria. The other two leptons pass the final selection criteria by definition of the Z_1. The presence of two jets is also required. This sample is denoted as *2 Prompt + 2 Fail + 2 jets* ($2P + 2F$) sample. It is expected to be populated with events that intrinsically have only two prompt leptons (mostly DY, with small fraction of $t\bar{t}$ and $Z\gamma$ events).

The second control sample is obtained by requiring one of the four leptons not to pass the final identification and isolation criteria. The other three leptons should pass the final selection criteria. Also in this case, the presence of two jets is required. This control sample is denoted as *3 Prompt + 1 Fail + 2 jets* ($3P + 1F$) sample. It is expected to be populated with the type of events that populate the $2P + 2F$ region, but with different relative proportions, with also a contribution from WZ events that intrinsically have three prompt leptons.

The control samples obtained in this way, orthogonal by construction to the signal region, are enriched with fake leptons and are used to estimate the reducible background in the signal region.

The expected number of reducible background events in the $3P + 1F$ region, N^{bkg}_{3P1F}, can be computed from the number of events observed in the $2P + 2F$ control region, N_{2P2F}, by weighting each event in the region with the factor $(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})$, where f_i and f_j correspond to the fake rates of the two loose leptons:

$$N^{bkg}_{\text{3P1F}} = \Sigma(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})N_{\text{2P2F}} \tag{4.3}$$

If the fake rates was measured in a sample that has exactly the same background composition as the $2P + 2F$ sample, the difference between the observed number of events in the $3P + 1F$ sample and the expected background predicted from the $2P + 2F$ sample would solely amount to the (small) WZ and $Z\gamma_{conv}$ contribution. Large differences arise because the fake rates used in eq. 4.3 do not properly account for the background composition of the $2P+2F$ control sample. The difference between the $3P + 1F$ observation and the prediction from $2P+2F$ to recover the missing contribution from photon conversion, and more generally, to correct for the fact that the fake rates do not properly account for the background composition of the $2P + 2F$ sample. More precisely, the expected reducible background in the signal region is given by the sum of two terms:

- a *2P2F component*, obtained from the number of events observed in the $2P + 2F$ control region, N_{2P2F}, by weighting each event in that region with the factor $(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})$, where f_i and f_j correspond to the fake rates of the two loose leptons

- a *3P1F component*, obtained from the difference between the number of observed events in the $3P + 1F$ control region, N_{3P1F}, and the expected contribution from the $2P + 2F$ region and ZZ processes in the signal region, $N^{\text{ZZ}}_{\text{3P1F}} + N^{\text{bkg}}_{\text{3P1F}}$. The $N^{\text{bkg}}_{\text{3P1F}}$ is given by the equation 4.3 and the $N^{\text{ZZ}}_{\text{3P1F}}$ is the contribution from ZZ which is taken from the simulation. The difference $N_{\text{3P1F}} - N^{\text{bkg}}_{\text{3P1F}} - N^{\text{ZZ}}_{\text{3P1F}}$, which may be negative, is obtained for each (p_T, η) bin for the F lepton, and is weighted by $\frac{f_i}{1-f_i}$, where f_i denotes the fake rate of this lepton. This *3P1F component* accounts for the contribution of reducible background processes with only one fake lepton (like WZ events), and for the contribution of other processes (e.g. photon conversions) that are not properly estimated by the *2P2F component*, because of the fake rates used.

Therefore, the full expression for the prediction of the reducible background in the signal region can be written as:

$$N^{\text{bkg}}_{\text{SR}} = \Sigma \frac{f_i}{(1-f_i)}(N_{\text{3P1F}} - N^{\text{bkg}}_{\text{3P1F}} - N^{\text{ZZ}}_{\text{3P1F}}) + \Sigma \frac{f_i}{(1-f_i)} \frac{f_j}{(1-f_j)} N_{\text{2P2F}} \tag{4.4}$$

The Z+X event yields in the signal region is obtained after applying the additional cuts of the selection; yields obtained for 2016 data are reported in Tab 4.13 - 4.14 - 4.15.

	4e	4μ	2e2μ
Z + X + 2jets	1.40 $\pm$ 0.67 (stat+syst)	0.79 $\pm$ 0.28 (stat+syst)	2.64 $\pm$ 1.10 (stat+syst)

Table 4.13: Reducible background yields for all the channels (2016 dataset).

	4e	4μ	2e2μ
Z + X + 2jets	0.52 $\pm$ 0.22 (stat+syst)	1.48 $\pm$ 0.48 (stat+syst)	2.00 $\pm$ 0.72 (stat+syst)

Table 4.14: Reducible background yields for all the channels (2017 dataset).

	4e	4μ	$2e2\mu$
Z + X + 2jets	0.72 ± 0.28 (stat+syst)	1.60 ± 0.50 (stat+syst)	2.58 ± 0.89 (stat+syst)

Table 4.15: Reducible background yields for all the channels (2018 dataset).

The same procedure is followed for 2017 and 2018 datasets. The BDT score is then extracted for all the events in the first (3P1F) and second (2P2F) control sample, then, the overall BDT shape of Z+X is computed using the previous equation. That shape will be used as one of the input of the statistical analysis.

Uncertainties on Reducible Background estimation

The uncertainty on the reducible background arises from the difference in composition of the reducible background processes in the region where the fake rate is measured and where it is applied. This uncertainty can be estimated by measuring the fake ratios for individual background processes in the Z + 1L region in simulation. The weighted average of these individual fake ratios is the fake ratio that we measure in this sample (in simulation). The exact composition of the background processes in the 2P+2F region where we plan to apply the fake ratios can be determined from simulation, and one can reweigh the individual fake ratios according to the 2P+2F composition. The difference between the reweighed fake ratio and the average one can be used as a measure of the uncertainty on the measurement of the fake ratios.

The uncertainties are estimated separately per 4-lepton final states and are reported in the table 4.16.

year	4μ	4e	$2e2\mu$
2016	30%	41%	35%
2017	30%	38%	33%
2018	30%	37%	33%

Table 4.16: Systematic uncertainties on the reducible background estimate for the 3 years.

4.6 Multivariate analysis

In order to better discriminate between signal and background events and to improve the sensitivity of the analysis, a Boosted Decision Tree is trained exploiting different kinematic variables.

The network is trained in a region enriched with signal events passing the following requirements:

- H $\rightarrow 4\ell$ full selection
- $|m_{4\ell} - 125| < 10$ GeV
- at least 2 jets in the event

After this selection the two jets with the highest b discriminator are selected to reconstruct the $Higgs \rightarrow b\bar{b}$.

The signal is trained against all the backgrounds which remain in the signal region: ttH, ttZ, ttW, SM Higgs, VBF, WH, ZH and ZZ. For each year, the strategy is to train three times the BDT, depending on the different Higgs decay mode: $H \rightarrow ZZ \rightarrow 4\mu$, $H \rightarrow ZZ \rightarrow 4e$ and $H \rightarrow ZZ \rightarrow 2e2\mu$; the shape of the BDT for all this 3 combination will be used as input in the statistical analysis.

4.6.1 2016 BDT configuration

A large number of studies were performed (see Appendix B (Sec. A.2)) to choose the best set of variables.
Looking at both the area under the ROC curve and the overtraining check as a figure of merit, the final choice for the set of variables is:

- p_T of the four leptons
- ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$
- Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4)
- p_T of the two jets with the highest b tag discrminator score
- two jets invariant mass

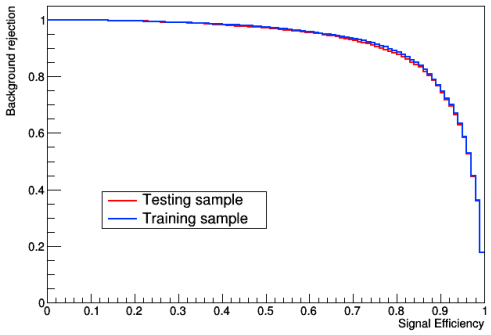


Figure 4.9: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4), p_T of the two jets, two jets invariant mass and the analysed dataset is 2016.

order	variable	ranking
1	$bdisc_{j1}$	$1.181e-01$
2	m_{jj}	$1.141e-01$
3	$bdisc_{j2}$	$1.098e-01$
4	ΔR_{HH}	$1.095e-01$
5	pT_{lep2}	$9.877e-02$
6	pT_{j1}	$9.584e-02$
7	pT_{lep4}	$9.351e-02$
8	pT_{lep1}	$9.194e-02$
9	pT_{lep3}	$8.688e-02$
10	pT_{j2}	$8.164e-02$

Table 4.17: Ranking variables for BDT used for 2016 dataset.

The area under the ROC curve is 0.92. As an overtraing check, in Figure 4.9 ROC curves for training and testing samples are show.

The ranking of the variables is in Table 4.17 .

For 2016 datasets the most discriminant variables is the invariant di-jet mass followed by the b tagging discriminator score of the jets and the ΔR between H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$; this was expected because none of the background samples contains $H \rightarrow b\bar{b}$ and thus informations such as the invariant di-jet mass, b discriminator score of the jets or the angular distance between H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$ become very powerful in distinguish backgrounds from signal sample.

The BDT input variables distributions are shown in Figure 4.10, while the correlation matrix for signal and background variables for the 4mu final state are shown in Figure 4.11. The BDT input variables in the four-lepton sidebands after requiring the presence of at least 2 jets are show in the appendix A (Sec A.1.1).

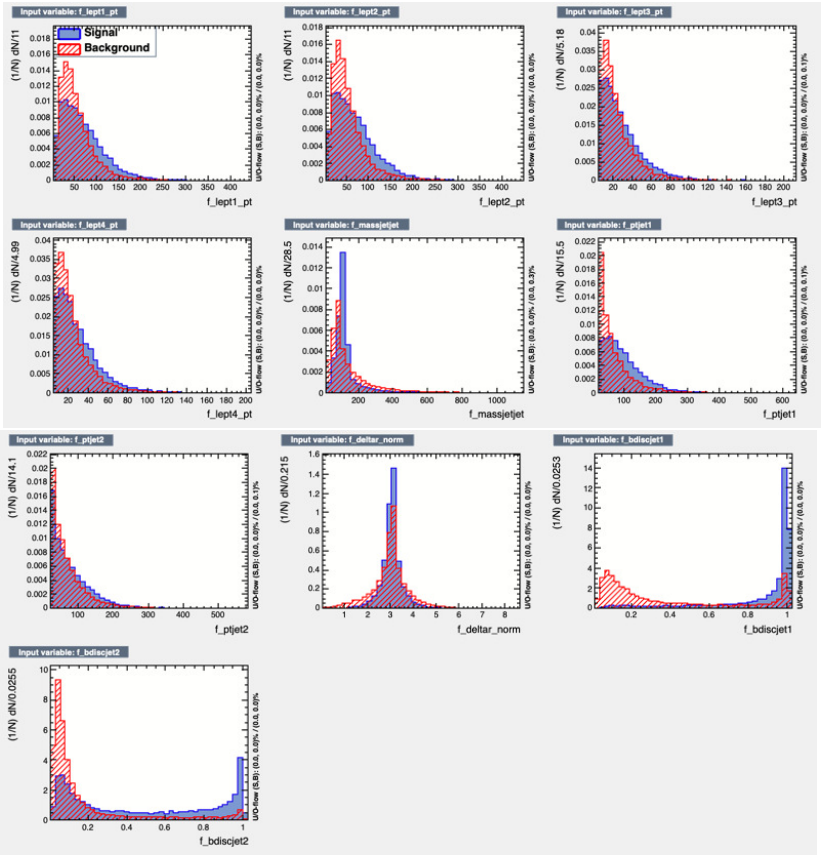


Figure 4.10: BDT input variables distributions for 2016.

The data / mc agreement of the variables in the signal region with the three channels (4e, 4μ and 2e2μ) summed together is shown in Figure 4.12 - 4.13. Even if the statistic is very low, a good agreement is observed for all the variables used in the BDT training.

In order to understand if the BDT response is well described by data, the score of the BDT is plotted in the 4 leptons sideband:

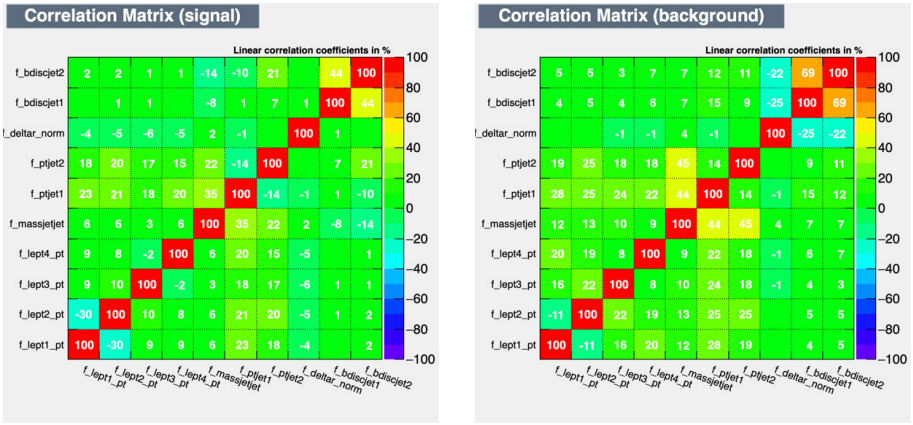


Figure 4.11: BDT correlation plots for 2016: correlation between background's variables (left), correlation between signal's variables (right).

- H $\rightarrow 4\ell$ full selection
- $|m_{4\ell} - 125| > 10$ GeV

as shown in Figure 4.14.

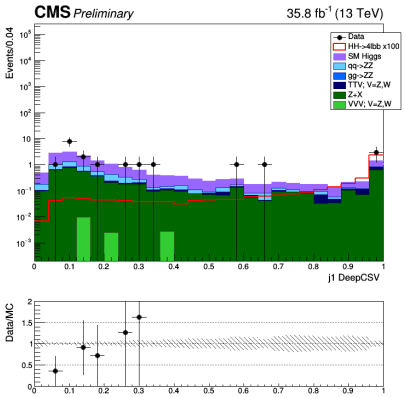
The BDT response in the sideband of the 4 leptons distribution shows a good data / mc agreement; this, together with the agreement shown in Fig 4.12 - 4.13, means that the BDT is under control and thus the shape of the BDT in signal region can be used to extract the upper limit on the signal stength.

4.6.2 2017 BDT configuration

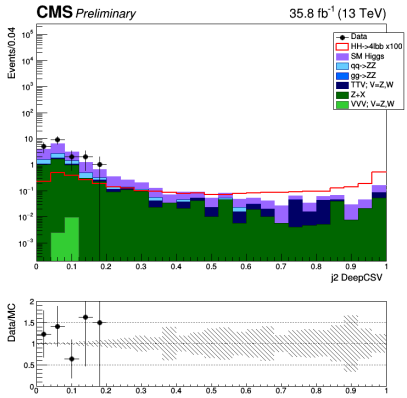
The same training used for 2016 is done for 2017.
The area under the ROC curve is 0.94.
As an overtraing check, in Figure 4.15 ROC curves for training and testing samples are shown.

The ranking of the variables is reported in Table 4.18.

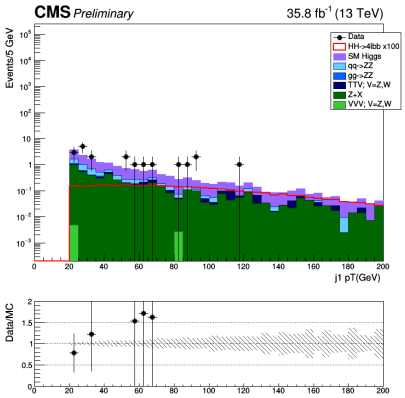
The most discriminant variables are those coming from $Higgs \rightarrow b\bar{b}$ production (b tagging information) and from the ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$, as expected (see Sec 4.6.1).
The BDT input variables distributions are shown in Figure 4.16, while the cor-



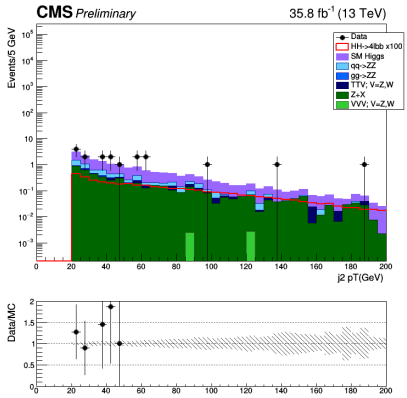
(a) B tagger score of the jet with the highest value of the deepCSV discriminant for 2016 dataset



(b) B tagger score of the jet with the second-highest value of the deepCSV discriminant for 2016 dataset



(c) p_T of jet with the highest value of the deepCSV discriminant for 2016 dataset

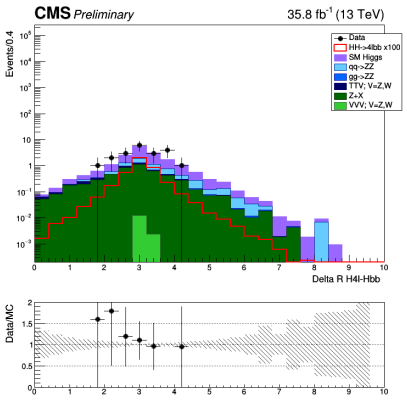


(d) p_T of jet with the second-highest value of the deepCSV discriminant for 2016 dataset

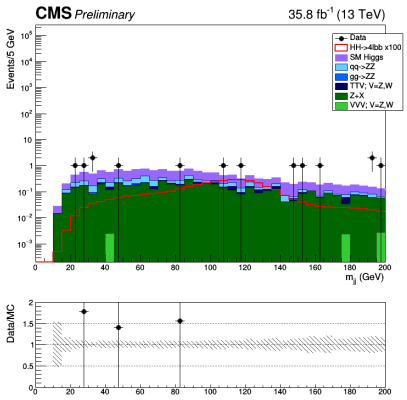
Figure 4.12: BDT input variables distributions for 2016

relation matrix for signal and background variables for the 4mu final state are shown in Figure 4.17.

The data-MC agreement of the variables in the signal region with the three channels (4e, 4μ and 2e2μ) summed together is shown in Figure 4.18.The BDT input variables in the four-lepton sidebands after requiring the presence of at least 2 jets are shown in the appendix A (Sec A.1.1). Even if the statistics is very low, a good agreement is observed for all the variables used in the BDT training.



(a) ΔR between the H $\to 4\ell$ and H $\to b\bar{b}$ for 2016 dataset



(b) Invariant di-jetmass built with the two jets with the highest value of the deepCSV discriminant for 2016 dataset

Figure 4.13: BDT input variables distributions for 2016.

order	variable	ranking
1	ΔR_{HH}	$1.261e-01$
2	$bdisc_{j1}$	$1.254e-01$
3	m_{jj}	$1.130e-01$
4	pT_{lep2}	$1.024e-01$
5	pT_{lep1}	$9.753e-02$
6	$bdisc_{j2}$	$9.504e-02$
7	pT_{lep4}	$9.431e-02$
8	pT_{j1}	$8.927e-02$
9	pT_{lep3}	$8.516e-02$
10	pT_{j2}	$7.180e-02$

Table 4.18: Ranking variables for BDT used for 2017 dataset.

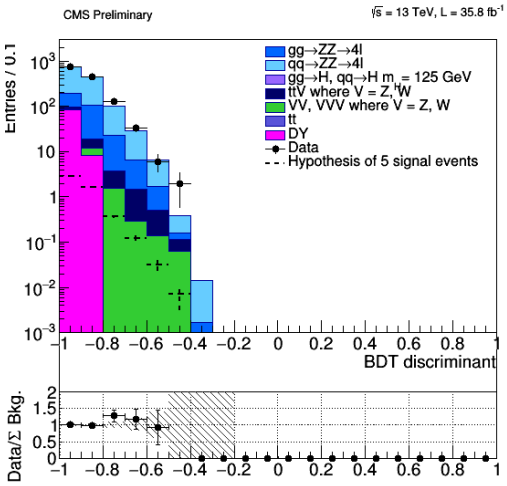


Figure 4.14: BDT score in the sideband of the 4 leptons distribution for 2016.

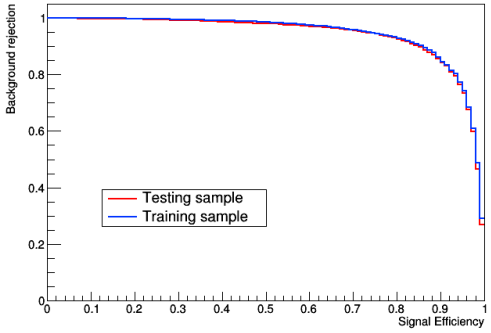


Figure 4.15: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4), p_T of the two jets, two jets invariant mass and the analysied dataset is 2017.

In order to understand if the BDT response is well described by data, the score of the BDT is plotted in the 4 leptons sideband:

- H $\rightarrow 4\ell$ full selection
- $|m_{4\ell} - 125| > 10$ GeV

as shown in Figure 4.24.

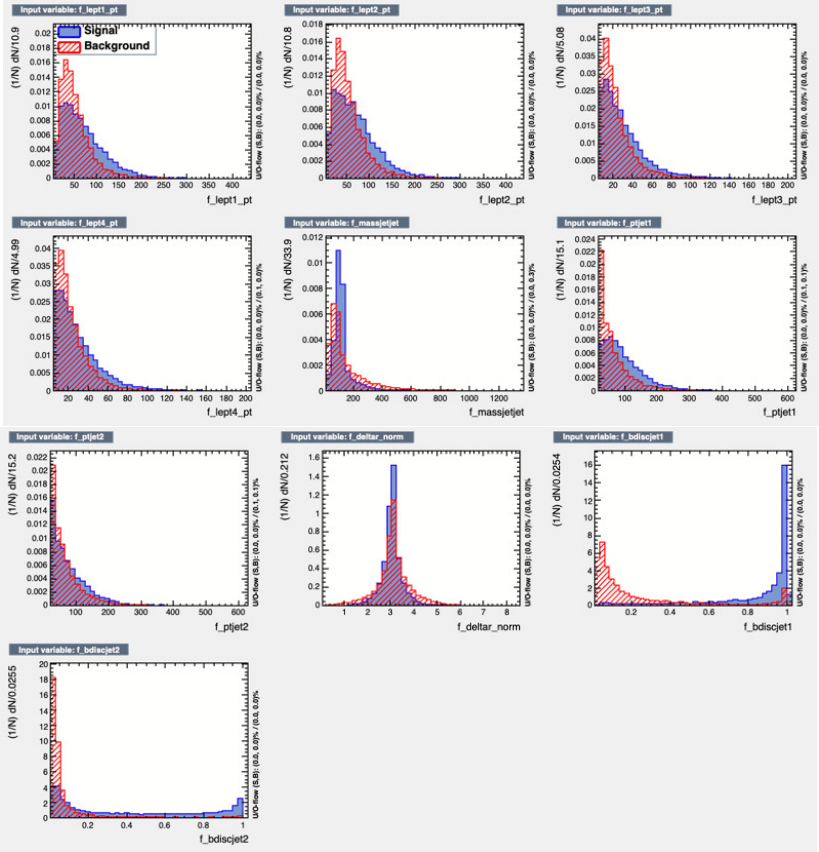


Figure 4.16: BDT input variables distributions for 2017.

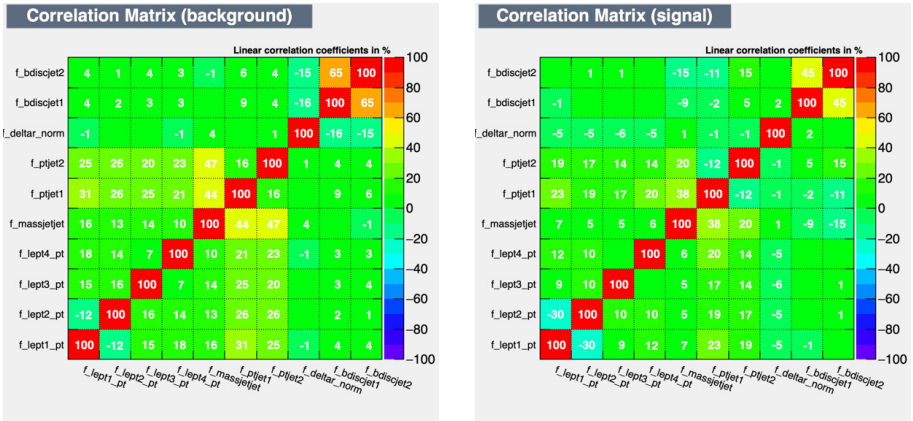
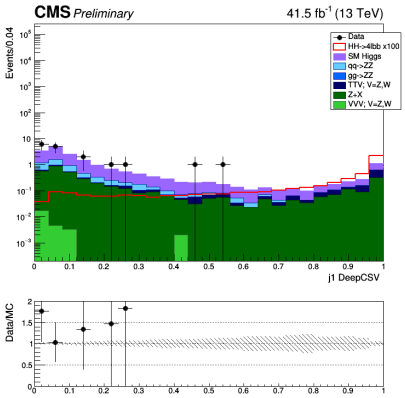
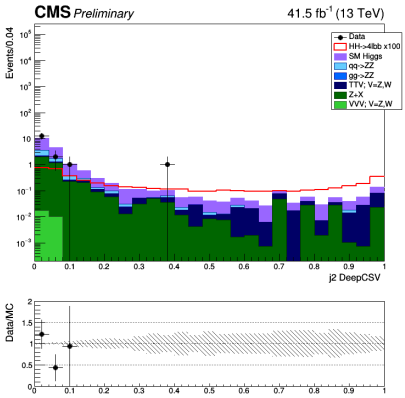


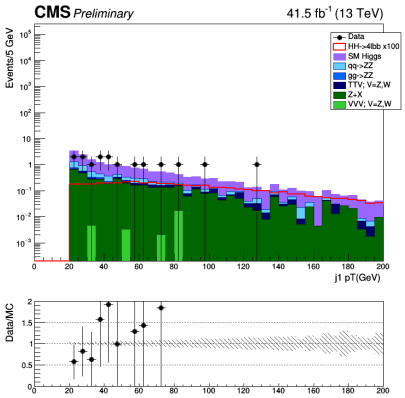
Figure 4.17: BDT correlation plots for 2017: correlation between background's variables (left), correlation between signal's variables (right).



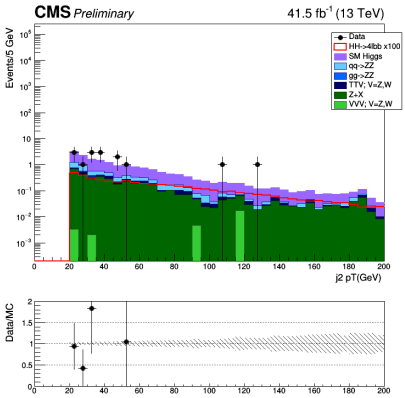
(a) B tagger score of the jet with the highest value of the deepCSV discriminant for 2017 dataset



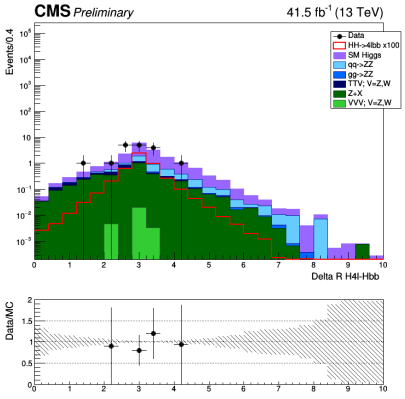
(b) B tagger score of the jet with the second-highest value of the deepCSV discriminant for 2017 dataset



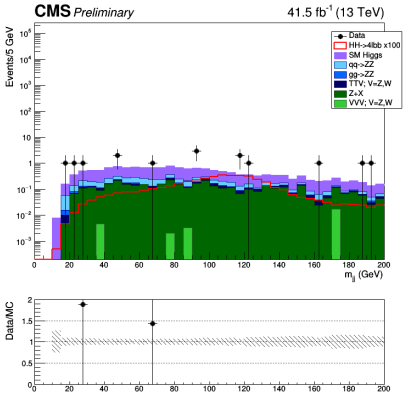
(c) p_T of jet with the highest value of the deepCSV discriminant for 2017 datase



(d) p_T of jet with the second-highest value of the deepCSV discriminant for 2017 dataset



(e) ΔR between the H $\to 4\ell$ and H $\to b\bar{b}$ for 2017 dataset



(f) Invariant di-jetmass built with the two jets with the highest value of the deepCSV discriminant for 2017 dataset

Figure 4.18: BDT input variables distributions for 2017

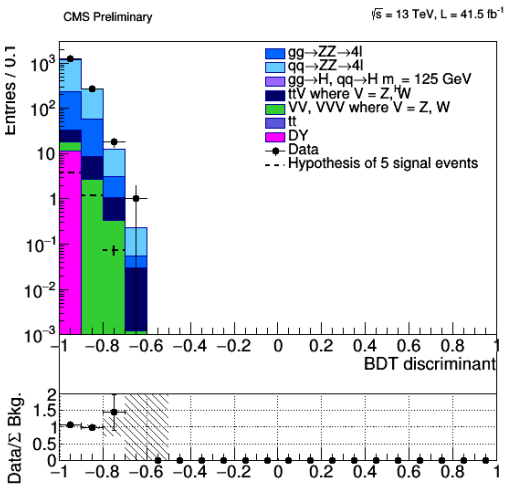


Figure 4.19: BDT score in the sideband of the 4 leptons distribution for 2017.

The BDT response in the sideband of the 4 leptons distribution shows a good data / MC agreement; this, with the agreement shown in Fig 4.18, means that the BDT of 2017 is under control (see Sec 4.6.1).

4.6.3 2018 BDT configuration

The same sets of variables used for 2016 and 2017 is used to train the BDT also for 2018.

The area under the ROC curve is 0.92.

As an overtraing check, in Figure 4.20 ROC curves for training and testing samples are show.

The ranking of the variables is in Table 4.19.

The most discriminant variables are those coming from from $Higgs \to b\bar{b}$ production (invarinat di-jet mass, b tagging information) and from the ΔR between the H $\to 4\ell$ and H $\to$ b$\bar{\text{b}}$.

The BDT input variables distributions are shown in Figure 4.21, while the correlation matrix for signal and background variables for the 4mu final state are shown in Figure 4.22

The data-MC comparison of the input variables in the signal region with the three channels (4e, 4μ and 2e2μ) summed together is shown in Figure 4.23.

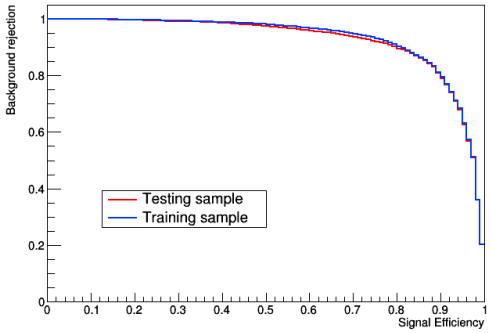


Figure 4.20: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4), p_T of the two jets, two jets invariant mass and the analysied dataset is 2018.

order	variable	ranking
1	$bdisc_{j1}$	$1.276e-01$
2	ΔR_{HH}	$1.111e-01$
3	$bdisc_{j2}$	$1.088e-01$
4	m_{jj}	$1.075e-01$
5	pT_{lep1}	$1.001e-01$
6	pT_{lep2}	$9.905e-02$
7	pT_{j1}	$9.376e-02$
8	pT_{lep4}	$9.262e-02$
9	pT_{lep3}	$8.229e-02$
10	pT_{j2}	$7.721e-02$

Table 4.19: Ranking variables for BDT used for 2018 dataset.

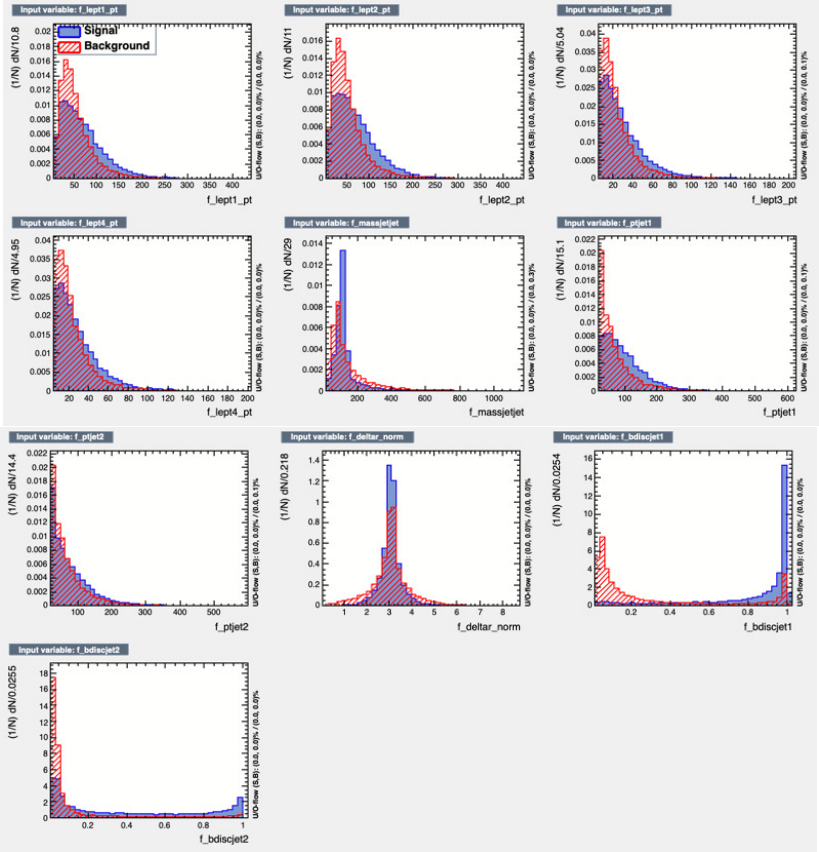


Figure 4.21: BDT input variables distributions for 2018.

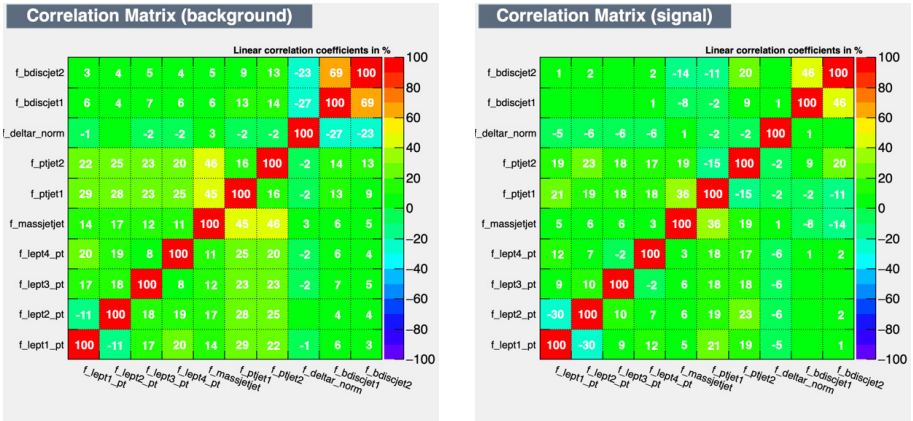


Figure 4.22: BDT correlation plots 2018: correlation between background's variables (left), correlation between signal's variables (right).

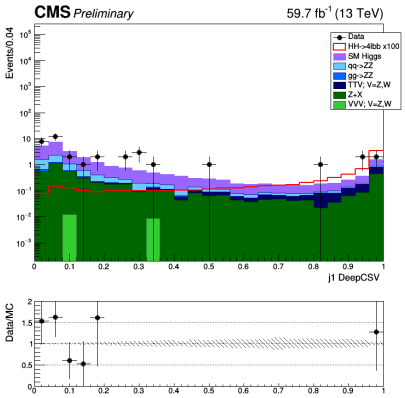
The BDT input variables in the four-lepton sidebands after requiring the presence of at least 2 jets are show in the appendix A (Sec A.1.1).Even if the statistic is very low, a good agreement is observed for all the variables used in the BDT training.

In order to understand if the BDT response is well described by data, the score of the BDT is plotted in the 4 leptons sideband:

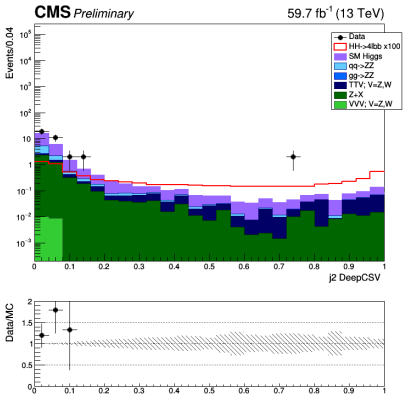
- H $\rightarrow 4\ell$ full selection
- $|m_{4\ell} - 125| > 10$ GeV

as shown in Figure 4.24.

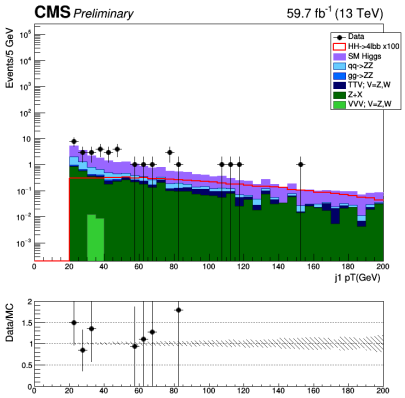
The BDT response in the sideband of the 4 leptons distribution shows a good data / mc agreement; this, with the agreement shown in Fig 4.23, means that the BDT of 2018 is under control (see Sec 4.6.1).



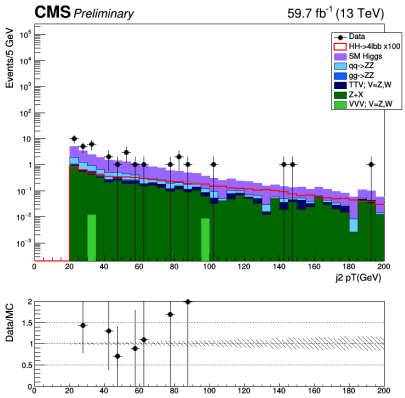
(a) B tagger score of the jet with the highest value of the deepCSV discriminant for 2018 dataset



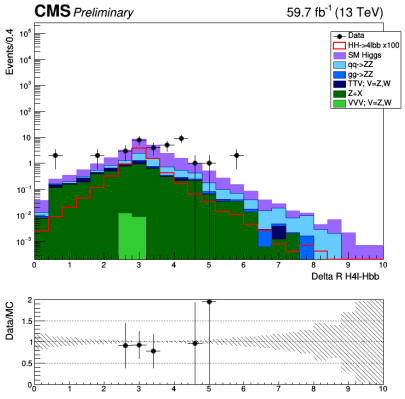
(b) B tagger score of the jet with the second-highest value of the deepCSV discriminant for 2018 dataset



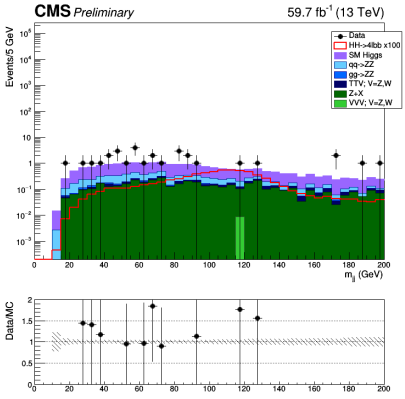
(c) p_T of jet with the highest value of the deepCSV discriminant for 2018 datase



(d) p_T of jet with the second-highest value of the deepCSV discriminant for 2018 dataset



(e) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$ for 2018 dataset



(f) Invariant di-jetmass built with the two jets with the highest value of the deepCSV discriminant for 2018 dataset

Figure 4.23: BDT input variables distributions for 2018

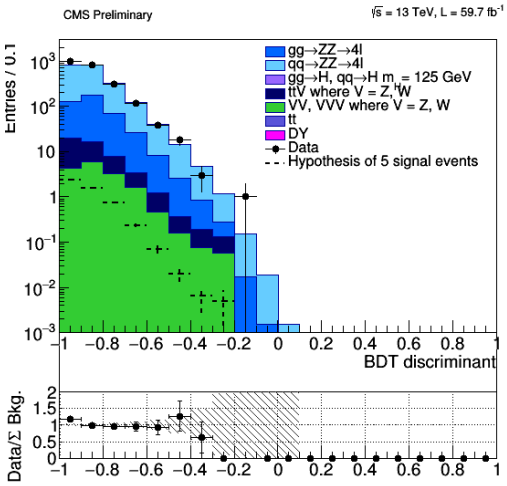


Figure 4.24: BDT score in the sideband of the 4 leptons distribution for 2018.

4.7 Systematic uncertainties

All the measurement in physics are affected by a uncertainties. These uncertainties depends on several sources: from the pure statistics fluctuations arising from the fact that a measurement is based on a finite set of observations, to uncertainties arising from inefficiencies of the detector or on the assumption of the scientist. Based on these difference we could distinguish two types of uncertainties: the statistical and the systematic uncertainty. The statistical uncertainty arises from the impossibility to have an infinite amount of data to analyse, examples of statistical uncertainties include the Poisson fluctuations associated with measurements involving finite sample sizes and random variations in the system one is examining. In the years we found a strong statistical theory to extrapolate results taking into account these sources.
Systematic uncertainties, on the other hand, arises from uncertainties associated with the nature of the detector itself or from the model that one uses to describe a physical phenomena. These uncertainties are usually correlated and we have a limited and incomplete theoretical framework in which we can interpret and accommodate these uncertainties in inference or hypothesis testing. Common sources of systematic uncertainties arises from the calibration or the resolution of the detector, an example of this are the uncertainties on the jet energy scale/resolution (JES and JER) that arise from how the information of the jets are extrapolated from the different subdetectors. The effect of the systematic uncertainties could be mainly of tow types: the uncertainty can affect

the only the final yield of the process, leading to a flat fluctuation up and down of this quantity and in this case the uncertainty follow a log normal distribution, or can affect the kinematic of the process itself, this is the case of JES and JER in which the up or down variation changes the momentum of the jet. In the latter case is not possible to accounted for this uncertainty with a 'flat' up or down variation, but the uncertainty affects the shape of the distributions bin by bin in a non uniform way (that one should calculate case by case rerunning and retuning the analysis for the different up and down variation of the uncertainty); we treat this uncertainties as a shape uncertainty.
In the case of CMS there is a strict policy on how to treat these uncertainties, in particular if we focus again on JES and JER, we end up with 11 sources of uncertainty some of them correlated or not to each other. In this analysis we follow the correlation scheme recommended by the collaboration.
Another source of systematic uncertainties are the theory uncertainties on how to treat the simulated processes. Usually we have a QCD scale uncertainty that is give by the QCD loops involved in the processes and the PDF uncertainty that is instead given from the choice of the PDF for a given process. More detail will be given in the next sections.

4.7.1 Experimental Uncertainties

There are different experimental systematic uncertainties sources that affect both signal and background processes:

- the uncertainty on the integrated luminosity, that results in a lnN contribution;
- the lepton identification and reconstruction efficiency, that results in a lnN contribution;
- b tagging scale factors related systematics, that result in 18 shape contributions due to the hadronic composition of the jet, the jet energy and the jet η;
- jet energy scale and jet energy resolution. This systematic uncertainties are computed by propagating the up and down variation of the jet energy through the event reconstruction chain up to the BDT; this uncertainty results in a shape contribution;

The uncertainty on the reducible background estimation Z+X (Section 4.5), is computed as explained in Section 4.5 and it results in a lnN contribution. The summary of experimental systematic uncertainties is reported in Table 4.20.

Experimental uncertainties			
type	2016	2017	2018
Luminosity	2.6%	2.3%	2.5%
Leptons ID and reco eff	$1.6 - 15.5\%$	$1.1 - 12.1\%$	$1.0 - 11\%$
b tagging SF	shape	shape	shape
Jet energy scale	shape	shape	shape
Jet energy resolution	shape	shape	shape
Z+X uncertainties	$30 - 41\%$	$30 - 38\%$	$30 - 37\%$

Table 4.20: Summary of experimental systematic uncertainties.

4.7.2 Theoretical Uncertainties

Theoretical uncertainties sources are the choice of PDF set, the uncertainty on α_s, the renormalization and factorization QCD scale. These uncertainties affect both signal and background processes. For the HH signal, in addition to the uncertainty sources just described, also an uncertainty related to missing finite top-quark mass effects gives a contribution. Values for the systematic uncertainties affecting the HH signal are taken from [72]. For the single Higgs backgrounds, the values of the systematic uncertainties are taken from [73]. For all the other backgrounds, the values of systematic uncertainties are computed by varying the QCD scale and the PDF set used for computing the sample cross section. An additional uncertainty of 10% on the k-factor is used for the gg $\rightarrow$ ZZ prediction and of 0.1% for the qq $\rightarrow$ ZZ prediction. The summary of theory systematic uncertainty is reported in Table 4.21. All experimental uncertainties are considered uncorrelated while all theoretical uncertainties are considered correlated among the three years.

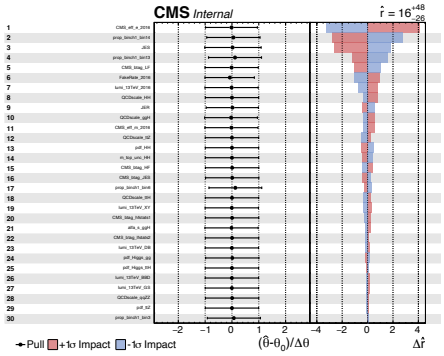
4.7.3 Impact of systematic uncertainties

The impacts of the systematic uncertainties for the three years of data taking are shown in Fig 4.25.
As the figure shows, the most important source of systematic uncertanty is the electron ID and reco efficiency (about 12%), followed by JES. The contribution

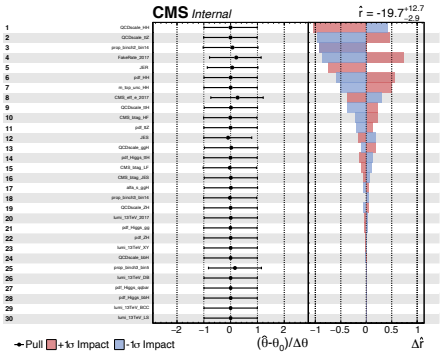
Theory uncertainties	
PDF set and α_s HH $\rightarrow 4\ell$b	3.0%
QCD scale HH $\rightarrow 4\ell$b	$2.2 - 5\%$
m_{top} unc HH $\rightarrow 4\ell$b	2.6%
PDF set ggH	1.8%
α_s ggH	$2.59 - 2.62\%$
QCD scale ggH	$4.27 - 6.49\%$
PDF set and α_s VBFH	2.1%
QCD scale VBFH	$0.3 - 0.4\%$
PDF set and α_s ZH	1.6%
QCD scale ZH	$2.7 - 3.5\%$
PDF set and α_s WH	1.3%
QCD scale WH	0.5%
PDF set and α_s bbH	3.2%
QCD scale bbH	$4.6 - 6.7\%$
PDF set and α_s ttH	3.6%
QCD scale ttH	$6.0 - 9.2\%$
PDF set and α_s qqZZ	$3.1 - 3.4\%$
QCD scale qqZZ	$3.2 - 4.2\%$
Electroweak correction qqZZ	0.1%
PDF set and α_s ttW	$25 - 37.5\%$
QCD scale ttW	$3 - 4\%$
PDF set and α_s ttZ	$7 - 14\%$
QCD scale ttZ	$2 - 3\%$
PDF set and α_s VVV	$2 - 17\%$
QCD scale VVV	3%
PDF set and α_s ggZZ	3.2%
QCD scale ggZZ	$4.6 - 6.7\%$
Electroweak correction ggZZ	10.0%

Table 4.21: Summary of theory systematic uncertainties.

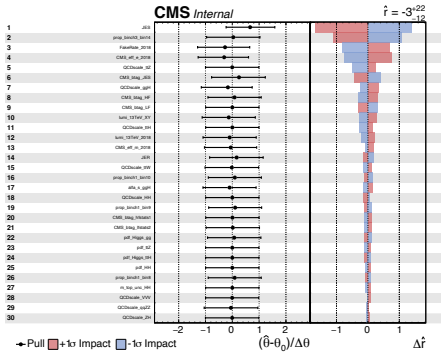
of b tagging scale factors is less important expecially in 2017 dataset. The pulls distributions look fine for all the three years.



(a) Impacts for 2016 dataset



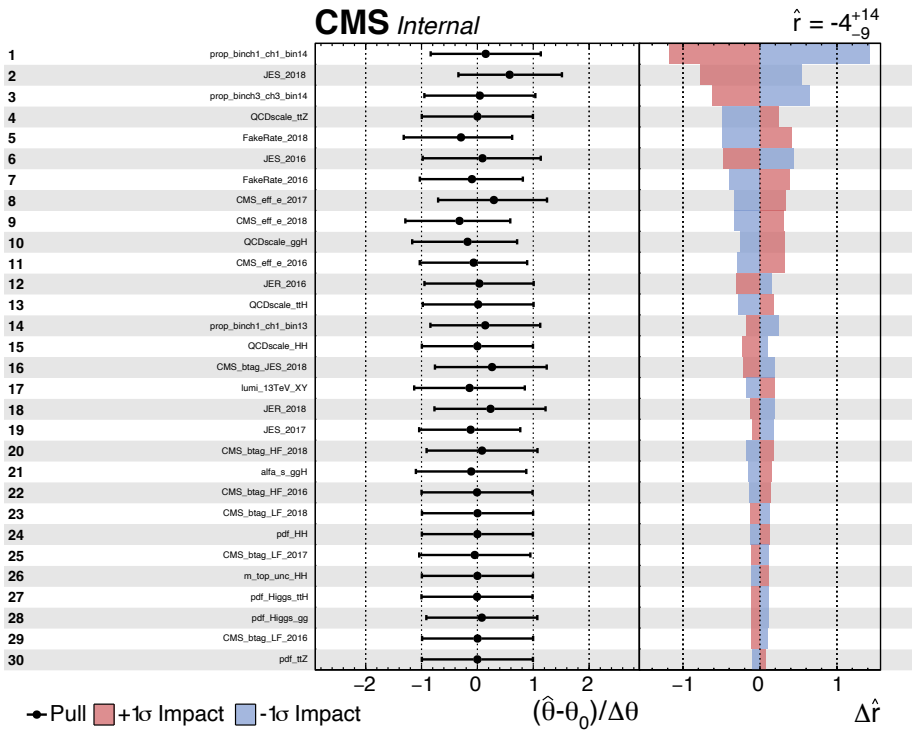
(b) Impacts for 2017 dataset



(c) Impacts for 2018 dataset

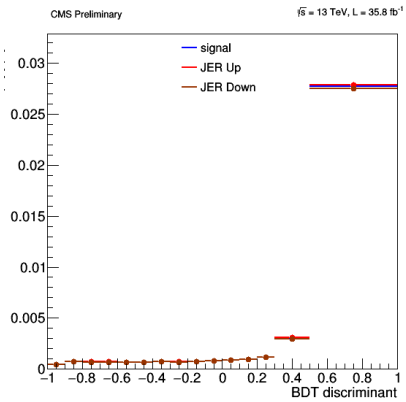
Figure 4.25: Impacts for full run2 analysis. The most important source of systematic uncertanty is the electron ID and reco efficiency(about 12%), followed by JES. The contribution of b tagging scale factors is less important expecially in 2017 dataset. The pulls distributions look fine for all the three years.

In Fig 4.26 the impacts combined for the full run2 datasets are shown. The most import source of systematic uncertanty is the statistical uncertainty on the last bin of BDT, this is due to the low statistic in the SR; JES and electron ID and reco efficiency followed in the ranking. The pulls distributions look fine, as shown in Fig 4.25.

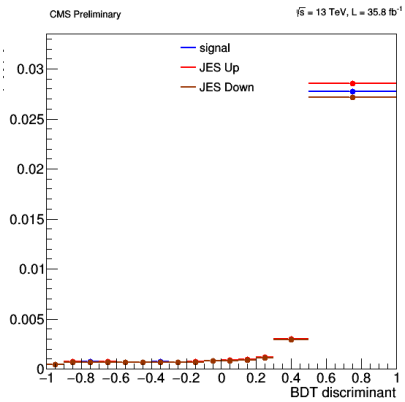


(a) Impacts for full run2 combination

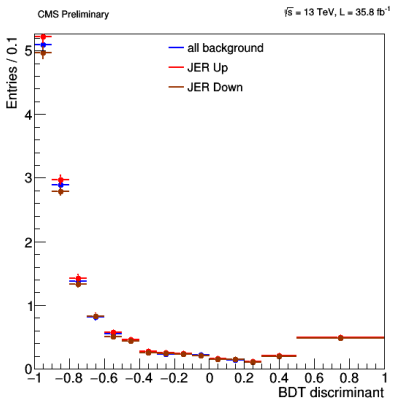
Figure 4.26: Impacts combined for the full run2 datasets. The most important source of systematic uncertantyis the statistical uncertainty on the last bin of BDT, this is due to the low statisticin the SR; JES and electron ID and reco efficiency followed in the ranking. The pulls distributions look fine.



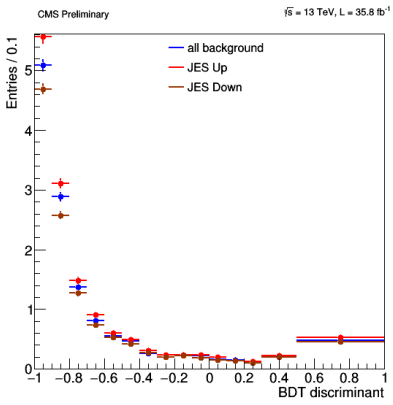
(a) JER for 2016 signal dataset



(b) JES for 2016 signal dataset

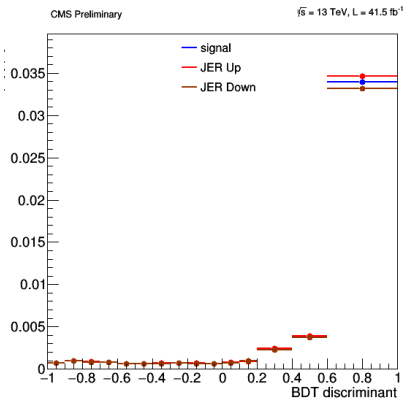


(c) JER for 2016 all background dataset in SR

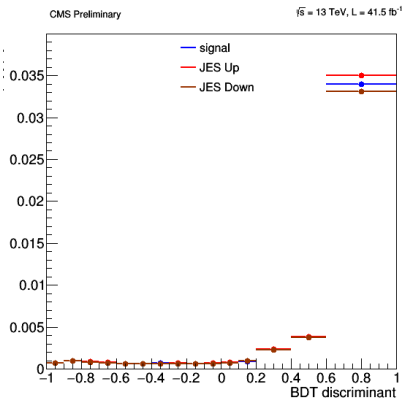


(d) JES for 2016 all background dataset in SR

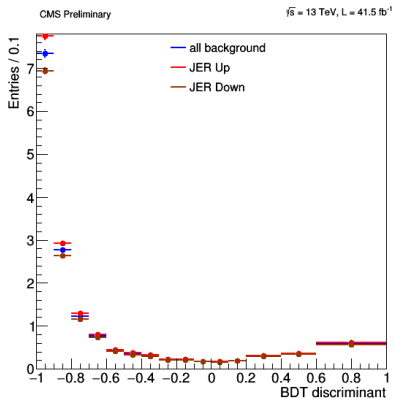
Figure 4.27: Distribution of signal and all background merged together (ggH, VBF, ttH, ttZ, VV, VVV, VH where V = Z, W) of JES and JER up and down variation for 2016 dataset



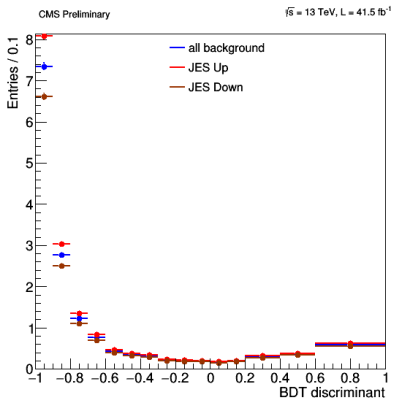
(a) JER for 2017 signal dataset



(b) JES for 2017 signal dataset



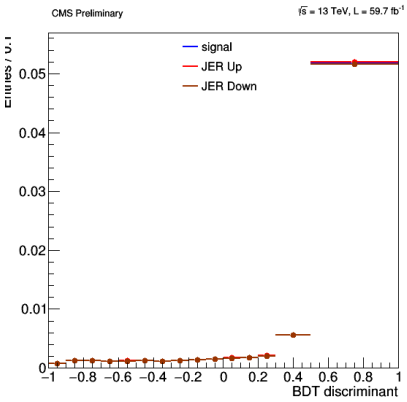
(c) JER for 2017 all background dataset in SR



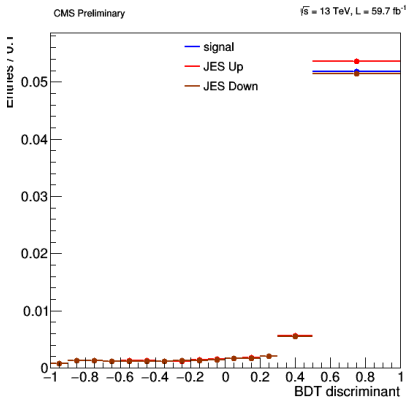
(d) JES for 2017 all background dataset in SR

Figure 4.28: Distribution of signal and all background merged together (ggH, VBF, ttH, ttZ, VV, VVV, VH where V = Z, W) of JES and JER up and down variation for 2017 dataset

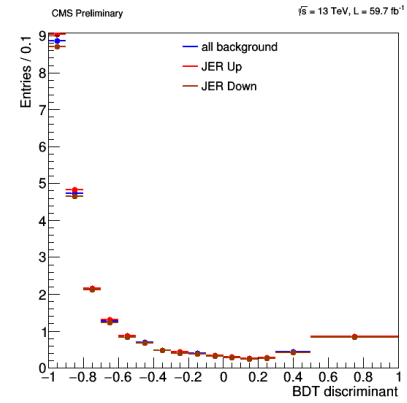
Distribution of signal and all background merged together (ggH, VBF, ttH, ttZ, VV, VVV, VH where V = Z, W) of JES and JER up and down variation are shown in Fig 4.27 - 4.28 - 4.29 for the three years.



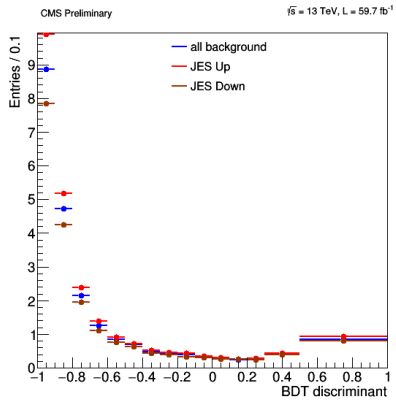
(a) JER for 2018 signal dataset



(b) JES for 2018 signal dataset



(c) JER for 2018 all background dataset in SR



(d) JES for 2018 all background dataset in SR

Figure 4.29: Distribution of signal and all background merged together (ggH, VBF, ttH, ttZ, VV, VVV, VH where V = Z, W) of JES and JER up and down variation for 2018 dataset

4.8 Yields and distribution

In this section we report the yields and distribution of the main kinematic variables in data and MC samples after the full selection (4 leptons + at least two isolated jets with respect to the leptons).

4.8.1 Signal Region Yields

The expected yields of the backgrounds and the HH signal in the $4\ell b\bar{b}$ final state for the three years are presented in Table 4.22 - 4.23 - 4.24, after the full event selection:

- H $\rightarrow 4\ell$ full selection
- $|m_{4\ell} - 125| < 10$ GeV
- at least two isolated jets (with respect to the leptons) in the event

Topology	Signal	ttZ	ttH	bbH	ZZ	Higgs+VBF	ZH	WH	others	Z+X	all bkg	all data
4μ	0.013	0.14	0.17	0.04	1.01	3.02	0.17	0.21	0.008	0.79	5.57	6
4e	0.007	0.07	0.09	0.02	0.38	1.61	0.10	0.11	0.01	1.40	3.78	3
$2e2\mu$	0.019	0.16	0.22	0.05	1.08	3.95	0.24	0.27	0.02	2.64	8.65	11

Table 4.22: Cut flow table after signal region selection for the 2016 dataset (35.8 fb^{-1}). The ZZ yields contain contributions from the processes $gg \rightarrow ZZ$ and $qq \rightarrow ZZ$, Higgs+VBF contains contributions from the single Higgs boson gluon fusion (ggH) and vector boson fusion (VBFH) production modes, others contains contributions from TTW, WWZ, WZZ and ZZZ, processes.

Topology	Signal	ttZ	ttH	bbH	ZZ	Higgs+VBF	ZH	WH	others	Z+X	all bkg	all data
4μ	0.016	0.18	0.20	0.05	1.06	3.73	0.21	0.27	0.01	1.48	7.21	9
4e	0.008	0.12	0.12	0.02	0.40	1.76	0.11	0.14	0.005	0.52	3.20	2
$2e2\mu$	0.024	0.30	0.27	0.06	1.09	4.57	0.29	0.35	0.04	2.00	9	6

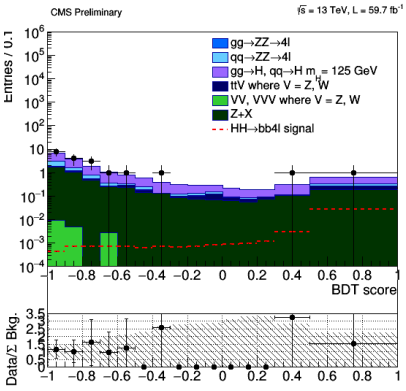
Table 4.23: Cut flow table after signal region selection for the 2017 dataset (41.5 fb^{-1}). The ZZ yields contain contributions from the processes $gg \rightarrow ZZ$ and $qq \rightarrow ZZ$, Higgs+VBF contains contributions from the single Higgs boson gluo fusion (ggH) and vector boson fusion (VBFH) production modes, others contains contributions from TTW, WWZ, WZZ and ZZZ processes.

The distribution of the BDT discriminat in the signal region for the three years is shown in Figure 4.30. The binning of the BDT is choosen in order to have (almost) the same statistical uncertainties in the last 4-5 bins, to flatten the statistical fluctuation that could lead to undesired behaviour in the impacts and in the final limits.

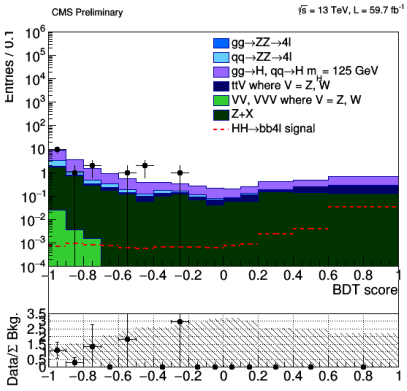
In Fig 4.34 the $m(4\ell)$, $m()$ and BDT score for the full Run2 datasets are shown, the data / mc agreement is good.

Topology	Signal	ttZ	ttH	bbH	ZZ	Higgs+VBF	ZH	WH	others	Z+X	all bkg	all data
4μ	0.025	0.24	0.30	0.07	1.65	5.14	0.32	0.40	0.02	1.60	9.76	14
4e	0.013	0.16	0.17	0.03	0.65	2.56	0.16	0.20	0.003	0.72	4.66	7
$2e2\mu$	0.036	0.46	0.39	0.08	1.68	6.55	0.41	0.50	0.05	2.58	12.73	16

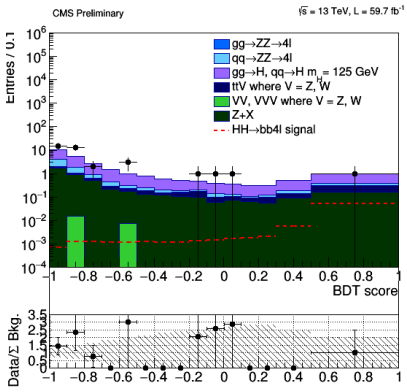
Table 4.24: Cut flow table after signal region selection for the 2018 dataset (59.7 fb^{-1}). The ZZ yields contain contributions from the processes $gg \rightarrow ZZ$ and $qq \rightarrow ZZ$, Higgs+VBF contains contributions from the single Higgs boson gluon fusion (ggH) and vector boson fusion (VBFH) production modes, others contains contributions from TTW, WWZ, WZZ and ZZZ processes.



(a) BDT discriminant in the signal region for 2016 dataset



(b) BDT discriminant in the signal region for 2017 dataset



(c) BDT discriminant in the signal region for 2018 dataset

Figure 4.30: BDT discriminant in the signal region for the three years

A good signal / background separation is observed for the three years, in particular $S/\sqrt{B}$ (where S is the signal and B is the background bin per bin) is higher in the last three bins, this means that the majority of the signal is concentrated at between 0.6 and 1, as expected for a well trained network.

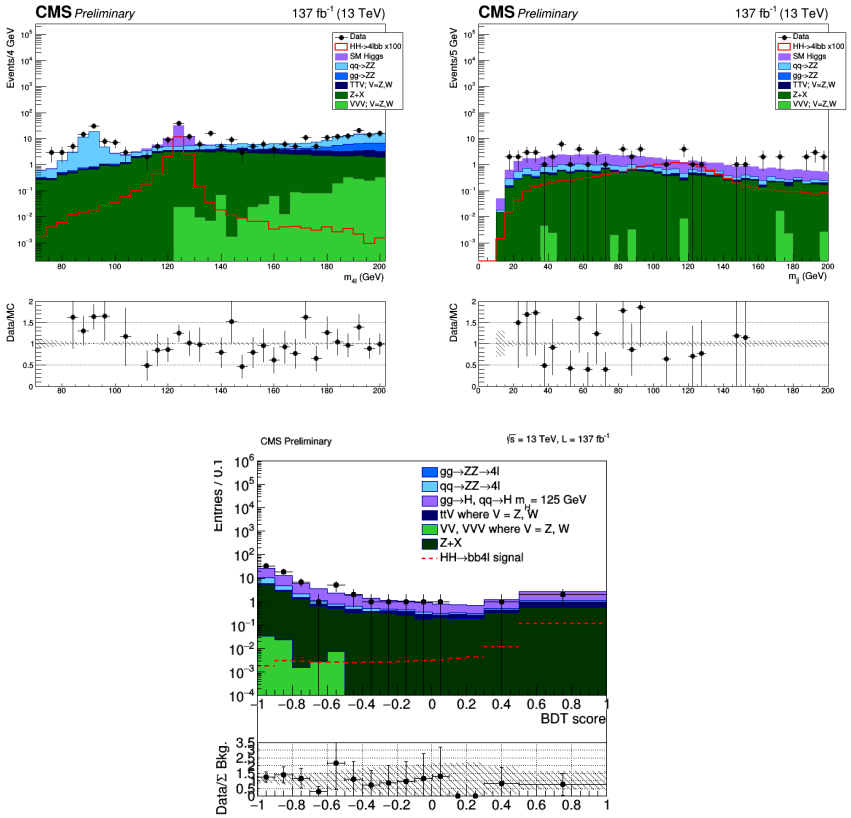


Figure 4.31: $m(4\ell)$ (left), $m(\bar{b}b)$ (right, considering only events in the mass window: 115 GeV $< m(4l) <$ 135 GeV) distributions and BDT score (bottom) in signal (scaled by an illustrative factor), estimated background components, and data. A good signal / background separation is observed for the three years, in particular $S\sqrt{B}$ (where S is the signal and B is the background bin per bin) is higher in the last three bins, this means that the majority of the signal is concentrated at between 0.6 and 1, as expected for a well trained network.

4.9 Results

A multi-dimesional fit is performed, using the shape of the BDT response, in order to extract the significance for the SM HH signal and the upper limit at the 95% confidence level (CL) on the signal strenght, $\mu = \sigma_{\mathrm{HH}}/\sigma_{\mathrm{HH}}^{\mathrm{SM}}$. The BDT response is taken in the signal region:

- H $\to 4\ell$ full selection
- $|m_{4\ell} - 125| < 10$ GeV
- at least two isolated jets (with respect to the leptons) in the event

For the Full Run 2 dataset ($137\mathrm{fb}^{-1}$) the expected yields are reported in Table 4.25.

Topology	Signal	ttZ	ttH	ZZ	Higgs+VBF	ZH	WH	others	Z+X	all bkg	all data
4μ	0.054	0.56	0.67	3.72	11.89	0.70	0.88	0.04	3.87	22.38	29
4e	0.028	0.35	0.38	1.43	5.93	0.37	0.45	0.02	2.64	11.60	12
2e2μ	0.079	0.92	0.88	4.93	15.07	0.94	1.12	0.11	7.22	31.36	33

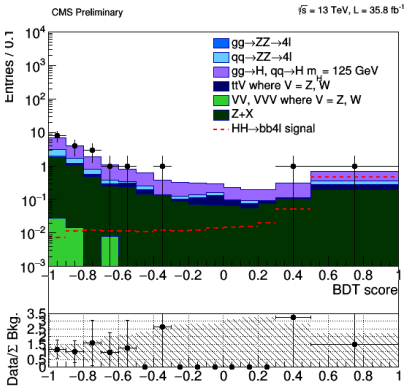
Table 4.25: Cut flow table after signal region selection. The ZZ yields contain contributions from the processes $gg \to ZZ$ and $qq \to ZZ$, Higgs+VBF contains contributions from the single Higgs boson gluon fusion (ggH) and vector boson fusion (VBFH) production modes, others contains contributions from TTW, WWZ, WZZ and ZZZ processes.

The postfit distribution of the BDT is shown in Fig 4.32 for each of the three years, and in Fig 4.33 for the full run2 combination.

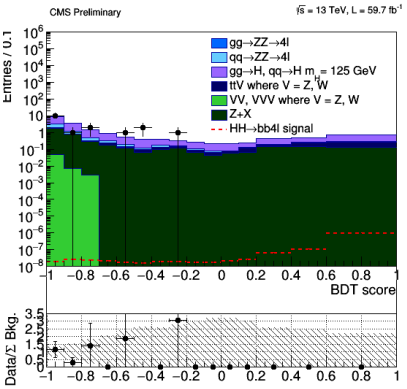
The analysis of the signal $HH \to \bar{b}b4l$ done with a shape analysis using the full Run2 datasets leads to an expected upper limit on the signal strength: $r < 38$ @ 95 % CL_s and an observed upper limit on the signal strenght: $r < 32$ @ 95 % CL_s as shown in Fig 4.34 - Tab 4.26.
The observed and the expected results are compatible in within 1 σ uncertainty.

4.9.1 k_λ scan

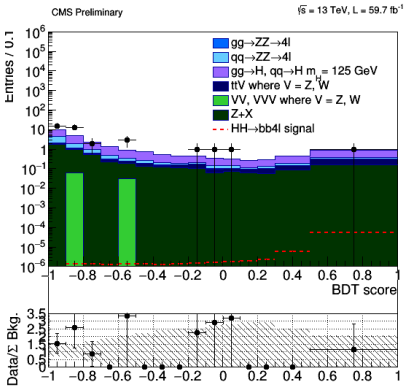
Under the assumption that no HH signal exists, $HH \to b\bar{b}4l$ CLs upper limits on the HH production cross section are derived as a function of $k_\lambda =$



(a) BDT discriminat in the signal region for 2016 dataset



(b) BDT discriminat in the signal region for 2017 dataset

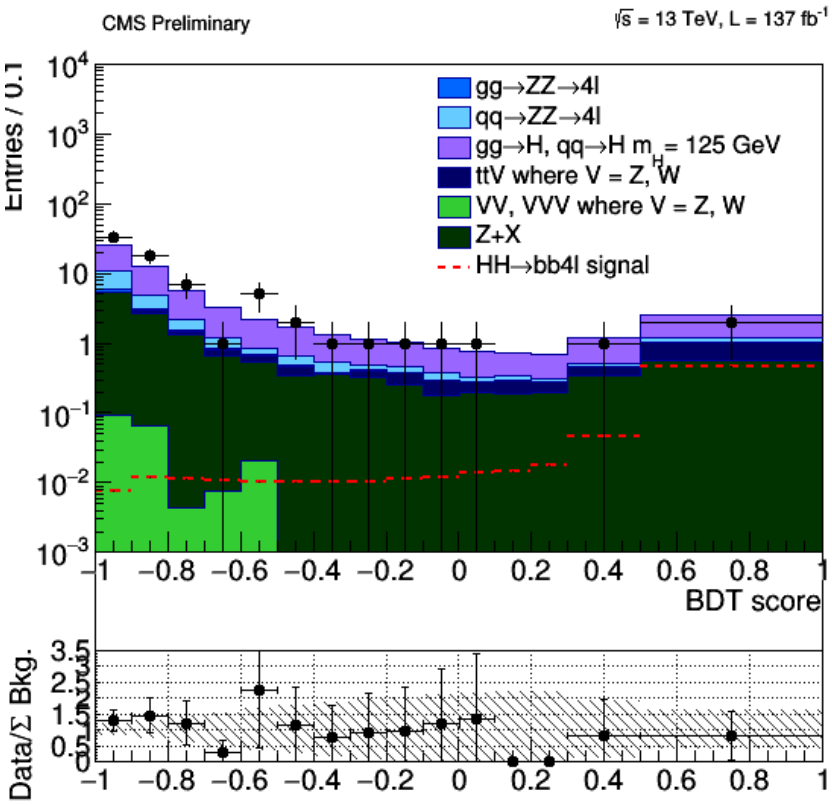


(c) BDT discriminat in the signal region for 2018 dataset

Figure 4.32: BDT discriminat in the signal region for the three years, postfit distribution

	2016	2017	2018	Combination
μ Observed	126	60	57	32
$\mu + 2\sigma$	44	38	27	18
$\mu + 1\sigma$	66	56	39	25
μ	105	90	62	38
μ - 1σ	175	149	101	59
μ - 2σ	280	239	159	90

Table 4.26: Table of limits.



(a) BDT discriminat full run2

Figure 4.33: BDT discriminat in the signal region for the full run2, postfit distribution

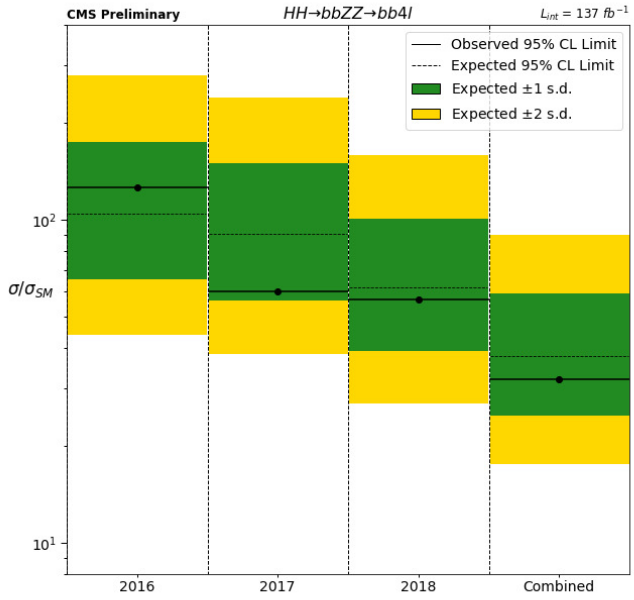
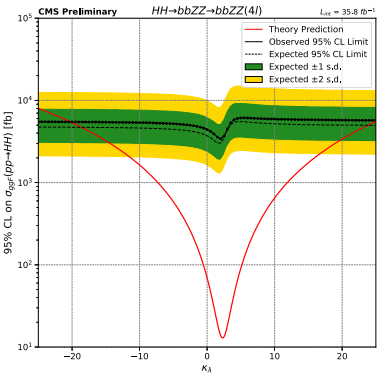
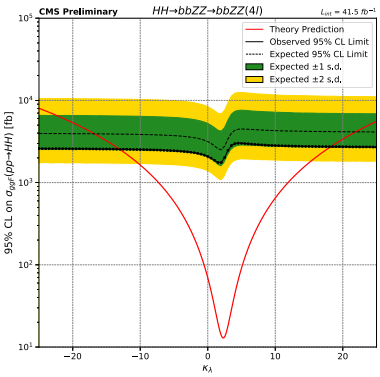


Figure 4.34: Upper limit on the signal strength for each dataset and for the full Run2 combination.

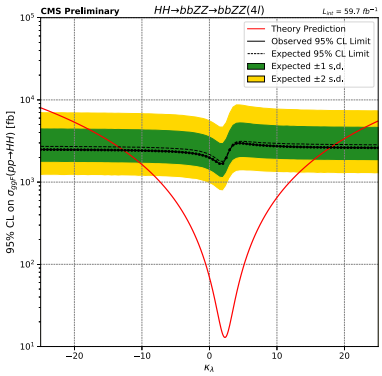
$\lambda_{\mathrm{HHH}}/\lambda_{\mathrm{SM}}$, where λ_{SM} denotes the SM prediction and λ_{HHH} denotes the measured value.
k_λ scan is performed for the full run2 dataset in the k_λ range [-20, 20] and the result is reported in Fig. 4.35 - 4.36. The red line in these plots represents the theoretical value of the cross section. All these results are obtained in the hypothesis where all the other possible couplings are assumed to be equal to the SM values.The analysis of the signal $HH \rightarrow \bar{b}b4l$ done with a shape analysis using the full Run2 is able to set limits on the trilinear Higgs boson coupling to the values between $-9 < k_\lambda < 14.5$ at 95% CL.



(a) Expected 95% CLs upper limit on the cross section as a function of k_λ for 2016 year

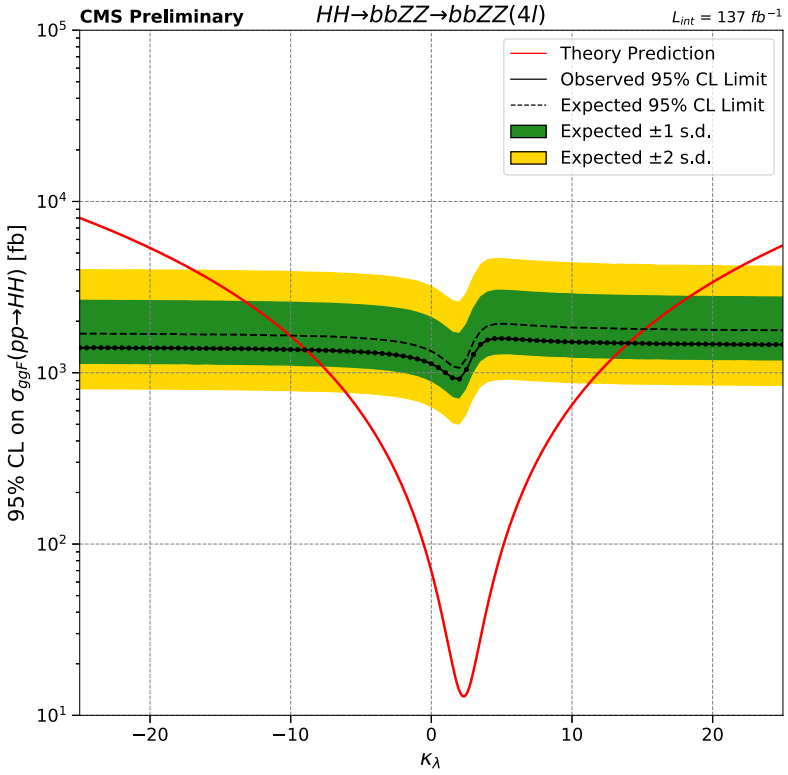


(b) Expected 95% CLs upper limit on the cross section as a function of k_λ for 2017 year



(c) Expected 95% CLs upper limit on the cross section as a function of k_λ for 2018 year

Figure 4.35: $HH \rightarrow b\bar{b}4l$ CLs upper limits on the HH production cross section derived as a function of $k_\lambda = \lambda_{\mathrm{HHH}}/\lambda_{\mathrm{SM}}$. The red line represents the theoretical value of the cross section.



(a) Expected 95% CLs upper limit on the cross section as a function of k_λ for the full run2

Figure 4.36: $HH \to b\bar{b}4l$ CLs upper limits on the HH production cross section derived as a function of $k_\lambda = \lambda_{\mathrm{HHH}}/\lambda_{\mathrm{SM}}$. The red line represents the theoretical value of the cross section.

4.10 Conclusion

Studying the production of di Higgs pairs is an important test for the standard model (SM) electroweak symmetry breaking (EWSB) sector, because it allows the extraction of Higgs self-coupling (λ_{HHH}), directly related to the potential structure of the Higgs field. Furthermore, any possible deviation in the Higgs boson self-coupling due to the beyond SM (BSM) effect could open the door for further study of the new physics scenario and provide important tests for SM validity. In fact, many BSM theories of the predict a higher di-Higgs production cross-section with respect to the one predicted by the SM, through the generation of a high mass resonance, which then decays into a pair of Higgs bosons. Hence, a parametrization of an anomalous coupling, $\lambda_{HHH} = k_\lambda \lambda_{SM}$ has been introduced, where k_λ is called self-coupling modifier. The challenge of these searches is undoubtedly the extremely tiny cross section (around 31 fb^{-1}), that results from the interference of the box and triangle production modes of this process Fig 4.37.

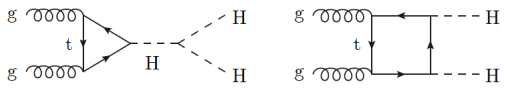


Figure 4.37: Gluon gluon fusion di Higgs production

A wide variety of HH decay channels and their combination were studied by the ATLAS [74–77] and CMS [78–84] Collaborations. The most recent published results on the HH combination [85, 86], with a data set corresponding to an integrated luminosity of about 36 fb^{-1}, set an observed (expected) upper limit at the 95% confidence level (CL) on the SM production cross section of 7 (10) and 22 (13) times the theoretical prediction respectively.
The analysis presented in this thesis focuses on HH pairs where one H decays to a Z boson pair, which in turn decays into 4l, and the other to a pair of b quarks, hadronizing into jets (bb). The final state forms a clear signature granted by the presence of the four leptons, while the high branching fraction of the bb decay channel partially compensates for the small branching fraction of the 4l channel.
The analysis flow starts from the reconstruction of the $H \to ZZ \to 4l$, requiring 4 cleaned and well defined leptons as explained in Sec 4.4. The $H \to \bar{b}b$ decay is taken into account by requiring the presence of two or more isolated

jets in each event (see Sec 4.3.4). Three categories, depending on the lepton flavor are studied: $\bar{b}b4\mu$, $\bar{b}b4e$ and $\bar{b}b2e2\mu$. The signal region is defined in a mass window of 10 GeV around the Higgs peak (see Sec 4.4). The treatment of non-prompt backgrounds, originated by hard processes that lead to non-prompt leptons, is derived from data as described in Sec 4.5.
A BDT is trained to disentangle signal versus background events. The variables fed in the BDT capture the full kinematic of the event (see Sec 4.6).
The signal extraction, described in Sec 4.9, relies on the shape of the BDT for each process; An observed (expected) upper limit on the HH production cross section is set at 32 (38) times the SM expected rate at the 95% confidence level. Possible modifications of the SM H trilinear coupling are investigated and constrained to be within the observed (expected) range -9 (-10) < k_λ < 14 (15.0) at 95% confidence level. This is the first analysis in CMS that investigates such peculiar signature and the results are complementary with those obtained by more sensitive channels Ref [78, 79].

4.11 Outlook

The High-Luminosity LHC (HL-LHC) will provide a unique opportunity to study HH production as predicted in the SM and identify possible deviations induced by BSM physics in the signal cross section or properties. Upgrades of the LHC machine will increase the peak instantaneous luminosity to 5-7.5 × 1034 $cm^{-2}s^{-1}$ and the CMS experiment will collect more than 3000 fb^{-1} over a decade of operation. The high instantaneous luminosity will lead to 140 to 200 additional interactions per bunch crossing. This pileup will constitute a formidable challenge for the experiment both in terms of event reconstruction and radiation damage. A comprehensive detector upgrade program is under development to maintain and improve the detector performance under these challenging conditions.
In the context of this upgrade, projections of the HH production have been done by the CMS collaboration: the study is performed using five decay channels of the HH system to 4b, $b\bar{b}\tau\tau$, $b\bar{b}WW$ (with both W decaying leptonically), $b\bar{b}\gamma\gamma$, and $b\bar{b}ZZ$ (with both Z decaying to a pair of electrons or muons). Assuming that a HH signal exists with the properties predicted by the SM, we expect a combined significance of 2.6σ and a determination of the λ_{HHH} coupling

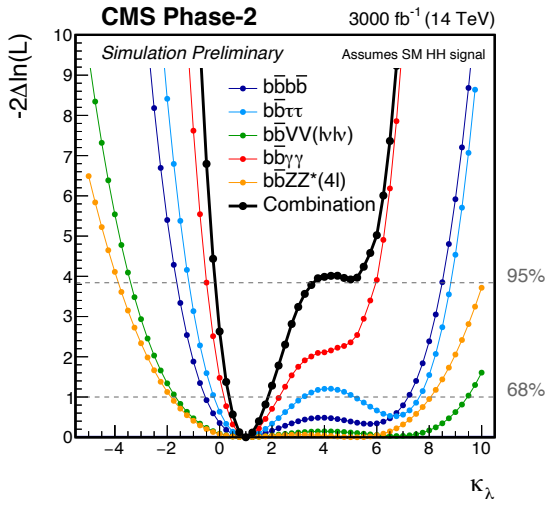


Figure 4.38: Expected likelihood scan as a function of κ_λ. The functions are shown separately for the three decay channels studied and for their combination [87]

corresponding to the interval [0.35, 1.9] at the 68% CL and to [0.18, 3.6] at the 95% CL. Results are shown in Fig 4.38 .

The peculiar HH likelihood function structure, characterised by a double minimum can be observed. This shape is explained recalling that the HH cross section has a the quadratic dependence on the k_λ with a minimum at $k_\lambda = 2.4$, that corresponds to the maximum interference of the box and triangle Feynman diagram of the HH production. Moreover, the kinematic differences for signals with κ_λ values symmetric around this minimum are relevant in the low region of m_{HH} spectrum. Consequently, a partial degeneracy can be observed between the $k_\lambda = 1$ value and a second k_λ value. The exact position and the height of this second minimum depends on the sensitivity of the analysis to the m_{HH} spectrum.

Speaking about the prediction for $HH \rightarrow b\bar{b}4l$ channel [87], the analysis techniques is slightly different with respect to the analysis that is done with the full Run2 dataset. The expected upgrade of the CMS detector (see Sec 2.5), will allow us to lowering the cuts for electrons and muon p_T, thus in this context, events are required to have at least four identified and isolated (isolation 0.7) muons (electrons) with p_T < 5 (7) GeV and $|\eta| < 2.8$, where muons (electrons) are selected if passing the Loose (Medium) Working Point identification. Furthermore, the presence of the Higgs decaying into $b\bar{b}$ is taken into account requiring least two (but not more than three) identified b jets with a B-Tag Medium working point. We have to recall that in the run2 analysis the

information of the b tagging is given as input in the BDT thus no further cut is applied in the analysis. Finally the signal extraction is done by fitting the invariant mass spectrum of the four leptons after the full event selection that is shown in Fig 4.39.

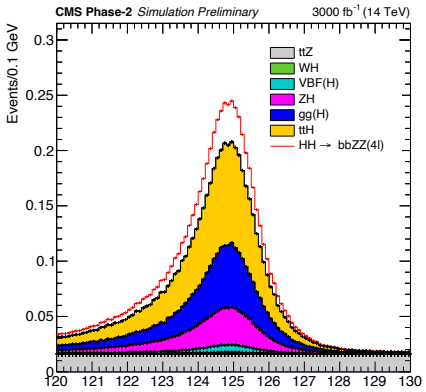


Figure 4.39: Invariant mass distribution of the four leptons final state.

The impact of the systematic uncertainties is computed taking into account the increasing of the luminosity and the upgrade of the CMS detector.
Considering both systematic and statistical uncertainties, the 95% CL limit on the signal strength μ is 6.6, and the correspondent signal sensitivity is 0.37. To evaluate the sensitivity of the analysis on the measurement of the Higgs self-coupling modifier at HL-LHC, a scan is performed for different values of k_λ. In order to model anomalous k_λ signals, the signal yields for various k_λ samples are fitted after the full selection by a quadratic function. The negative log-likelihood on the self-coupling modifier k_λ is shown in Figure 4.40 for the inclusive $b\bar{b}4l$ final state. The projected confidence intervals on the Higgs self-coupling for the $b\bar{b}4l$ final state corresponds to [-3.9, +9.9] at 95% CL, considering both systematic and statistical uncertainties.

The results obtained by the projection are compatible with the results obtained by the full run2 analysis.

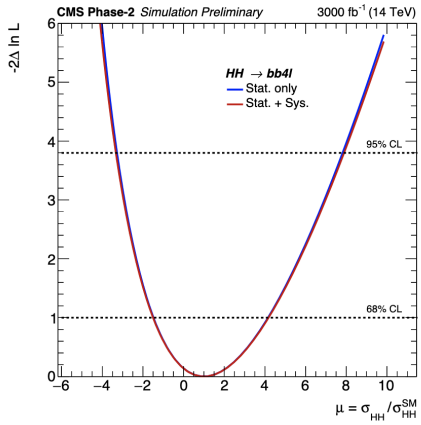


Figure 4.40: Projected signal strength uncertainty in two different scenarios (with and without systematics) for the $b\bar{b}4l$ final state

Chapter 5

HNL analysis

5.1 Introduction

I am currently working in a Heavy Neutral Lepton search that looks for right-handed neutrinos with Majorana masses below the electroweak scale. A more clear overview of the theoretical motivation for this analysis is given in Sec 1.6.
The addition of right-handed neutrinos is able to generate both the light neutrino masses and the baryon asymmetry of the universe via low scale leptogenesis. In the history, lots of different phase space of masses and life time were studied to find evidence of such peculiar particles.

Searches for HNLs have been conducted by several experiments exploring a mass range from few to several hundred GeV [16]. Figure 5.1 shows the current limits on the HNL mixing parameters with three lepton families (denoted as $|\mathrm{V}_{\mathrm{N}\ell}|^2$) and their masses ($m_N$), as well as the projected sensitivity that can be achieved in planned future facilities.

HNLs are singlets in each of the SM gauge groups, which means they can't interact with the SM particle through electroweak or strong interactions. They can, however, mix with SM neutrinos, and this mixing may be used at the LHC to look for such states. The search described hereafter will probe the direct production of HNLs in the decays of W bosons, where the SM neutrino oscillates into a HNL, and the HNL afterwards decay into a W boson and a charged lepton. In this analysis hadronic decays of W boson are considered, leading to a final state composed by two leptons and two quarks (which will fragment into hadrons or lead to jets).

A Feynman diagram of the process considered is shown in Figure 5.2. In case the HNL is a Majorana particle, ℓ_1 and ℓ_2 can either have the same chirality, lepton-number violation (LNV) decay, or opposite chirality, lepton number conservation (LNC) decay. A LNV decay can lead to final states with

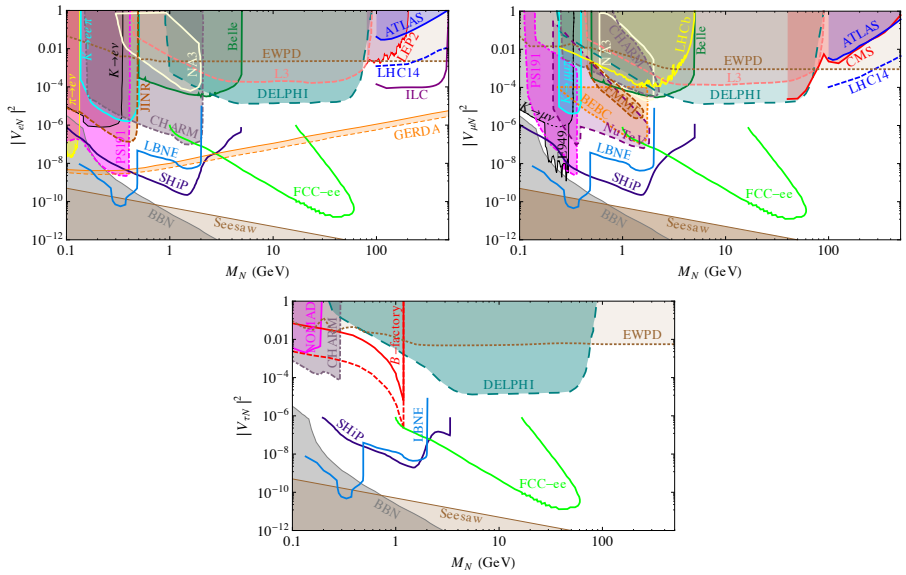


Figure 5.1: Current limits on the HNL masses and their couplings to the three lepton families, and projected sensitivity achievable in planned future facilities [16].

no opposite-sign, same-flavour lepton pairs (no-OSSF), such as $e^{\pm}e^{\pm}q\bar{q}'$ or $\mu^{\pm}\mu^{\pm}q\bar{q}'$. Such final states have relatively low SM background rates, providing a characterizing signature for the performed HNL search.

The HNL can couple exclusively to a single lepton-neutrino family or to multiple families. In the former case, only one of $|V_{eN}|$, $|V|$, or $|V|$ is nonzero and ℓ_1 and ℓ_2 always belong to the same lepton generation, conserving the lepton flavour (LFC). In the latter case instead, at least two of $|V_{eN}|$, $|V|$, and

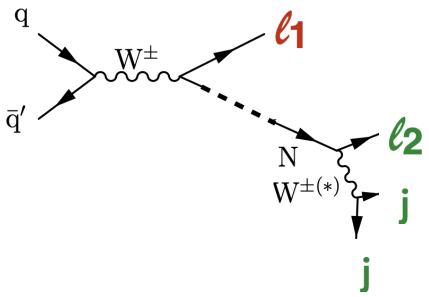


Figure 5.2: Typical diagrams for the production of a HNL at the LHC (N) through its mixing with a SM neutrino, leading to a final state with two charged leptons and two jets.

$|V|$ are nonzero at the same time and the lepton flavour can be violated (LFV). In this search, both LFC and LFV cases are considered.

The lifetime of a HNL is inverse proportional to m_N and $|\mathrm{V_{N\ell}}|^2$: $\tau_\mathrm{N} \propto \mathrm{m_N^{-5}}|\mathrm{V_{N\ell}}|^{-2}$. This means that HNLs with masses less than roughly 20 GeV can have long lifetime, so their kinematics and acceptance can be affected: its decay products arise from a secondary vertex that is spatially separated from the process's primary vertex and hence identifiable from it.

The production rates of HNLs depend on their mass m_N and on the squared mixing parameter $|\mathrm{V_{N\ell}}|^2$. The final results of this analysis will be presented as a function of both m_N and $|\mathrm{V_{N\ell}}|^2$, for each flavour ℓ separately.

There have been several searches for HNLs in CMS, ATLAS and LHCb. The CMS experiment reported on a search for HNLs using events with two same-sign leptons and at least one jet is searched for using data collected during 2016 in proton-proton collisions at a center-of-mass energy of 13 TeV and corresponding to an integrated luminosity of 35.9 fb^{-1}.
The limits shown in Fig 5.3 are written as function of the mass of the HNL and of the coupling with leptons and are the most restrictive direct limits up to now [17].

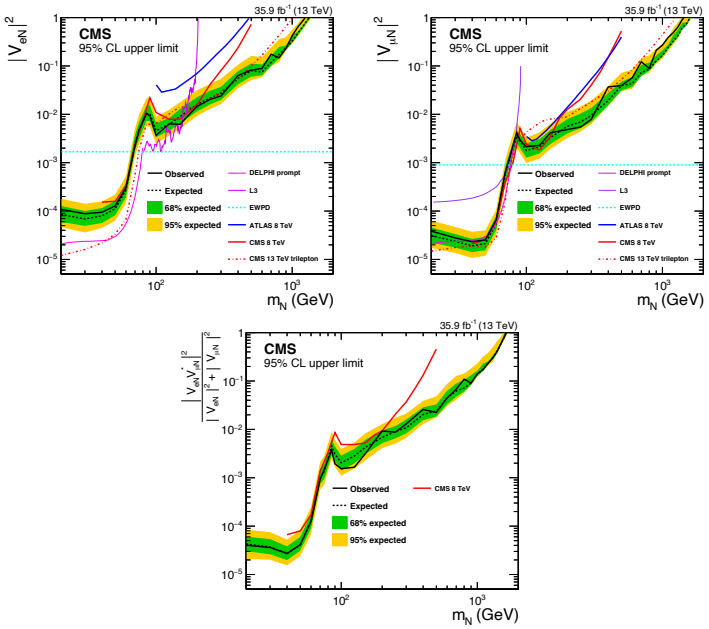
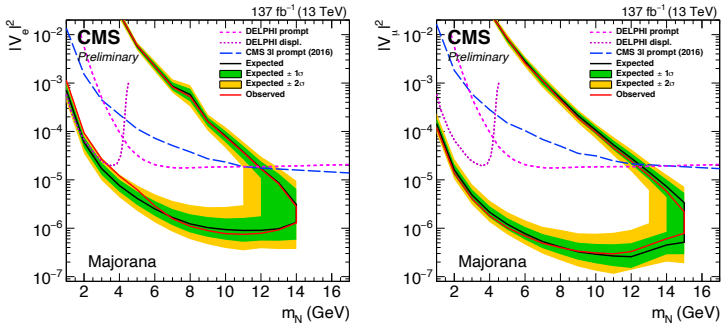


Figure 5.3: Exclusion region at 95% CL in the $|V_{eN}|$ (top-left), $|V|$ (top-right) and, $|V_{\mathrm{N}e}V^{*}_{\mathrm{N}\mu}|^2/(|V_{\mathrm{N}}|^2 + |V_{\mathrm{N}}|^2)$ (bottom) vs. m_{N} plane [17].

A search for heavy neutral leptons (HNLs), the right-handed Dirac or Majorana neutrinos, is performed in final states with three charged leptons (electrons or muons) using proton-proton collision data collected by the CMS experiment at 13 TeV at the CERN LHC Ref [88]. The data correspond to an integrated luminosity of 138 fb^1. The decay length of these particles can be large enough so that the secondary vertex of the HNL decay can be resolved with the CMS silicon tracker. The selected final state consists of one lepton emerging from the primary proton-proton collision vertex, and two leptons forming a displaced, secondary vertex. In Fig 5.4 - 5.5 results are provided in the m_N - $|\mathrm{V}_{\mathrm{N}\ell}|^2$plane.



Limits on $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) and mixed coupling $|V_{Nl}|^2$ ($l = e, \mu$)

Figure 5.4: Exclusion region at 95% CL in the $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) as a function of m_N for a Majorana HNL. Results from the Delphi [89] and the CMS [90, 91] Collaborations are shown for reference.

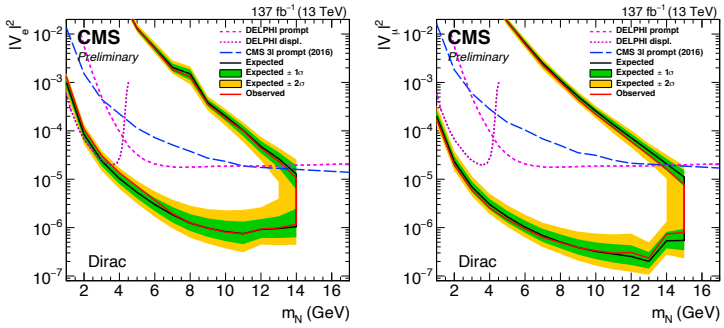


Figure 5.5: Exclusion region at 95% CL in the $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) as a function of m_N for a Dirac HNL. Results from the Delphi [89] and the CMS [90, 91] Collaborations are shown for reference.

No significant deviations from the standard model expectations are observed, and constraints are obtained on the HNL mass and coupling strength parameters, excluding previously unexplored regions of parameter space in the mass range 1-20 GeV and squared mixing parameter values as low as 107.
Complementary to this search, in this chapter I will present a HNL search in mass range between 1 and 15 GeV that can decay in the tracker volume up to 60 cm. The difference with the 3 lepton HNL search is that we will span the 2 leptons phase space as shown in Fig 5.2. Such displaced search is highly peculiar and challenging analysis at the LHC in high demand for dedicated data reconstruction tools in order to extend their sensitivity. The signature of the search is a a prompt lepton, a displaced lepton, a displaced jet and a secondary

displaced vertex. Details on the analysis strategy and limits as a function of both m_N and $|\mathrm{V_{N\ell}}|^2$, for each flavour ℓ separately will be presented in this chapter.

5.1.1 Publication plans and my contribution

In the context of this analysis, I pursued the effort on the publication doing some studies mainly on the signal topology and characterization: I took care of reweigthing the signal samples to the different couplings, SR optimization studies and studies to understand the composition of the most important backgrounds. Furthermore I helped in the validation of the ABCD methods used for the data-driven background estimation.
This analysis is currently under the process of approval within the CMS collaboration.

5.2 Data, Monte Carlo Simulation and Trigger selection

This analysis uses the pp collisions collected at the center-of-mass energy of 13 TeV by the CMS experiment in 2016, 2017 and 2018, corresponding to an integrated luminosity of 35.92 fb^{-1}, 41.53 fb^{-1} and 59.74 fb^{-1}, respectively. Two datasets are used, depending on the flavour of the prompt lepton produced in association with the HNL:

- SingleMuon;
- SingleElectron, called EGamma in 2018.

The choice of these datasets, rather than the DoubleMuon (DoubleElectron) samples, is driven by the optimization of the trigger selection in CMS. The leptonic trigger is optimized for prompt leptons identification. Since using the displaced lepton as a trigger object would lead to inefficiencies, it has been decided to not include double-muon and double-electron triggers in the analysis, as well as the corresponding data samples. The list of triggers and dataset used is summarized in Table 5.1.

Table 5.1: Summary of the trigger used in the analysis.

Dataset	Trigger		
	2016	2017	2018
SingleMuon	HLT_Iso(TK)Mu_24	HLT_IsoMu_24(27) + HLT_IsoMu_24_eta2p1	HLT_IsoMu_24
SingleElectron	HLT_Ele27_WPTight_Gsf	HLT_Ele32_WPTight_Gsf	
EGamma			HLT_Ele32_WPTight_Gsf

All background samples, reported in Table 5.2, 5.3 and 5.4, were generated centrally. The simulated events are re-weighted to match the distribution of in-time pileup interactions per bunch crossing to that of data following the recommendations of the Physics Validation Group.

Table 5.2: Simulated background samples with 2016 data-taking conditions and their effective cross sections.

Sample	No. events	Cross-sec. [pb]
WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	86916347	61334.9
WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	261093184	61334.9
DYJetsToLL_M-10to50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	108345066	18610
DYJetsToLL_M-10to50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	35114799	18610
DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	146280262	6077.22
TTJets_DiLept_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	27601593	87.315
TTJets_SingleLeptFromT_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	60461873	182.175
TTJets_SingleLeptFromTbar_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	56887791	182.175
WGToLNuG_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	27511829	585.8
ZGTo2LG_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	16679829	123.9
WWTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8	5246469	49.997
WWTo2L2Nu_13TeV-powheg	1999000	12.178
WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	26517270	5.60
WZTo3LNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	11928705	4.4297
WZToLNu2Q_13TeV_powheg_pythia8	497702	10.71
ZZTo2L2Nu_13TeV_powheg_pythia8_ext1	48442234	1.256
ZZTo2L2Q_13TeV_powheg_pythia8	496436	3.28
ZZTo4L_13TeV_powheg_pythia8_ext1	96451071	1.256
ST_s-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8	9811800	3.68
ST_t-channel_antitop_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUETP8M1	38811009	80.95
ST_t-channel_top_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUETP8M1	67105872	136.02
ST_tW_antitop_5f_NoFullyHadronicDecays_13TeV-powheg_TuneCUETP8M1	8681541	35.85
ST_tW_top_5f_NoFullyHadronicDecays_13TeV-powheg_TuneCUETP8M1	8681495	35.85
QCD_Pt-15to20_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	4141251	3819570
QCD_Pt-20to30_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	31878737	2960198
QCD_Pt-30to50_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	29664457	1652471
QCD_Pt-50to80_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	19662174	437504
QCD_Pt-80to120_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	23415442	106033
QCD_Pt-120to170_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	19809958	25190
QCD_Pt-170to300_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	37139899	8654
QCD_Pt-300to470_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	49005966	797
QCD_Pt-470to600_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	19489273	45.832
QCD_Pt-600to800_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	19909525	25.095
QCD_Pt-800to1000_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	19940740	4.707
QCD_Pt-1000toInf_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8	13540768	1.621
QCD_Pt-20to30_EMEnriched_TuneCUETP8M1_13TeV_pythia8	9241498	4833200
QCD_Pt-30to50_EMEnriched_TuneCUETP8M1_13TeV_pythia8	11508842	6850000
QCD_Pt-50to80_EMEnriched_TuneCUETP8M1_13TeV_pythia8	45789054	1900000
QCD_Pt-80to120_EMEnriched_TuneCUETP8M1_13TeV_pythia8	77800199	478520
QCD_Pt-120to170_EMEnriched_TuneCUETP8M1_13TeV_pythia8	78578408	68592
QCD_Pt-170to300_EMEnriched_TuneCUETP8M1_13TeV_pythia8	11540162	20859
QCD_Pt-300toInf_EMEnriched_TuneCUETP8M1_13TeV_pythia8	7380341	1350
QCD_Pt_15to20_bcToE_TuneCUETP8M1_13TeV_pythia8	2685602	254596
QCD_Pt_20to30_bcToE_TuneCUETP8M1_13TeV_pythia8	10987945	328999
QCD_Pt_30to80_bcToE_TuneCUETP8M1_13TeV_pythia8	15342782	405623
QCD_Pt_80to170_bcToE_TuneCUETP8M1_13TeV_pythia8	14851986	38104
QCD_Pt_170to250_bcToE_TuneCUETP8M1_13TeV_pythia8	9862069	2635.8
QCD_Pt_250toInf_bcToE_TuneCUETP8M1_13TeV_pythia8	9861592	711.92

Table 5.3: Simulated background samples with 2017 data-taking conditions and their effective cross sections.

Sample	No. events	Cross-sec. [pb]
WJetsToLNu_0J_TuneCP5_13TeV-amcatnloFXFX-pythia8	170861262	5547
WJetsToLNu_1J_TuneCP5_13TeV-amcatnloFXFX-pythia8	259753064	8018
WJetsToLNu_2J_TuneCP5_13TeV-amcatnloFXFX-pythia8	199304579	1940
WJetsToLNu_TuneCP5_13TeV-madgraphMLM-pythia8	107708702	61334.9
DYJetsToLL_M-10to50_TuneCP5_13TeV-madgraphMLM-pythia8	79002200	18610
DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8	200661434	6077.22
TTJets_DiLept_TuneCP5_13TeV-madgraphMLM-pythia8	28308163	87.315
TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	9596464	87.315
TTToHadronic_TuneCP5_13TeV-powheg-pythia8	42357942	687.1
TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	43732445	364.350
ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8	9883805	3.68
ST_t-channel_antitop_4f_inclusiveDecays_TuneCP5_13TeV-powhegV2-madspin-pythia8	3675910	80.95
ST_t-channel_top_4f_inclusiveDecays_TuneCP5_13TeV-powhegV2-madspin-pythia8	5982064	136.02
ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	5635539	35.85
ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	4955102	35.85
WWTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8	5054021	49.997
WWTo2L2Nu_NNPDF31_TuneCP5_13TeV-powheg-pythia8	2000000	12.178
WWToLNuQQ_NNPDF31_TuneCP5_13TeV-powheg-pythia8	8782524	49.997
WW_DoubleScattering_13TeV-pythia8_TuneCP5	999976	2.02
WZTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8	19086372	10.71
WZTo1L3Nu_13TeV_amcatnloFXFX_madspin_pythia8_v2	4994394	40.58
WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	27582163	5.60
WZTo3LNu_TuneCP5_13TeV-amcatnloFXFX-pythia8	20574696	4.4297
ZZTo2L2Nu_13TeV_powheg_pythia8	6366194	1.256
ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	22309069	3.28
ZZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8	62172311	4.416
ZZTo4L_13TeV_powheg_pythia8	119607848	1.256
ZGToLLG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	30306208	123.9
WGToLNuG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	25918959	585.8
QCD_Pt-15to20_MuEnrichedPt5_TuneCP5_13TeV_pythia8	5859837	3819570
QCD_Pt-20to30_MuEnrichedPt5_TuneCP5_13TeV_pythia8	28213498	2960198
QCD_Pt-30to50_MuEnrichedPt5_TuneCP5_13TeV_pythia8	29030190	1652471
QCD_Pt-50to80_MuEnrichedPt5_TuneCP5_13TeV_pythia8	24068592	437504
QCD_Pt-80to120_MuEnrichedPt5_TuneCP5_13TeV_pythia8	23248987	106033
QCD_Pt-120to170_MuEnrichedPt5_TuneCP5_13TeV_pythia8	20774842	25190
QCD_Pt-170to300_MuEnrichedPt5_TuneCP5_13TeV_pythia8	45892445	8654
QCD_Pt-300to470_MuEnrichedPt5_TuneCP5_13TeV_pythia8	17532738	797
QCD_Pt-470to600_MuEnrichedPt5_TuneCP5_13TeV_pythia8	24243589	45.832
QCD_Pt-600to800_MuEnrichedPt5_TuneCP5_13TeV_pythia8	17263676	25.095
QCD_Pt-800to1000_MuEnrichedPt5_TuneCP5_13TeV_pythia8	17114527	4.707
QCD_Pt-1000toInf_MuEnrichedPt5_TuneCP5_13TeV_pythia8	11596693	1.621
QCD_Pt-15to20_EMEnriched_TuneCP5_13TeV_pythia8	11215174	1345920
QCD_Pt-20to30_EMEnriched_TuneCP5_13TeV_pythia8	11212740	4833200
QCD_Pt-30to50_EMEnriched_TuneCP5_13TeV_pythia8	14765960	6850000
QCD_Pt-50to80_EMEnriched_TuneCP5_13TeV_pythia8	10477134	1900000
QCD_Pt-80to120_EMEnriched_TuneCP5_13TeV_pythia8	9104852	478520
QCD_Pt-120to170_EMEnriched_TuneCP5_13TeV_pythia8	8515107	68592
QCD_Pt-300toInf_EMEnriched_TuneCP5_13TeV_pythia8	2874295	1350
QCD_Pt_20to30_bcToE_TuneCP5_13TeV_pythia8	10270541	328999
QCD_Pt_30to80_bcToE_TuneCP5_13TeV_pythia8	16072968	405623
QCD_Pt_80to170_bcToE_TuneCP5_13TeV_pythia8	15999454	38104
QCD_Pt_170to250_bcToE_TuneCP5_13TeV_pythia8	9847659	2635.8
QCD_Pt_250toInf_bcToE_TuneCP5_13TeV_pythia8	9996885	711.92

Table 5.4: Simulated background samples with 2018 data-taking conditions and their effective cross sections.

Sample	No. events	Cross-sec. [pb]
WJetsToLNu_TuneCP5_13TeV-madgraphMLM-pythia8	71026850	61334.9
WJetsToLNu_0J_TuneCP5_13TeV-amcatnloFXFX-pythia8	192288248	5547
WJetsToLNu_1J_TuneCP5_13TeV-amcatnloFXFX-pythia8	171715343	8018
WJetsToLNu_2J_TuneCP5_13TeV-amcatnloFXFX-pythia8	98362049	1940
DYJetsToLL_M-10to50_TuneCP5_13TeV-madgraphMLM-pythia8	39392055	18610
DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8	194213216	6077.22
DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8	100194586	6077.22
JpsiToMuMu_JpsiPt8_TuneCP5_13TeV-pythia8	244777941	874800
TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	63990000	87.315
TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	101550000	364.350
TTToHadronic_TuneCP5_13TeV-powheg-pythia8	133808000	313.9
WWToLNuQQ_NNPDF31_TuneCP5_13TeV-powheg-pythia8	19199100	49.997
WWTo2L2Nu_DoubleScattering_13TeV-pythia8	871500	2.02
WWTo2L2Nu_NNPDF31_TuneCP5_13TeV-powheg-pythia8	7758900	12.178
WZJJToLNu_EWK_QCD_TuneCP5_13TeV-madgraph-pythia8	476600	0.4332
WZTo3LNu_mllmin01_NNPDF31_TuneCP5_13TeV_powheg_pythia8	89479397	4.4297
WZTo3LNu_TuneCP5_13TeV-amcatnloFXFX-pythia8	21997587	4.4297
WZTo3LNu_TuneCP5_13TeV-powheg-pythia8	1976600	4.4297
WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	28193647	5.60
WZTo1L3Nu_13TeV_amcatnloFXFX_madspin_pythia8	1690064	40.58
ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	27900469	3.28
ZZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8	57728992	4.416
ZZTo2L2Nu_TuneCP5_13TeV_powheg_pythia8	56428600	1.256
ZZTo4L_TuneCP5_13TeV_powheg_pythia8	105698893	1.256
WGToLNuG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	27933663	585.8
ZGToLLG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	33809449	123.9
ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-madgraph-pythia8	39917000	3.68
ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	154307600	136.02
ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	79090800	80.95
ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	8722734	35.85
ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	6909815	35.85
QCD_Pt-15to20_MuEnrichedPt5_TuneCP5_13TeV_pythia8	4765928	3819570
QCD_Pt-20to30_MuEnrichedPt5_TuneCP5_13TeV_pythia8	30612338	2960198
QCD_Pt-30to50_MuEnrichedPt5_TuneCP5_13TeV_pythia8	35622532	1652471
QCD_Pt-50to80_MuEnrichedPt5_TuneCP5_13TeV_pythia8	20268872	437504
QCD_Pt-80to120_MuEnrichedPt5_TuneCP5_13TeV_pythia8	25652280	106033
QCD_Pt-120to170_MuEnrichedPt5_TuneCP5_13TeV_pythia8	21490842	25190
QCD_Pt-170to300_MuEnrichedPt5_TuneCP5_13TeV_pythia8	36033125	8654
QCD_Pt-300to470_MuEnrichedPt5_TuneCP5_13TeV_pythia8	29488563	797
QCD_Pt-470to600_MuEnrichedPt5_TuneCP5_13TeV_pythia8	20495750	45.832
QCD_Pt-600to800_MuEnrichedPt5_TuneCP5_13TeV_pythia8	16618977	25.095
QCD_Pt-800to1000_MuEnrichedPt5_TuneCP5_13TeV_pythia8	16749914	4.707
QCD_Pt-1000toInf_MuEnrichedPt5_TuneCP5_13TeV_pythia8	10719790	1.621
QCD_Pt-15to20_EMEnriched_TuneCP5_13TeV_pythia8	15955389	1345920
QCD_Pt-20to30_EMEnriched_TuneCP5_13TeV_pythia8	14255377	4833200
QCD_Pt-30to50_EMEnriched_TuneCP5_13TeV_pythia8	23489939	6850000
QCD_Pt-50to80_EMEnriched_TuneCP5_13TeV_pythia8	10798233	1900000
QCD_Pt-80to120_EMEnriched_TuneCP5_13TeV_pythia8	9648791	478520
QCD_Pt-120to170_EMEnriched_TuneCP5_13TeV_pythia8	9964143	68592
QCD_Pt-170to300_EMEnriched_TuneCP5_13TeV_pythia8	3712174	20859
QCD_Pt-300toInf_EMEnriched_TuneCP5_13TeV_pythia8	2901355	1350
QCD_Pt_15to20_bcToE_TuneCP5_13TeV_pythia8	4316068	254596
QCD_Pt_20to30_bcToE_TuneCP5_13TeV_pythia8	10561226	328999
QCD_Pt_30to80_bcToE_TuneCP5_13TeV_pythia8	15177630	405623
QCD_Pt_80to170_bcToE_TuneCP5_13TeV_pythia8	14903409	38104
QCD_Pt_170to250_bcToE_TuneCP5_13TeV_pythia8	9654492	2635.8
QCD_Pt_250toInf_bcToE_TuneCP5_13TeV_pythia8	10191317	711.92

5.3 Signal simulation

The optimisation of the search as well as evaluation of the sensitivity is performed using simulated HNL signal. The samples used for this analysis are privately produced.
They are generated using the MadGraph5_aMCatNLO generator [92] at leading order (LO) using the `heavyN` model described in Ref. [93–95]. We choose this model because it can extend the SM with up to 3 right handed neutrinos which are singlet to the SM gauge symmetry, since we are looking for a simplified SM extension, we will add only 1 right-handed neutrino.
The parameter of interested of this search are the HNL mass and its couplings. We simulated HNLs in a mass range

$$m_N$$

= 1−−15 GeV$GeV which couple exclusively to one of the three SM neutrino flavour with different. Given the mass and the coupling it is possible to derive the lifetime of the particle; namely, the mean life$
$\mathrm{m}_N{}^{-5}|V_{\mathrm{N}\ell}|^{-2}$. Clearly, producing signal samples with a very fine grid of m_N and $|\mathrm{V}_{\mathrm{N}\ell}|^2$is not possible due to computing limitations. Fortunately, estimating the search sensitivity for each mass and coupling parameter can also be done with a careful reweighting procedure, reducing the amount of parameter points that need to be generated. The reweighing procedureus allows us to move along the $|\mathrm{V}_{\mathrm{N}\ell}|^2$axis of the $(\mathrm{m}_N,|\mathrm{V}_{\mathrm{N}\ell}|^2)plane.$
$Figure\ 5.6 shows all the full signal samples available to this analysis is in the \mathrm{m}_N$
and $|\mathrm{V}_{\mathrm{N}\ell}|^2$plane. Practically, the goal is to have maximal statistical coverage at each $|\mathrm{V}_{\mathrm{N}\ell}|^2$. This can be achieved by combining the available samples to a single sample and appropriately reweighting each event such that we again arrive at a physically valid HNL sample with a certain $|\mathrm{V}_{\mathrm{N}\ell}|^2$.
The applied weights need to satisfy two conditions for the full merged sample: the cross section needs to be correct and the HNL lifetime distribution needs to be a falling exponential distribution with mean lifetime τ_{N}.
A general formula can be derived that can give the weight based on the proper decay time ct of the specific event, the $|\mathrm{V}_{\mathrm{N}\ell}|^2$that we want to arrive at and parameters that describe the original samples $c\tau(|\mathrm{V}_{\mathrm{N}\ell}|^2_{\mathrm{i}})$ and N^i_{tot}(the total number of unweighted MC events in the sample). The weight can be found as:

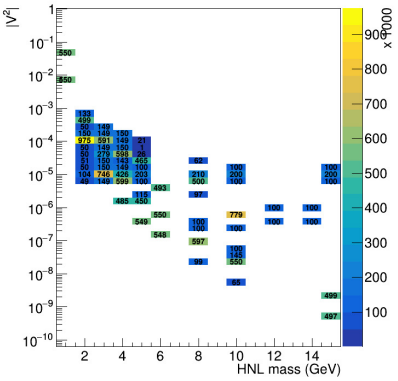
$$w\left(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}, ct\right) = \sigma\left(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}\right) \cdot \mathcal{L} \cdot \frac{\frac{1}{c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}})} exp\left(\frac{-ct}{c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}})}\right)}{\sum_{i=1}^{n} \frac{N^i_{tot}}{c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{i}})} exp\left(\frac{-ct}{c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{i}})}\right)} \quad (5.1)$$

where the fraction of exponentials ensures that we arrive at a correct lifetime distribution for the full sample, with an integrated weight of 1, which we then multiply by $\sigma\left(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}\right) \cdot \mathcal{L}$ to also arrive at the correct cross section for the sample.
The lifetime distribution of the sum of original samples without reweighting can be described by the sum of exponentials in the numerator, a fact that is also validated in Figure 5.7 where 2 samples and their sum are shown, as well as the resulting summed sample when they have been reweighted. Finally the values $\sigma\left(|\mathrm{V_{N\ell}}|^2\right)$ and $c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}})$ are rescaled from a reference value we take from one of the original samples

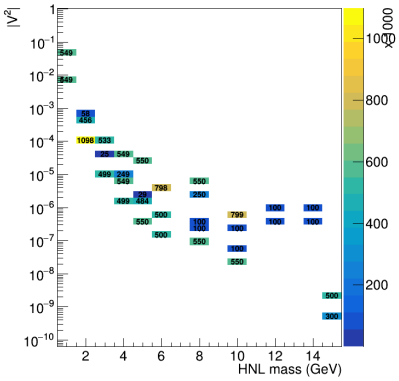
$$\sigma\left(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}\right) = \sigma\left(|\mathrm{V_{N\ell}}|^2_{0}\right) \cdot \frac{|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}}{|\mathrm{V_{N\ell}}|^2_{0}}$$
$$c\tau(|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}) = c\tau(|\mathrm{V_{N\ell}}|^2_{0}) \cdot \frac{|\mathrm{V_{N\ell}}|^2_{0}}{|\mathrm{V_{N\ell}}|^2_{\mathrm{f}}}$$

We also perform validations of this procedure versus several kinematic observables at the reconstruction level.
In Figure 5.8 we show various kinematic distributions. More descriptions for these observables will be given in the further sections.
From the reasonably good agreement between each reweighted sample we conclude that the procedure is acceptable to the limit of the available statistics of the samples we have available.

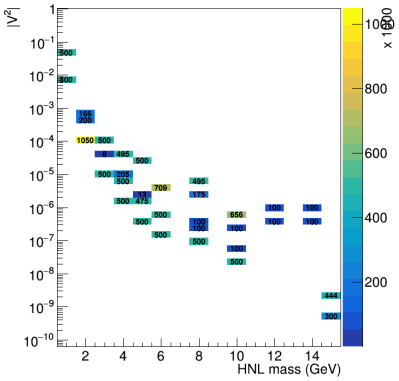
My contribution to the signal simulation samples is the validation of the reweighing procedure.



(a) HNL samples available for 2016.



(b) HNL samples available for 2017.



(c) HNL samples available for 2018.

Figure 5.6: HNL signal samples available to this analysis.

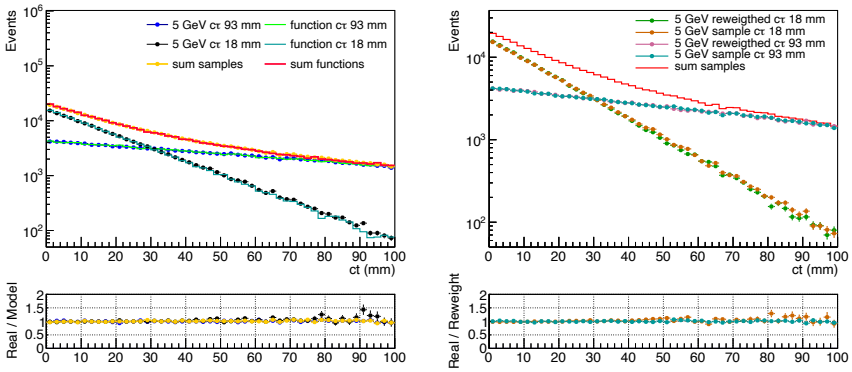


Figure 5.7: HNL signal sample reweighting illustrated for two samples of 5 GeV with $c\tau = 18$mm and $c\tau = 93$mm. The left plot shows that each separate sample and their sum can be described by exponential functions that use as only parameters $c\tau(|\mathrm{V_{N\ell}}|^2_1)$ and N^i_{tot}. On the right, the original samples are shown together with the summed and reweighted sample. The red histogram shows the unweighted sum of the original samples. Finally in green and pink, the summed sample has been reweighted event-by-event to represent each of the initial samples again, but with increased statistics.

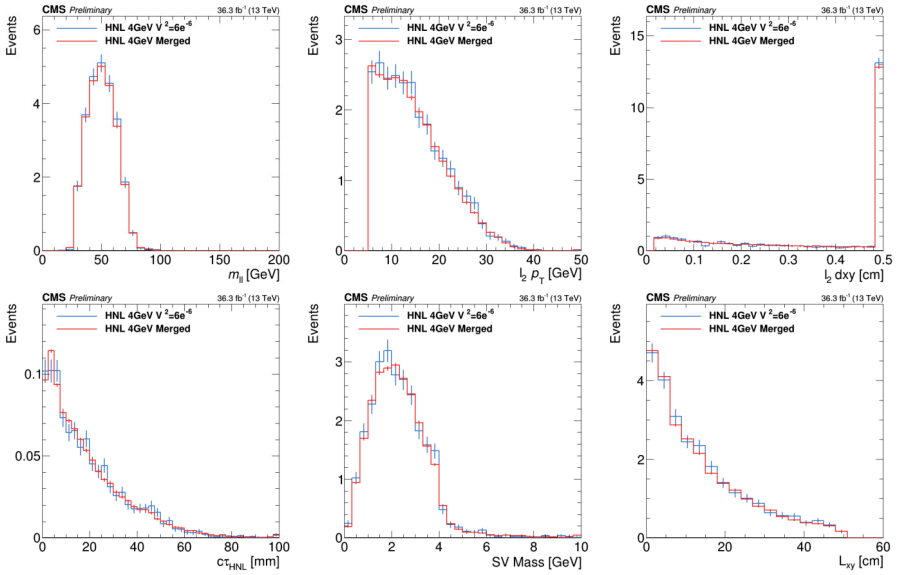


Figure 5.8: Several kinematic distributions in the $\mu\mu$ final state comparing an original HNL sample that has not been reweighted with the fully merged HNL sample reweighted to the same $|V|^2 = 6e^{-6}$. The applied selection can be found in table 5.10.

5.4 Object Selection

We should remember that the signature of this analysis is a prompt lepton originating from the W decay and a displaced lepton plus jets coming from the hypothetical production of an HNL. Displaced leptons are originated by the so-called secondary vertex and along with the additional hadronic decay products of the HNL.
The reconstruction of SV becomes then crucial, together with the leptons and jets informations, to the proper reconstruction of the final state.
The simulation of these vertexes is done with an algorithm called Inclusive Vertex Finder [96] that is designed to find and fit all likely secondary vertices in an event based on the full set of PF particles.
The primary vertex (PV) is chosen among all reconstructed primary proton-proton collision vertices as the one with the highest scalar sum of transverse energy of particles associated with it. Two quantities related to the displacement of a particle in relation to the PV are d_{xy} and d_z.
They are respectively the transverse and longitudinal distance from the PV of the particles track at its point of closest approach (in 3D) to the PV. d_{xy} above a certain value is the criterion used to identify a displaced lepton. I_{rel}, the relative isolation, is defined for leptons as the scalar p_T sum of charged hadrons originating from the PV, neutral hadrons, and photons within a cone of $\Delta R < 0.3$ around the lepton, divided by the lepton p_T. A correction is applied to the relative isolation to mitigate the contribution of neutral hadrons and photons coming from pileup.
My contribution in this point was to tune the cuts of d_{xy} and d_z for electrons and muons to maximise the significance, more details in the following sections.

5.4.1 Electrons

Electron reconstruction is based on the combination of tracker and ECAL information in a Gaussian Sum Filter (GSF) track [97], which accounts for possible bremsstrahlung from the electron.
Electrons are reconstructed within the geometrical acceptance of the CMS tracking system, $|\eta| < 2.5$. Identification criteria based on the electromagnetic shower shape, track quality, track impact parameters with respect to the primary vertex, and isolation are used to select signal electrons and reduce the rate of mis-identified and background electrons (referred to as “fake electrons” hereafter). The electron selection criteria are summarized in Table 6.1.

Prompt electrons are further identified with the help of a multi-variate discriminator which includes observables such as the presence of bremsstrahlung along the electron trajectory, the geometrical and momentum-energy matching between the electron trajectory and the associated cluster in the ECAL.

The displaced electron reconstruction and ID efficiency, measured in HNL signal samples, as a function of the lepton p_T and of the gen-level SV displacement in 2D are shown in figures 5.9 and 5.10.
The reconstruction efficiency is calculated compared to having a truth-level lepton with $p_T > 7$ GeV, $|\eta| < 2.5$ and a 2D displacement of the HNL decay vertex of less than 50 cm. No ID-requirements are put on the reconstructed lepton.
It can be seen that the efficiency drops quickly with displacement. This happens because the standard electron reconstruction algorithm [97] used for the electron has been optimized for the prompt leptons. The ID efficiency in Figure 5.10 is measured, instead, compared to having a reconstructed electron with no ID requirements coming from HNL. The same truth-level acceptance as for the reconstruction efficiency are applied.
The efficiency stays relatively good over the displacement range, a result of removing the ID cuts from the cut-based Loose ID that would interfere with the displacement.
My contribution to this section is to a dedicated study that was performed to tune the d_{xy} and d_z cut for the prompt and the displaced electron: we measured the efficiency and the background rejection for each point of the d_{xy} (d_z) space and then, based on these numbers, we choose the working point. The corresponding efficiency and background rejection are summarized in Tab 5.6.

5.4.2 Muons

Muons are reconstructed by combining the information of the tracker and of the muon spectrometer [98, 99].
The geometric compatibility between these separate measurements is used in the further selection of muons. They are required to have $|\eta| < 2.4$ to fall inside the geometric acceptance of the muon detector.
Muon selection criteria based on reconstructed properties are summarized in Table 6.2. The displaced muon reconstruction and ID efficiency as a function

Table 5.5: Requirements for an electron to pass each of the defined selection working points. Where three values are given for a single variable, they correspond to electrons with $|\eta| < 0.8$, $0.8 < |\eta| < 1.479$, and $1.479 < |\eta| < 2.5$

Selection name	prompt	displaced
$\|\eta\|$	< 2.5	< 2.5
p_T	$>$ 30–34 GeV	> 7 GeV
$\|d_{xy}\|$	< 0.02 cm	> 0.02 cm
$\|d_z\|$	< 0.04 cm	< 10 cm
I_{rel}	< 0.1	—
$\sigma_{i\eta i\eta}$	—	$< (0.11, 0.11, 0.0314)$
H/E	—	$< (0.298, 0.298, 0.101)$
$\Delta\eta_{\text{in}}$	—	$< (0.00477, 0.00477, 0.00868)$
$\Delta\varphi_{\text{in}}$	—	$< (0.05, 0.05, 0.1)$
$1/E - 1/p$	—	$< (0.241, 0.241, 0.14)$
MVA estimator	$> f(\eta, \text{p}_T)$	—

Table 5.6: Efficiency and background rejection for dxy and dz cuts for prompt and displaced electrons

Prompt electron ID		
Variable	Efficiency	Background rejection
dxy	99%	0.02%
dz	99%	0.08%
Displaced electron ID		
dxy	60%	83%

of the lepton p_T and of the gen-level 2D displacement of the vertex are shown in figures 5.12 and 5.13. They are measured in HNL signal samples. The reconstruction efficiency is calculated compared to having a truth-level lepton with $p_T > 5$ GeV, $|\eta| < 2.4$ and a 2D displacement of the HNL decay vertex of less than 50 cm. No ID-requirements are put on the reconstructed lepton. It can be seen that the efficiency drops with displacement, though it stays relatively high compared to displaced electrons. This will translate to much higher signal efficiency for HNL coupling to muons than for electrons and ultimately, a better sensitivity to $|V|$ than $|V_{eN}|$. Next, the ID efficiency in figure 5.13 is measured compared to having a reconstructed muon with no ID requirements coming from HNL. The same truth-level acceptance as for the reconstruction efficiency are applied. The efficiency of the full displaced ID is measured, ex-

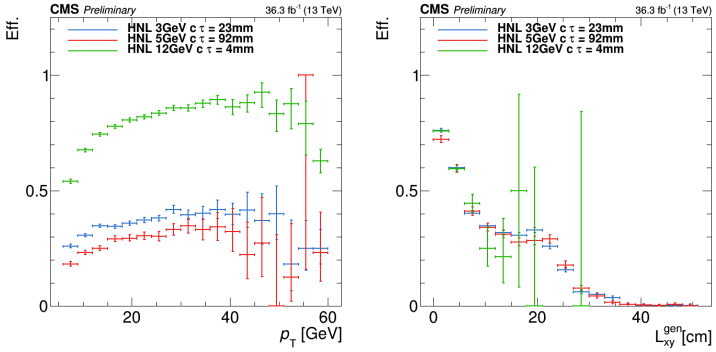


Figure 5.9: Displaced electron reconstruction efficiency measured compared to having a truth-level electron coming from the HNL decay, with $p_T > 7$ GeV, $|\eta| < 2.5$ and a 2D displacement of the HNL decay vertex of less than 50 cm. No ID requirements are put on the reconstructed lepton.

cept for the impact parameter cut $d_{xy} > 0.02$ cm. The ID efficiency is very good for the full displacement range.

Table 5.7: Requirements for a muon to pass each of the defined selection working points.

Selection name		prompt	displaced
$\|\eta\|$		< 2.4	< 2.4
p_T		$>$ 25–28 GeV	> 5 GeV
$\|d_{xy}\|$		< 0.01 cm	> 0.02 cm
$\|d_z\|$		< 0.1 cm	< 10 cm
I_{rel}		< 0.1	—
Loose ID		True	True
Fraction of valid tracker hits		> 0.8	—
5*Global muon	Global muon fit	True	True
	Global track χ^2/dof	< 3	—
	Track–muon matching χ^2/dof	< 12	< 12
	"Kink finder" estimator	< 20	< 20
	Segment-compatibility estimator	> 0.303	> 0.303
Tracker muon	Segment-compatibility estimator	> 0.451	> 0.451

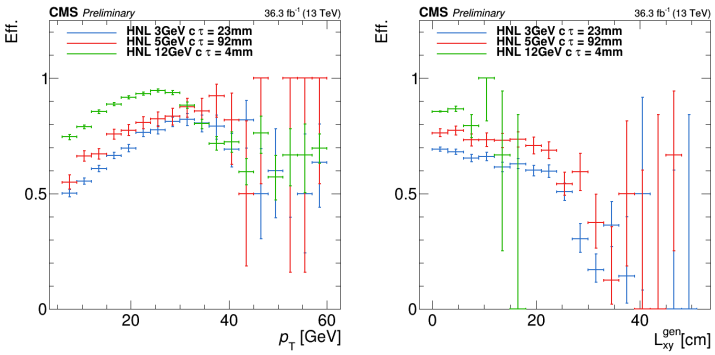


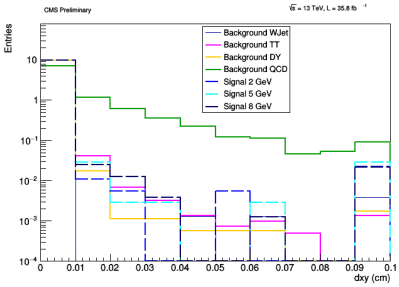
Figure 5.10: Displaced electron ID efficiency measured compared to having a reconstructed electron with no ID coming from the HNL decay, with acceptance cuts on the truth-level lepton: $p_T > 7$ GeV, $|\eta| < 2.5$ and a 2D displacement of the HNL decay vertex of less than 50 cm. The efficiency of the full displaced ID as mentioned in table 6.1 is measured, except for the $d_{xy} > 0.02$ cm requirement. The $d_{xy} > 0.02$ cm cut is studied separately.

Table 5.8: Efficiency and background rejection for dxy and dz cuts for prompt and displaced muons

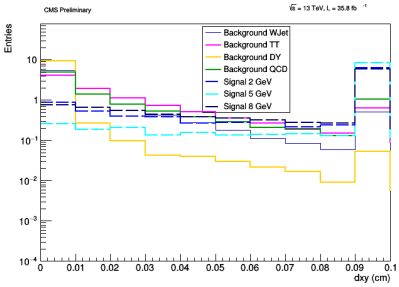
Prompt muon ID		
Variable	Efficiency	Background rejection
dxy	99%	0.008%
dz	99%	0.02%
Displaced muon ID		
dxy	85%	80%

5.4.3 Jets

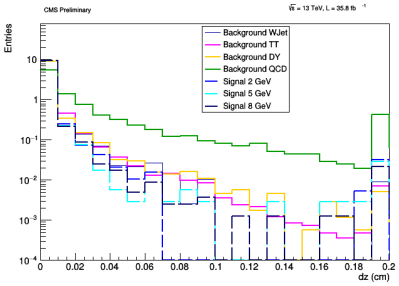
Jets are created by clustering PF candidates using the anti-k_T algorithm [100], where the standard value of 0.4 for the distance parameters is used.
A few quality requirements are applied to identify genuine jets: we are following the recommendation of the CMS collaboration using the Tight Jet ID [63] that are specific to each year. Two further requirements are placed on jets to ensure the quality, requiring $p_T > 20$ GeV and $|\eta| < 2.4$.
Figure 5.14 shows the efficiency of finding such a jet within $\Delta R < 0.7$ of the displaced lepton. This cut will leave the signal events practically untouched but removing significant background.
Jet cleaning from leptons is applied for counting jets for the jet veto. Jets are not counted when they have an angular overlap $\Delta R < 0.4$ with an electron or



(a) dxy distribution for 3 signal points and the main backgrounds for the prompt muon.



(b) dxy distribution for 3 signal points and the main backgrounds for the displaced muon.



(c) dz distribution for 3 signal points and the main backgrounds for the prompt muon.

Figure 5.11: dxy and dz distribution for prompt and displaced muons. Plots are done with 3 rapresentative signal points: 2-5-8 GeV with a lifetime of 74.22-69.65-5.46 mm. The backgrounds are: Drell-Yan, $t\bar{t}$, WJets and QCD.

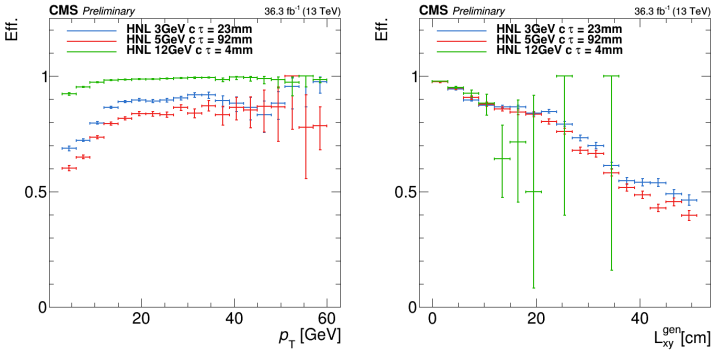


Figure 5.12: Displaced Muon reconstruction efficiency measured compared to having a truth-level muon coming from the HNL decay, with $p_T > 5$ GeV, $|\eta| < 2.4$ and a 2D displacement of the HNL decay vertex of less than 50 cm. No ID requirements are put on the reconstructed muon.

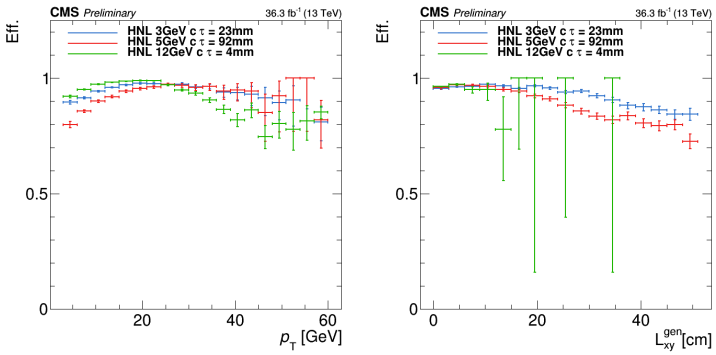


Figure 5.13: Displaced muon ID efficiency measured compared to having a reconstructed muon with no ID coming from the HNL decay, with acceptance cuts on the truth-level muon: $p_T > 5$ GeV, $|\eta| < 2.4$ and a 2D displacement of the HNL decay vertex of less than 50 cm. The efficiency of the full displaced ID as mentioned in table 6.2 is measured, except for the $d_{xy} > 0.02$ cm requirement. This last cut is studied separately.

muon (that passes either the prompt or tight ID requirements).
An additional cut, aimed mostly at reducing the $t\bar{t}$ background, sets a veto on more than one cleaned jet passing the tight ID.
Jet Energy Corrections [64] are applied to correct some knonw effects, as reconstructed energy of jets, including pileup, non-uniformities of the detector response that cause discrepancies between simulation and data.

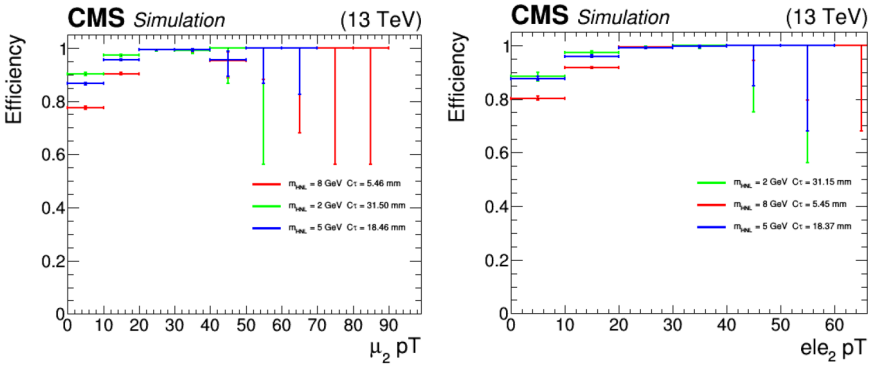


Figure 5.14: event-level efficiency of finding a jet passing the jet ID that is within an angular distance of $\Delta R < 0.7$ of the identified displaced lepton. The efficiency is measured in events with a displaced muon (left) and a displaced electron (right).

5.4.4 Secondary Vertex

The Inclusive Vertex Finder (IVF) [96] is a vertex fitting algorithm that reconstructs all potential secondary vertices in an event, based on the full set of Particle Flow (PF) tracks. We are actually only interested in a single secondary vertex, the decay vertex of the HNL.
The reconstruction of the displaced vertex in this case is not straightforward because we have multiple tracks, coming from the hadronic activity of the event, plus a lepton track and we don't know how to properly associate and combine them.
The IVF solved these problems by, starting from the set of PF tracks that pass the requirements in Table 5.9, groupping the tracks into clusters based on seed tracks that are at least slightly displaced from the PV. Tracks are added to a cluster if they are more compatible with the seed track and their point of closest approach than with the PV. The resulting clusters of tracks are each fit into one or more vertices using the Adaptive Vertex Reconstruction (AVR) [101] algorithm, discards incompatible tracks until it has a good vertex and does the same again for the discarded set of tracks.
When the vertex is properly reconstructed, other quality criteria, based on the significance of the PV-SV distance, the normalized χ^2 of the fit and the alignment of the PV-SV distance to the sum of the four-momenta of the tracks, are applied.
The reconstructed vertices are then validated by comparing their properties to the truth-level properties of the HNL decay vertex and the reconstruction efficiency of the SV is measured and shown in Fig. 5.15 as a function of displacement. The reconstruction efficiency is then defined as events that have a secondary vertex containing the displaced lepton, and the 3D distance of that SV from its gen-level position to be within 10% of the distance of the SV from the PV.

Table 5.9: Requirements for a PF track to pass the IVF preselection.

Variable	Requirement
p_T	> 0.8 GeV
Valid track hits	> 6
Only for Seed tracks:	
3DIP	> 0.005 cm
3DIPSig	> 1.2

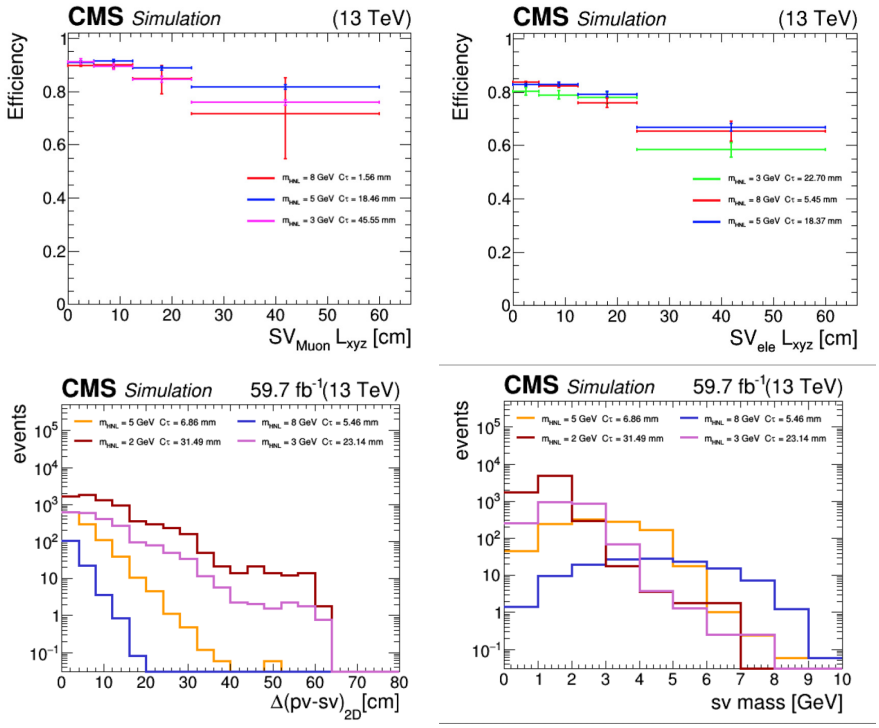


Figure 5.15: Secondary vertex reconstruction efficiency in HNL decay process as a function of flight distance d_{3D} for vertices with a muon(upper left) and with an electron(upper right). The bottom plots show the 2D decay distance from the PV(lower left) and the reconstructed mass of the SV(lower right).

The reconstruction efficiency drops with the displacement because the quality of the tracks deteriorates accordingly.

5.5 Event Selection

We perform a baseline selection on all datasets to select signal-like events. HNL signature is reconstructed from events that have one $\ell_1$and one $\ell_2$where $\ell_2$is included in a secondary vertex.
Since in this analysis we want to determine the strength of the HNL coupling to each lepton flavor, the events are further splitted into categories of lepton charge and flavor.
The charge classification on the charge will be used to prove HNL scenarios in which the two leptons can have the same electric charge (SS) or opposite electric charge (OS) known as Majorana scenario or cases in which the leptons

Table 5.10: Full selection applied to all datasets.

Variable	Requirement
ℓ_1	Prompt ID
ℓ_2	Displaced ID
SV	$\ell_2$part of SV
$\Delta R(\ell_2, jet)$	< 0.7
$M(\ell_1, \ell_2)$	> 10 GeV
$\Delta\varphi(\ell_1, \ell_2)$	> 0.4
$N_{jets,cl}$	≤ 1
N_l	$= 2$

are always OS known as Dirac scenario. In total we analyze 8 event categories in each data-taking period: ee, $\mu\mu$, $e\mu$ and μe for OS and SS charge configuration. The event selection is briefly summarized in table 5.10.
The first step of the analysis is the identification of the prompt lepton (ℓ_1) together with the displaced (ℓ_2) lepton by applying the prompt or displaced ID. If more then one lepton passes the prompt (displaced) cuts, the lepton with the highest p_T is selected.
The SV is then chosen among the list of vertices provided by IVF as the vertex that is associated to the displaced lepton track, if there is any. The displaced jet is chosen as the jet closest to the displaced lepton, where overlap between jet and lepton is allowed, to be efficient for the boosted nature of the HNL.
The selection criteria mentioned above are not sufficient to fully capture the complex signature of out HNL, the use of machine will then become crucial and it will be described in the next sections: we can say that these selection criteria are only used as baseline to select a reasonably good amount of MC events for the model to be trained and evaluated on, thus the applied cuts are deliberately kept loose.
Only a cut on the ML model output can be tune and defined to reach an optimal signal-to-background ratio region that will be defined as signal region.

5.5.1 Remaining backgrounds and Control regions

After the full event selection four main sources of backgrounds remain: top quark pair production, W+jets and the Drell-Yan and QCD.
Their relative contributions vary across the channels: in the OS and SS ee channels there is a large contribution from DY due to the poor reconstruction

of the displaced vertex, thus both the electrons from Z are selected.
The backgrounds from $t\bar{t}$ and W+jets typically have one prompt lepton and a second lepton in a jet that is reconstructed as a secondary vertex.
The QCD background is undoubtedly the most complicated background to deal with in the analysis. This background is poorly simulated from MC samples and contributes with two "fake" leptons (usually jets mis-reconstructed as leptons). In principle it can be reduced by appling more stringent cuts on the ℓ_1. Tightening the cuts will reduce a lot the sensitivity of the analysis, thus the job of properly disentangle QCD from signal events is left to to our ML network.
All these background are obtained based on MC simulation which suffers a lot from statistics and from detector mis-modeling effects, thus a proper estimation of the background contribution is done with a data-driven estimation based on the ABCD method described in Sec 5.7.
To validate both the ML techiques and the ABCD method several control region are defined:

- **Two prompt lepton control region:** This control region is formed by requiring two high p_T prompt leptons in each channel ee, $\mu\mu$, $e\mu$ both with OS and SS. We see a good agreement across all observables.

- **Control region with two or more jets:** The HNL production via W boson decays is not expected to contain multiple jets. Therefore asking at least two jet keeping all other selection identical forms a region with representative background processes in the signal region but with nearly no signal. We see a good agreement across all observables.

5.6 Particle Flow Network

This chapter deals with the applied Machine Learning model. In section 3.7 a more wide view of machine learning is given.
I want to highlight that I haven't worked on the architecture or the validation of the Particle Flow Network. In this section I'll give a brief overview of this promising technique showing only some of the validation plots that had been used to validate this procedure.
It was clear from the beginning that the HNL signature is too complicated, especially for the presence of the displaced vertex, to be captured using only cuts on the object of interest. A machine learning technique is then used to have a better understanding on how the signal looks like.
Energy Flow Networks [102], with the extension of Particle Flow Networks (PFN), have recently been introduced as an application of the Deep Sets Theorem [103]. Deep Sets are defined as a fundamental architecture that can deal with sets that are invariant under permutations as inputs. An example of the invariant under permutation definition is a a collider event (or part of an event) that is invariant by definition under permutations of the individual particles.
In this network all the features of each particles are combined together (step 1) to give the so called "event representation", then a function transforms (step 2) this representation into the desire output. PFNs represent each of the two steps by DNNs. With the summation of the output of the first transformation implemented, both DNNs can be combined into a single network that can be trained together. A skatch of the architecure of the PFN is give in Fig 5.16.

The particle-level information fed to the first part of the network consists of all Particle Flow (PF) particles in the identified displaced jet. In this analysis, the PFN only uses information regarding the displaced part of the event to be fully orthogonal to the prompt lepton information; this orthogonality between the prompt and the displaced part of the events will be used in the ABCD method to define the desire region (Sec 5.7).
All MC samples are used to train the PFN.
The PFN is also trained for each of the 2 displaced lepton flavour (μ and e) to take into account the difference in the reconstruction. A further splitting is made into low mass HNL region (1-5 GeV) and high mass HNL region ($>$ 5 GeV) and, finally, since the pixel tracker detector had been changed in 2017, a further split on the years, considering 2016 and 2017-18 is made. In total in this analysis we are dealing with 8 PFNs.

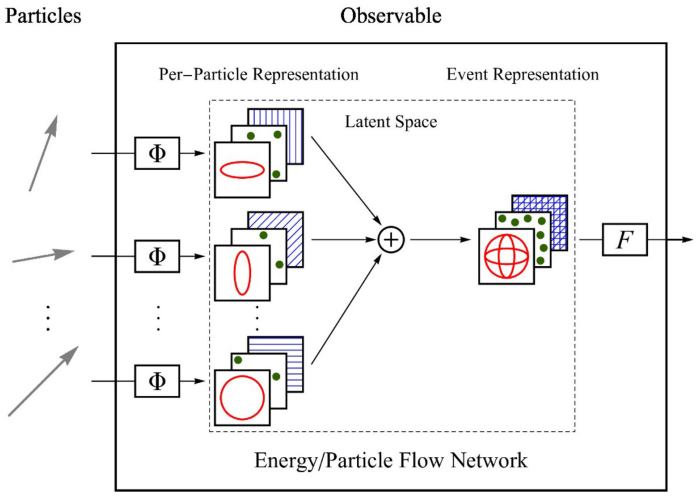


Figure 5.16: Visualization of the Particle Flow Network architecture. Individual particles are transformed by the function Φ to a latent space representation, generally of a high dimension, where they are summed together to arrive at an event representation At this stage, precomputed high-level variables can be added to the event representation. The function F finally transforms the event representation to the desired output observable. Figure taken from [102].

The results of training for 2017-18 years all PFNs can be found in figure 5.17, showing the metrics evaluated on the training and validation datasets. The overtraining (see Sec 3.7) had also been checked.

The AUC (see 3.7) is good in all the studied configurations; this ensure a very nice signal to background separation and makes the PFN output the most sensitive variable of this analysis.

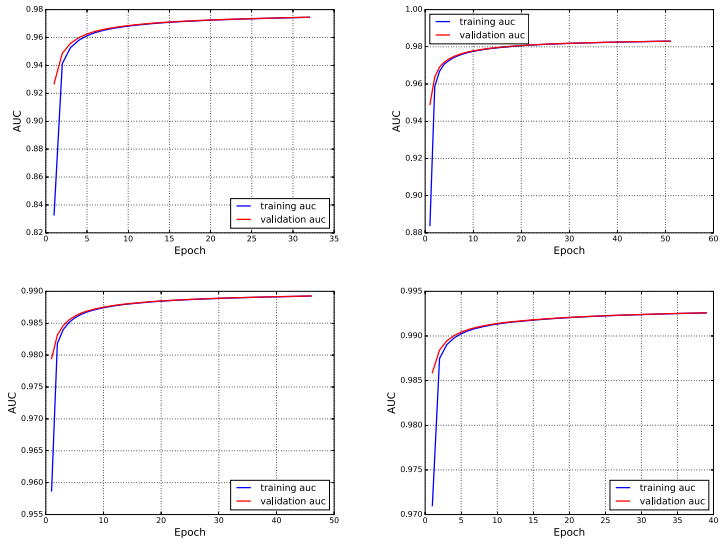


Figure 5.17: AUC plots for 2017-18 PFNs, split between validation and training set. Top left is electron low mass PFN, bottom left is muon low mass, top right is electron high mass, bottom right is muon high mass.

5.7 Background Estimation: ABCD method

In this section I will describe the ABCD method used for the data-driven background estimation.
I want to highlight that I haven't worked on the study of the ABCD regions but I worked on the study of possible source of systematic uncertainties derived from this method. For the clarity of this thesis, a brief overview on the ABCD method is given.
As mentioned in the previous section, the PFN output is the most discriminating variable of this analysis, this means that by cutting not he PFN score, is possible to define a region enriched with signal and with only few tens of background events. Given the fact that the most important backgrounds of this analysis, as W+jets and QCD, have a very high cross section (and thus a very high impact of the analysis), if we use only a cut on the PFN score to define our signal region we will suffer from a very large statistical uncertainties arising from the few MC background events that enter in the region with high weight. This will give us a poor estimation of the real background contamination and a worsening of the sensitivity of the analysis.
It is however possible to have a more precise way of estimating the remaining background in the signal region using a data-driven method, where we use the background events in certain signal region sidebands to calculate a prediction in the signal region.
A robust and straightforward method for this task is the ABCD method.
The ABCD method works with the definition of four regions: A, the signal region and B,C,D background enriched control regions that are defined by cutting on the two most sensitive observables. In our case we will used the PFN output and the mass of the invariant mass of $\ell_1$and the SV (m(ℓ_1, SV)). Before prediction the background in region A, we should also make sure of two things:

- the chosen variables are uncorrelated, and this comes by definition since the PFN was made agnostic to ℓ_1.
- the signal is confined to region otherwise we would have to a contamination of the prediction. This is also granted thanks to the powerful signal background separation of the PFN output.

Checked those two requirements, it holds a relation as:

$$\frac{A_{bkg}}{B_{bkg}} = \frac{C_{bkg}}{D_{bkg}} \tag{5.2}$$

thus the estimation of background in region A is:

$$A_{bkg}^{pred} = B_{bkg}\frac{C_{bkg}}{D_{bkg}} \tag{5.3}$$

A sketch of the ABCD region is in Fig 5.18.

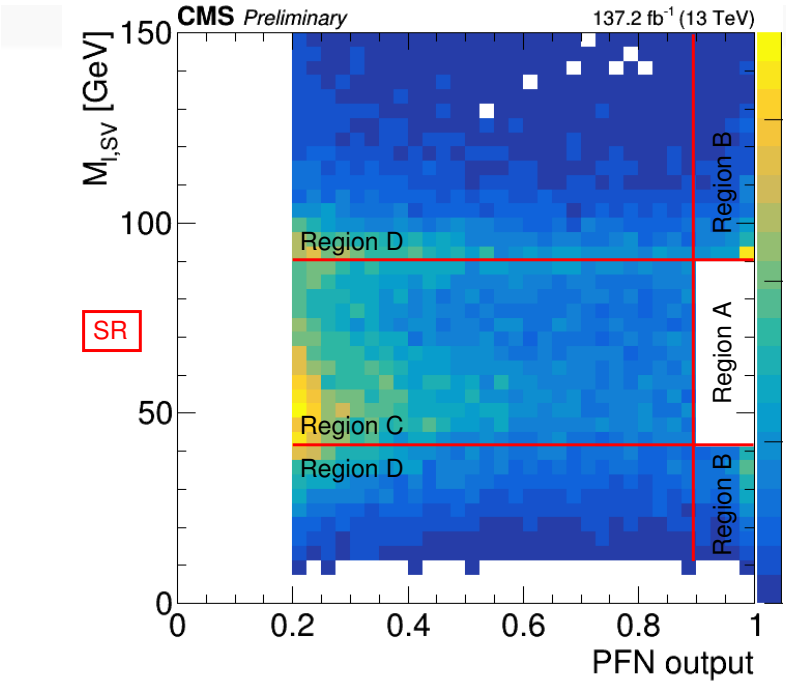


Figure 5.18: A visualization of the definition of the ABCD regions based on the PFN output and $m(\ell_1, SV)$on 2016 OS $\mu\mu$ data. Region A is kept blinded as it contains the HNL signal. The actual ABCD prediction uses a much tighter cut on the PFN output than shown here at 0.8.

There are some baseline cuts that apply to the full ABCD region:

- PFN score > 0.2: this gets rid of the majority of the background that's peaked at 0. This way, it is ensured that we make a prediction based on background that is at least somewhat 'signal-like'.

- $m(\ell_1, SV) > 10$ GeV: A baseline cut that has little impact, but is applied anyway as a safety cut.

Divisions along the two variables are made to create a normal region and an inverted region:

- $m(\ell_1, SV)$normal region: $m(\ell_1, SV) \in [50, 85]$ GeV

- $m(\ell_1, SV)$inverted region: $m(\ell_1, SV) \in [10, 50]$ GeV + $m(\ell_1, SV) > 85$ GeV

- PFN normal region: PFN $>$ x, where x differs for each PFN and flavor final state

- PFN inverted region: PFN $<$ x

The cut value differs for each PFN, as the PFN output distributions are not inherently the same. Optimal cutting points are currently defined such that the background is kept small but non-zero. In 5.18 the PFN cut at 0.9 is just the value that has been used to keep the data blinded. The signal region is divided into bins based on the relative charge of the leptons (OS and SS), the SV 2D displacement and the SV mass. This division is meant to improve the overall sensitivity of the analysis.
The validation on the ABCD method goes thought a closure test, where for closure test it is meant to compare the prediction can with the actual observed yields in a region where we don't have signal contamination. Several control region are investigated:

- CR1: we predict the background events with PFN score $\in [0.7, 0.9]$. The B,C and D regions are also shifted along to the left. Closure test look generally good and, I'll report only a couple of plots to show the data-prediction agreement for CR1: Fig 5.19.

- CR2: is defined in the inverted $m(\ell_1, SV)$region, with the same PFN cuts as are applied for the final signal region. The normal requirement of one or zero jets in the event is removed in order to reach enough statistics to perform a meaningful test. Background events are predicted

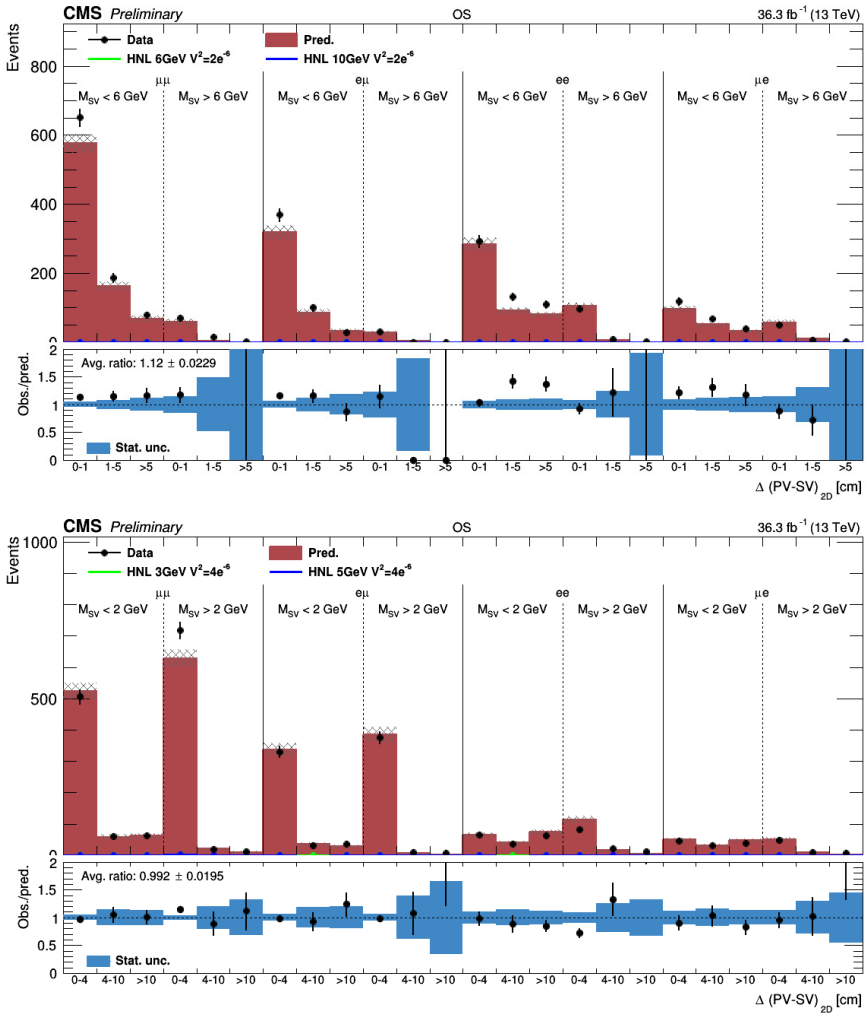


Figure 5.19: Comparison of the predicted and observed yields in CR1 for opposite-sign 2016 events. We show the prediction for the high-mass PFNs (top) and the low-mass PFNs (bottom).

in the region $m(\ell_1, SV) \in [85, 110]$ GeV. Regions A and B for this control region are poorly populated this the closure test has been done by combining data for all 3 years. Even in this region, the prediction works, I'll report only a couple of plots to show the data-prediction agreement for CR2: Fig 5.20.

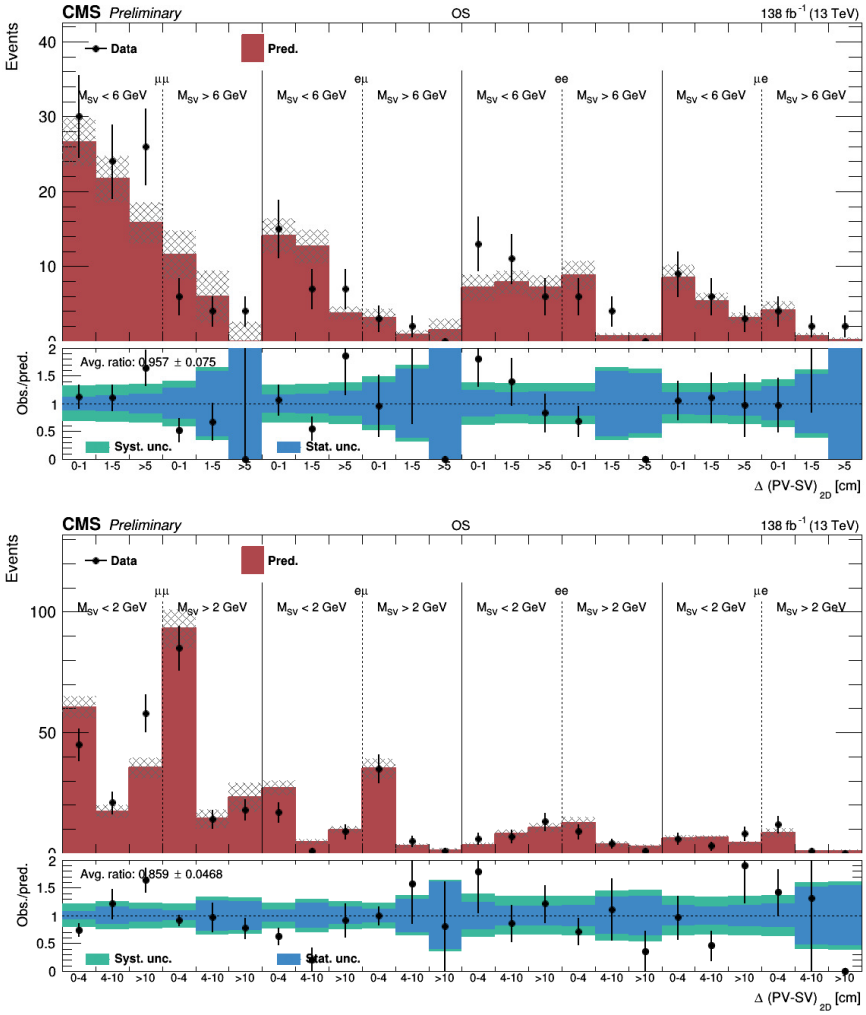


Figure 5.20: Comparison of the predicted and observed yields in CR2 from Run 2 data. Events of all 3 years are added together before performing the ABCD prediction. We show OS regions in these plots. High Mass regions are on the top and Low Mass regions on the bottom.

I'll end this section with the estimation of the systematic uncertainties coming from the ABCD method.

5.7.1 ABCD pulls

To further validate the ABCD methods and to take into account all the possible source of systematic uncertainties, pulls for CR1 and CR2 are studied.
The pulls are made separately for each lepton flavor combination. They are done as a function of the displacement Fig. 5.21 - 5.22 Fig. 5.25, in which case we sum bins of SV mass and of lepton charge. They are also done as a function of the lepton charge Fig. 5.23 - 5.24 Fig. 5.26, in which case we sum bins of SV mass and displacement.

The method shows no large systematic deviation between the predicted yields and the observed data within the statistical uncertainties. We use all deviations in the closure tests to estimate the systematic uncertainties: a 20% uncertainty on same-sign channels to be large enough. For the Low Mass search regions, we similarly apply a 30% uncertainty on opposite-sign channels with a displaced electron and 20% on opposite-sign channels with a displaced muon and on same-sign channels. I worked on this estimation.

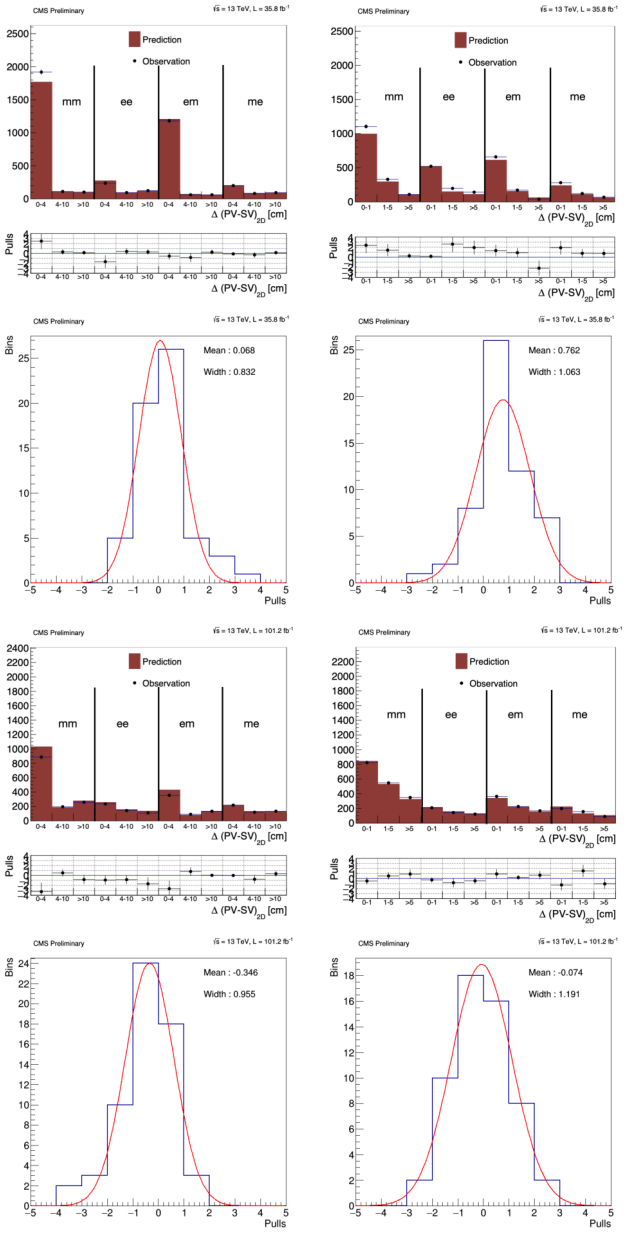


Figure 5.21: Pulls for CR1 as a function of displacement for each flavor. We show the prediction for 2016 separately (top) and for 2017+2018 (bottom), with plots for the low mass PFNs on the left and the high mass PFNs on the right, OS and SS categories are merged together.

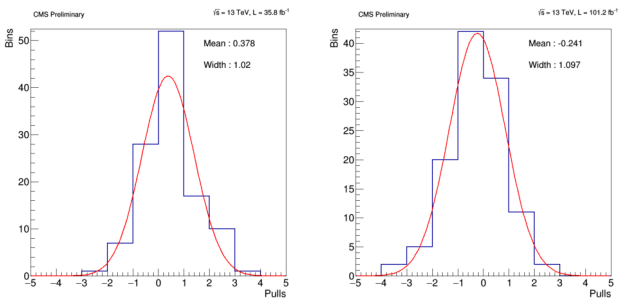


Figure 5.22: Pulls for CR1 as a function of displacement and for each flavor, divided in bins low mass PFNs and high mass PFNs are combined together. We show the prediction for 2016 separately (left) and for 2017+2018 (right).

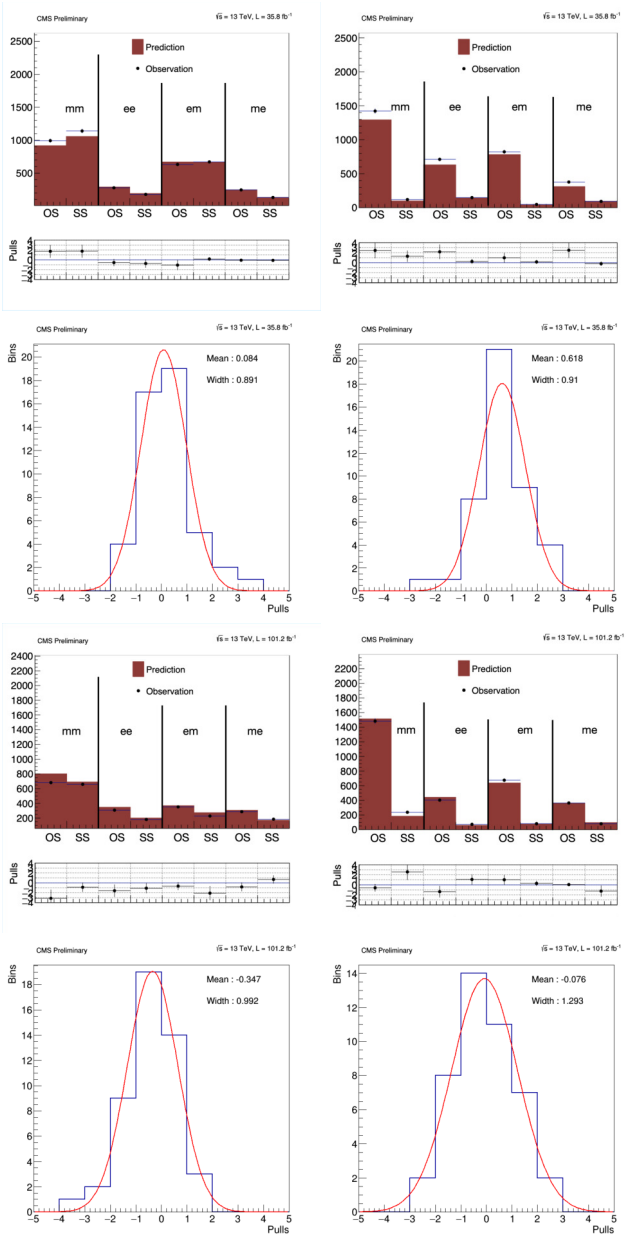


Figure 5.23: Pulls for CR1 as a function of the lepton's charge and divided in bins. We show the prediction for 2016 separately (top) and for 2017+2018 (bottom), with plots for the low mass PFNs on the left and the high mass PFNs on the right, displacemenet bins are merged together.

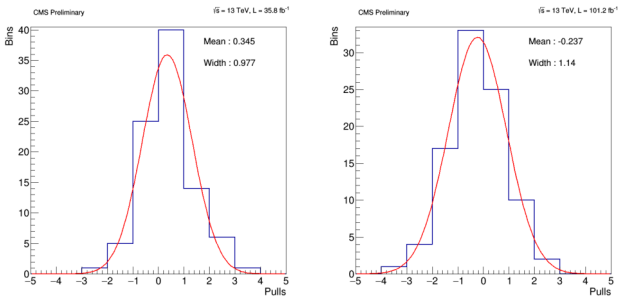


Figure 5.24: Pulls for CR1 as a function of the lepton's charge divided in bins, low mass PFNs and high mass PFNs are combined together. We show the prediction for 2016 separately (left) and for 2017+2018 (right).

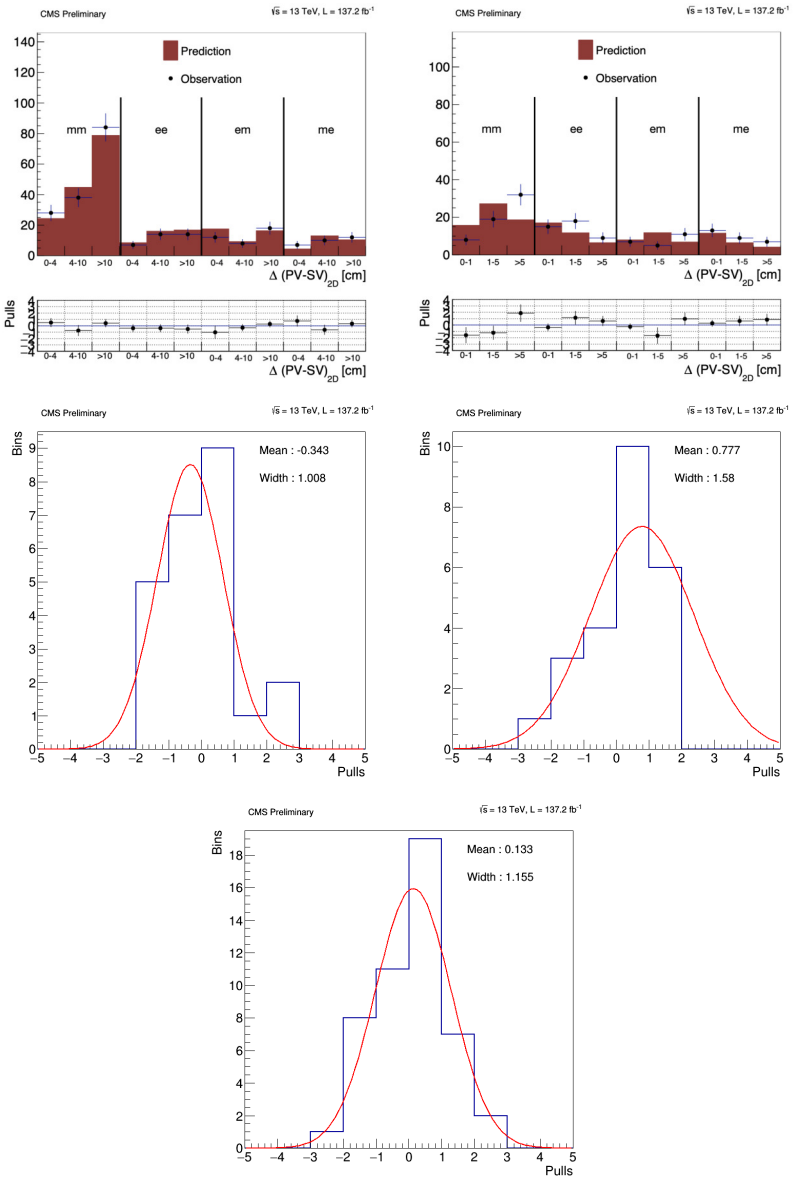


Figure 5.25: Pulls for CR2 as a function of displacement for each flavor. We show the prediction for the full runII, with plots for the low mass PFNs on the left and the high mass PFNs on the right, OS and SS categories are merged together. Bottom plot is for low mass PFNs and high mass PFNs are combined together.

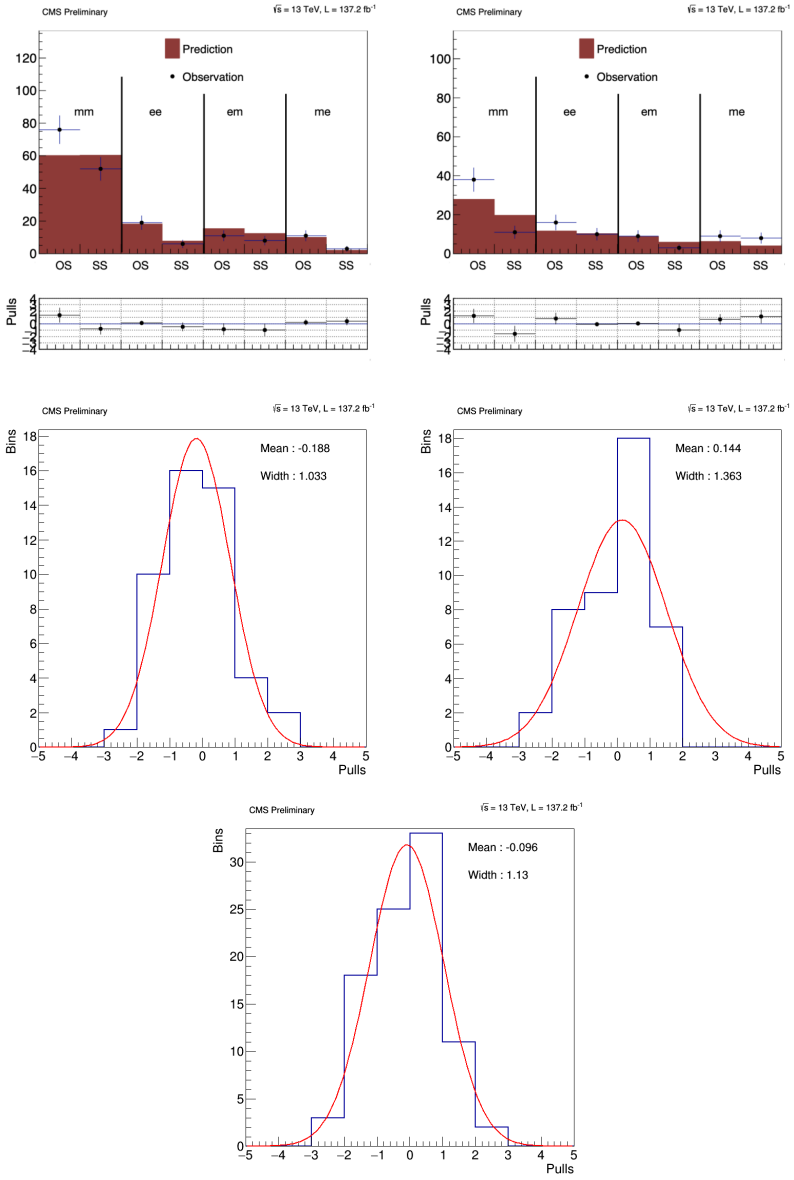


Figure 5.26: Pulls for CR2 as a function of the lepton's charge. We show the prediction for the full runII, with plots for the low mass PFNs on the left and the high mass PFNs on the right, displacemenet bins are merged together. Bottom plot is for low mass PFNs and high mass PFNs are combined tog ether.

5.8 Systematics

Details on systematics uncertainties are reported in appendix Sec B.1.
The most important source of systematic uncertainty comes from the theory assumption on the HNL model. Then we considered also uncertainties coming from pile-up luminosity and trigger, jets corrections, PFN and ABCD method as mentioned in the previous section. Systematic uncertainties are also assigned to the SV determination.

5.9 Results

In this section I'll present the results for this analysis. I haven't worked on these limits.

An exclusion of HNL signal scenarios by computing the excluded cross section for each $(\mathrm{m}_N, |\mathrm{V}_{\mathrm{N}\ell}|^2) hypothesis is given. Limits are recalculated with the modified frequentist construct$ with a binned profile likelihood test statistics, using the bins of Figs. 5.27 — 5.30. This fit is performed in the asymptotic approximation [105]. The exclusion limits are evaluated from a simultaneous fit of predicted signal and background events in the considered displacement regions, separately in the ee and $\mu\mu$ channels. We plan to add an interpretation for HNLs coupling to both muons and electrons as well, where we use the additional $e\mu$ and μe signal regions that open up in this scenario. The excluded regions of the HNL mass and the corresponding mixing parameter are presented in Figs. 5.31 — 5.32. Figure 5.33 shows the combined full Run 2 limits.

The analysis is currently under the process of unblinding data in the signal region. Dirac interpretation of the results will also come soon.

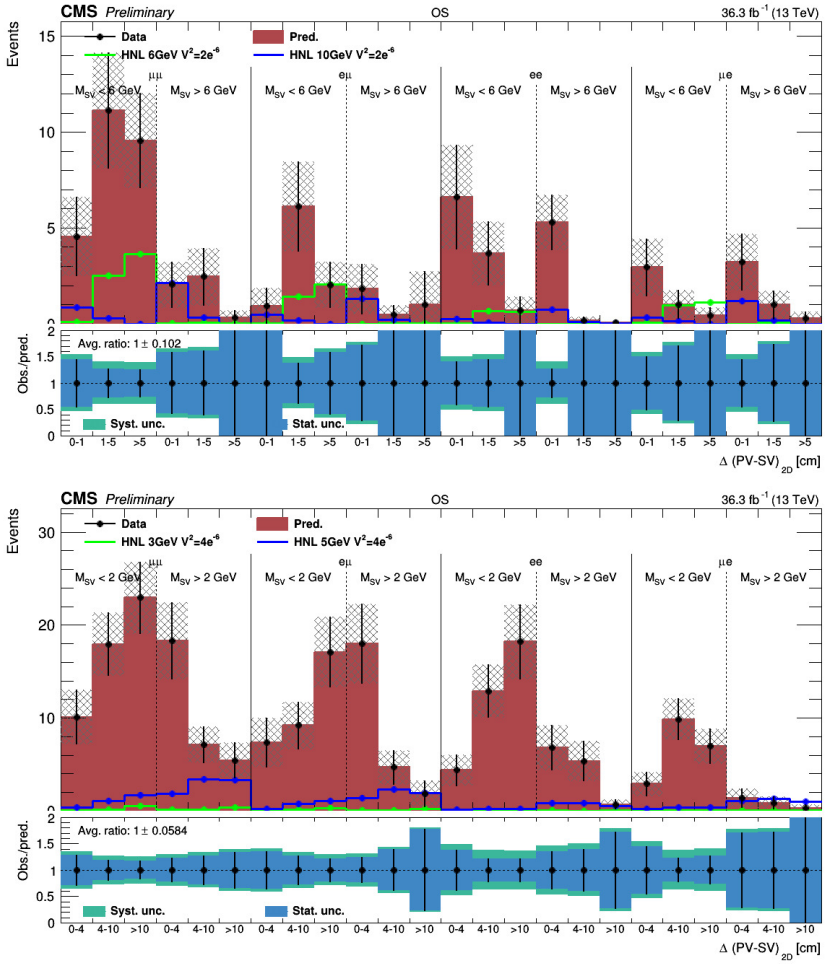


Figure 5.27: The predicted background and signal yields in the signal region for opposite-sign 2016 events. We show the prediction for the high-mass PFNs (top) and the low-mass PFNs (bottom).

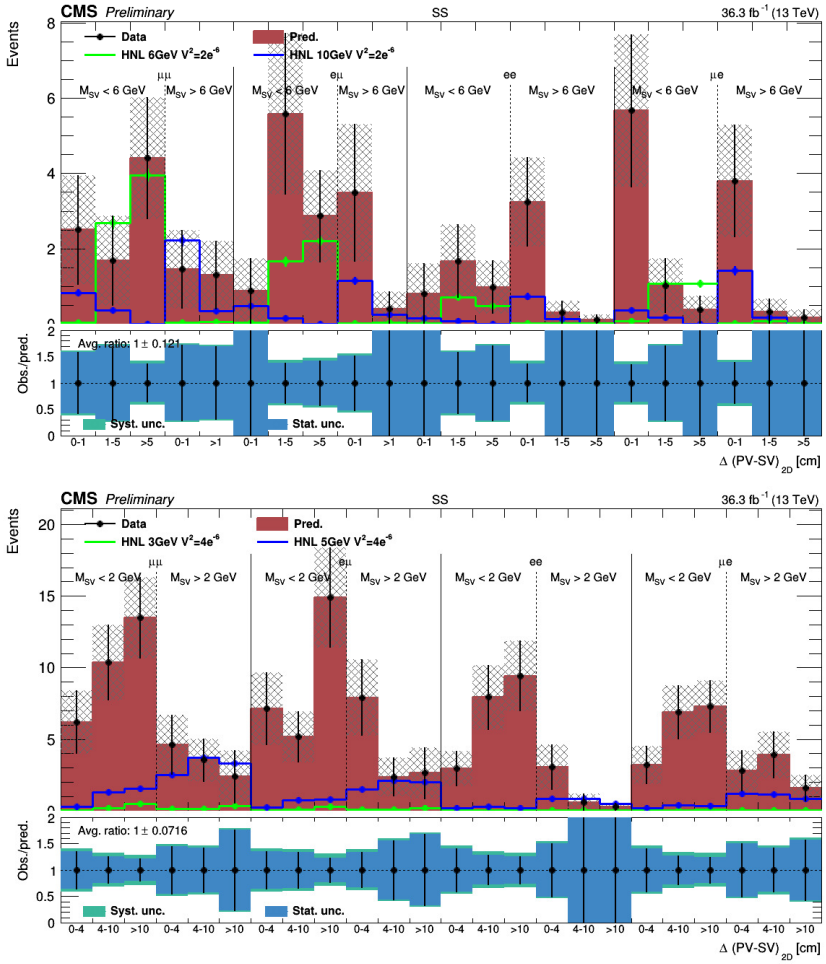


Figure 5.28: The predicted background and signal yields in the signal region for same-sign 2016 events. We show the prediction for the high-mass PFNs (top) and the low-mass PFNs (bottom).

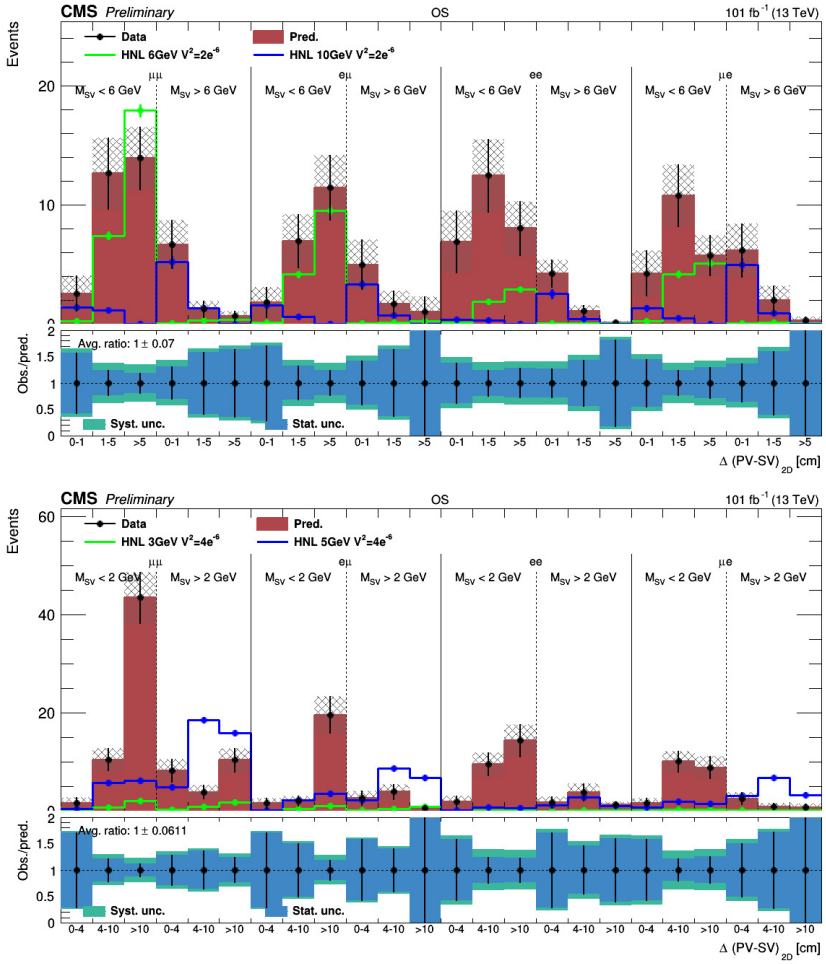


Figure 5.29: The predicted background and signal yields in the signal region for opposite-sign 2017+2018 events. We show the prediction for the high-mass PFNs (top) and the low-mass PFNs (bottom).

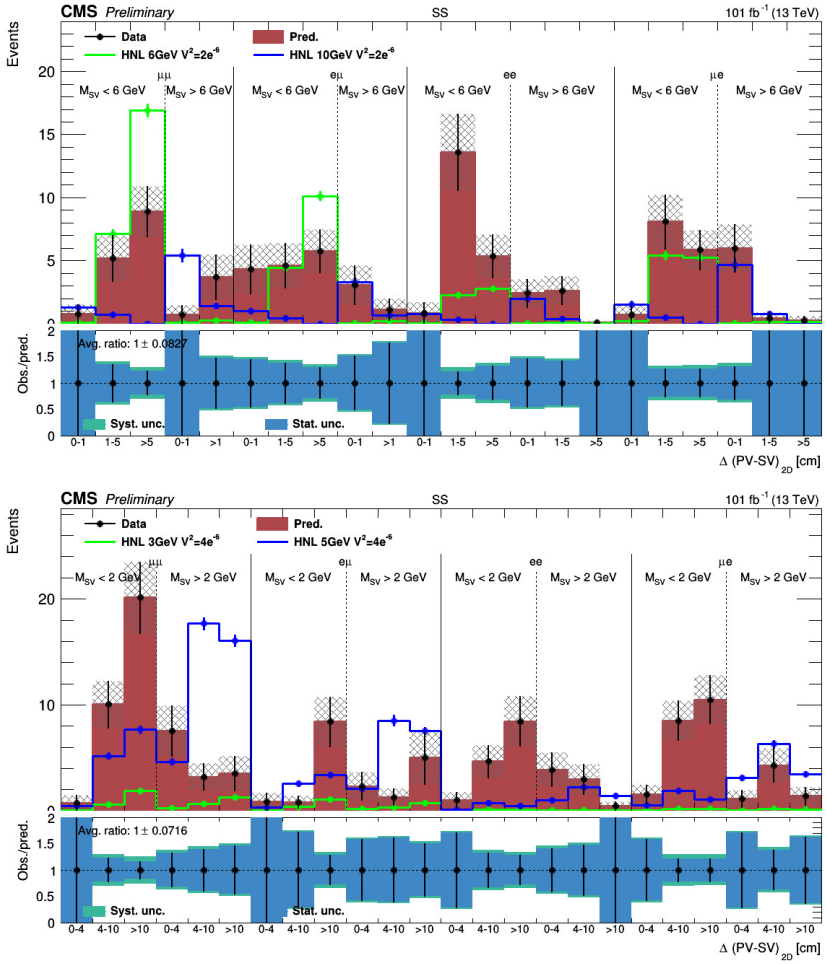


Figure 5.30: The predicted background and signal yields in the signal region for same-sign 2017+2018 events. We show the prediction for the high-mass PFNs (top) and the low-mass PFNs (bottom).

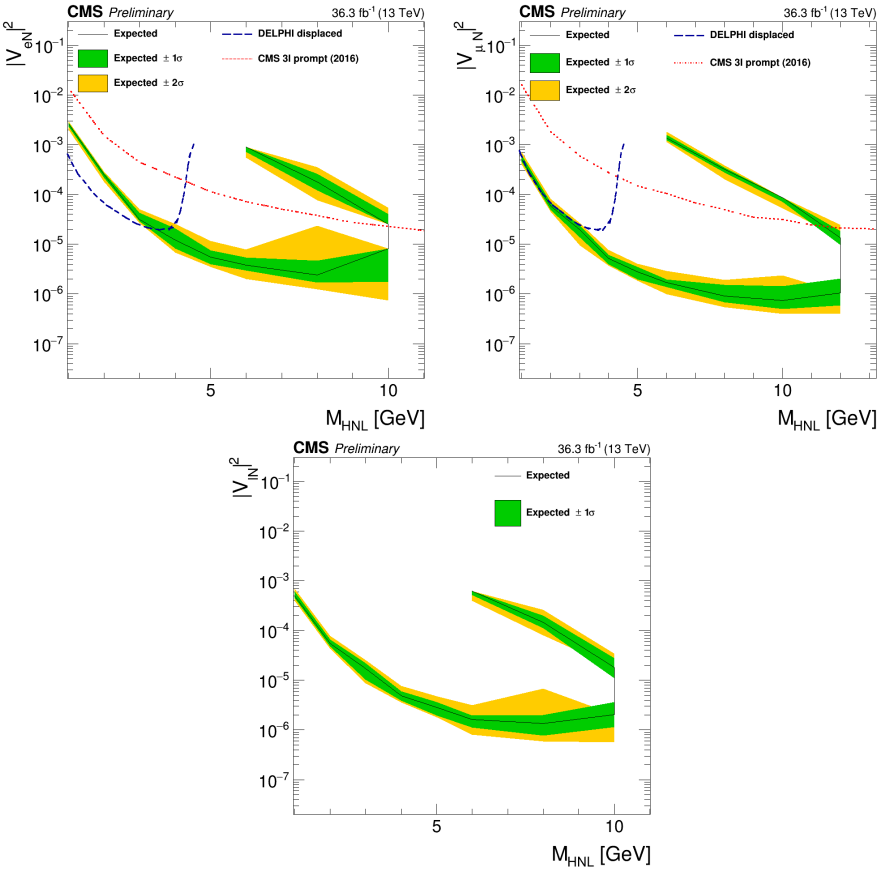


Figure 5.31: Limits on $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) and mixed coupling $|V_{Nl}|^2$ ($l = e$, μ) (below) as a function of M_N for a Majorana HNL using the 2016 data set. The excluded region is situated between an upper and lower limit line. Above the upper limit, we are no longer sensitive to HNLs because they are no longer long-lived

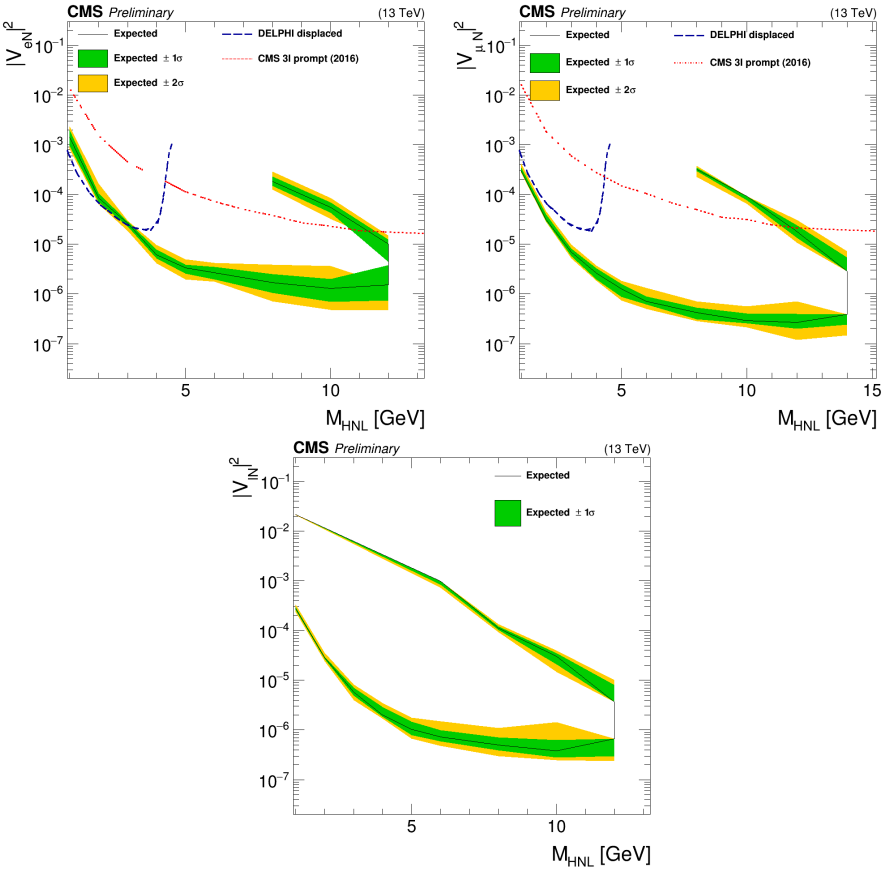


Figure 5.32: Limits on $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) and mixed coupling $|V_{Nl}|^2$ ($l = e$, μ) (below) as a function of M_N for a Majorana HNL using the 2017 and 2018 data set. The excluded region is situated between an upper and lower limit line. Above the upper limit, we are no longer sensitive to HNLs because they are no longer long-lived

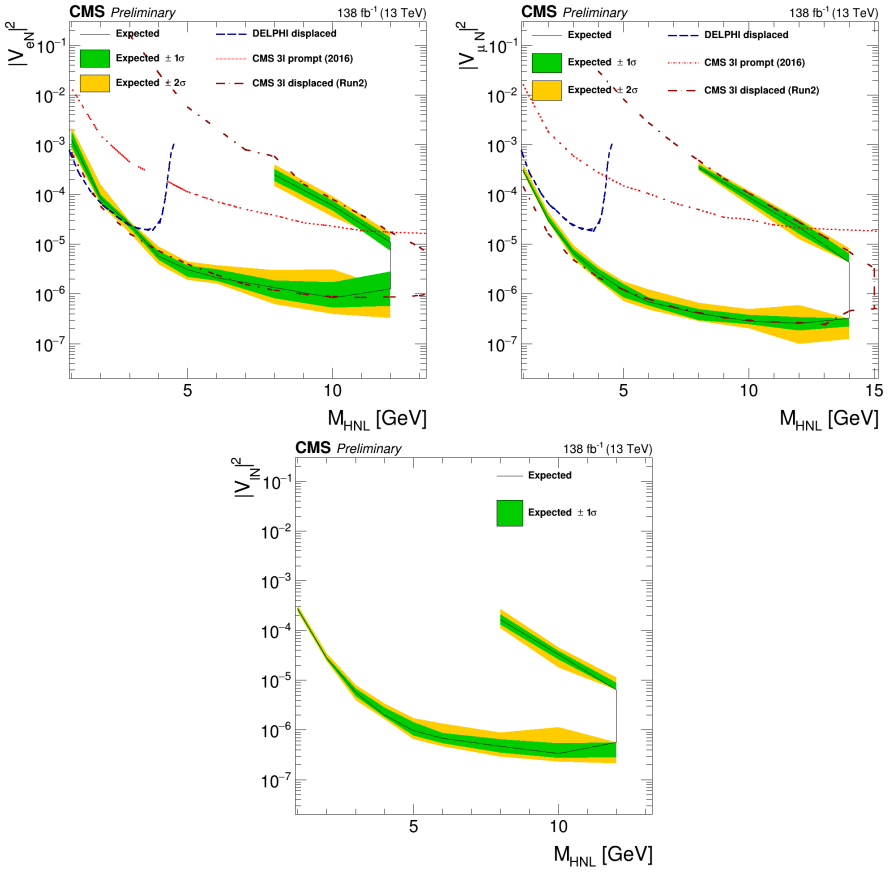


Figure 5.33: Limits on $|V_{Ne}|^2$ (left), $|V_{N\mu}|^2$ (right) and mixed coupling $|V_{Nl}|^2$ ($l = e$, μ) (below) as a function of M_N for a Majorana HNL using the full Run 2 data set. The excluded region is situated between an upper and lower limit line. Above the upper limit, we are no longer sensitive to HNLs because they are no longer long-lived

5.10 Conclusion

The evidence of neutrino oscillation was one of the first observation of new physics not incorporated in the SM theory, thus the understanding of the neutrino behavior is extremely important to shed a light on BSM theories. Therefor, for the LHC experiments, the investigation of these peculiar signature become crucial.

We are interested in studying the properties of the right-handed neutrino. The two parameter that we want to determine are the mass and the mixing parameter $|V_{\alpha N}|^2$, with the SM neutrino of flavor α, related to the Yukawa coupling $F_{\alpha N}$. The results that we provided are express in terms of the $|V_{\alpha N}|^2$ as a function of m_N for a given flavor α.

Only results with the Majorana scenario are presented in this thesis, we are working on the Dirac scenario as well.

The signature that we are searching for is a prompt lepton, a displaced lepton, a displaced jet, and a secondary displaced vertex. The signal is very difficult to capture for two main reasons: the displaced vertex and the lack of SM processes that can be used to described the expected background. For the former, the search looks for a displaced vertex containing a lepton track thus a more accurate algorithm for displaced vertex had been studied (see Sec 5.4.4) for the latter, the lack of background processes that can be simulated with the MC led us to evaluate the background contamination using the ABCD data driven techniques (see Sec 5.7). Last but not least having a set of kinematics cuts that can enhance the signal versus the background is extremely complicated and almost impossible without loosing signal efficiency thus the use of a machine learning technique to better capture the signal structure is unavoidable. We used a PFN network (described in Sec 5.6) that is trained for the different flavor and for 2016 and 2017-18 separately; the network has been made agnostic on the information of the prompt lepton to enter in the ABCD variables. Limits are extracted in the $(m_N, |\mathrm{V_{N\ell}}|^2)$ plane by fitting the predicted signal and background events in the considered displacement regions see Sec .

Results are provided for the full run2 with an integrated luminosity of 137 fb^{-1}.

The results of this analysis are very competitive with the other CMS searches in the same mass-coupling plane. The main limitation of this analysis is the lack of MC statistic both for the signal simulation and for the background prediction. The first straightforward improvement for the next version of this analysis would be the increasing of the MC statistic for both signal and background

samples. The other possible improvement comes from a better description of the secondary vertex: with new machine learning techniques we can better capture the inner structure of the secondary vertex at large displacement, we don't have to forget that with the increasing of the displacement we loose an important factor of muon (electron) reconstruction efficiency. Furthermore a possible improvement of the ABCD method could be investigated to improve the results.

5.11 Outlook

From the dissertation done in this chapter 5, it is absolutely clear how the HNL plays a major role in probing physics beyond the SM. If they happen to exist, their Dirac or Majorana behavior, their masses and their mixing parameters with the SM particles will have to be determine. For this reason lots of effort has been put into having projection for this particles at higher energy than the one accessible now, probing heavy neutral leptons at MeV or TeV mass scales. With the start of the High-Luminosity program at LHC, Ref [23], the CMS detector will go under an ambitious detector upgrade described in Sec 2.5. The long lived searches will benefit from the addition of new layers in the tracker, increasing the coverage in the forward region, from the MIP detector, that will help in the reconstruction of the secondary vertex with a better timing and spatial resolution, from the new track triggers that will be performed for the HL-LHC.
Some projection has been done for the displaced searches at 3 ab^{-1}. Referring to [106] the authors probe the sensitivity of long-lived HNLs for CMS, ATLAS and LHCb detectors. It is proposed to extend the analysis acceptance through the reconstruction of the SV using the muon system; in this way even the HNLs decaying outside the tracker volume are accounted. Results are shown in Fig 5.34. The extrapolation of the limits in the mass-coupling plane are found to be three order of magnitude better then the results obtained by DELPHIS.

A more clear overview of the projections for the HNL searches is given in Fig 5.35. The exclusion reach using Run3 and 3 ab^{-1} (HL-LHC) data is shown for long-lived HNL analysis performed at CMS or ATLAS [24].

The focused searches at the high energy electron-positron circular collider, FCC-ee, will provide the next breakthrough moment in the far future: HNL sensitivities shows exclusion limits down to $|V_{lN}^2| \simeq 10^{-12}$ covering N masses

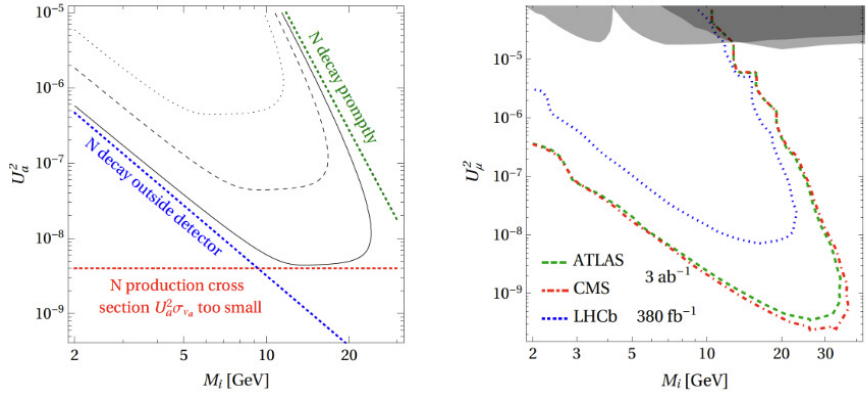


Figure 5.34: Left: a simplified illustration of the three main obstacles in improving the sensitivity. Right: exclusion reach of ATLAS, CMS with 3 ab^{-1} and LHCb with 380 fb^{-1} for the HL-LHC under ideal conditions for pure electron and pure muon mixing. Plots are taken from [106]

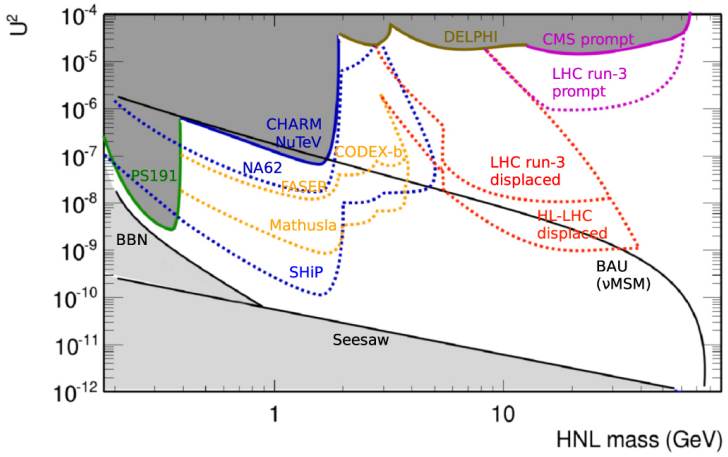


Figure 5.35: Summary of projected experimental sensitivities to HNLs in various experiments, in the coupling strength versus mass plane. Plot is taken from [24]

between 10 and 80 GeV [24]. The estimates are based on $Z \to \nu N$ decays, with N decaying further into $N \to \mu^+ W^- \to \mu^+ q\bar{q}$. Long-lived HNL scenarios with decay lengths ranging from 0.01 cm to 500 cm are investigated [107].

Chapter 6

ttH analysis

6.1 Introduction

In this chapter I will present the ttH differential-EFT analysis, based on the ttH inclusive analysis that was public in 2019 [108].
With this analysis we are targeting a differential cross section measurement using the transverse momentum of the Higgs as figure of merit and an EFT interpretation of the results.
A brief theoretical introduction follows, more details in Sec 1.5.2 and in Chapter 1.
The observation of a scalar boson by the CMS and ATLAS experiments in 2012 [109, 110] opened a new field for explorations in the realm of particle physics: the properties of the newly observed boson must be measured in detail to ascertain that it is compatible with the Higgs boson predicted by the SM. For a given Higgs mass, the SM provides accurate predictions for all its properties, among which the couplings to other SM particles.
The SM Yukawa couplings y_{f} of Higgs to fermions are proportional to the fermion mass m_{f}, namely $y_{\mathrm{f}} = \sqrt{2}m_{\mathrm{f}}/v$, where $v \approx 246$ GeV denotes the vacuum expectation value of the Higgs field. The top quark, with a mass of $m = 173.34 \pm 0.76$ GeV [111], is by far the heaviest fermion to date: its Yuwaka coupling y is expected to be of the order one. The large mass of the top quark may indicate that it plays a special role in the mechanism of electroweak symmetry breaking [112–114]; deviations of y from the SM prediction would unambiguously indicate the presence of new physics beyond the SM, and thus the determination of y is of a special interest in the study of the Higgs boson.
The Yukawa couplings of the Higgs to fermions have been determined with a large overall uncertainty during the LHC Run 1. During the LHC Run 2, the Yukawa coupling y of the Higgs to the top quark divided by its SM expectation, $\kappa_t := y/y^{SM}$, has been determined by the CMS Collaboration

to be within $-0.9 < \kappa_t < -0.7$ or $0.7 < \kappa_t < 1.1$, at 95% confidence level [108] by probing the associated production of the Higgs in association with a top quark pair and subsequent decay to leptonic final states. This measurement has been improved via more aggressive event categorization in Ref. [84], yielding the following best fit: $\kappa_t = 1.01^{+0.11}_{-0.11}$. Several analysis had been done by the CMS collaboration to measure the ttH process; depending on the Higgs decay, three different searches can be distinguished: ttH where $H \to b\bar{b}$ which is the channel with the highest branching fraction but also with the highest background contamination [115], ttH where $H \to \gamma\gamma$ [116] which has the clearest signature but also the lowest branching fraction and ttH where $H \to W^+W^-, ZZ, \tau^+\tau^-$ with the advantage of having reasonably high branching fraction and clear final state given by the vector boson (taus) production [117]. For the purpose of this thesis, we will focus on the ttH production with a multileptonic final state, where H decays via $H \to W^+W^-, ZZ, \tau^+\tau^-$ and top quarks decay either leptonically or hadronically.

The measured production rates for the ttH and tH signals in the multilpetonic final state has been measured in [117] amount to 0.92±0.19(stat)+0.170.13(syst) and 5.7±2.7(stat)±3.0(syst) times their respective standard model (SM) expectations. Assuming that the H boson coupling to the τ lepton is equal in strength to the values expected in the SM, the coupling y_t of the H boson to the top quark is constrained, at 95% confidence level, to be within $0.9 < y_t < 0.7$ or $0.7 < y_t < 1.1$ times the SM expectation for this coupling. Following the stream of [117] we performed a differential and EFT analysis in ttH with a multileptonic final state.

The measurements are based on the data recorded by the CMS experiment in pp collisions at $\sqrt{s} = 13$ TeV center-of-mass energy during LHC Run 2, corresponding to an integrated luminosity of 137.2 fb^{-1}. The measurements are done in final states with electrons, muons, and hadronically decaying tau (τ_h leptons. In particular, the following final states are considered: $2\ell\text{ss} + 0\tau_h$, $2\ell\text{ss}+1\tau_h$ and $3\ell+0\tau_h$, where ℓ denotes light leptons (e, μ), "ss" stands for the "same sign". As in the previous analysis, the separation of the ttH and tH signals from backgrounds is improved with machine learning techniques, mainly Boosted Decision Trees (BDTs) and Artifical Neural Networks (ANNs), as well as with matrix element methods (MEM) [118, 119]. The differential analysis observables have been reconstructed using either simple combinations of final-state objects (for m_{ttH}) or via ANNs-based regression. Complementary

to the differential and EFT analysis, a CP search in the same final state has also being published by the CMS experiment [120].

6.1.1 Publication plans and my contribution

I started to work on this analysis around one year ago and I will be the responsible person.
My main contributions are the differential analysis and the unfolding procedure together with the EFT validation and extractions of the results.
The analysis is still under the process of the review from the CMS collaboration.

6.2 Analysis strategy

The event selection is designed to detect the rare ttH and tH (tHq and tHW are considered) signals in multilepton final states. To achieve this, three event categories are defined, selecting events with large lepton and τ_h multiplicity: $2l0\tau$, $2l1\tau$ and $3l0\tau$. These regions correspond to the most sensitive ones in the inclusive measurement of ttH and tH production in the same channel and are described in Section 6.5. Additionally, these regions correspond to those that contain enough ttH events to be sensitive to the signal kinematic properties, which is a necessary aspect in a search for anomalous values of EFT couplings. The number of events in the $2l0\tau$, $2l1\tau$, and $3l0\tau$ categories is not enough to allow for sub-categorization on top of the use of a differential distribution in some kinematic observable. The contribution of signal and background events is briefly described in Section 6.6. Signal and irreducible backgrounds are estimated using samples of simulated events, while reducible backgrounds are largely suppressed by the lepton and τ_h selection described in Section 6.4, and are estimated using data-driven methods. This selection allows to gain sensitivity to the inclusive ttH and tH production. We then reconstruct kinematic observables related to the Higgs boson, to use them to extract differential cross sections for ttH production (SM analysis) and set constraints on anomalous values of several effective field theory (EFT) operators (BSM analysis), as well as on the value of the Higgs boson self-coupling. We regress the Higgs transverse momentum using a deep neural network, although we also document alternative techniques, based on simple combinatorics requirements and on the solution of the equations for the kinematics of the final state by means of χ^2 minimization. We perform a so-called “SM

analysis", where we unfold to particle level the Higgs boson transeverse momentum. For the Higgs transeverse momentum, we use the latest Les Houches agreement on the simplified template cross section (STXS) binning of that observable. We also perform a so-called "BSM analysis", where we use the kinematic observables mentioned above to set confidence regions on the values of EFT operators describing possible new-physics effects. We build a statistical model encompassing all the operators that have an effect on top-related signals and backgrounds (ttH, ttW, ttZ, tHq, and tZq) as well as on all QCD-related processes. The statistical model encompasses all possible EFT operators and is based on Ref. [121]. Despite the selection mentioned above, the regions are still dominated by background events. We are considering the use multivariate methods, namely artificial neural networks, to separate the contribution from the various signal and background species. These discriminators were developed for the inclusive measurement and take as an input variables related to the event kinematics. The neural networks have 3 (4) output nodes in the $2l1\tau$ and $3l0\tau$ ($2l0\tau$) categories, which estimate the probability for each event to correspond to ttH, tHq or another background (to ttH, tHq, ttW or other background). The extra node is added in the $2l0\tau$ category to gain sensitivity to the irreducible ttW background.

6.3 Data samples and Monte Carlo simulation

The analyzed data were collected in pp collisions at $\sqrt{s} = 13$ TeVcenter-of-mass energy and with a 25 ns bunch crossing period. The events were recorded using a combination of single, double, and triple lepton triggers and by triggers based on the presence of two τ_h or a lepton and a τ_h. Only data-taking periods where all detector systems were fully operational are included in the analysis. The integrated luminosity of the analyzed dataset amounts to 35.9 fb^{-1} in 2016, 41.5 fb^{-1} in 2017 and 59.7 fb^{-1} in 2018, totalling 137.2 fb^{-1} in the full LHC Run 2. Approximately 30 inelastic pp interactions (pileup) occurred per bunch crossing, on average.

Samples of ttH, tHq, and tHW signals and of background events, produced by Monte Carlo simulation, are used for the purpose of estimating signal and background yields in the analysis and to train machine learning algorithms. The following processes are simulated: Z+jets, W+jets, single top, top quark pairs, diboson (WW, WZ, ZZ, $W\gamma$, and $Z\gamma$), triboson ($WPWW$, WWZ, WZZ, ZZZ, and $WZ\gamma$) production, the production of a single Higgs boson by gluon-gluon fusion (ggF) and by vector-boson fusion (VBF), the associ-

ated production of a Higgs boson with vector bosons (WH and ZH), double-Higgs production (HH), as well a few selected "rare" processes. The exotic processes, such as $tttt$ and the production of same-sign WW boson pairs, typically have very small cross sections, but may nevertheless yield non-negligible background contributions in some event categories. All the simulated samples are produced setting the Higgs boson mass to $m_{Higgs} = 125$ GeV.

The samples that are produced from leading-order (LO) matrix elements are generated with the MADGRAPH5_AMCATNLO [122] event generator. The implementation of next-to-LO (NLO) matrix elements in the simulation is performed using the MADGRAPH5_AMCATNLO and POWHEG $v2$ [123–125] generators. Below we detail which generator for used for each of the simulated processes. Parton shower, hadronization processes, and decays of leptons—including polarization effects—are modelled using PYTHIA. The event generator tunes are based on the *Monash* tune [126]. The samples produced by PYTHIA with *CUETP8M1* tune use the NNPDF2.3LO set of parton distribution functions (PDFs); the samples produced using MADGRAPH and POWHEG, or using PYTHIA with the *CUETP8M2** tune, use the NNPDF3.0 PDF set; finally, the samples produced using PYTHIA with the *CP5* tune use the NNPDF3.1 PDF set [127–129].

The contribution from the ttH signal and the backgrounds arising from $t\bar{t}$ production in association with W and Z bosons (ttW and ttZ), from triboson (WWW, WWZ, WZZ, ZZZ, and $WZ\gamma$) production, as well as from the production of four top quarks ($tttt$) are generated at NLO using MADGRAPH5_AMCATNLO [122]. POWHEG $v2$ is used to simulate at NLO: the backgrounds arising from +jets, tW, and diboson (WW, WZ, and ZZ) production; the backgrounds arising from the production of single top quarks; the backgrounds arising from SM Higgs boson production via gluon fusion (ggH) and vector boson fusion (qqH) processes; and the backgrounds arising from the production of SM Higgs bosons in association with W and Z bosons (WH, ZH) and with W and Z bosons along with a pair of top quarks ($t\bar{t}WH$, $t\bar{t}ZH$). The tH and the ttH signal samples, as well as the $tt\gamma$, tZ, $ttWW$, W+jets, Drell–Yan (DY), W/γ, and Z/γ backgrounds, are generated at LO accuracy using MADGRAPH5_AMCATNLO.

For what concerns the simulation of EFT effects, each of the signal and background processes that is affected by the value of EFT operators is generated with a set of event weights corresponding to different values of the relevant EFT operators. We give the details of this procedure in Section 6.11.4.

The Z+jets, W+jets, and $t\bar{t}$+jets processes are normalized to cross sections computed at next-to-next-to-leading-order (NNLO) accuracy [130–132]. The cross sections of single top quark [133–135], ttW, and ttZ production, as well as ttH, tHq, and tHW signal production [136], are computed at NLO accuracy. Special care is taken when scaling the ttZ background to the NLO cross-section of Ref. [136]: there are subtle differences between the process generated in the CMS simulated samples and the process for which the NLO calculation was performed.

In the case of the ttW background, besides the NLO MADGRAPH5_AMCATNLO samples, we use additional privately produced samples to improve the modeling of this background in the three years. In these samples, the ttW process is modeled including the higher order NLO3+NLO4 electroweak corrections, as recently derived in [137], and the associated cross section to this correction amounts to 0.0162562 pb.

All simulated events are overlaid with minimum-bias events generated with PYTHIA, according to the luminosity profile of the analyzed data and for a pp inelastic cross section of 69.2 mb.

All generated events are passed through a detailed simulation of the CMS apparatus, based on GEANT4 [138], and are reconstructed using the same version of the CMS event reconstruction software as used for data.

In order to improve the modeling of the data, we apply corrections to simulated events, which we denote as "data-to-Monte Carlo" corrections. The difference between data and Monte Carlos simulation refer to the following calibrations:

- pileup reweighting;
- trigger efficiency;
- e and μ identification and isolation efficiency;
- τ_h identification efficiency;
- τ_h energy scale;
- -tag efficiency and mistag rate;
- E_T resolution and response;
- prefiring probability of Level-1 ECAL trigger;
- reweighting of Drell-Yan events;

- reweighting of $t\bar{t}$ events.

6.4 Object selection

The object selection is mutated from the inclusive ttH search [108] and in this section I will briefly describe the objects that we used.

6.4.1 Object selection

Lepton selection

Topologies with prompt leptons are one of the main distinctive features of signal in this analysis. The lepton selection of this analysis aims to efficiently select prompt leptons and rejecting leptons coming from other sources, such as leptons produced in the decay of heavy flavor hadrons or due to detector mismeasurements that may contaminate the search region. As described in Sec 4.5, this source of background goes under the name of 'reducible source of background' and originate from processes that contain one or more non prompt lepton. To estimate the reducible background we will follow a data-driven approach very similar to the one used in Sec 4.5 that will be briefly described in Sec 6.6.1. Depending on the object that we are targeting, three increasingly selection criteria are applied for leptons:

- loose leptons are required to have at least 5 GeV and to be within the acceptance $|\eta| < 2.4(2.5)$ for muons (electrons). Additionally, they are required to pass mild requirements on identification and isolation.

- fakeble leptons will be used to estimate the amount of fakable object in the signal region. Fakeable leptons are required to pass the loose identification criteria. In order to reduce potential biases of the background estimation procedure, the p_T of leptons that pass the fakeable but fail the tight lepton selection criteria is set to 0.90 times the p_T of the associated jet, when the distance between lepton and associated jet satisfies the requirement $\Delta R < 0.4$.

- tight selection is used to achieve maximum prompt lepton purity. Leptons in this category are required to have a prompt-lepton MVA score, defined in [108], greater than 0.85 (0.8) for electrons (muons). Muons are additionally required to pass the medium identification criteria (loose

muon see 3.2 with additional track-quality and muon-quality requirements [139]). The working point (WP) of the cut on the prompt-lepton MVA has been carefully tuned to achieve maximum sensitivity in the analysis.

All the selection criteria in electrons and in muons are summarized in tables 6.1 and 6.2, respectively.

Electrons			
Observable	Loose	Fakeable	Tight
Cone-p_T	$> 7\ GeV$	$> 10\ GeV$	$> 10\ GeV$
$\|\eta\|$	< 2.5	< 2.5	< 2.5
$\|d_{xy}\|$	< 0.05 cm	< 0.05 cm	< 0.05 cm
$\|d_z\|$	< 0.1 cm	< 0.1 cm	< 0.1 cm
d/σ_d	< 8	< 8	< 8
I_e	$< 0.4\times$p$_T$	$< 0.4\times$p$_T$	$< 0.4\times$p$_T$
$\sigma_{i\eta i\eta}$	—	$<$ { 0.011 / 0.030 }[1]	$<$ { 0.011 / 0.030 }[1]
H/E	—	< 0.10	< 0.10
1/E - 1/p	—	> -0.04	> -0.04
Conversion rejection	—	✓	✓
Missing hits	≤ 1	$= 0$	$= 0$
Jet relative isolation[1]	—	< 0.7 (—) †	—

[1] Defined as $1/{p_T}^{ratio}$-1 if the electron is matched to a jet within $\Delta R < 0.4$ or as the PF relative isolation with ΔR=0.4 otherwise.
† Fails (passes) the requirement prompt-e MVA > 0.80.

Table 6.1: Loose, fakeable and tight selection criteria for electrons. A long dash (—) indicates selection criteria that are not applied.

τ_h selection

τ_h are reconstructed using the hadron-plus-strips (HPS) algorithm [140]. Similarly to the selection of light leptons and with the same purposes, three levels of τ_h identification are used in the analysis: loose, fakeable, and tight. We use the "Deep Tau v2.1" discriminator [141], which we refer to as DeepTau, together with other variables for these selections. The criteria for each selection are described in Table 6.3.

Muons			
Observable	Loose	Fakeable	Tight
p_T	$> 5\ GeV$	$> 10\ GeV$	$> 10\ GeV$
$\|\eta\|$	< 2.4	< 2.4	< 2.4
$\|d_{xy}\|$	< 0.05 cm	< 0.05 cm	< 0.05 cm
$\|d_z\|$	< 0.1 cm	< 0.1 cm	< 0.1 cm
d/σ_d	< 8	< 8	< 8
I_μ	$< 0.4\times p_T$	$< 0.4\times p_T$	$< 0.4\times p_T$
Jet relative isolation[1]	—	<0.5 (—) †	—

[1] Defined as 1/`jetPtRatio`-1 if the muon is matched to a jet within $\Delta R < 0.4$ or as the PF relative isolation with ΔR=0.4 otherwise.
† Fails (passes) the requirement prompt-μ MVA > 0.85.

Table 6.2: Loose, fakeable and tight selection criteria for muons. A long dash (—) indicates selection criteria that are not applied.

Jet selection and b tagging

Jets are reconstructed using particle-flow candidates from the primary vertex using the anti-kt algorithm [100] with a distance parameter $R = 0.4$. Jets are required to pass the loose (tight) working point of the Particle Flow jet identification criteria in 2016 (2017 and 2018) [63]. The energy of reconstructed jets is calibrated as a function of jet p_T and η [142].The jet energy scale (JES) systematic uncertainties are evaluated by shifting the JES applied to the reconstructed jets up and down by one standard deviation, following the recom-

Hadronic τ			
Observable	Loose	Fakeable	Tight
p_T	$> 20\ GeV$	$> 20\ GeV$	$> 20\ GeV$
$\|\eta\|$	< 2.3	< 2.3	< 2.3
$\|d_z\|$	< 0.2 cm	< 0.2 cm	< 0.2 cm
Decay mode finding	New	New	New
Decay modes	All	All except 2-prong(+π^0)[1]	All except 2-prong(+π^0)[1]

[1] Tau CMS reconstruction recommends [141] to use the 2-prong and 2-prong+π^0 decay modes only in analysis containing high-p_T τ_h's.

Table 6.3: Loose, fakeable and tight selection criteria for hadronic τ decays. A long dash (—) indicates selection criteria that are not applied.

mendations of CMS for the jet energy scale uncertainties [64]. The events are then re-analyzed, including the re-application of the jet-based selection and the computation of all relevant kinematic quantities, to derive the varied kinematic distributions, which enter as shape uncertainties (including rate effects) in the final fit.

To discriminate jets produced by heavy flavor quarks from those coming from light flavor quarks and gluons ("light jets") we make use of the DeepJet discriminator [65, 143]. In this analysis we use both the loose and medium b tagging WPs.

Samples of simulated events in the analysis are corrected by data-to-MC correction factors that account for the lepton and τ_h selection and reconstruction efficiency, jet energy response, and the efficiency and mistag rate of the b tagging methods.

Missing transverse energy

The missing transverse momentum vector is computed as the negative sum of the transverse momentum vectors of all the Particle-Flow candidates reconstructed in the event.

6.5 Event selection

The event selection aims at targeting events where the Higgs boson is produced in association to a pair of top quarks (ttH) and decays into a pair of W or Z bosons, or τ leptons. The W and Z bosons can subsequently decay either hadronically or into electrons or muons, while the τ can decay to electrons, muons, or τ_h. Events with a pair of loose leptons with an invariant mass smaller than 12 GeVare rejected, as they are not well modeled by the simulation. Moreover, events are required to contain at least one jet of $p_T > 25$ GeV. A combination of single-lepton triggers and triggers based on the presence of a lepton and a τ_h (also known as lepton+τ_h "cross-triggers") are used to record events in the channels containing one lepton. A combination of single-lepton and dilepton triggers are used to record events in the channels containing two leptons, where the inclusion of single-lepton triggers boosts the events acceptance thanks to the absence of a p_T threshold on a second lepton. Similarly, a mix of single-lepton, dilepton, and trilepton triggers is used to record events in the channels containing at least three leptons. For further details and trigger paths see appendix C.1.

6.5.1 Event channels

The events are analyzed in mutually exclusive categories, also referred to as "channels", based on the multiplicity of light leptons and τ_h, with requirements on the charge of the light leptons in some cases. The three different channels covered in this analysis are $2lss + 0\tau_h$, $2lss + 1\tau_h$ and $3l + 0\tau_h$, where l refers to an electron or muon and the τ_h refers to a hadronic τ, while the "ss" ("os") refers to same sign (opposite sign) electrons or muons. Details on the selections applied to each channel are shown in what follows, and a summary can be found in Tables 6.4 and 6.5.

$2lss + 0\tau_h$ category

The $2lss + 0\tau_h$ category targets ttH or tH signal events in which the Higgs boson decays into a pair of W bosons, one of which decays leptonically while the other decays hadronically. In the ttH case, one top quark decays to leptons and the other one decays to hadrons, while in the tH case, the single top quark decays leptonically. Additionally, the lepton from the W boson and the lepton from the top quark have the same sign. Selected events are therefore required to contain two leptons of the same charge and passing the tight object selection criteria. The lepton of highest (lowest) p_T is required to have $p_T > 25GeV$ ($> 15GeV$). The requirement on both leptons to be of the same charge cuts half of the ttH signal events away, but removes almost all of the large $t\bar{t}$+jets background. The background from $t\bar{t}Z$ production is suppressed by requiring that the event contains no pair of same-flavor opposite-sign loose leptons with mass close to the mass of the Z boson, i.e. $|m_{ll} - m_Z| < 10$ GeV, where $m_Z = 91.2$ GeV [144]. In the case where the two selected leptons in the event are electrons, the invariant mass of electron pair is required not to be close to the mass of Z boson, i.e. $|m_{ee} - m_Z| < 10$ GeV, and the event must satisfy the condition $E_T^{miss} > 30$ GeV. For what concerns the jet multiplicity, we select any event containing at least one of the two following topologies, which are expected from either the ttH or the tH process, respectively:

- To target the ttH signal, we select events that have at least three jets of $p_T > 25$ GeVand $|\eta| < 2.4$, among which at least two satisfy the loose WP of the b-tagging discriminator [143] or that at least one satisfies the medium WP.

- To target the tH signal, we select events that have at least one jet of $p_T > 25$ GeVand $|\eta| < 2.4$ which passes the medium WP of the b-tagging discriminant and at least one light jet as defined in Section 6.4.1.

Events containing more than two tight leptons or a loose τ_h passing the very-loose WP of the τ_h identification discriminant are vetoed. The former avoids overlap with the $3l + 0\tau_h$ category, while the latter avoids overlap with the $2lss + 1\tau_h$ category.

$2lss + 1\tau_h$ category

The $2lss + 1\tau_h$ category targets ttH or tH signal events in which the Higgs boson decays into a pair of τ leptons, where one decays leptonically while the other decays into a τ_h. In the ttH case, one top quark decays to leptons and the other one decays to hadrons, while in the tH case, the single top quark decays leptonically. Additionally, the lepton from the τ decay and the lepton from the top quark decay have the same sign. The event selection criteria applied in the $2lss + 1\tau_h$ category are identical to those applied in the $2lss + 0\tau_h$ category, except that events selected in this category are required to contain one τ_h passing the very-loose WP of the τ_h identification discriminant, and there is no requirement on the mass of the selected electron pair m_{ee}. Also, if the second highest p_T lepton is a muon, the p_T requirement is relaxed to $p_T > 10$ GeV. The charge of the τ_h is required to be opposite to the charge of the leptons. Overlap with the $2l + 2\tau_h$ category is avoided by applying a modified τ_h veto requiring that the events selected in the $2lss + 1\tau_h$ category do not contain two loose τ_h passing the medium WP of the τ_h identification discriminant.

$3l + 0\tau_h$ category

The $3lss + 0\tau_h$ category targets ttH or tH signal events in which the Higgs boson decays into a pair of W bosons. In the ttH case, two final states are covered: those where only one top quark and the two W bosons decay leptonically, and those where both top quarks but only one W boson decays leptonically. In the tH case, the single top quark and both W bosons decay leptonically. Selected events are required to contain exactly three leptons passing the tight object selection criteria. The lepton of highest, second-, and third-highest p_T is required to have $p_T > 25$ GeV, > 15 GeVand > 10 GeV, respectively. The charge sum of the leptons is required to be either $+1$ or -1. The back-

ground from $t\bar{t}Z$ production is suppressed by vetoing events containing a pair of loose leptons of the same flavor and opposite charge with invariant mass $|m_{\ell\ell} - m_Z| < 10$ GeV. For what concerns the jet multiplicity, we select any event containing at least one of the two following topologies, which are expected from either the ttH or the tH process, respectively:

- To target the ttH signal, we selected events that have at least two jets, among which at least two satisfy the loose WP of the b-tagging discriminant or at least one satisfies the medium WP. Events containing less than four jets are required to satisfy the condition $E_T^{miss} > 45$ GeVin case the event contains a pair of loose leptons of the same flavor and opposite charge, and $E_T^{miss} > 30$ GeVif the event does not contain such lepton pair. If the event contains four or more jets, no requirement on E_T^{miss} is applied, as the contributions of background processes are negligible in this case.

- To target the tH signal, we select events that have at least one jet passing medium WP of the b-tagging discriminant and at least one light jet as defined in Section 6.4.1.

Events containing a loose τ_h passing the very-loose WP of the τ_h identification discriminant are vetoed, as are events containing two pairs of loose leptons of the same flavor and opposite charge which satisfy the condition $m_{\ell\ell\ell\ell} < 140$ GeV. While the first condition avoids overlap with the $3l + 1\tau_h$ category, the second condition avoids overlap with the $t\bar{t}H$-tagged category of the $H \to ZZ^* \to 4\ell$ analysis [145].

6.6 Background estimation

A quick summary of the background estimation is given in this section. I want to highlight that the background estimation is mutuated from the inclusive ttH search.
We differentiate between three main types of reducible backgrounds (details in Sec 4.5). The first one is called “misidentified leptons”, which means that either at least one of the reconstructed electrons or muons stem from a non-prompt lepton or hadron, or that at least one of the reconstructed τ_{h} originates from the misidentification of a quark or gluon jet. The main contribution to this background arises from tt + jets production, due to the large cross section

Selection	$2lss + 0\tau_h$	$2lss + 1\tau_h$
Targeted ttH decays	$t \to bl\nu, t \to bqq,$ $H \to WW \to l\nu$	$t \to bl\nu, t \to bqq,$ $H \to \tau\tau \to l\tau_h + \nu$'s
Targeted tH decays	$t \to bl\nu,$ $H \to WW \to l\nu qq$	$t \to bl\nu,$ $H \to \tau\tau \to l\tau_h + \nu$'s
Trigger	Single- and double-lepton triggers	
Lepton p_T	p_T > 25/15 GeV	p_T > 25/15(e)or10 GeV(μ)
Lepton η	$\eta < 2.5$ (e) or 2.4 (μ)	
τ_h p_T	—	$p_T > 20$ GeV
τ_h η	—	$\eta < 2.3$
Charge requirements	2 same-sign leptons and charge quality requirements $\sum_{l,\tau_h} q = \pm 1$	
Jet multiplicity and b-tag†	≥3 jets, ≥1 medium -tagged jet or ≥2 loose -tagged jets	
Light jet and b-tag†	≥1 light jets, ≥1 medium -tagged jet	
Missing transverse momentum	$E_T^{miss} > 30$ GeV **	
Dilepton mass	$m_{ll} > 12$ GeV *, $m_{ll} - m_Z > 10$ GeV ***, $m_{ee} - m_Z > 10$ GeV **	

* Applied on all pairs of leptons that pass loose selection.

** If both leptons are electrons.

*** Applied on all same flavor opposite sign pairs of leptons that pass loose selection.

† Pass at least one of these two jet and b-tag cuts.

‡ If the event contains a same flavor opposite sign lepton pair and $N_{\mathrm{j}} \leq 3$.

Table 6.4: Event selections applied in the $2lss + 0\tau_h$ and $2lss + 1\tau_h$ categories.

of this process and the presence of b jets in the final state. The second one is named "asymmetric photon conversions", which entails the conversion of a photon into electrons. It is possible that the electron produced in a photon conversion carries almost all the energy of the original converted photon, while the other electron carries too little energy to be detected. The contribution of these backgrounds is typically dominated by production. The third one is denoted as "charge flips". This background contributes to the 2l0τ and 2l1τ channels and originates from a mismeasurement of the lepton charge. The electron charge mismeasurement rate is very small, below the percent order for the relevant energy ranges, however, since the backgrounds yielding to opposite-sign lepton pairs have much higher cross section than signal, this corresponds to a significant background. The charge misidentification rate is negligible in muons. Most of the irreducible background is made up by $ttWW$, ttZ and diboson backgrounds, from SM Higgs boson production via the processes ggH, qqH, WH, $ttWH$ and $ttZH$ and so-called "rare" backgrounds. Rare backgrounds include processes as tZ production, the production of same-sign WW pairs, triboson and $tttt$ production. Irreducible backgrounds are modeled using samples of simulated events with the highest precision available. We use

Selection	$3l + 0\tau_h$
Targeted ttH decays	$t \to bl\nu$, $t \to bl$, $H \to WW \to l\nu qq$
	$t \to bl\nu$, $t \to bq$, $H \to WW \to l\nu l\nu$
Targeted tH decays	$t \to bl\nu$, $H \to WW \to l\nu l\nu$
Trigger	Single-, double- and triple-lepton triggers
Lepton p_T	p_T > 25 / 15 / 10 GeV
Lepton η	$\eta < 2.5$ (e) or 2.4 (μ)
τ_h p_T	—
τ_h η	—
Charge requirements	$\sum_l q = \pm 1$
Jet multiplicity**	$\geq$2 jets
tagging requirements**	$\geq$1 medium -tagged jet or $\geq$2 loose -tagged jets
Light jet and b-tag***	$\geq$1 light jets, $\geq$1 medium -tagged jet
Missing transverse**	No requirement if $N_\text{j} \geq 4$
momentum	$E_T^{miss} > 45$ GeV †
	$E_T^{miss} > 30$ GeV otherwise
Dilepton mass	$m_{ll} > 12$ GeV * and $m_{ll} - m_Z > 10$ GeV ‡
Four-lepton mass	$m_{4l} > 140$ GeV §

* Applied on all pairs of leptons that pass loose selection.
** In $3l$ category, if events do not pass these cuts, *** is required.
† If the event contains a same flavor opposite sign lepton pair and $N_\text{j} \leq 3$.
‡ Applied to all same flavor opposite sign lepton pairs.
§ Applied only if the event contains 2 same flavor opposite sign lepton pairs.

Table 6.5: Event selections applied in the $3l + 0\tau_h$ category.

NLO samples for the main backgrounds: ttZ, ttW, and WZ. We also make use of dedicated ttW samples including α^3 terms and their NLO corrections in perturbative QCD, proportional $\alpha^3\alpha_s$, which are found to have a significant effect [146, 147]. The normalization of ttZ, ttW and WZ is parameterized as freely-floating nuisance parameters that are profiled in the signal extraction fit.

6.6.1 Estimation of misidentified leptons background

To estimate the background originating from the misidentification of leptons the fake-factor (FF) method [148] is used. It is separately applied to each event category. In this method, a sample of events is selected, where all electrons and muons only have to pass the relaxed selection criteria ("fakeable objects"), while otherwise the usual selection criteria for the corresponding category are

Table 6.6: Summary of the reducible backgrounds for all channels. The dagger symbol † denotes, when a background is estimated with data driven methods, while otherwise they are estimated using MC simulations.

Process	$2ss0_h$	$2ss1_h$	30_h
Fakes†	✓	✓	✓
Flips†	✓	✓	—
Conversions	✓	✓	✓

applied. To avoid overlap with the signal region (SR), events where all leptons pass the tight selection criteria are vetoed. These event samples are referred to as the "application region" (AR) of the FF method. It is then possible to obtain an estimate of the contribution of the misidentified lepton background to the SR by applying appropriately chosen weights to the events selected in the AR where the weights themselves are derived from the so called fake-factors. The weights depend on the probability f_i for a misidentified electron or muon to pass the relaxed or fakeable selection criteria but fail the tight selection criteria. Depending on the multiplicity of the leptons passing the relaxed and the multiplicity of the ones passing the nominal selection criteria the weights differ. In equation 6.1 we list the expressions for two or three fakeable objects.

$$\begin{aligned} N_{\text{pp}}^{\text{fake}} &= \sum_{\text{fp}} F_1 + \sum_{\text{pf}} F_2 + \sum_{\text{ff}} F_1 F_2 \\ N_{\text{ppp}}^{\text{fake}} &= \sum_{\text{fpp}} F_1 + \sum_{\text{pfp}} F_2 + \sum_{\text{ppf}} F_3 \\ &\quad - \sum_{\text{ffp}} F_1 F_2 - \sum_{\text{fpf}} F_1 F_3 - \sum_{\text{pff}} F_2 F_3 + \sum_{\text{fff}} F_1 F_2 F_3 \end{aligned} \tag{6.1}$$

In equation 6.1 the $N_{\text{pp}}^{\text{fake}}$ and $N_{\text{ppp}}^{\text{fake}}$ symbols correspond to the contributions of the fake background of events containing two or three electrons or muons in the signal region. The F_i symbol corresponds to $F_i \equiv \frac{f_i}{1-f_i}$. The labels "p" and "f" show if a lepton either passes or fails the nominal selection criteria and ordered with the highest p_T on the left to the lowest on the right. To further illustrate, the label "fpp" denotes a summation over all event where the highest p_T lepton fails the nominal selection criteria, while the other two lower p_T leptons pass said selection criteria. The events in the sum are then weighed by F_1 and the sign of the weights alternates with the number of leptons failing the tight selection criteria, where an even number results in a negative sign. To address the contamination from irreducible backgrounds with prompt leptons and

genuine $_{\mathrm{h}}$ and to avoid double counting, we estimate their contribution in the AR through MC simulation and subtract them from the estimate. The f_i used in F_i are measured separately for electrons, muons and $_{\mathrm{h}}$ and are parametrised as a function of η and cone-p_T of the lepton. So called "measurement regions" (MR) were used to measure the FF and serve as control regions. In the MR the selection criteria are chosen so that the relative fractions of non-prompt leptons and hadrons (quark jets and gluon jets) are similar between AR and MR. Similarly, to reduce the difference between the FF for non-prompt leptons and hadrons the fakeable lepton criteria are chosen accordingly, which guarantees an unbiased estimate of the fake background in the SR. The events in the MR are collected using a set of prescaled single lepton triggers and are required to have a reconstructed lepton and a jet separated in ΔR from it. Such a region is expected to be dominated by multijet events. Residual contributions from processes with prompt leptons (mostly +jets) are subtracted by performing a fit to the $m_{\mathrm{T}}^{\mathrm{fix}}$ distribution. I report here its definition for completeness:

$$m_{\mathrm{T}}^{\mathrm{fix}} = \sqrt{2\, p_{\mathrm{T}}^{\mathrm{fix}}\, E_T\, (1 - \cos\Delta\varphi)}\,, \tag{6.2}$$

where the symbol $p_{\mathrm{T}}^{\mathrm{fix}}$ denotes a constant of value 35 GeV. As performed in the inclusive analysis, in the the $2ss1_{\mathrm{h}}$ channel we only estimate the contribution of nonprompt leptons using the FF method, since signal has a significant contribution from misidentified τ_h. In the $2ss1_{\mathrm{h}}$ channel, the contributions of events with misidentified τ_h mostly originate from ttW, ttWW, ttZ and diboson production processes. In order to be able to take into account these events in the signal extraction, we do not include the τ_h in the FF method. Instead, we estimate the contribution from this processes using MC simulation. In order to account for mismodeling of the FF in simulations, we apply a correction factor on the efficiency of those misreconstructed τ_h equal to the data/MC ratio of the τ_h FF measured in a $t\bar{t}$-enriched MR.

6.6.2 Estimation of flips background

The so-called charge flips background is mostly relevant for events with one or two reconstructed electrons in the $2ss0_{\mathrm{h}}$ and $2ss1_{\mathrm{h}}$ channels and is dominated by $+$jets events with two prompt leptons. Only events where at least one of the leptons is an electron are considered. They are usually produced in the decay of $t\bar{t} \to bW^+bW^- \to b\bar{l}\nu\bar{b}l\bar{\nu}$ with one of the charges of either prompt lepton mismeasured. Using a similar strategy than for the fake factor, the background

contribution is measured from data. A control region, containing events that pass all selection criteria but where the two leptons are required to be of opposite charge, is implemented to gain an estimate on the flips background. However, in the $2ss1\tau_{\mathrm{h}}$ channel only the lepton having the same charge as $_{\mathrm{h}}$ is considered, as only those events satisfy the condition $\sum_{l,\tau} q = \pm 1$ applied in the SR after mismeasuring of the charge of said lepton. The sum of the probabilities to mismeasure the charge of either lepton is then applied as an event weight. Using $Z/\gamma^* \rightarrow ee$ events, we measured the charge misidentification rate for electrons and parameterized it as a function of p_T and η of the electron.

6.7 Background control regions

To test the Monte Carlo modeling of irreducible background contributions in the SR, several control regions are defined:

- 3 ℓ and 4 ℓ control region, defined by inverting the cuts on the Z boson veto and not appling the b tagging requirement on jets (4 ℓ control region)
- 3j, WZ, ttZ control region, built to check to test the agreement of the variables used as input to the discriminants. These region are described in appendix Sec C.2 (see Sec 6.8.2)

These control region are the same control regions used in the ttH inclusive analysis [108] and in the following section a quick overview is given.

6.7.1 3 lepton control region

The three lepton control region is defined by inverting the Z boson veto present in the 3l signal region defined in 6.4. The selection criteria on jets and b tagged jets are dropped in this region. Besides, the $E_T^{miss} > 45$ GeV criteria is applied in case the event contains a pair of loose leptons of the same flavor and opposite charge, if the event does not contain such lepton pair the criteria is $E_T^{miss} > 30$ GeV and no requirement on $E_T^{miss} > 30$ is applied at all if the event contains four or more jets.

Events are then classified as the number of jets and b tagged jets in 12 bins: 0 b tagged jets with 1, 2, 3 or more than 3 jets; 1 b tagged jet with 2, 3, 4 or more than 4 jets, and more than 1 b tagged jets with 2, 3, 4 or more than 4 jets. This allow to separate ttZ and WZ backgrounds. Additionally, events

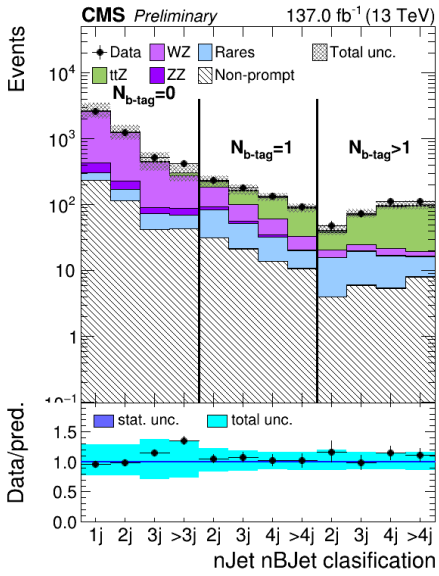


Figure 6.1: 3 lepton control region

are categorized according to the flavor of the leptons. The obtained distribution is shown in figure 6.1.

6.7.2 4 lepton control region

A four lepton control region is defined by events with four leptons, two of which form a Z boson candidate. Moreover, the jet and b tagging discriminator selections are not applied in this case. Events in this region are classified according to the number of Z boson candidates and the jet and b tagged jet multiplicity, four categories are created using that information:

- Events with two pairs of opposite-sign same flavor leptons with masses compatible with the Z boson.
- Events with exactly one pair of opposite-sign same flavor leptons with a mass compatible with the Z boson and no jets.
- Events with exactly one pair of opposite-sign same flavor leptons with a mass compatible with the Z boson and exactly 1 medium WP b tagged jet.

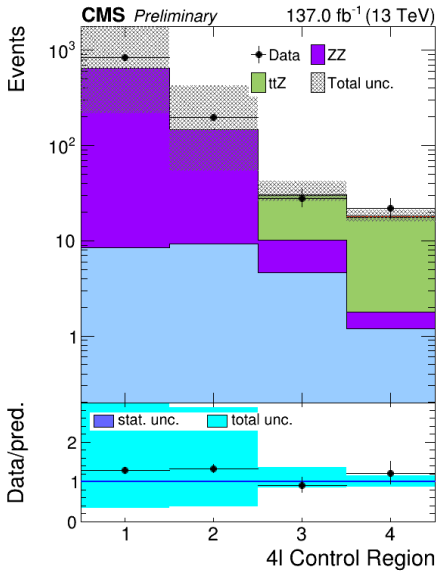


Figure 6.2: 4 lepton control region

- Events with exactly one pair of opposite-sign same flavor leptons with a mass compatible with the Z boson and more than one medium WP b tagged jet.

The obtained distribution is shown in figure 6.2. The sensitivity to ttZ and background process is driven by the 3 ℓ control region, however, 4 ℓ control region provides a reasonably sensitive alternative measurement and is enriched in ZZ events.

6.8 Standard Model Analysis

The measurement of the ttH differential cross section as a function of the Higgs boson transverse momentum, p_T^H, allows to disentangle the effects of modified Higgs boson self-coupling values from other effects such as the presence of anomalous top-Higgs couplings [149]. Moreover, such measurements are proven to provide a strong handle on relevant Effective Field Theory (EFT) dimension-six operators through the constraint of the corresponding Wilson coefficients [11]. Differential results within a Simplified Template Cross Section (STXS) framework are also very important in that they permit the integration of top-Higgs coupling results in the global fits for the properties of the Higgs(125) boson. Preliminary results on the H cross-section measurements

are therefore of uttermost importance, as they anticipate strategic decisions within the HEP community.

6.8.1 Available final states

The decision on which final state must be investigated first is driven by two main considerations: the amount of events expected in each final state, and the possibility of performing a full or approximately full kinematic reconstruction of each final state. The ability of performing an approximately full kinematic reconstruction of the final state is crucial to disentangle the top-quark-related portion of the event from the Higgs-boson-related part of the event. The $2lss + 0\tau_h$ final state is characterized by the largest expected yields. Its kinematics is determined by 40 free parameters:

$$\begin{aligned} P(t) = P(b) + P(W) \qquad & P(W) \quad = P(\ell_1) + P(\nu_1) \\ P(\bar{t}) = P(b) + P(W) \qquad & P(W) \quad = P(q_1) + P(q_2) \\ P(H) = P(W) + P(W) \qquad & P(W) \quad = P(\ell_2) + P(\nu_2), \\ & P(W) \quad = P(q_3) + P(q_4), \end{aligned} \tag{6.3}$$

where $P(\cdot)$ indicates the four-momentum of a given particle. This final state can be determined by using 39 measurements and constraints, as outlined in Table 6.7. A last constraint can be added using experimental hypotheses, such as adding a longitudinal momentum hypothesis for the neutrinos, or splitting the MET into components assigned to the two neutrinos.

Object	Parameter	Type	N_{pars}	Total (incremental)
Charged leptons	$(E, \vec{p})$	measurement	2×4	8
Jets	$(E, \vec{p})$	measurement	6×4	32
Neutrinos	MET	measurement	2	34
W (had)	window on/off shell	constraint	2×1	36
Top mass	$m_t = 174.3$ GeV, $m_t = m_{\bar{t}}$	constraint	2	38
Higgs mass	$m_H = 125.1$ GeV	constraint	1	39

Table 6.7: Measurements and constraints for the 2lss + 0τ

The $3l + 0\tau_h$ final state is also characterized by quite large expected yields. Its kinematics is described by 40 parameters:

$$\begin{aligned} P(t) = P(b) + P(W) \qquad & P(W) \quad = P(\ell_1) + P(\nu_1) \\ P(\bar{t}) = P(b) + P(W) \qquad & P(W) \quad = P(q_1) + P(q_2) \\ P(H) = P(W) + P(W) \qquad & P(W) \quad = P(\ell_2) + P(\ell_3), \\ & P(W) \quad = P(q_3) + P(q_4), \end{aligned} \tag{6.4}$$

and the measurements and constraints that can be set result in an overconstrained system, as outline in Table 6.8. A big difficulty lies however in the problem of assigning each object to its parent particle, due to combinatorics effects. This final state is therefore not examined in this version of the analysis.

Object	Parameter	Type	N_{pars}	Total (incremental)
Charged leptons	$(E, \vec{p})$	measurement	3×4	12
Jets	$(E, \vec{p})$	measurement	6×4	36
Neutrinos	MET	measurement	2	38
Z (lep)	window on/off shell	constraint	1	39
Z (had)	window on/off shell	constraint	1	40
W (had)	window on/off shell	constraint	1	41
Top mass	$m_t = 174.3$ GeV, $m_t = m$	constraint	2	43
Higgs mass	$m_H = 125.1$ GeV	constraint	1	44

Table 6.8: Measurements and constraints for the $3l + 0\tau_h$ final state.

The $4l + 0\tau_h$ final state is characterized by very low expected yields, making it not useful for a differential measurement with the available luminosity. Furthermore, it is still determined by 40 parameters:

$$\begin{aligned} P(t) = P(b) + P(W) \qquad & P(W) \quad = P(\ell_1) + P(\nu_1) \\ P(\bar{t}) = P(b) + P(W) \qquad & P(W) \quad = P(\ell_2) + P(\nu_2) \\ P(H) = P(W) + P(W) \qquad & P(W) \quad = P(\ell_3) + P(\nu_3), \\ & P(W) \quad = P(\ell_4) + P(\nu_4), \end{aligned} \tag{6.5}$$

Object	Parameter	Type	N_{pars}	Total (incremental)
Charged leptons	$(E, \vec{p})$	measurement	4×4	16
Jets	$(E, \vec{p})$	measurement	2×4	24
Neutrinos	MET	measurement	2	26

Table 6.9: Measurements and constraints for the $4l + 0\tau_h$ final state.

or:

$$\begin{aligned} P(t) &= P(b) + P(W) & P(W) &= P(\ell_1) + P(\nu_1) \\ P(\bar{t}) &= P(b) + P(W) & P(W) &= P(\ell_2) + P(\nu_2) \\ P(H) &= P(Z) + P(Z) & P(Z) &= P(\ell_3) + P(\ell_4), \\ & & P(W) &= P(q_1) + P(q_2). \end{aligned} \tag{6.6}$$

This time, however, the system is severely underconstrained as outline in Table 6.9. Even when imposing constraints on the mass of the parent objects the system remains severely underconstrained. The combination of the low expected yields and of the lack of enough constraints led to the decision of not considering this final state for this preliminary differential analysis.

Final states with hadronic taus result in more favourable combinations (because of the distinguishability of the tau from the light leptons, and are analyzed via the DNN regression approach described in Section 6.8.2.

6.8.2 Regressing the Higgs transverse momentum with a Deep Artificial Neural Network

We regress the Higgs transverse momentum using a dedicated Deep Artificial Neural Network (DNN), separately trained for each signal category ($2lss + 0\tau_h$, $3l$, $2lss + 1\tau_h$). It was found that performing a dedicated training for each year separately did not significantly improve performance, and so the MC events for all three years were combined for the training and validation. For all the channels, the network is trained on the 70% and validate on the 30% of events. For recap on what training and validation means for a neural network please refer to Sec 3.7. The target of the regression is the generator-level Higgs transverse momentum p_T. A Keras frontend [150] is used on top of a Tensorflow backend [151].

$2lss + 0\tau_h$, $3l$ final states

Input Variables A variety of input parameters were tried during the optimization of the DNN. The final list of features is shown in Table 6.10. The l_i are the top leptons sorted by p_T. The t_{had} variables, including the score, come from the Resolved Hadronic Top Tagger algorithm based on neural networks that is able to identify the probability for a jet to be a top jet. To capture the jet structure of the average signal event, we include the sum of the top 5 jets sorted by p_T, as this is the most common number of jets in a signal region event, both in the $2lss + 0\tau_h$ and $3l$ channels. We subsequently discovered that including the sum of the remaining jets as an additional input slightly but significantly improved the performance of the network. We found that the summed jet variables were sufficient to allow the network to learn about the jet structure of the event, and that including the quadrimomenta of individual jets yielded no additional benefit. In addition, the inclusion of δR between final state objects did not improve network performance. Including the mass component of the quadrimomenta also added no benefit. Finally, the final variable which sums all jets and fakeable leptons, while being similar to MET, improves the network performance above MET alone because it is less sensitive to event-by-event variation in the pile-up.

Input	Number of Variables	Which Channels
$l_1(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
$l_2(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
$l_3(\mathrm{p}_T,\eta,\varphi)$	3	$3l$
$t_{had}(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
t_{had} BDT Score	1	$2lss$ & $3l$
MET	1	$2lss$ & $3l$
φ_{MET}	1	$2lss$ & $3l$
$\sum_{n=1}^{5} j_n(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
$\sum_{n>5} j_n(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
$\sum_n j_n + \sum_n l_n(\mathrm{p}_T,\eta,\varphi)$	3	$2lss$ & $3l$
Total	21	-

Table 6.10: Input variables to DNN used for p_T regression.

Loss Function Choosing the correct loss function is important for allowing the DNN to correctly model the true Higgs p_T distribution. Many loss functions were tried, all of which suffered from the same mismodelling effect - The DNN consistently predicted a Higgs p_T close to the mean of the true p_T spectrum, resulting in a peak in the predicted p_T distribution with a much smaller spread than the true p_T distribution. An example of this can be seen in Figure 6.3, where the Mean-Squared-Error loss function is used. Similar peaks were observed when using the Mean-Squared-Logarithmic-Error, Mean-Absolute-Error, and Mean-Absolute-Percentage-Error.

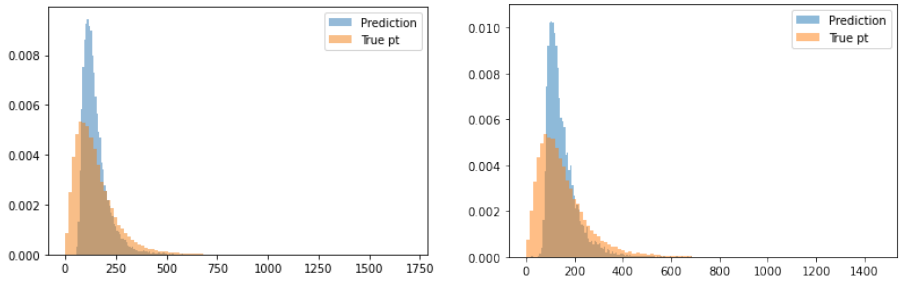


Figure 6.3: The Higgs p_T distribution predicted by the DNN when using the Mean-Squared-Error loss function, overlayed on the True Higgs p_T distribution. For the year 2016, in 2lss (left) category and 3l (right) category, other years are similar. Generated using simulated ttH events.

To remedy the problem, we introduced a term into the loss function which penalizes differences in the variance of the predicted and true p_T distributions in the batch of events being processed, in addition to the penalty for the deviation between the true and predicted p_T for each individual event that is present in traditional loss functions. The final loss function is given by Equation 6.7. The first term in the product is the standard Mean-Squared-Error loss, and the second term penalizes differences in variance between the two distributions. The result of the regression using this new loss function is shown in figure 6.4, were much better modelling of the true p_T distribution is achieved.

$$L = \frac{1}{N_{batch}} \sum [$$

$$\mathrm{p}_T{}^{true} - \mathrm{p}_T{}^{pred}]^2 |\sigma^2_{true} - \sigma^2_{pred}| \quad (6.7)$$

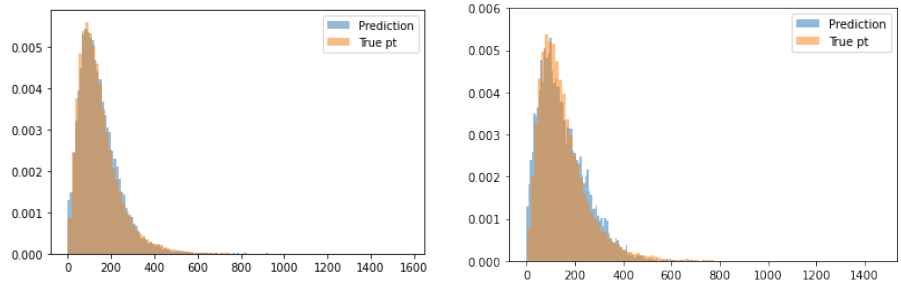


Figure 6.4: The Higgs p_T distribution predicted by the DNN when using Equation 6.7 as the loss function, overlayed on the True Higgs p_T distribution. For the year 2016, in 2lss (left) category and 3l (right) category, other years are similar. Generated using simulated ttH events.

Network Hyperparameters The performance of the DNN was found to be only weakly dependent on the chosen network hyperparameters. Four fully connected hidden layers with $N_{nodes} = (512, 256, 128, 64)$ were used. Each of the hidden layers used a Scaled Exponential Linear Unit (SELU) activation function, and the final single-node layer used the absolute value for an activation function. To prevent overtraining, a dropout layer with a drop rate of 10% was added just before the final single-node layer. A batch size of 128 was used, chosen to balance the need to have enough events per batch to allow a meaningful calculation of variance, against having as many batches with different variances in the training set as possible. The network is trained for 100 epochs with a learning rate of 10^{-4}. The loss function converges rapidly to its minimum - an example of the loss as a function of epoch for 2016 is shown in Figure 6.5. Losses for the other years are similar.

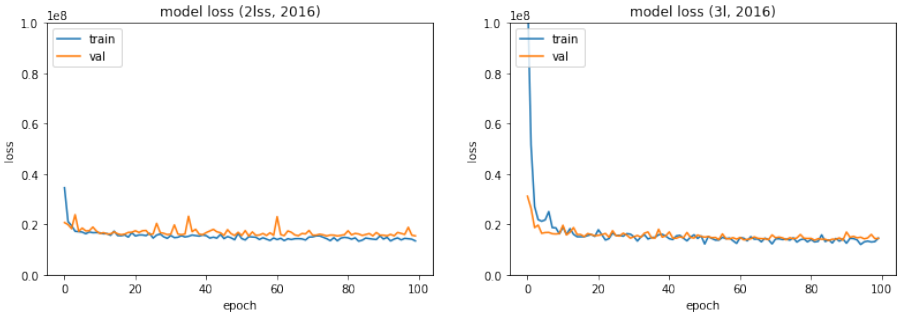


Figure 6.5: The loss function for the year 2016 in 2lss (left) category and 3l (right) category. Loss functions for other years are similar.

Network Performance The figure of merit that we used to check the performances of the regression of the Higgs p_T is the response matrix. The response matrix is a matrix in which each row is normalized to unity, so that each cell corresponds to the conditional probability that an event which originates in truth bin i will be reconstructed in the reconstructed bin j. The response matrices for the final DNN for each category/year are shown in Figures 6.6 to 6.7. It was found that DNN did not benefit from year-specific training, and so the same network is used for all years within a category. In addition, the same network structure is used for all categories, with the exception that the input variables vary slightly depending on what is available in the category. These differences are detailed in Section 6.8.2.

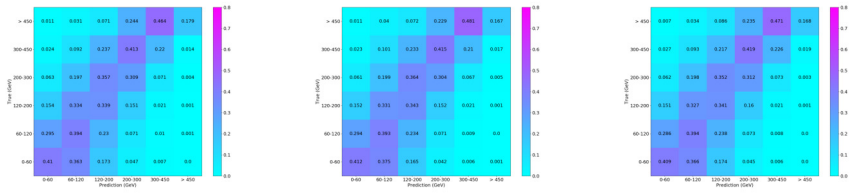


Figure 6.6: The response matrix of the final trained DNN for 2lss category, for 2016 (left), 2017 (middle), and 2018 (right). The p_T bins are chosen to correspond to the latest STXS conventions (see Sec 6.10.1). Each row is normalized to unity, so that each cell (i, j) corresponds to the conditional probability that an event which originates in truth bin i will also be reconstructed in the reconstructed bin j.

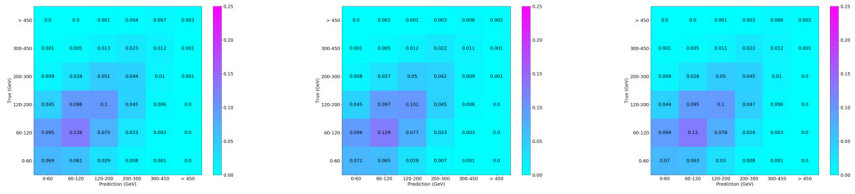


Figure 6.7: The response matrix of the final trained DNN for 2lss category, for 2016 (left), 2017 (middle), and 2018 (right). The p_T bins are chosen to correspond to the latest STXS conventions (see Sec 6.10.1). The full matrix is normalized to unity, so that each cell (i, j) corresponds to the probability that an event will both originate in truth bin i and will be reconstructed in the reconstructed bin j.

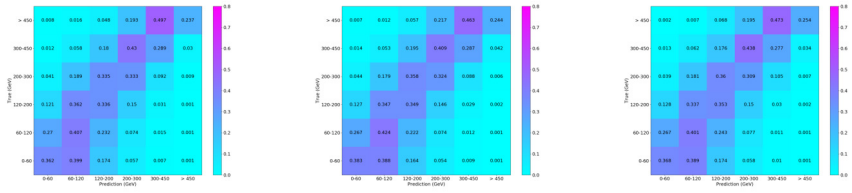


Figure 6.8: The response matrix of the final trained DNN for 3l category, for 2016 (left), 2017 (middle), and 2018 (right). The p_T bins are chosen to correspond to the latest STXS conventions. Each row is normalized to unity, so that each cell (i, j) corresponds to the conditional probability that an event which originates in truth bin i will also be reconstructed in the reconstructed bin j.

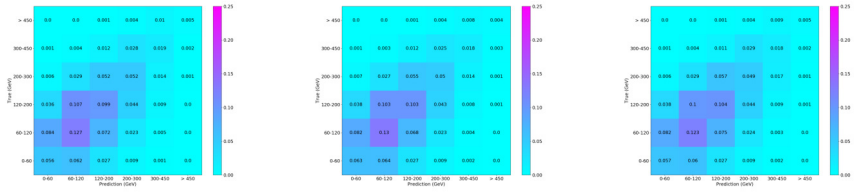


Figure 6.9: The response matrix of the final trained DNN for 3l category, for 2016 (left), 2017 (middle), and 2018 (right). The p_T bins are chosen to correspond to the latest STXS conventions. The full matrix is normalized to unity, so that each cell (i, j) corresponds to the probability that an event will both originate in truth bin i and will be reconstructed in the reconstructed bin j.

$2lss + 1\tau_h$ final state

The 2lss1τ channel is the most sensitive channel of the analysis and a slightly different architecture is used.

Input variables Various input parameters were tested to optimize the DNN. After careful optimization the 29 input variables listed in table 6.11 were chosen as inputs. We found a combination of the jet, lepton and tau variables as well as some higher level variables like the average delta r between jets (listed as avg_dr_jet in table 6.11) to be optimal.

Network Hyperparameters The performance of the DNN was not strongly dependant on the amount of layers and the number of nodes in them. However adding batch normalization layers increased the performance of the DNN. The final architecture of the network can be seen in fig 6.10. It consists of four

Table 6.11: Table showing the variables used for the 2lss1tau channel$2lss + 1\tau_h$

Input	Number of variables	Description
$SelJet1(\mathrm{p}_T,\eta,\varphi)$	3	Kinematic variables of the jet with highest p_T
$SelJet2(\mathrm{p}_T,\eta,\varphi)$	3	Kinematic variables of the jet with second highest p_T
$SelJet1,2(isFromHadTop)$	2	Flag if the jet is originating from a hadronic top
$SelJet1,2(btagDeepFlavB)$	2	Score of how likely the jet originates from a b quark
$Lep1(m_T,\mathrm{p}_T,\eta,\varphi)$	4	Kinematic variables of one of the leptons
$Lep2(m_T,\mathrm{p}_T,\eta,\varphi)$	4	Kinematic variables of one of the leptons
$Tau(\mathrm{p}_T,\eta,\varphi)$	3	Kinematic variables of one of the τ_h
$nSelJets$	1	Number of jets in the event
MET	1	Missing transverse energy
HTT_score	1	Highest BDT score of jet triplet from t [108]
$visHiggs(\mathrm{p}_T,\eta)$	2	Kinematic variables of the visible Higgs
Hj_tagger_hadTop	1	Higgs jet tagger
avg_dr_jet	1	Average dR distance among all jets
$m_H\_2lss1tau$	1	Invariant mass of the ttH system
Total	29	

fully connected dense hidden layers with $N_{nodes} = (80, 80, 40, 20)$ with an Exponential Linear Unit (ELU) activation function and an additional batch normalization layer between each of them. Additionally, a dropout layer with a dropout rate of 0.3 was added between the last two hidden layers to help counteract overtraining. The final layer consists of a single node with an ELU activation function. During training we used a batch size of 500 as this guaranteed fast training without compromising variance within the batches. To estimate the loss we used the Mean Squared Error (MSE) and started with a learning rate of 10^{-4} in the Adam optimizer. The use of the customised loss function is under investigation and will be added in the next steps of the analysis. To further counteract overtraining we implemented and Early Stopping function with a patience of 15 and a delta of 10^{-5}.

Network performance One can see the predicted Higgs p_T distribution in Figures 6.11 and 6.12. The predicted Higgs p_T distribution tends to be more centered around the average true Higgs p_T. However if we look at Figure 6.12, we can see that the predicted and true Higgs p_T are strongly correlated (Pearson correlation $\rho_{X,Y} = 0.7$). A condition number of 186 of the Higgs p_T prediction points to a the need of regularization (see Sec 6.10) afterwards which isn't problematic as the correlation between true and predicted Higgs p_T is rather high. We are still optimising this channel, we will include in the

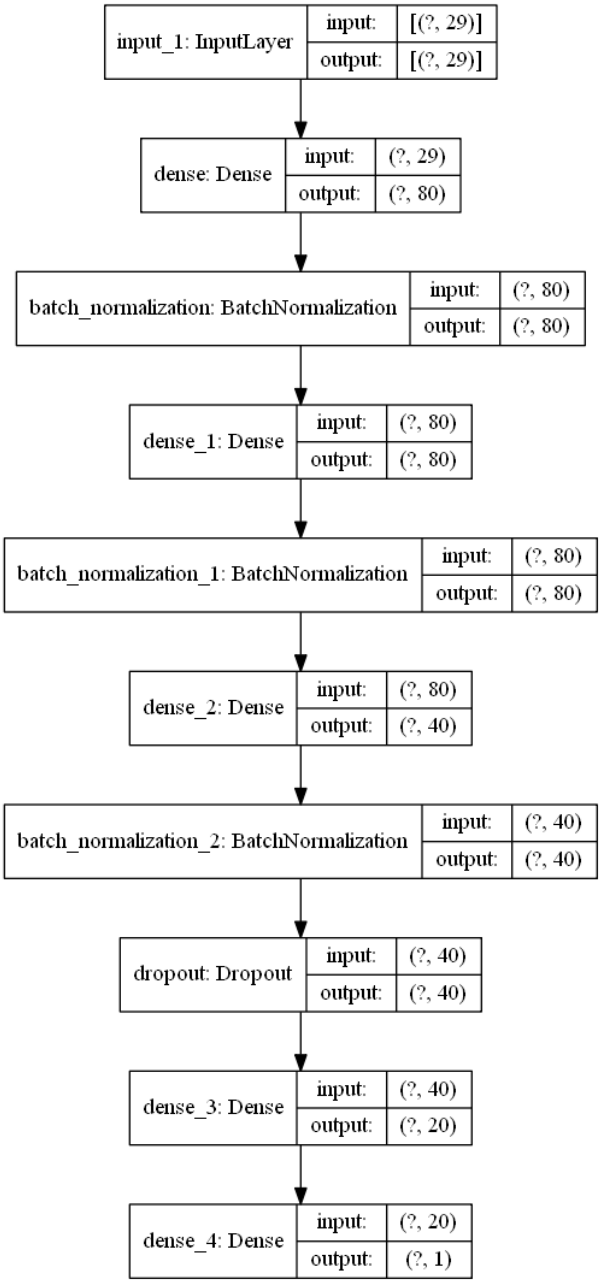


Figure 6.10: Architecture of the DNN used for the Higgs p_T regression in the $2lss + 1\tau_h$ channel

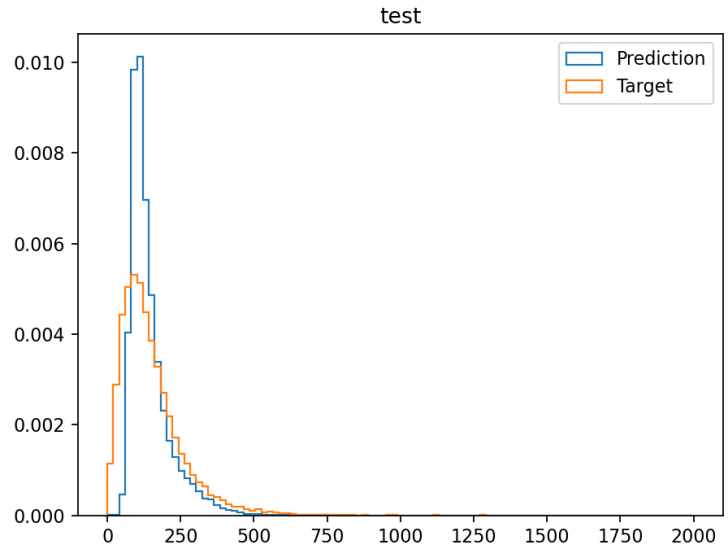


Figure 6.11: The Higgs p_T distribution predicted by the DNN overlaid with the true Higgs p_T distribution for the $2lss + 1\tau_h$ channel

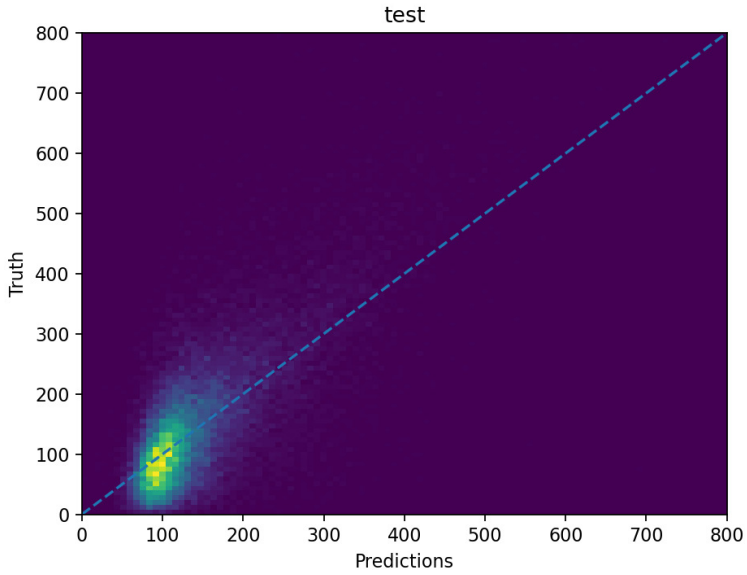


Figure 6.12: A heat map showing the predicted Higgs p_T distribution against the true Higgs p_T distribution for the $2lss + 1\tau_h$ channel

next steps of the analysis the customised loss function added in 2lss+0tau and 3l channels.

6.9 Systematic uncertainties

The systematic uncertainties taken into account in the following sections are summarised in appendix C.3.
We considered systematics on luminosity, pileup, trigger selection, lepton and jets identification and background estimation.

6.10 Unfolding

The unfolding is the procedure thought which is possible to disentangle the effect of the detector reconstruction from the underling truth of the studied process.
In more simple word, if we think of any reconstructed physics observable, as the Higgs p_T in the ttH system, we should remember that what is plotted is a convolution of the physics process that we want to study plus the detector effects (interaction with the material, resolution ecc) that is not trivial to estimate. What unfolding does is to disentangle these two effects.
In particular, signal events from one particular generator-level bin, can contribute to multiple reconstruction level bins. This feature is well displayed in Fig 6.6 where each row is normalized to unity, so that each cell corresponds to the probability that an event in that cell's truth bin will be reconstructed in a certain reconstructed bin. If the response matrix was found to be diagonal, this means that the detector effect reconstruction is negligible.
The unfolding problem is essentially solving the linear relationships in n = Rμ+b where n, μ and b are vectors of the observed event yields, of the signal strength modifiers, and of the reconstructed background, respectively, and R is the response matrix (see Sec 6.8.2). The response matrix element R_{ij} where i denotes the bin of the observed histogram and j the bin of the true histogram, is not necessarily symmetric. The simultaneous fit maximizes the following likelihood function:

$$\mathcal{L} = \prod_{i=1} Poisson(n_i; \sum_j [R_{ij}\mu_j] + b_i)\mathcal{C}(\theta_i(\mu)) \tag{6.8}$$

where θ_i represents the nuisance parameters in bin i, Poisson(n,γ) is the Poisson probability for observing n events under the given expectation of λ and C corresponds to the constraint of the nuisance paramenters θ_i. The signal strengths μ_i are fully correlated between final states since similar phase spaces are selected with the fiducial region definitions and the main difference be-

tween final states lies in the well-known branching fraction of the τ lepton.
The unfolding procedure is sensitive to statistical fluctuations in the observed distribution. To attenuate this feature it is possible to add to the likelihood function a multiplicative penalty term, called the regularization factor, that aims to increase the smoothness between the measurements in adjacent ranges of the observables to avoid unphysical fluctuations. This factor can be also written as:

$$\mathcal{P} = \tau A(v(\mu)) \tag{6.9}$$

where A is the curvature matrix:

$$A_{M,M-1} = \begin{pmatrix} -1 & 1 & \cdots & & \\ 1 & -2 & 1 & \cdots & \\ 0 & 1 & -2 & 1 & \cdots \\ \cdots & & & & \end{pmatrix} \tag{6.10}$$

where M is the number of bins in the true histogram.
The parameter τ, is called strength of the regularisation, and is determined separately for each observable as the value that minimizes the mean of the global correlation coefficient in Asimov pseudo-data.
The regularization done with the above curvature matrix is called Tikhonov regularisation, results with this regularisation are shown in the next section.

6.10.1 Final categorisation and results

In theories BSM, not only the simple scale of couplings can give hints of new physics, also the kinematic dependence of the Higgs observable is crucial to sick for deviation from the SM. Fiducial cross sections (FXS) measurement give a large model-independent way to test the deviation from the SM in the kinematics distributions of a certain phase space; however theory assumptions coming both from theoretical uncertainties and from the underlying physics model, are folded into FXS. A way that we have to reduce these effect and to finally have a model-independet measurement is to carefully choose the phase space that we want to study. To achieve this puropose the Simplified Template Cross Section (STXS) bin schemes have been adopted by the LHC experiments as a common framework for Higgs measurements. The bin splitting goes as:

- $0 < p_{T Higgs} < 60 GeV$
- $60 < p_{T Higgs} < 120 GeV$
- $120 < p_{T Higgs} < 200 GeV$
- $200 < p_{T Higgs} < 300 GeV$
- $300 < p_{T Higgs} < 450 GeV$
- $pT_{Higgs} > 450 GeV$

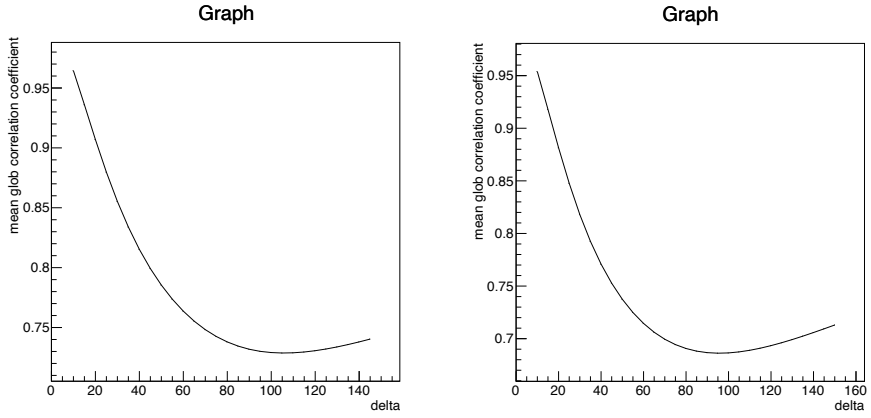
A simultaneous fit on the Higgs p_T regressed from the DNN has been performed to obtain the signal strength for each of the studied bins. For the lack of statistics, in order to minimize the errors in each bin, a further constraint on the 3l and 4l control regions (see 6.7) has been added to the fit. Results are shown in Tab 6.12. The effect of the Tikhonov regularisation is to drastically reduced the uncertainties on each bin.

Table 6.12: Unfolding results with different hypothesis.

	No regularisation	No regularisation + constraints on bkg	Tikhonov regularisation + constraints on bkg
$0 < p_T^H \leq 60$ GeV	+1.000 -6.734/+6.743	+1.000 -1.549/+1.516	+1.000 -1.393/+1.362
60 GeV $< p_T^H \leq 120$ GeV	+1.000 -4.879/+4.939	+1.000 -0.586/+0.576	+1.000 0.518/+0.509
120 GeV $< p_T^H \leq 200$ GeV	+1.000 -3.461/+3.431	1.000 -0.589/+0.588	+1.000 -0.520/+0.521
200 GeV $< p_T^H \leq 300$ GeV	+1.000 -3.178/+3.344	+1.000 -0.971/+0.981	+1.000 -0.866/+0.881
300 GeV $< p_T^H \leq 450$ GeV	+1.000 -4.196/+4.209	+1.000 -2.030/+2.059	+1.000 -1.817/+1.855
450 GeV	+1.000 -3.090/+4.696	+1.000 -2.950/+3.995	+1.000 -2.874/+3.789

The strength of the regularisation, τ, was tuned separately for the different configuration in Tab 6.12 and it is shown in Fig 6.13. In particular for the configuration without any constrain on the CR, the value of τ is 105, while for the configuration with the constrain on the CR it is found to be 95.

The unfolding study is still ongoing, in particular we are estimating the bias of the regularisation and we are adding a subsequent event categorisation on the flavour of leptons.



(a) **(b)**

Figure 6.13: On the left: strength tuning for the full RunII dataset without adding any constrain on the CRs. The minimum of the global correlation coefficient is obtained for $\tau = 105$. On the right: strength tuning for the full RunII dataset adding the constraints on the CRs. The minimum of the global correlation coefficient is obtained for $\tau = 95$

6.11 Beyond Standard Model Analysis

6.11.1 Theory overview

The associated production of a Higgs boson with a top quark pair at the LHC can serve as a measurement of the interaction of both the top and the Higgs. The latter being the two heaviest elementary particles in the Standard Model (SM) to date, their interaction can hint to evidence for physics beyond the SM. The SM prediction for the $t\bar{t}H$ process is known at NLO in QCD [152–159].

Precise measurements at the LHC can potentially unveil deviation from the SM expectations and therefore hint to the existence of new physics. A powerful approach to encapsulate and interpret these deviations is to employ an effective field theory approach [160–162]. One choice is to use the Standard Model Effective Field Theory (SMEFT). The SMEFT Lagrangian is obtained through augmenting the SM one with higher dimensional operators that respect the SM gauge symmetries. These SMEFT operators are scaled by a dimensionless coefficient, Wilson Coefficient (WCs), C_i, and suppressed by the scale of new physics Λ which is large enough compared to the LHC energies. The WCs can be constrained from experimental data. All operators of odd dimension violate baryon and lepton number [163], and therefore, the EFT series is restricted to dimension-6 operators.

$$\mathcal{L} = \mathcal{L}_{SM} + \sum i \frac{\mathcal{C}_i}{\Lambda^2} O_i + \mathcal{O}(\Lambda^{-4}) \tag{6.11}$$

A generic observable in SMEFT can be expressed as

$$\sigma = \sigma_{SM} + \sum_i \frac{C_i^{(6)}}{\Lambda^2}\sigma_i + \sum_{ij} \frac{C_i^{(6)} C_j^{(6)}}{\Lambda^4}\sigma_{ij} \tag{6.12}$$

where the second term is the interference contributions of the dimension-6 operators with the SM, and the last term is the squared contributions of the dimension-6 operators, *i.e.* pure EFT contributions.

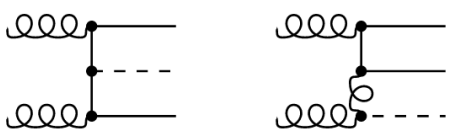


Figure 6.14: Representative Feynman diagrams showing the $\bar{t}tH$ production in the SM

The anomalous interactions between SM particles may lead to energy growth of the scattering amplitudes through the introduction of new Lorentz structures or via spoiling delicate unitarity cancellation in the SM amplitudes. This energy growth can impact the kinematic distributions significantly, a feature that emphasizes the importance of the EFT interpretations of differential measurements and its capability to outperform inclusive ones. The bottom-up approach of EFT is a model-independent approach which renders the EFT framework a vigorous tool to maximize the reach of LHC searches, and to interpret the vast number of existing measurements.

6.11.2 The dim6top model

Samples for the $t\bar{t}Z$, $t\bar{t}W$ and $t\bar{t}H$ processes are generated privately with the the variations from the SMEFT operators accounted for. The `dim6top` model [164] is used to simulate those effects.

6.11.3 Relevant operators

In this analysis, we only consider diagrams with at most one EFT insertion at the production-level. The relevant SMEFT operators impacting the $t\bar{t}H$ process are those which modify the $t\bar{t}H$, $t\bar{t}g$ and ggH vertices [11]. These operators establish a connection between the top-quark and the Higgs-boson sectors in the SMEFT at dimension-six. The red modification in Fig 6.15 is

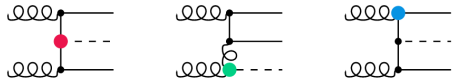


Figure 6.15: Example diagrams for the insertion of the SMEFT relevant operators in the $t\bar{t}H$ production. (a) $O_{t\varphi}$, (b) $O_{\varphi G}$ and (c) O_{tG}

through the $O_{t\varphi}$ operator which rescales the top Yukawa coupling in the SM and introduce to a new $ttHH$ coupling. The blue modification is the from the chromomagnetic dipole operator O_{tG} which gives rise to a dipole interaction in the $t\bar{t}g$ vertex and introduces $ggtt$, $gttH$, and $ggttH$ vertices. In principle, the green modification via $O_{\varphi G}$ is relevant for $t\bar{t}H$ for it being a loop- induced interaction between the gluon and Higgs fields but it is not included in the `dim6top` model and so it is not included in our analysis. The operators $O_{t\varphi}$

Name	Operator	Comments
ctp	$_{ij}\tilde{\varphi}\left(\varphi^\dagger\varphi\right)$	Effects on tHq
cpt	$\left(\varphi^\dagger i\overleftrightarrow{D}_\mu\varphi\right)\left(_i\gamma^\mu_j\right)$	Effects on ttH, ttW, ttZ, and tZq
cptb	$\left(\tilde{\varphi}^\dagger iD_\mu\varphi\right)\left(_i\gamma^\mu_j\right)$	Effects on tHq and tZq
ctG	$\left(_i\sigma^{\mu\nu}T^A_j\right)\tilde{\varphi}G^A_{\mu\nu}$	Effects on every process with a top quark
cpG	$\left(\varphi^\dagger\varphi\right)G^A_{\mu\nu}G^{A\mu\nu}$	Effects on every QCD process

Table 6.13: Summary of the EFT operators relevant to ttH production. The third column highlights which processes characteristic of a ttH multilepton analysis are affected by each operator.

and O_{tG} read

$$O_{t\varphi} = y_t^3(\varphi^\dagger\varphi)(\bar{Q}t)\tilde{\varphi} \tag{6.13}$$

$$O_{tG} = y_t g_s(\bar{Q}\sigma^{\mu\nu}T^A t)\tilde{\varphi}G^A_{\mu\nu}. \tag{6.14}$$

The main background processes to the $t\bar{t}H$ process are $t\bar{t}W$ and $t\bar{t}Z$. Considering only the two-heavy-quarks `dim6top` operators, both backgrounds are affected by the O_{tG} operator as it modifies the gtt interaction. $t\bar{t}Z$ is also affected by three more operators O_{tZ}, $O^-_{\varphi q}$ and $O_{\varphi t}$ [165].

6.11.4 Search for new physics in the EFT framework

We perform a search for non-zero values of Wilson Coefficients (WCs) associated to an EFT theory. To do so, we measure the yields in regions enriched in signal events binned as a function of certain kinematic variables that can be of interest in such theories. We consider the cpt, ctp and ctG operators.
Since these operators also affect the production rate of some of our backgrounds, we consider the effect on ttH, tH, ttZ and ttW.

In order to perform inference on the values of coefficients, we parametrize the expected number of signal and background events in the bins of the measured distribution as a funcion of the WCs. The number of events x in a given

bin as a function of n Wilson Coefficients $\{C_i\}_{i=0}^{n}$ can be parametrized as a multidimensional quadratic function as

$$x = x^{\mathrm{SM}} + \sum_{i=0}^{n} A_i C_i + \sum_{i,j=0}^{n} B_{ij} C_i c_j, \tag{6.15}$$

where x^{SM} is the expected contribution in the SM, and A_i, B_{ij} represent the contribution to the event yield from BSM diagrams. The terms A_i represent the contribution from the interference between SM and BSM diagrams, while B_{ij} represent the contribution only from BSM diagrams. We determine the parametrization in equation 6.15 by reweighting the leading-order MC simulations for ttH, ttW and ttZ processes. The associated weight is computed using MadGraph_2_7_3 and the dim6top model using the generator-level information in the nanoAOD dataformat[1].

Validation of reweighing method

The reweighting procedure is only valid if the support of the distributions of the BSM scenario is included in the support of the distributions under the SM hypothesis. In order to check that hypothesis, we compare generator-level distributions generated using MadGraph_2_7_3 under BSM hypotheses with events generated under the SM hypothesis but then weighted to the BSM hypothesis.

These distributions are shown in figures 6.16, 6.17 and 6.18, for ttH, ttW, and ttZ. In the plots four different hypothesis are displayed:

- blue line the SM hypothesis
- cyan line the response of Madgraph using dim6top for a given operator
- yellow line the response of Madgraph using SMEFT-LO for a given operator
- pink line the reweighing procedure starting from MadGraph SM to a particular dim6top operator

There are 2 things that we want to look at:

- the agreement between the pink and the cyan line, this tells us how and if it's working the reweighing procedure

[1] We use the `LHEPart` collection, which is synchronized with the output of the generator

- the agreement between the yellow and cyan lines that tells us whether the response of dim6top and SMEFT-LO is different for a particular operator

Looking at the plots we can see that the reweighing procedure gives a resonable response. Furthermore the agreement between dim6top and SMEFT is almost perfect but for some differences in the ctG operator that arise on how the settings of Madgraph for the two models are specified (for whom is familiar with Madgraph this difference is due to the difference in the param card of the two models, if the same param card is used the difference that you see in the plot will disappear).

Signal extraction

In order to measure potential BSM values of the WCs we need to both disentangle signal from the various background and also to explore kinematic variables that are sensitive to potential EFT signals.

In order to maximize the sensitivity with this dataset we also consider the distribution of the m_{ttH} variable. Figure 6.19 shows the distribution of the m_{ttH} variable in ttH simulated events in the 2lss0tau category under the SM hypothesis and setting each of the WC of interest to 1. All the coefficients under study induce a significant effect in the sample normalization, while only `ctG` shows a non-negligible effect in the event kinematics. Out of the distributions explored at generator level, the mass of the ttH system is the one with the stronger effect, and this effect gets translated to the m_{ttH} variable at reconstruction level, at shown in figure 6.19.

This plot is a reconstruction level plot. The effect on the Higgs p_T shape of each of the 3 operators is displayed.
From the ratio plot, we could notice that the effect of the ctG operator on the signal grows with the Higgs p_T, as expected from the theory, while it remains flat on the backgrounds, this happens because we have no Higgs production in the ttW or ttZ process and the ctG operator has an effect on the shape only if an Higgs is produced in the final state. The effect of the the ctp and cpt operator is instead flat on all the processes.
Figure 6.20 shows the likelihood scan as a function of `ctp`, `cpt` and `ctG`,

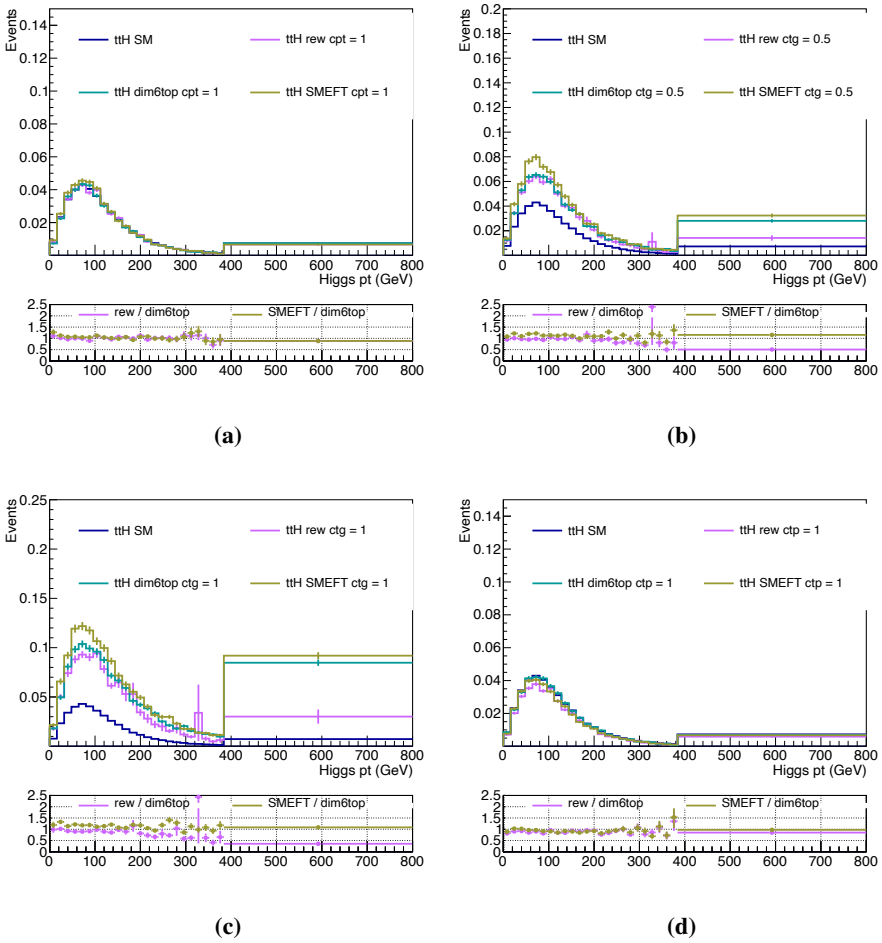


Figure 6.16: Comparison of the reweighted ttH MC signal distribution with the response of dim6top and SMEFT-LO for (a) cpt = 1, (b) ctG = 0.5, (c) ctG =1 and (d) ctp = 1

assuming the other two are fixed to the SM value, while figure 6.21 shows the 2-dimensional confidence regions as a function of pairs of those coefficients.

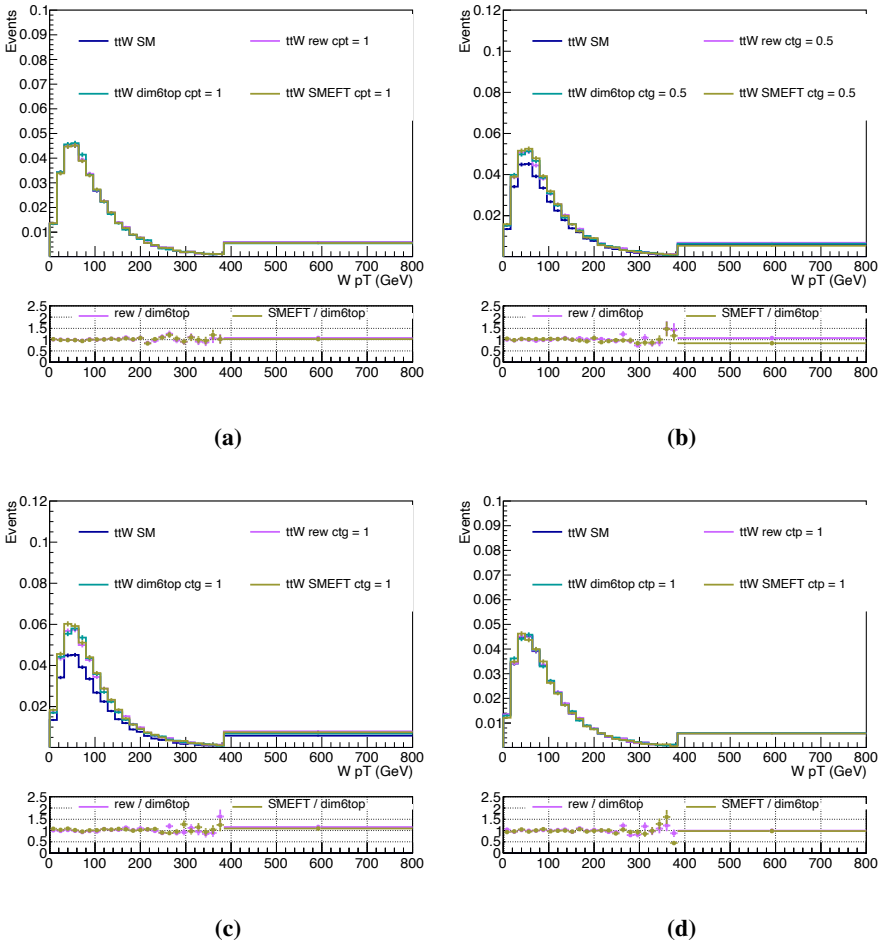
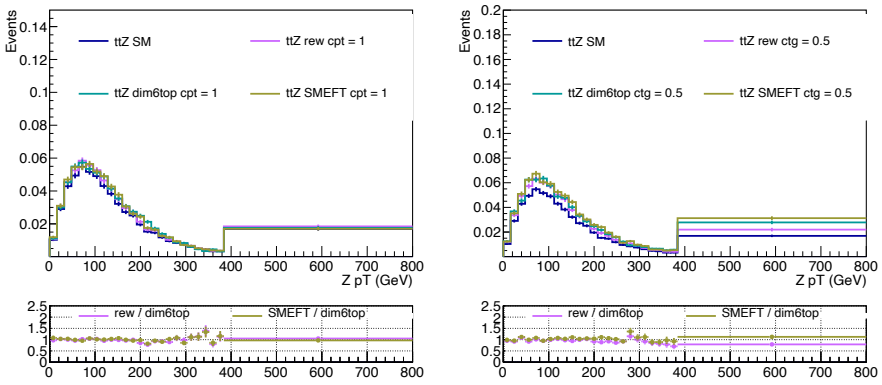
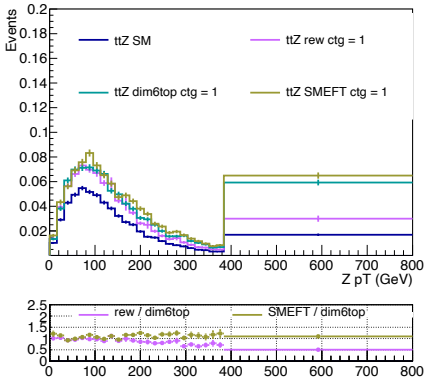


Figure 6.17: Comparison of the reweighted ttW MC distribution with the response of dim6top and SMEFT-LO for (a) cpt = 1, (b) ctG = 0.5, (c) ctG =1 and (d) ctp = 1

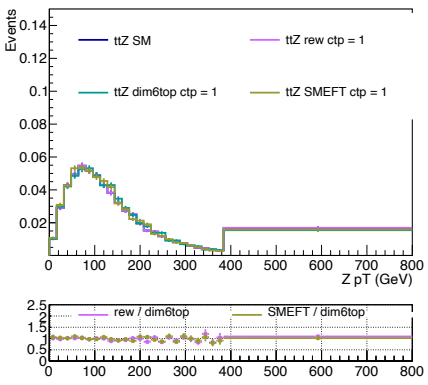


(a)

(b)



(c)



(d)

Figure 6.18: Comparison of the reweighted ttZ MC distribution with the response of dim6top and SMEFT-LO for (a) cpt = 1, (b) ctG = 0.5, (c) ctG =1 and (d) ctp = 1

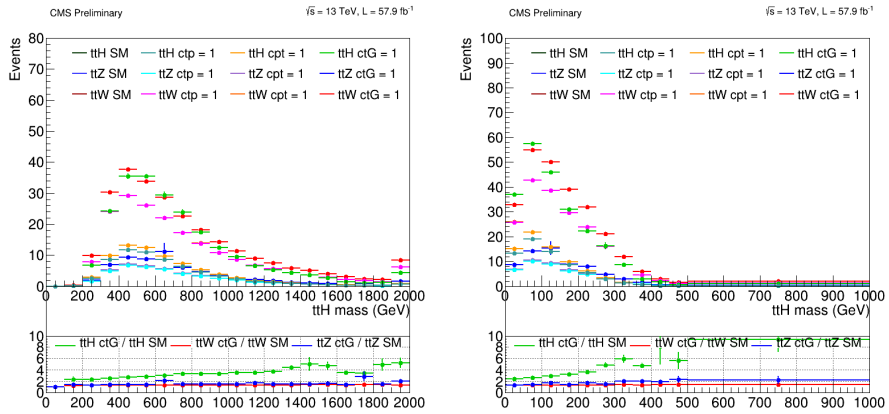


Figure 6.19: On the left distribution of the m_{ttH} variable in the 2lss0tau category under the SM hypothesis and setting each of the WC of interest to 1. On the right distribution of the Higgs p_T in the 2lss0tau category under the SM hypothesis and setting each of the WC of interest to 1.

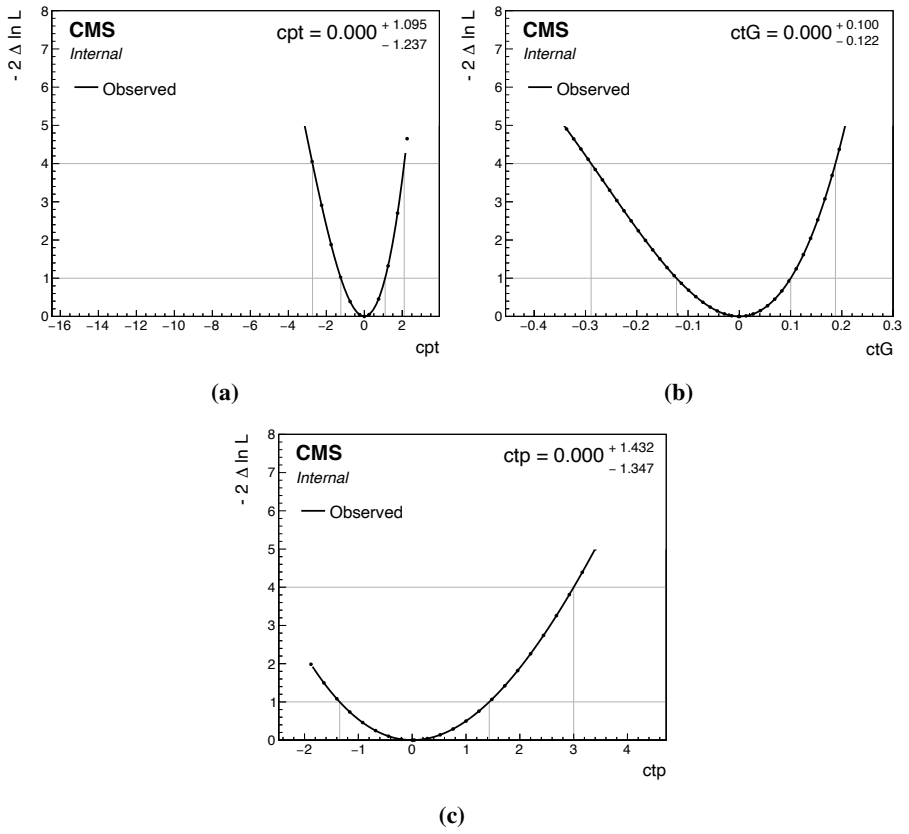
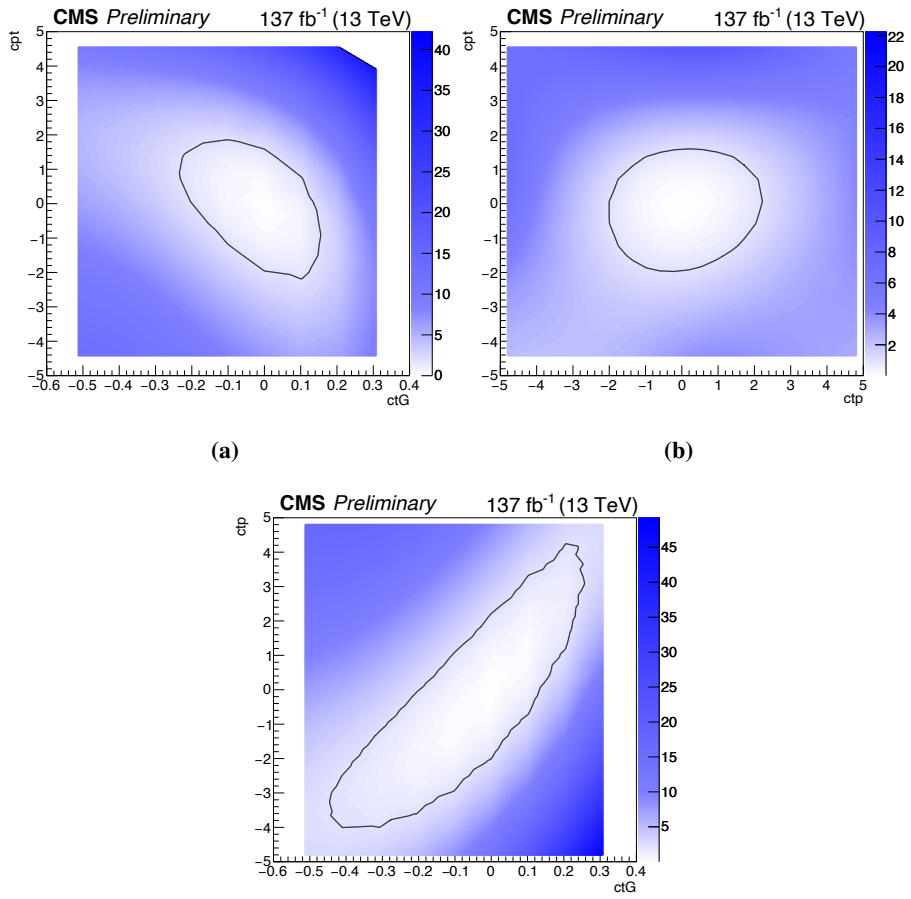


Figure 6.20: Likelihood scan as a function of `ctp` (top left), `cpt` (top right), `ctG` (bottom) coefficients, fixing the others to their SM values.



(a) **(b)**

(c)

Figure 6.21: Likelihood scan as a function of `cpt` and `ctG` (top left), `cpt` and `ctp` (top right), and `ctp` and `ctG`, fixing the third parameter to its SM value.

6.12 Conclusion

The discovery of a scalar boson by the CMS and ATLAS experiments in 2012 [109, 110] opened a new era of particle physics research: the characteristics of the newly discovered boson must be examined in detail to see if it is consistent with the Higgs boson predicted by the Standard Model. The SM Yukawa couplings y_{f} of Higgs to fermions are proportional to the fermion mass m_{f}, namely $y_{\mathrm{f}} = \sqrt{2}m_{\mathrm{f}}/v$, where $v \approx 246$ GeV denotes the vacuum expectation value of the Higgs field. With a mass of m = 173.34, the top quark is by far the heaviest fermion yet discovered: its Yukawa y is expected to be of the order one. The most precise measurement of this parameter was done in Ref. [84], yielding the following best fit: $\kappa_t = 1.01^{+0.11}_{-0.11}$. Several analysis had been done by the CMS collaboration to measure the ttH process; depending on the Higgs decay, three different searches can be distinguished: ttH where $H \to b\bar{b}$ which is the channel with the highest branching fraction but also with the highest background contamination [115], ttH where $H \to \gamma\gamma$ [116] which has the clearest signature but also the lowest branching fraction and ttH where $H \to W^+W^-, ZZ, \tau^+\tau^-$ with the advantage of having reasonably high branching fraction and clear final state given by the vector boson (taus) production [117]. For the purpose of this thesis, we will focus on the ttH production with a multileptonic final state, where H decays via $H \to W^+W^-, ZZ, \tau^+\tau^-$ and top quarks decay either leptonically or hadronically.

The measured production rates for the ttH and tH signals in the multilpetonic final state has been measured in [117] amount to 0.92±0.19(stat)+0.170.13(syst) and 5.7±2.7(stat)±3.0(syst) times their respective standard model (SM) expectations. Assuming that the H boson coupling to the τ lepton is equal in strength to the values expected in the SM, the coupling y_t of the H boson to the top quark is constrained, at 95% confidence level, to be within $0.9 < y_t < 0.7$ or $0.7 < y_t < 1.1$ times the SM expectation for this coupling. Following the stream of [117] we performed a differential and EFT analysis in ttH with a multileptonic final state.

The analysis aims at obtaining unfolded differential cross section results assuming the SM, and on constraining effective field theory (EFT) parameters that, if found to deviate from the SM, may be a direct evidence for new physics. A key difficulty in this analysis is that the presence of missing transverse energy in the signal process makes very hard to properly regress the Higgs transverse momentum that is used as input in the unfolding analysis. That's why we

developed a Deep Neural Network to regress the Higgs transverse momentum starting from well-understood signal observables.
The regression of the Higgs transverse momentum together with the ttH system mass are the two main observable that are unfolded to measure the ttH differential cross section. Regarding the EFT interpretation, we studied the most important Wilson Coefficients that affect the signal (and also some of the backgrounds) we ended up with ctp, ctg and cpt (see Sec 6.11.3) that are constraint in Sec 6.11.4. Complementary to the differential and EFT analysis, a CP search in the same final state has also being published by the CMS experiment [120].

6.13 Outlook

In this section I will discuss the possible improvements on the Higgs boson precision measurmente for HL-LHC. I will focus my attention mainly on the differential measurement, I'm referring to a ttH + tH, $H \rightarrow \gamma\gamma$ search [166].
The large number of events expected in the HL-LHC period, will enable us to measure the Higgs boson characteristics with extreme precision: from the determination of the production cross section to the precise determination of its couplings to gauge bosons and fermions. Furthermore, in many measurement channels, the experimental and theorical uncertainties will be compatible with the overall expected statistical uncertainties. We have to remember that the main limitation that we have nowadays for the complete understanding and measurement of many Higgs properties is the limited statistics of the data.
In this seciton, I will describe the strategy needed for a possible upgrade on the differential p_T cross section for the Higgs boson production in association with at least one top quark. We will analyze the most sensitive decay channel: ttH + tH, $H \rightarrow \gamma\gamma$; this final state provides an optimal reconstruction of the decay of the Higgs boson, and thus a direct measurement of the p_T differential cross-section can be done.
The expected precision of the analysis is determined based on simulated proton-proton (pp) events, at a centre of mass energy of 14 TeV. Event selection, applied to both signal and background, follow a similar strategy of the CMS full Run2 analysis:

- $|\eta_\gamma| < 2.4$, excluding the region $1.44 < |\eta| < 1.57$
- $100 < m_{\gamma\gamma} < 180$ GeV where the leading-p_T (subleading-p_T) satisfies $p_{T\gamma}/m_{\gamma\gamma} > 1/3(1/4)$

- $\Delta R_{\gamma\gamma} > 0.4$

Photons are required to be isolated by asking that the sum of charged transverse momentum in a cone of radius $\Delta R_{\gamma} = 0.4$, centred on the photon direction, is less than 0.3.
For events where more than one photon pair passes the selection, then the pair with $m_{\gamma\gamma}$ closest to the Higgs boson mass is chosen.
To suppress the single Higgs production background, all events are required to have at least one b-tagged jet. Two orthogonal categories, based on the decay products of the top quark are defined: hadronic category and a leptonic category.
Further requirements on the number of jets and on their kinematic are applied to separate the two categories. For the leptonic category, muons and electrons satisfy requirements on the η detector acceptance and on the isolation, defined as the sum of all reconstructed particles p_T, inside a cone of radius ΔR = 0.4, excluding the lepton itself. In addition, for electrons, the invariant mass of pairs formed from the electron and either selected photon is required to be greater than 95 GeV to reduce contamination from $Z \to e^+e^-$ decays. Events passing the leptonic category selection are excluded from the hadronic selection to maintain orthogonality of the two categories. A BDT is trained separately in each category to disentangle signal-like events from background-like events, the BDT is fed with the input variables related with the kinematic of the event, such as the lepton and jet momenta and η, and the scalar sum of transverse momentum of all final state objects in the event.
A cut on the BDT output is tuned to provide the best sensitivity to BSM scenarios. The hadronic category is further divided into two different region of the BDT output to reduce the contamination of the gluon gluon fusion Higgs production. Finally, the events are further divided into six bins of $p_{\gamma\gamma T}$, given in Tab. 30, making a total of 17 categories.
The dominant source of uncertainties is the reconstruction and identification efficiencies for photons and b jets as well as the energy scale and resolution of reconstructed jets (see Sec 4.7).
The differential cross-section is determined from a simultaneous maximum likelihood fit to an Asimov data set [163] corresponding to 3 ab^1, and assuming SM Higgs boson production in each category. Results are shown in Fig 6.22.

The results shown are unfolded back to a fiducial region which is common to both the hadronic and leptonic selections, and shown using only the hadronic

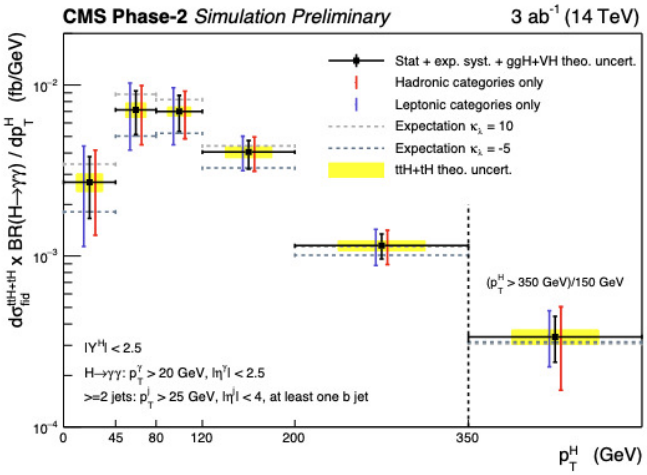


Figure 6.22: The expected p_T^H differential ttH + tH cross sections times branching ratio, along with their uncertainties. The error bars on the black points include the statistical uncertainty, the experimental systematic uncertainties and the theoretical uncertainties related to the ggH and VH contamination, which is subtracted in the fit. The cross section for p_T^H > 350 GeV is scaled by the width of the previous bin. The expected ttH + tH cross sections for anomalous values of the Higgs boson self-coupling (k_λ = 10 and k_λ = -5) are shown by the horizontal dashed lines. Plot is taken from [166]

or leptonic categories, and their combination. The precision obtained for each of the six bins analysed shows a great improvement with respect to the expectation of the run2 analysis [116]. A similar improvement is expected in all the ttH final states. To conclude this section is worth to mentioned prospects done by CMS and ATLAS collaborations for a scenario with 6000 fb^{-1} [166]. With this luminosity it will be possible to further split the Higgs p_T distribution ending up with 9 bins as shown in Fig 6.23.

Very few expectation on the EFT constrain of the ttH production can be find in [166], studies are still ongoing.

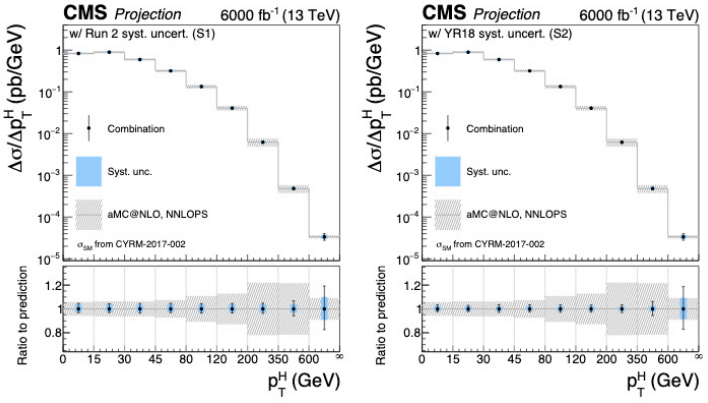


Figure 6.23: Projected differential cross section for p_T^H an integrated luminosity of 6000 fb^{-1} (represent-ing the sensitivity achievable by an eventual ATLAS and CMS combination) [166]

Conclusion

In my scientific career I mainly participated in activities related to physics analysis of the CMS experiment. Presently I am a key person in the ttH multileptonic differential analysis, I had a leading role in the double Higgs production in $b\bar{b}4l$ analysis and I have an active role in the Heavy Neutral Lepton analysis. In parallel to the analysis activities, I have a strong interest in hardware optimization: within the CMS experiment I was in charge of quality control tests for the wire bonding of the Outer Tracker system for Phase II upgrade and I was the responsible for the new derivation of 2016 UL gain payloads. On the hardware side, I also took part of a test beam for future colliders working on a drift chamber. Currently I am the CMS ttH Montecarlo Generator Contact for the Higgs Group and I am responsible for the production of all the samples of simulated events for all the CMS ttH analyses. Below I describe the results achieved during my research activity and the status of my current work.
I worked on the reconstruction of $HH \rightarrow b\bar{b}ZZ^* \rightarrow b\bar{b}4l$ channel in order to estimate an upper limit on the signal strength of the HH production 4. This analysis explored a phase-space not yet explored at LHC that could lead to new physical phenomena due to the very high purity of the final state; the search is also particularly challenging due to very high background contamination. I led this analysis from the beginning and I touched all the aspect: from the simulation of the signal process to the evaluation of the upper limit on different observables. This analysis was meant to target both the SM by setting an upper limit on the signal strength modifier μ and anomalous values of the Higgs trilinear coupling. I was also involved in the ttH multileptonic differential analysis 6. The analysis aims at obtaining unfolded differential cross section results assuming the SM, and on constraining effective field theory (EFT) parameters that, if found to deviate from the SM, may be a direct evidence for new physics. I was involved from the beginning in all the aspects of the analysis: I am a ttH Montecarlo Generator Contact for the Higgs Group, that is the responsible of generating the samples that are used for the analysis, and I set up all the framework for the differential cross section measurement. A key difficulty in this analysis is that the presence of missing transverse energy

in the signal process makes very hard to properly regress the Higgs transverse momentum that is used as input in the unfolding analysis. That's why I developed a Deep Neural Network to regress the Higgs transverse momentum starting from well-understood signal observables. The regression of the Higgs transverse momentum together with the ttH system mass are the two main observable that are unfolded to measure the ttH differential cross section. Once I finalised the unfolding part, I started to look at the EFT interpretation, I studied the most important Wilson Coefficients that affect the signal (and also some of the backgrounds) and for this purpose I learned some advanced features of the MadGraph generator that are specific to EFT studies. I'm now working on decoupling the EFT effects from the SM and on finalizing the full run2 analysis. While working on ttH and to further investigate BSM scenarios with displaced particles, I am currently working in a Heavy Neutral Lepton search that looks for right-handed neutrinos with Majorana masses below the electroweak scale 5. This addition is able to generate both the light neutrino masses and the baryon asymmetry of the universe via low scale leptogenesis. In the history, lots of different phase space of masses and life time were studied to find evidence of such peculiar particles; this search is focused on a phase space that is still uninvestigated from the all the other experiment: mass range between 1 and 15 GeV that can decay in the tracker volume up to 60 cm. Such displaced search is highly peculiar and challenging analyses at the LHC in high demand for dedicated data reconstruction tools in order to extend their sensitivity. The signature of the search is a a prompt lepton, a displaced lepton, a displaced jet and a secondary displaced vertex. In the context of this analysis, I pursued the effort on the publication doing some studies mainly on the signal topology and characterization: I took care of reweigthing the signal samples to the different couplings, SR optimization studies, SF measurement and studies to understand the composition of the most important backgrounds. The analysis is in advanced state and is also aiming for winter conferences.

Acknowledgement

I would like to thank my UCLouvain group: Claudio, Jessica and in particular my supervisor Giacomo, I could not have undertaken this journey without you all. Thanks a lot for the support, the work, the effort and the confidence placed in me.
I would like to express my deepest appreciation to all the people in my department, for giving me the best environment to grow up as a physicist.
I'm extremely grateful to all the members of the thesis jury for helping me in improving this work. Special thanks goes to my family and friend for the support and love.

Appendix A

Appendix HH

A.1 ZX control plots validation

A.1.1 Sidebands control region

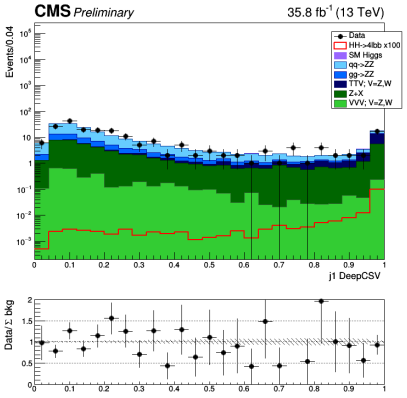
In this section we report the control plots done in the four-lepton sidebands after requiring the presence of at least 2 jets in the event (4ljj sidebands). Events selected in these plots are contained in the region outside the $m_{4\ell}$ region of the Higgs boson peak: events contained in the region $115 < m_{4\ell} < 135$ GeV are not included in the following plots.

2016 plots

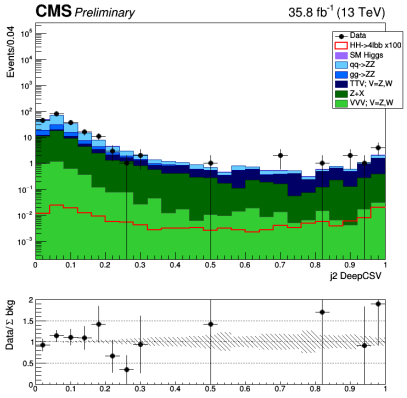
In this section we report the control plots done with 2016 datastes in the four-lepton sidebands after requiring the presence of at least 2 jets in the event. In Figure A.1 and A.2 we report the distributions obtained with 2016 samples. The yields obtained in this region are reported in Tab A.1.
A good data/mc agreement is observed.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.00042	0.48	35.54	8.61	4.42	17.90	1.11	68.49	75
4e	0.00045	0.36	20.81	5.48	3.42	6.92	0.91	39.46	36
$2e2\mu$	0.0008	0.74	49.55	13.60	7.89	24.14	1.85	97.78	96

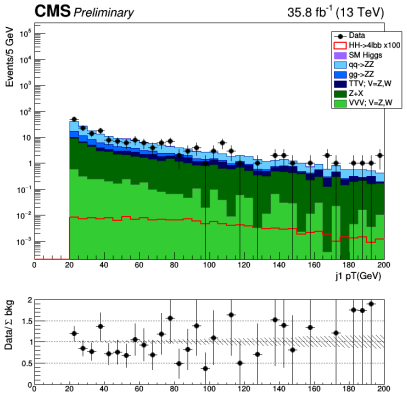
Table A.1: Yields four-lepton sidebands after requiring the presence of at least 2 jets in the event for 2016 dataset.



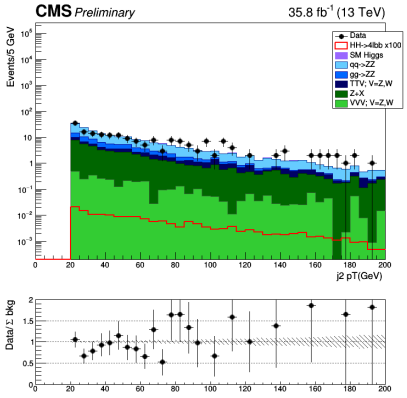
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant

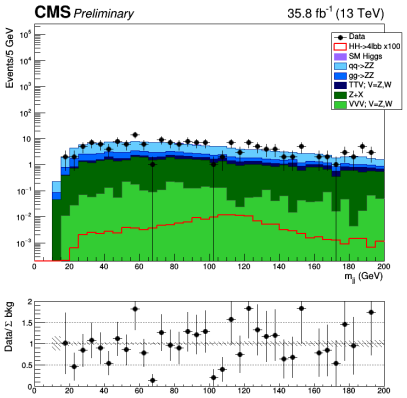


(c) p_T of jet with the highest value of CSV discriminant for 2016 dataset

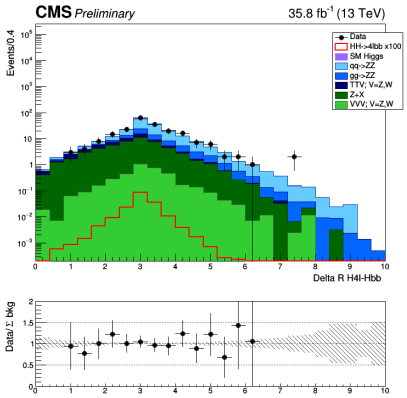


(d) p_T of jet with the second highest value of deepCSV discriminant

Figure A.1: Four-lepton + 2jets sideband plots for 2016 datasets.



(a) Invariant mass built from the two selected jets



(b) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$

Figure A.2: Four-lepton + 2jets sideband plots for 2016 datasets.

2017 plots

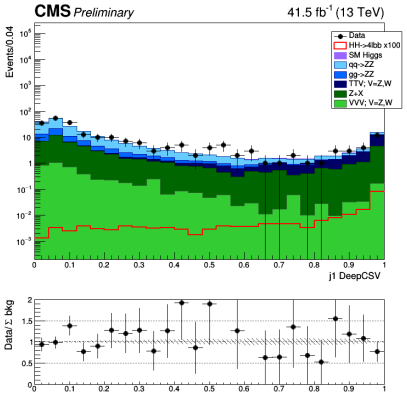
In this section we report the control plots done in the four-lepton sidebands after requiring the presence of at least 2 jets in the event (4ljj sidebands) for 2017 dataset. In Figure A.3 and A.4 we report the distributions obtained with 2017 samples. The yields obtained in this region are reported in Tab A.2.
A good data/mc agreement is observed.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.0005	0.55	38.35	9.06	5.62	19.26	1.49	74.33	83
4e	0.0004	0.38	22.42	5.63	4.07	5.68	1.03	39.21	47
$2e2\mu$	0.0009	0.84	53.61	14.60	8.34	23.50	2.27	103.17	99

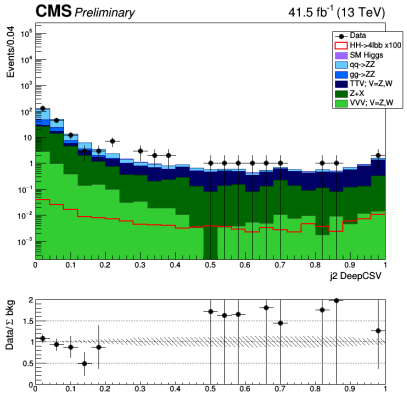
Table A.2: Yields four-lepton sidebands after requiring the presence of at least 2 jets in the event for 2017 dataset.

2018 plots

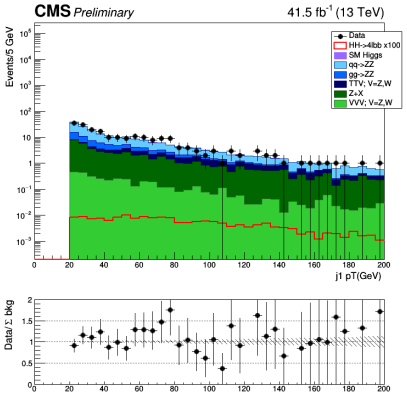
In this section we report the control plots done in the four-lepton sidebands after requiring the presence of at least 2 jets in the event (4ljj sidebands) for 2018 dataset. In Figure A.5 and A.6 we report the distributions obtained with



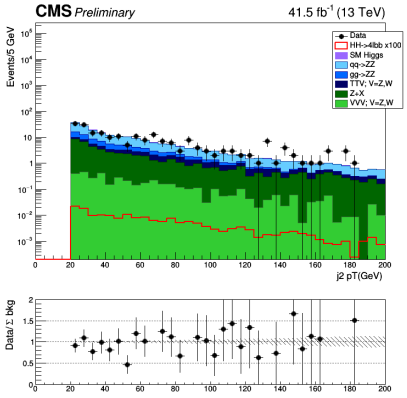
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant

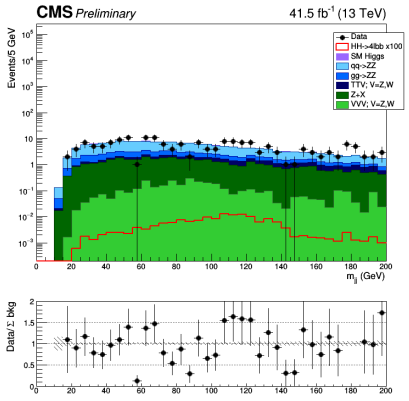


(c) p_T of jet with the highest value of CSV discriminant for 2017 dataset

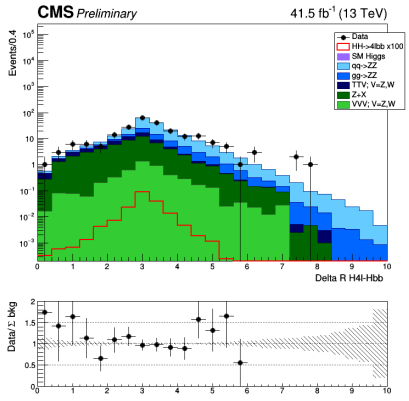


(d) p_T of jet with the second highest value of deepCSV discriminant

Figure A.3: Four-lepton + 2jets sideband plots for 2017 datasets.



(a) Invariant mass built from the two selected jets



(b) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$

Figure A.4: Four-lepton + 2jets sideband plots for 2017 datasets.

2018 samples. The yields obtained in this region are reported in Tab A.3. A good data/mc agreement is observed.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.0008	0.84	59.13	13.88	8.65	30.24	2.28	115.02	114
4e	0.0008	0.58	33.66	8.48	5.54	8.28	1.11	57.65	70
$2e2\mu$	0.0014	1.25	81.71	21.99	12.31	33.62	3.43	154.33	169

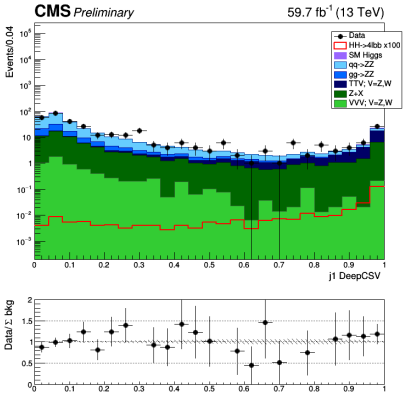
Table A.3: Yields four-lepton sidebands after requiring the presence of at least 2 jets in the event for 2018 dataset.

A.1.2 Reduced sidebands control region

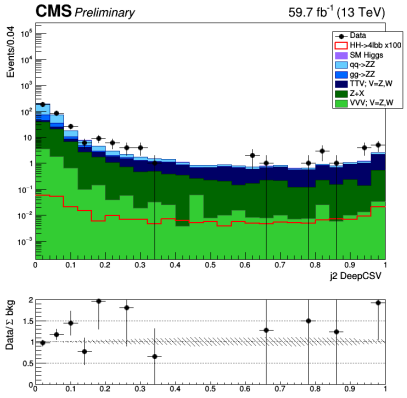
In this section we report the control plots done in the four-lepton reduced sidebands after requiring the presence of at least 2 jets in the event (4ljj sidebands). Events selected in these plots are contained in the regions $95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV, after requiring the presence of at least 2 jets in the event.

2016 plots

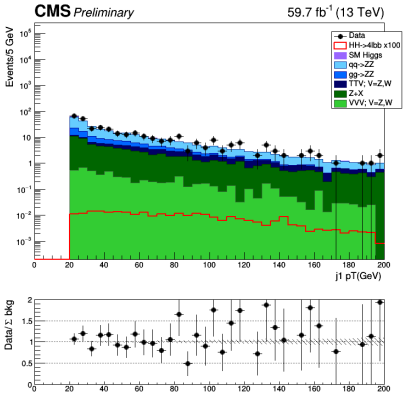
In this section we report the control plots done with 2016 dataset in the four-lepton reduced sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV)



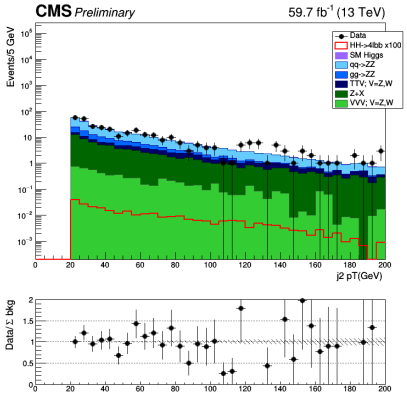
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant

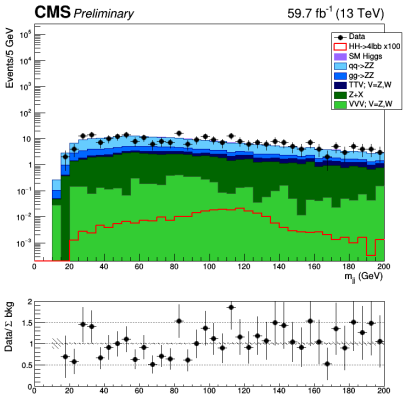


(c) p_T of jet with the highest value of CSV discriminant for 2018 dataset

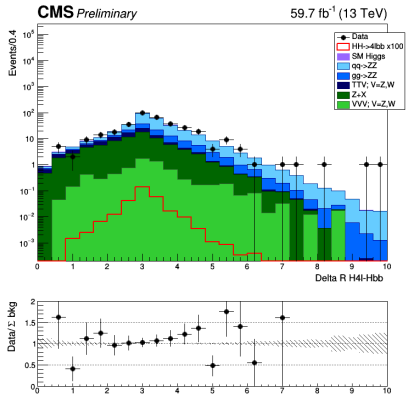


(d) p_T of jet with the second highest value of deepCSV discriminant

Figure A.5: Four-lepton + 2jets sideband plots for 2018 datasets.



(a) Invariant mass built from the two selected jets



(b) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$

Figure A.6: Four-lepton + 2jets sideband plots for 2018 datasets.

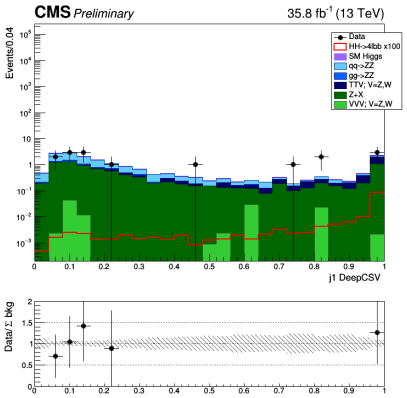
after requiring the presence of at least 2 jets in the event. In Figure A.7 and A.8 we report the distributions obtained with 2016 samples. The yields obtained in this region are reported in Tab A.4.
A good data/mc agreement is observed.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.0003	0.21	2.71	0.27	0.61	3.53	0.004	7.33	6
4e	0.0004	0.17	1.53	0.20	0.42	1.04	0.031	3.39	4
2e2μ	0.0007	0.35	2.79	0.37	1.11	4.47	0.081	9.17	6

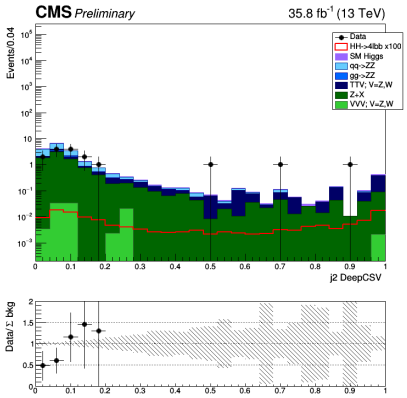
Table A.4: Yields four-lepton sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) after requiring the presence of at least 2 jets in the event for 2016 dataset.

2017 plots

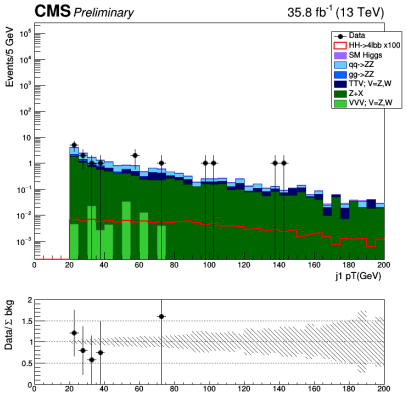
In this section we report the control plots done with 2017 dataset in the four-lepton reduced sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) after requiring the presence of at least 2 jets in the event. In Figure A.9 and A.10 we report the distributions obtained with 2017 samples. The yields ob-



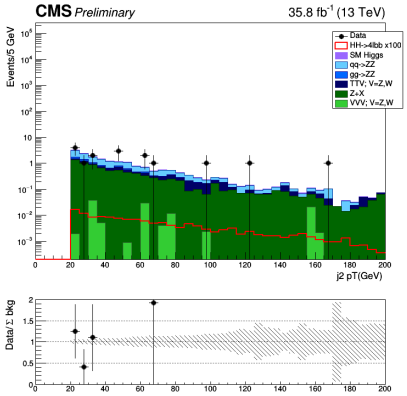
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant



(c) p_T of jet with the highest value of CSV discriminant for 2016 dataset

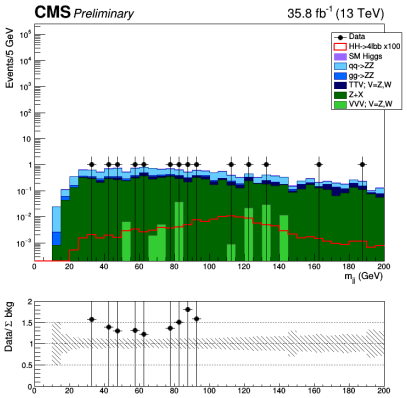


(d) p_T of jet with the second highest value of deepCSV discriminant

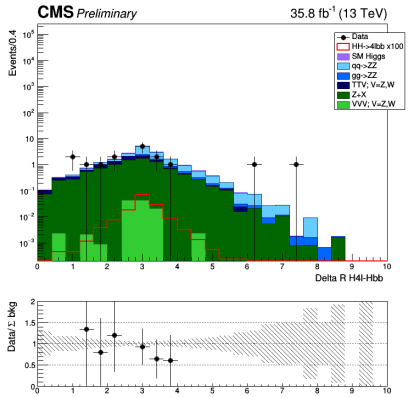
Figure A.7: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2016 datasets.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.0004	0.23	2.88	0.33	0.89	3.73	0.007	8.07	11
4e	0.0004	0.17	1.33	0.21	0.48	0.89	0.026	3.11	5
$2e2\mu$	0.0008	0.39	3.10	0.38	1.27	4.70	0.011	9.85	8

Table A.5: Yields four-lepton sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) after requiring the presence of at least 2 jets in the event for 2017 dataset.



(a) Invariant mass built from the two selected jets



(b) ΔR between the H $\to 4\ell$ and H $\to$

Figure A.8: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2016 datasets.

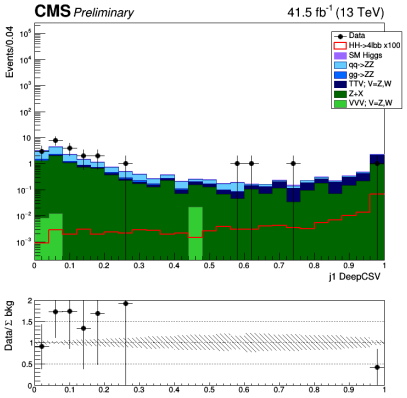
tained in this region are reported in Tab A.5.
A good data/mc agreement is observed.

2018 plots

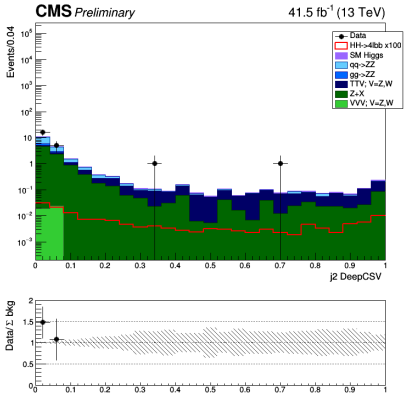
In this section we report the control plots done with 2018 dataset in the four-lepton reduced sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) after requiring the presence of at least 2 jets in the event. In Figure A.11 and A.12 we report the distributions obtained with 2018 samples. The yields obtained in this region are reported in Tab A.6.
A good data/mc agreement is observed.

Topology	Signal	SM Higgs	qqZZ	ggZZ	TTV	Z+X	VVV	All background	Data
4μ	0.0006	0.35	4.46	0.53	1.16	5.84	0.07	12.4	12
4e	0.0006	0.27	2.07	0.30	0.67	1.31	0.03	4.67	6
2e2μ	0.0012	0.57	4.87	0.62	2.00	6.15	0.05	14.3	14

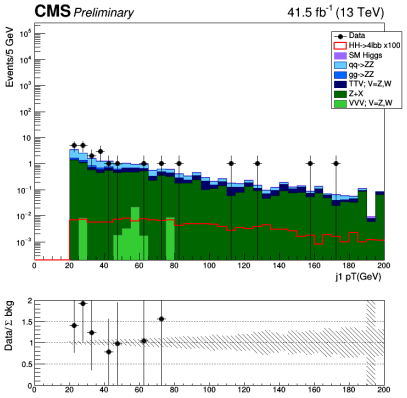
Table A.6: Yields four-lepton sidebands ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) after requiring the presence of at least 2 jets in the event for 2018 dataset.



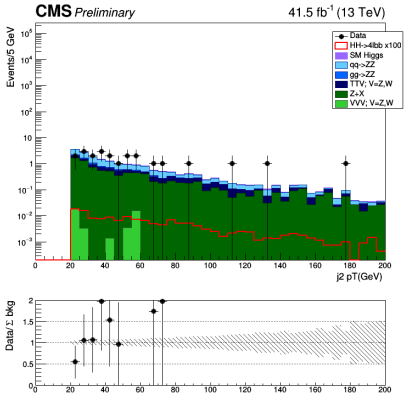
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant

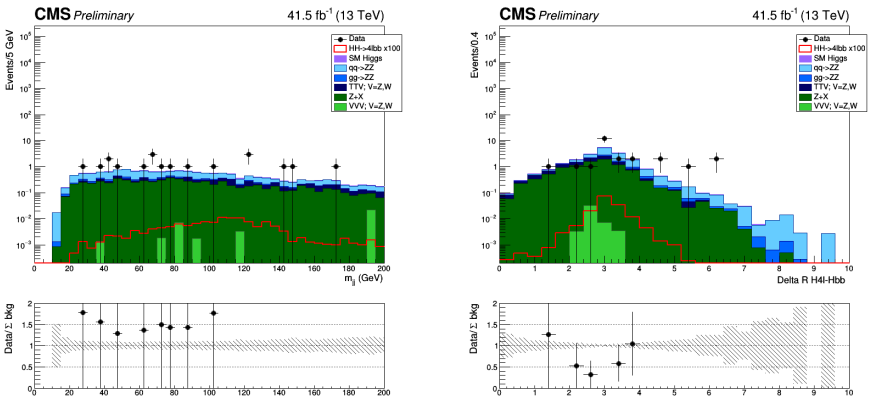


(c) p_T of jet with the highest value of CSV discriminant for 2017 dataset



(d) p_T of jet with the second highest value of deepCSV discriminant

Figure A.9: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2017 datasets.



(a) Invariant mass built from the two selected jets

(b) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$

Figure A.10: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2017 datasets.

A.2 Appendix B: test on BDT configuration

A.2.1 Tests on BDT configurations

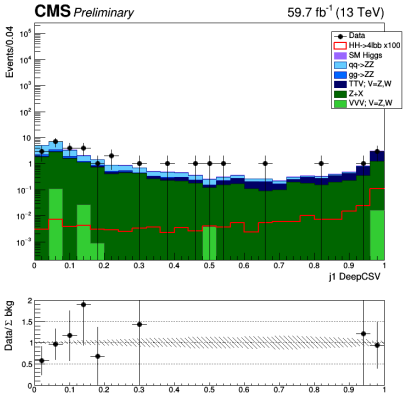
Different input variables were tested to find the most discriminant combination. All the studies were done for 2016 datasets, for the final state $H \rightarrow ZZ \rightarrow 4\mu$.

Tests on BDT configurations: first set of variables

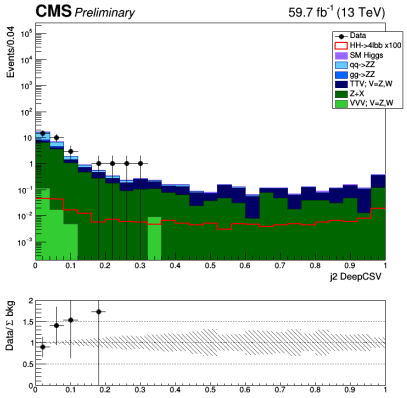
The first set of variables tryed was:

- p_T of the four leptons
- ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$
- $\Delta\varphi$ between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$
- Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4)

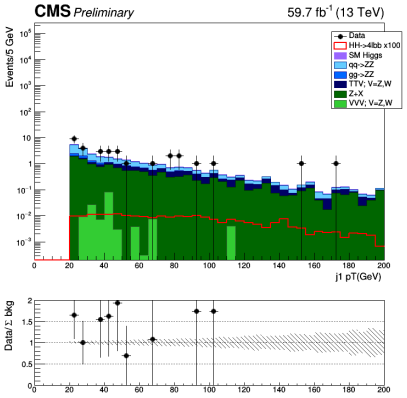
The area under the ROC curve is 0.894. As an overtraing checking, in Figure A.13 ROC curves for training and testing samples are show.



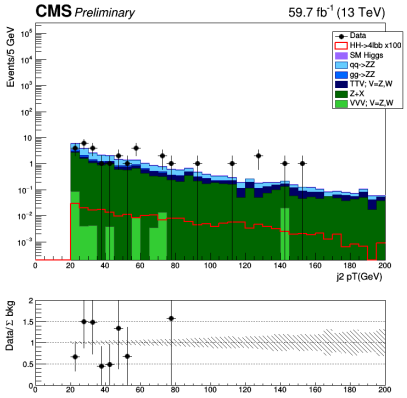
(a) B tagger score of the jet with the highest value of the deepCSV discriminant



(b) B tagger score of the jet with the second highest value of the deepCSV discriminant

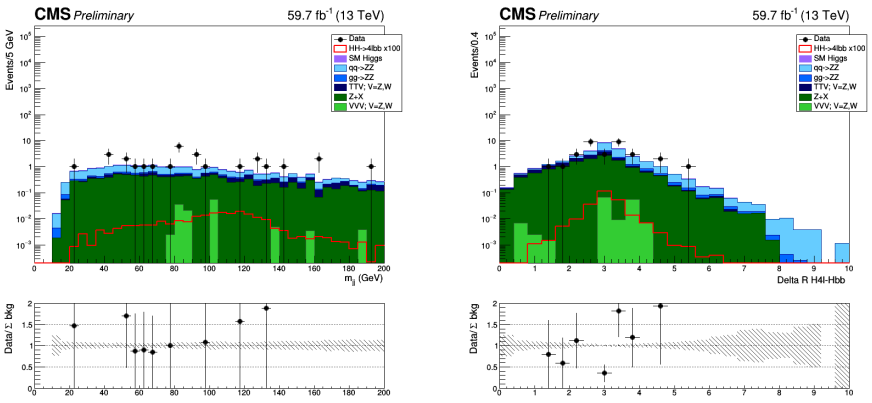


(c) p_T of jet with the highest value of CSV discriminant for 2018 dataset



(d) p_T of jet with the second highest value of deepCSV discriminant

Figure A.11: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2018 datasets.



(a) Invariant mass built from the two selected jets

(b) ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$

Figure A.12: Four-lepton + 2jets sideband ($95 < m_{4\ell} < 115$ GeV and $135 < m_{4\ell} < 170$ GeV) plots for 2018 datasets.

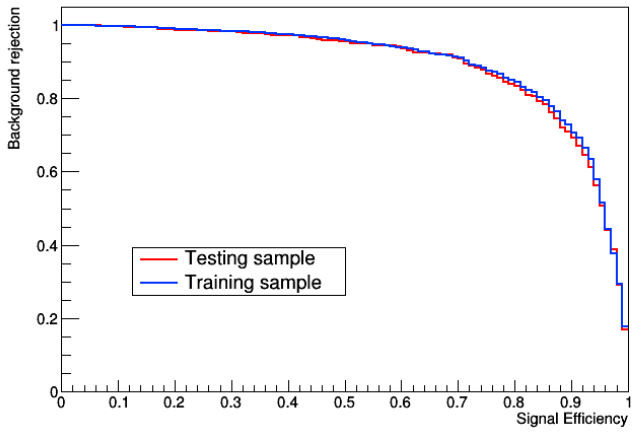


Figure A.13: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, $\Delta\varphi$ between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4) and the analysied dataset is 2016.

Tests on BDT configurations: second set of variables

In order to suppress the background from ttZ of ttH the missing energy is considered:

- p_T of the four leptons

- Missing Energy Transverse (MET) in the event
- ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$
- $\Delta\varphi$ between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$
- Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4)

The area under the ROC curve now is 0.901. As an overtraing check, in Figure A.14 ROC curves for training and testing samples are show.

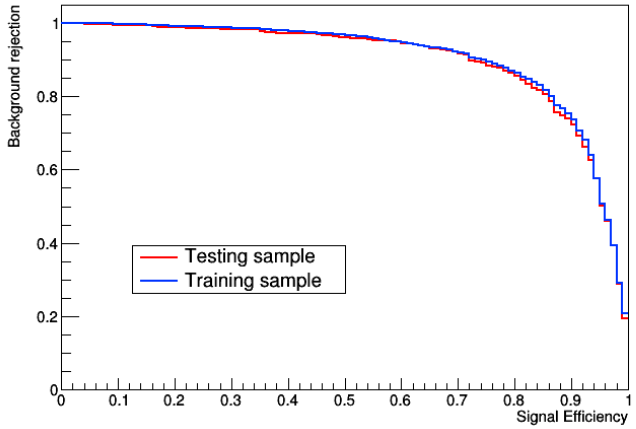


Figure A.14: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, Missing Energy Transverse (MET) in the event, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$, $\Delta\varphi$ between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$, value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4) and the analysied dataset is 2016.

Tests on BDT configurations: third set of variables

As the $\Delta\varphi$ between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$ is isotropic and, to take into account the presence of the $Higgs \rightarrow b\bar{b}$, we removed the $\Delta\varphi$ and we added the $p_T s$ of the jets:

- p_T of the four leptons
- Missing Energy Transverse (MET) in the event
- ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow$ b$\bar{\text{b}}$

- Value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4)
- p_T of the two jets

The area under the ROC curve now is 0.904. As an overtraing check, in Figure A.15 ROC curves for training and testing samples are show.

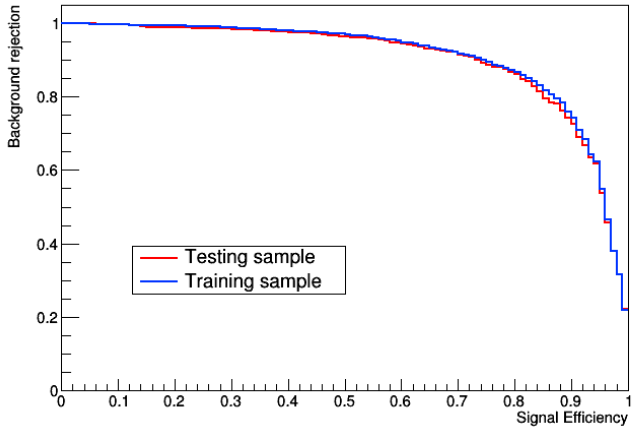


Figure A.15: ROC curve for training and testing samples. The BDT was trained with: p_T of the four leptons, Missing Energy Transverse (MET) in the event, ΔR between the H $\rightarrow 4\ell$ and H $\rightarrow b\bar{b}$, value of the b tagging algorithm score of the two jets with the highest b tagger score (see Section 4.4), p_T of the two jets and the analysied dataset is 2016.

Trying to find the best set input variables of the BDT, more observables were tested and the performance of the network is sumarized in Tab A.16.

	pt letptons	eta letptons	phi leptons	pt jets	eta jets	phi jets	b disc jets	delta R HH	MET	delta phi HH	mass jets		ROC
1													0.894
2													0.901
3													0.901
4													0.904
5													0.905
6													0.904
7													0.901
8													0.917
9													0.917
10													0.917
11													0.911
12													0.909
13													0.917

Figure A.16: List of input variables tryed for training the BDT

Appendix B

Appendix HNL

B.1 Systematics

The systematic uncertainties related with this analysis are briefly discussed below. I want to highlight that I haven't worked on estimating them.

B.1.1 Uncertainty on signal MC cross section

The heavy Neutrino model used for generation of Monte Carlo HNL events [93–95] does not allow for NLO QCD calculations. The simulation of the HNL events is at LO, this leads to a large theoretical uncertainties on the cross section (up to 15%) that will affect the final result. To don't rely on this uncertainties a general correction factor for the cross section from LO to NNLO can be derived based on the SM production of $W \to l\nu$, where the only difference is that in the HNL production the ν is exanged with the HNL; furthermore the effect of the mass and coupling of the HNL is not affected by the PDF and scale variations thus the dominant effect of these variations comes from the W boson, therefore it is reasonable to estimate the systematic uncertainties on the HNL production from the $W \to l\nu$ process. The recommended NNLO value for the SM process is $61526.7^{+497.1}_{-264.6} \pm 2312.7$ pb where the quoted uncertainties are respectively scale and PDF uncertainties. Assuming uncorrelated uncertainties and taking the maximum of the two asymmetric errors, the combined uncertainty is 61526.7 ± 2365.5 pb, an effect of 3.86%. This gives a final scale factor of 1.089 ± 0.042.

NLO corrections on transverse momentum spectra

In addition to the reweighting of the inclusive cross section, the impact of NLO corrections on the transverse momentum distributions of the generated

Table B.1: Simulated $W \rightarrow l\nu$ samples with 2016 data-taking conditions.

Sample	No. events
/WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/MINIAODSIM	29514020
/WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3_ext2-v2/MINIAODSIM	57402435
/WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v1/MINIAODSIM	24120319
/WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3_ext2-v1/MINIAODSIM	237263153

particles are studied. Also in this case we can use the $W \rightarrow l\nu$ SM process and compare the LO and NLO accuracy of the samples listed in Table B.1.

For each event of each sample we selected the W boson decayin into lepton and neutriono and we derive the impact of the NLO correction by comparing the LO/NLO variation. For each distribution, the envelope of six matrix element (ME) scale variations is estimated to assign a theory uncertainty; the six variations are for μ_{R} and μ_{F} being varied up and down by a factor of 2 separately and coherently; the statistical uncertainties of the MC prediction is also taken into account.

In Fig. B.1, the transverse momentum spectrum of the generated W boson is shown. Both at LO and NLO, the uncertainty from ME scale variations is small for transverse momenta below about 20 GeV, and increases afterwards. In the NLO/LO comparison, deviations of less than 10% are observed, especially around 20 GeVwhere the deviations are larger than the scale uncertainty. Similar conclusion can be done both for the transverse momentum spectrum of the generated lepton and for the transverse momentum spectrum of the generated neutrino (the different scale choices agree very well up to a transverse momentum of about 45 GeV). It is reasonable to apply these uncertaities only if we deal with high enough transverse momenta object, but since the bulk of $W \rightarrow l\nu$ events is generated with very small transverse momenta, the NLO correction on the transverse momenta can be neglected. Thus, no additional uncertainty will be assigned.

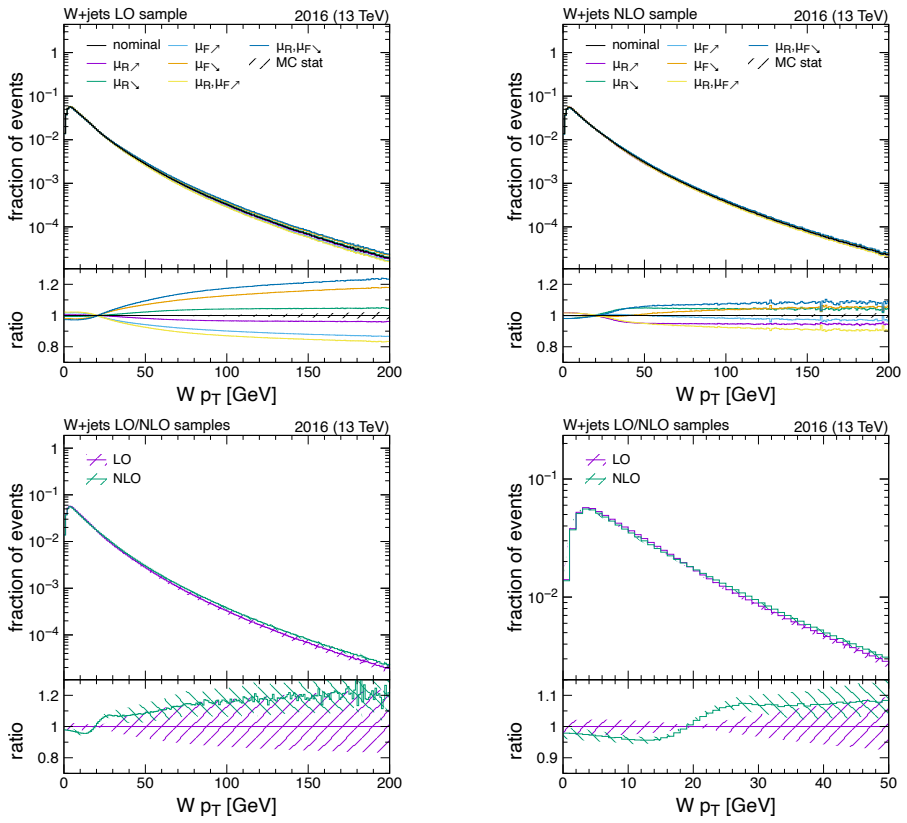


Figure B.1: Generated W boson transverse momentum spectrum predicted at LO accuracy (upper left) and NLO accuracy (upper right), and the comparison of both predictions (lower left, lower right zoomed in).

B.1.2 Particle Flow Network

Since the PFN has been trained on MC, it is sensitive to data-MC differences in his input variables. We perform validation of the PFN output score in data and MC in two validation regions: k decays, which can mimic the HNL signature and $t\bar{t}$ dominated region where we are more sensitive to differences in the PFN Background efficiency.
Based on these two regions seems that the mismodeling of the input variables does not have a significant impact on the PFN output shape for signal. To be anyway conservative, we derived an uncertainties based on data-MC ratios that covers any possible PFN mismodeling. The value that is determine for this uncertainties is 10%.

B.1.3 Pileup

The pileup distribution in MC samples is different from the one in data and a per-event reweigthing is usually applyed to cover this discrepancy. The expected pileup profile in data is calculated using the total inelastic cross section of 69.2mb and the luminosity measurement. The uncertainty in the total inelastic cross section is 4.6%, and its impact on the results is evaluated by varying the pileup reweighting scale factors according to the variation of the cross section.

B.1.4 Integrated Luminosity

The measured integrated luminosity for CMS has an uncertainty of 1.2% (2016) [167], 2.3% (2017) [168], 2.5% (2018) [169]. Taking the correlations between the years into account, the uncertainty for the full Run 2 dataset is 1.6%. Based on the CMS recommendations we evaluated and splitted the correlated and uncorrelated parts of the uncertaintis for 2016 and 2017+2018.

B.1.5 Trigger Efficiency

The analysis employs the respective unprescaled single electron and single muon triggers for each year to select events. The prompt lepton is geometrically matched to the trigger lepton. A tag-and-probe technique is used to measure a scale factor between data and MC in dilepton events. The probe is a fully identified and isolated prompt electron or muon, the other lepton is used as the tag to measure the trigger efficiency with. The difference in efficiency

is taken as the data-MC scale factor. The uncertainties from the tag-and-probe fits are propagated to get a systematic uncertainty on the obtained scale factors. This happens in the same way as described for prompt lepton identification. They are less than 1% in size for both electron and muon triggers.

B.1.6 Prompt Lepton identification efficiency

Prompt electron and muon identification and isolation efficiencies are measured in data and simulation using a tag-and-probe method, applied to samples of inclusive Z boson events. The data-to-MC scale factors are measured as a function of the lepton p_T and η. The ID scale factors are provided by the CMS collaboration [139]. The results are presented in Fig. B.3. For the prompt electrons, the SFs are obtained following the CMS prescriptions using the same tag and probe techniques as the one for the muons. The results as a function of transverse momentum or pseudorapidity are shown in Fig B.2.

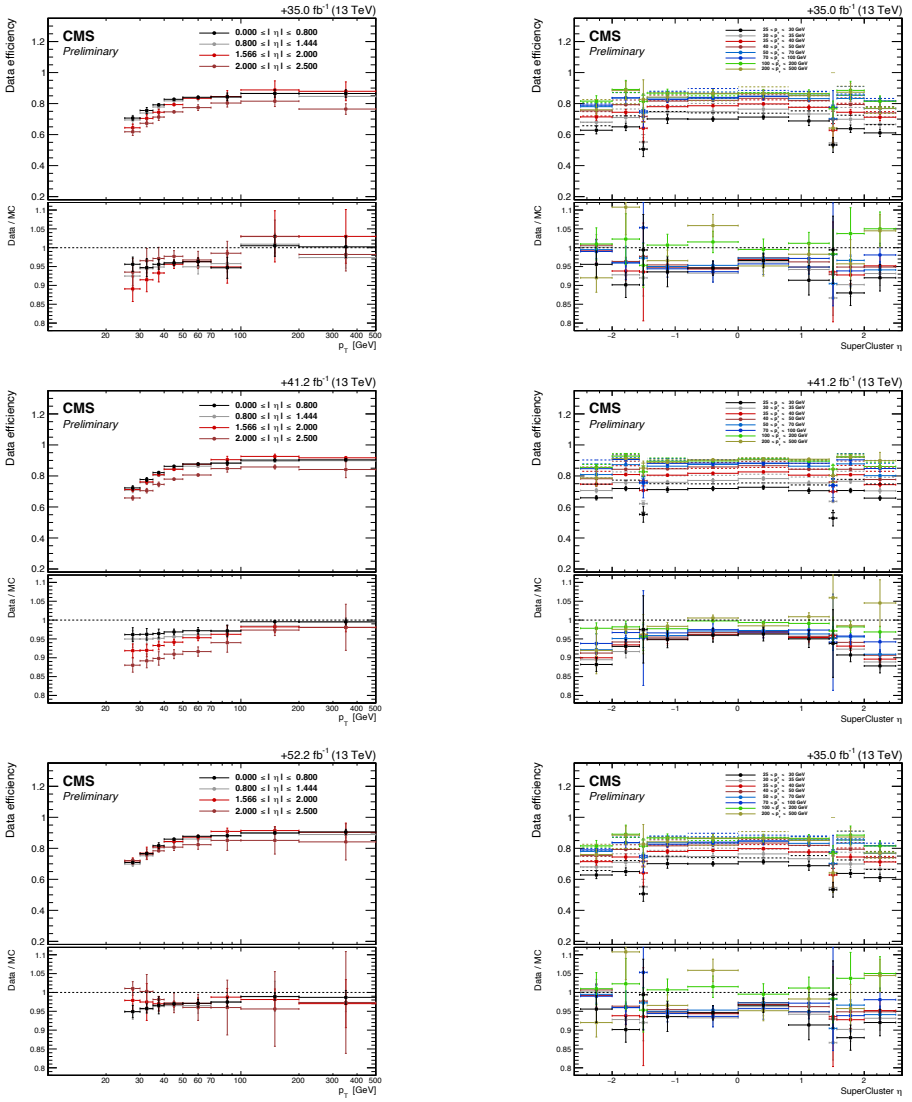


Figure B.2: Prompt electron efficiencies and data-MC scaling factors as a function of transverse momentum (left) or pseudorapidity (right). SF for 2016/17/18 (top, middle, bottom).

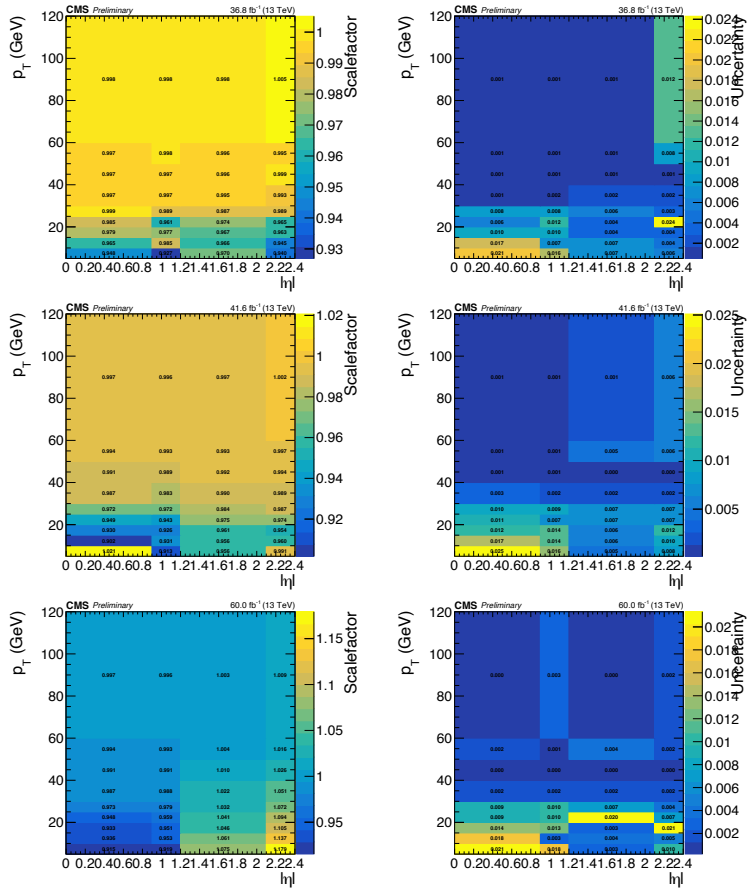


Figure B.3: Data/MC efficiency scale factors (left) and associated systematic uncertainty (right) for impact parameter and isolation requirement efficiency for prompt muons.

B.1.7 Displaced lepton identification efficiency

In this section we will use the tag-and-probe method to assign a systematic uncertainties on the leptons identification, the systematic uncertainties on the track reconstruction efficiency will be discussed in the next section.

Muons

The displaced muon identification efficiency is computed using the tag-and-probe method in $B^{\pm} \rightarrow J/\Psi K^{\pm} \rightarrow \mu^{-}\mu^{+}K^{\pm}$ event. We choose to work with $J/\Psi \rightarrow \mu^{-}\mu^{+}$ produced in B-hadron decays, thus displaced, because is the process that best mimic the HNL signature.
The samples used for this study are listed in Tab B.2.

Table B.2: Data sets and MC samples used in the $\rightarrow$ tag-and-probe study to assess displaced muon efficiencies.

Year	Data set or MC sample
3*2016	/Charmonium/Run2016[B-H]-17Jul2018-v1/MINIAOD /BuToJpsiK_BMuonFilter_SoftQCDnonD_TuneCUEP8M1_13TeV-pythia8-evtgen/ RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM
2*2017	/Charmonium/Run2017[B-F]-31Mar2018-v1/MINIAOD /BuToJpsiK_BMuonFilter_SoftQCDnonD_TuneCP5_13TeV-pythia8-evtgen/ RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v3/MINIAODSIM
3*2018	/Charmonium/Run2018[A-C]-17Sep2018-v1/MINIAOD /Charmonium/Run2018D-PromptReco-v2/MINIAOD /BuToJpsiK_BMuonFilter_SoftQCDnonD_TuneCP5_13TeV-pythia8-evtgen/ RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/MINIAODSIM

To retain enough events, we use specific triggers that require a muon with $p_T > 7.5$ GeV and a track with $p_T > 2$, 3.5, or 7 GeV, which together have an invariant mass compatible with a J/Ψ. Furthermore, events are selected with the following offline requirements:

- a muon and a tracker track (trk$_\mu$) with opposite charges, matched geometrically to the two legs of the trigger and with the same p_T thresholds, with an invariant mass compatible with a J/Ψ;
- a third track (trk$_K$, i.e.the kaon candidate) that makes a secondary vertex with the muon and the track above, with a χ^2 probability of the fit

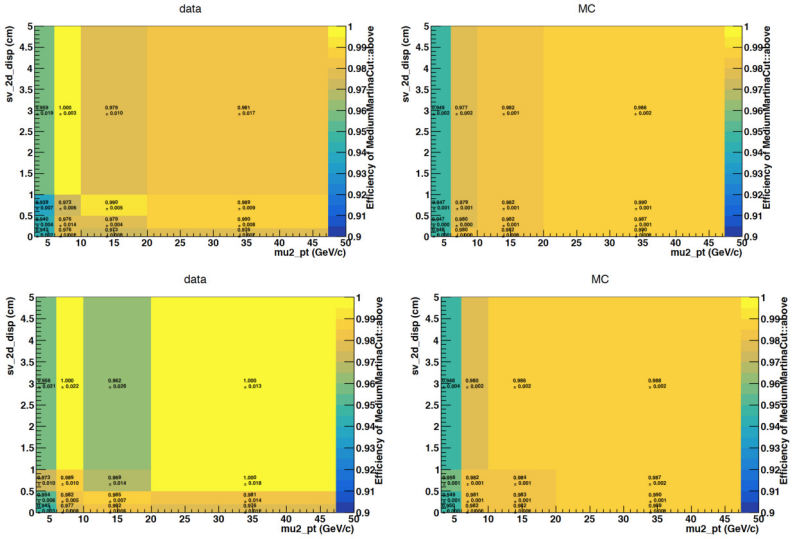


Figure B.4: Efficiency for displaced muon identification (as defined in Table 6.2), as a function of the muon p_T and the vertex transverse displacement, 2017 (top), and 2018 (bottom) data and simulations. It is measured in $B^{\pm} \to J/\Psi K^{\pm} \to \mu^{-}\mu^{+}K^{\pm}$ events.

$P(\chi^2_{\text{vtx}}) > 5 \times 10^{-3}$, a back-pointing angle $\cos\theta_{\text{SV,B}} > 0.99$, and a three-particle invariant mass compatible with a $B^{\pm}$ meson.

The muon is used as the tag and the trk$_{\mu}$ track as the probe. The latter can be used to measure the efficiency of the displaced tight muon identification (see Table 6.2), as a function of the muon p_T and the transverse displacement of the vertex. Figure B.4 shows the efficiency for the data and simulations. Figure B.5 shows the data-to-simulation efficiency scale factors for the three data sets. They are fairly close to unity (range 1-2%) and are thus not used as corrections, but rather, half of the difference between the SF and unity are taken as systematic uncertainties on the displaced muon efficiencies.

Electrons

In order to assess systematics due to the displaced electron identification efficiencies, we select asymmetric photon conversions in events ($Z \to \ell^{-}\ell^{+}\gamma \to \ell^{-}\ell^{+}e^{\pm}(e^{\mp})$,) where $(e^{\mp})$ represents a very-low-p_T electron that fails reconstruction and/or identification [170]. The potential systematics are assessed based on the displacement proxy variable based on the transverse impact pa-

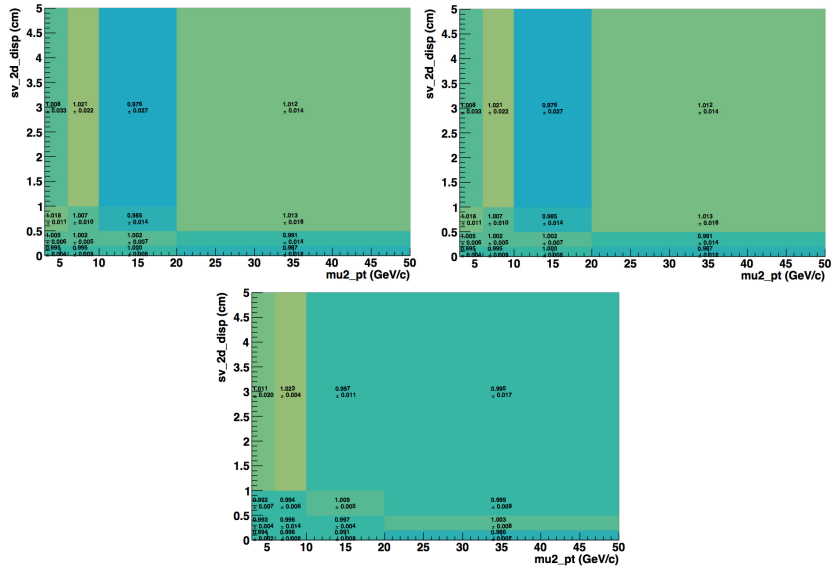


Figure B.5: Efficiency scale factors for displaced tight muon identification (as defined in Table 6.2), as a function of the muon p_T and the vertex transverse displacement, for 2016 (top), 2017 (middle), and 2018 (bottom) data and simulations. It is measured in $B^{\pm} \to J/\Psi K^{\pm} \to \mu^{-}\mu^{+}K^{\pm}$ events.

rameter of the electron track and the track curvature radius in the magnetic field. The systematics are taken as the envelope of all shape differences seen between data and simulation in the description of this variable. The assessed conservative value of this uncertainty is currently taken as 20%. The results are shown in Fig. B.6. The derived correction factors aim at displaced electron track reconstruction and also take partially into account the effect associated with electron ID efficiency differences in simulation with respect to data.

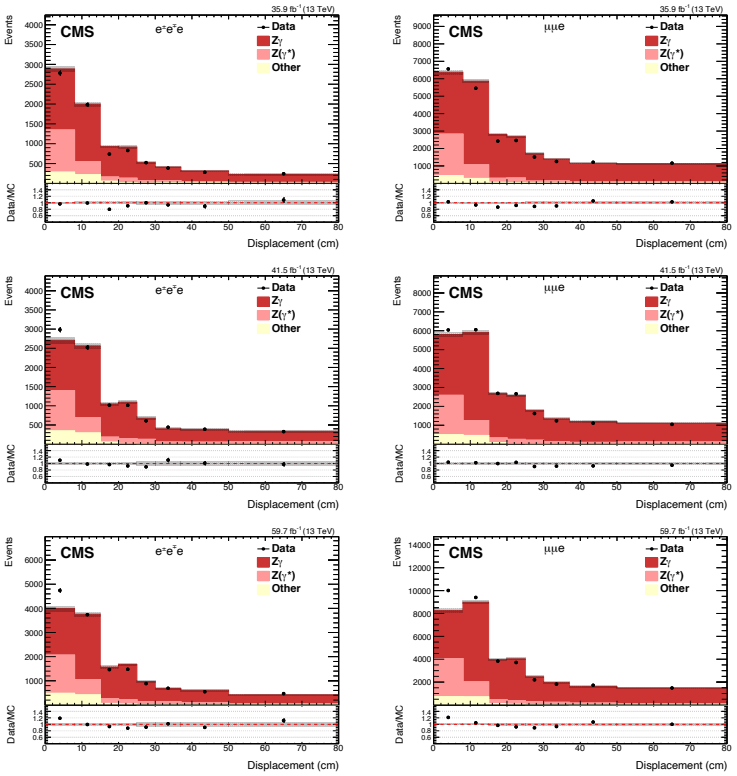


Figure B.6: Comparison between data and simulation for electron displacement in eee (left) and $\mu\mu e$ (right) channels. From top to bottom: 2016, 2017, and 2018.

B.1.8 Displaced track identification efficiency

To estimate the data-to-MC corrections for the displaced vertex and tracks we use $K_s^0 \rightarrow \pi^+\pi^-$ decays to two charged particles, giving a signature of two displaced tracks coming from a common vertex. The results of the method are presented in full detail in Ref. [171]. The outcome of the study is that the inefficiency for reconstruction derives almost entirely from the track reconstruction inefficiency.
For the displaced electron, we have tag and probe SF that cover its full systematics. For the displaced muon, we have tag and probe systematics that cover only its ID systematics. But for the muon tracking reconstruction systematics, we use the K_s^0 study. We calculate the single muon tracking efficiency as the square root of the full systematics that were calculated for a K_s^0 decay, because the calculated systematic factor represents 2 tracks in a vertex. Considering that the SV reconstruction efficiency is near to 100%, we can ignore the systematic uncertainty related to specifically reconstructing the SV. Furthermore, since in the HNL vertex usually we have more the 2 tracks in a vertex, we can assign a systematic uncertainties using the K_s^0 study where only 2 tracks are in the SV considering that the presence of more trakcs will only improve our reconstruction.
The systematics are parametrized as a function of track p_T and thus the choice is made to use the track with the highest p_T among all the tracks for evaluating the systematics. We assess the associated systematics both in the case of the displaced muon and of the displaced tracks by taking the uncertainty as 50% of the difference of the measured scale factor from 1.

B.1.9 Displaced momentum scale and resolution

The study of K_s^0 decays (Section B.1.8) also provides useful information about the momentum scale and resolution of the displaced tracks, from Gaussian fits to the K_s^0 mass profiles. Details of this study can be found in Ref. [170] (Section 9.1.6). Table B.3 reports the relative data-MC difference in momentum scale, while Table B.4 reports the data-MC difference (in quadrature) in resolution. In both cases, the difference is at the permil level, thus negligible.

Table B.3: Relative difference in momentum scale between data and simulation.

$2*\Delta_{2D}$[cm]	Data-MC scale difference [%]		
	2016	2017	2018
< 0.5	-0.092	-0.074	-0.088
0.5–1.5	-0.088	-0.110	-0.096
1.5–4.0	-0.086	-0.098	-0.104
> 4.0	-0.074	-0.090	-0.098

Table B.4: Difference in quadrature in momentum resolution between data and simulation.

$2*\Delta_{2D}$[cm]	Data-MC resolution difference [%]		
	2016	2017	2018
< 0.5	0.20	0.37	0.33
0.5–1.5	0.19	0.26	0.15
1.5–4.0	0.25	0.26	0.22
> 4.0	0.29	0.19	0.22

B.1.10 JEC and JER variations

The Jet Energy Corrections and Jet Energy Resolution variations have been propagated fully and are taken into account. The versions of the JECs used for each year and for data and MC are summarized in table Their systematic uncertainty is taken as the difference in signal region yields as a result of the up and down variations in the relevant instances. These variations lead to an effect of generally less than 1%. It is clear that they are not among the dominant uncertainties in the analysis.

B.1.11 Statistical uncertainty of MC samples

Since the MC signal samples are limited in size, their statistical uncertainty needs to be taken into account. The statistical uncertainty on the event yield in each search region bin corresponds to the sum in quadrature of the MC event weights.

B.1.12 Systematic uncertainty treatment for background

The background estimation has beend done with the ABCD method as documented in Sec 5.7. We use the closure in CR2 (see Sec 5.7) and a control

region defined by inverting the jet vet (the nominal signal region selection allows for a maximum of 1 cleaned jet see Sec 5.4.3) to determine a systematic uncertainty for the background prediction. These regions are kinematically closest to the signal region and therefore we consider them most important. In order to cover practically all deviations in the closure tests the following systematic uncertainties are deemed appropriate. As said in 5.7.1, we've determined a 30% uncertainty on opposite-sign channels and a 20% uncertainty on same-sign channels to be large enough. For the Low Mass search regions, we similarly apply a 30% uncertainty on opposite-sign channels with a displaced electron and 20% on opposite-sign channels with a displaced muon and on same-sign channels.

Appendix C

Appendix ttH

C.1 Trigger

The HLT paths used to record events are given in Table C.1. A combination of single-lepton triggers and triggers based on the presence of a lepton and a τ_h (also known as lepton+τ_h "cross-triggers") are used to record events in the channels containing one lepton. A combination of single-lepton and dilepton triggers are used to record events in the channels containing two leptons, where the inclusion of single-lepton triggers boosts the events acceptance thanks to the absence of a threshold on a second lepton. Similarly, a mix of single-lepton, dilepton, and trilepton triggers is used to record events in the channels containing at least three leptons. For some triggers, we use a combination of HLT paths with different p_T thresholds or a combination of paths with and without a d_z requirement applied to the leptons. In the case of the double muon triggers we further use a mix of HLT paths with and without a requirement on the invariant mass of the di-muon pair, namely $m_{\mu\mu} > 3.8\,\mathrm{GeV}$. The motivation for choosing such a mix is to use the trigger with highest efficiency whenever available and to employ triggers of lower efficiency whenever the most efficient ones are disabled or prescaled. Events recorded in any data-taking period as well as in simulated samples are selected in case they pass any of the HLT paths that are required for each category.

Single lepton triggers	HLT_Ele25_eta2p1_WPTight_Gsf (✓/–/–)	
	HLT_Ele27_WPTight_Gsf (✓/–/–)	
	HLT_Ele27_eta2p1_WPLoose_Gsf (✓/–/–)	
	HLT_Ele32_WPTight_Gsf (–/✓/✓)	
$2\ell ss + 0\tau_h$	HLT_Ele35_WPTight_Gsf (–/✓/–)	
$2\ell ss + 1\tau_h$	HLT_IsoMu22 (✓/–/–)	
	HLT_IsoTkMu22 (✓/–/–)	
	HLT_IsoMu22_eta2p1 (✓/–/–)	
$3\ell + 0\tau_h$	HLT_IsoTkMu22_eta2p1 (✓/–/–)	
	HLT_IsoMu24 (✓/✓/✓)	
	HLT_IsoTkMu24 (✓/–/–)	
	HLT_IsoMu27 (–/✓/✓)	
Lepton+τ_h cross−triggers	HLT_Ele24_eta2p1_WPLoose_Gsf_LooseIsoPFTau20 (✓/–/–)	
	HLT_Ele24_eta2p1_WPLoose_Gsf_LooseIsoPFTau20_SingleL1 (✓/–/–)	
$1\ell + 1\tau_h$	HLT_Ele24_eta2p1_WPLoose_Gsf_LooseIsoPFTau30 (✓/–/–)	
	HLT_Ele24_eta2p1_WPTight_Gsf_LooseChargedIsoPFTau30_eta2p1_CrossL1 (–/✓/✓)	
	HLT_Ele24_eta2p1_WPTight_Gsf_LooseChargedIsoPFTauHPS30_eta2p1_CrossL1 (–/–/✓)	
	HLT_IsoMu19_eta2p1_LooseIsoPFTau20_SingleL1 (✓/–/–)	
	HLT_IsoMu20_eta2p1_LooseChargedIsoPFTau27_eta2p1_CrossL1 (–/✓/✓)	
	HLT_IsoMu20_eta2p1_LooseChargedIsoPFTauHPS27_eta2p1_CrossL1 (–/–/✓)	
Double lepton triggers	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ (✓/–/–)	
	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL (–/✓/✓)	
$2\ell ss + 0\tau_h$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL (✓/–/–)	
$2\ell ss + 1\tau_h$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ (✓/–/–)	
	HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL (✓/–/–)	
	HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ (✓/–/–)	
$3\ell + 0\tau_h$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8 (–/✓/–)	
	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 (–/✓/✓)	
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL (✓/–/–)	
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ (✓/✓/✓)	
	HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL (✓/–/–)	
	HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL_DZ (✓/–/–)	
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ (✓/✓/–)	
	HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ (✓/✓/–)	
	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL (✓/✓/–)	
	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ (–/✓/–)	
Triple lepton triggers	HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL (✓/✓/✓)	
	HLT_TripleMu_12_10_5 (✓/✓/✓)	
$3\ell + 0\tau_h$	HLT_Mu8_DiEle12_CaloIdL_TrackIdL (✓/✓/✓)	
	HLT_DiMu9_Ele9_CaloIdL_TrackIdL (✓/–/–)	
	HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ (–/✓/✓)	

Table C.1: Triggers (2016/2017/2018) used to record events. A hyphen (−) indicates the trigger is not used in the corresponding data-taking year.

C.2 Control regions

The control regions described in this section are used to check the data to MC agreement for the variables that will be used as input in the DNN training (see Sec 6.8.2). These control region are the same control regions used in the ttH inclusive analysis [108].

C.2.1 3j control region

The 3j control region is built by selecting events in the 2lss0τ region that contain exactly three jets, that can be forward jets. This enlargement of the acceptance allows us to check the modeling of the variables in the forward region, that is relevant for the measurement of tHq production. Such a region is enriched in ttW production and events with misidentified leptons, mostly due to tt events.

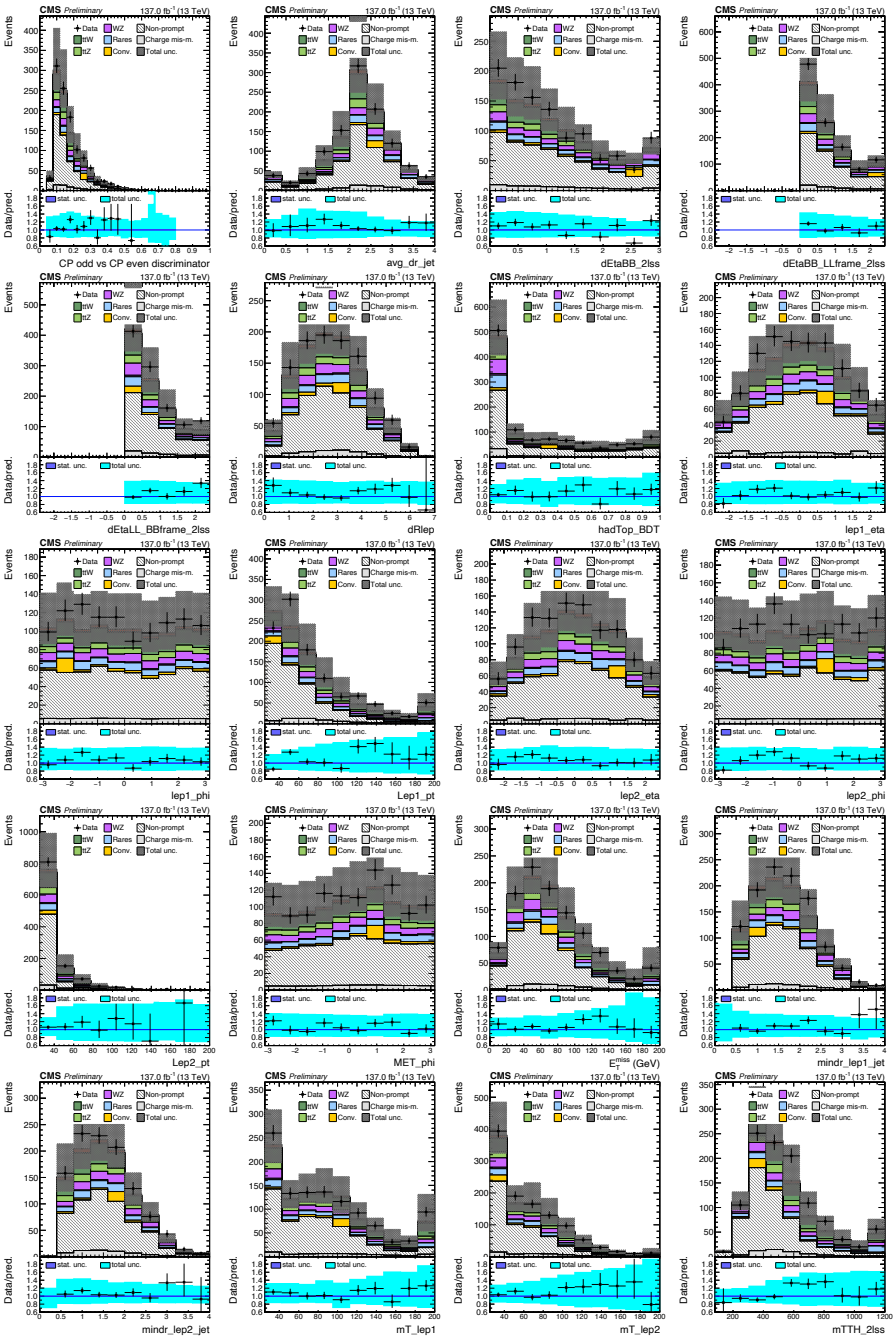


Figure C.1: Control region plots for the "3 jet" (3j) control region

C.2.2 ttZ control region

This control region is defined by reversing the Z veto in the 3l region and selecting only events with at least two central jets.

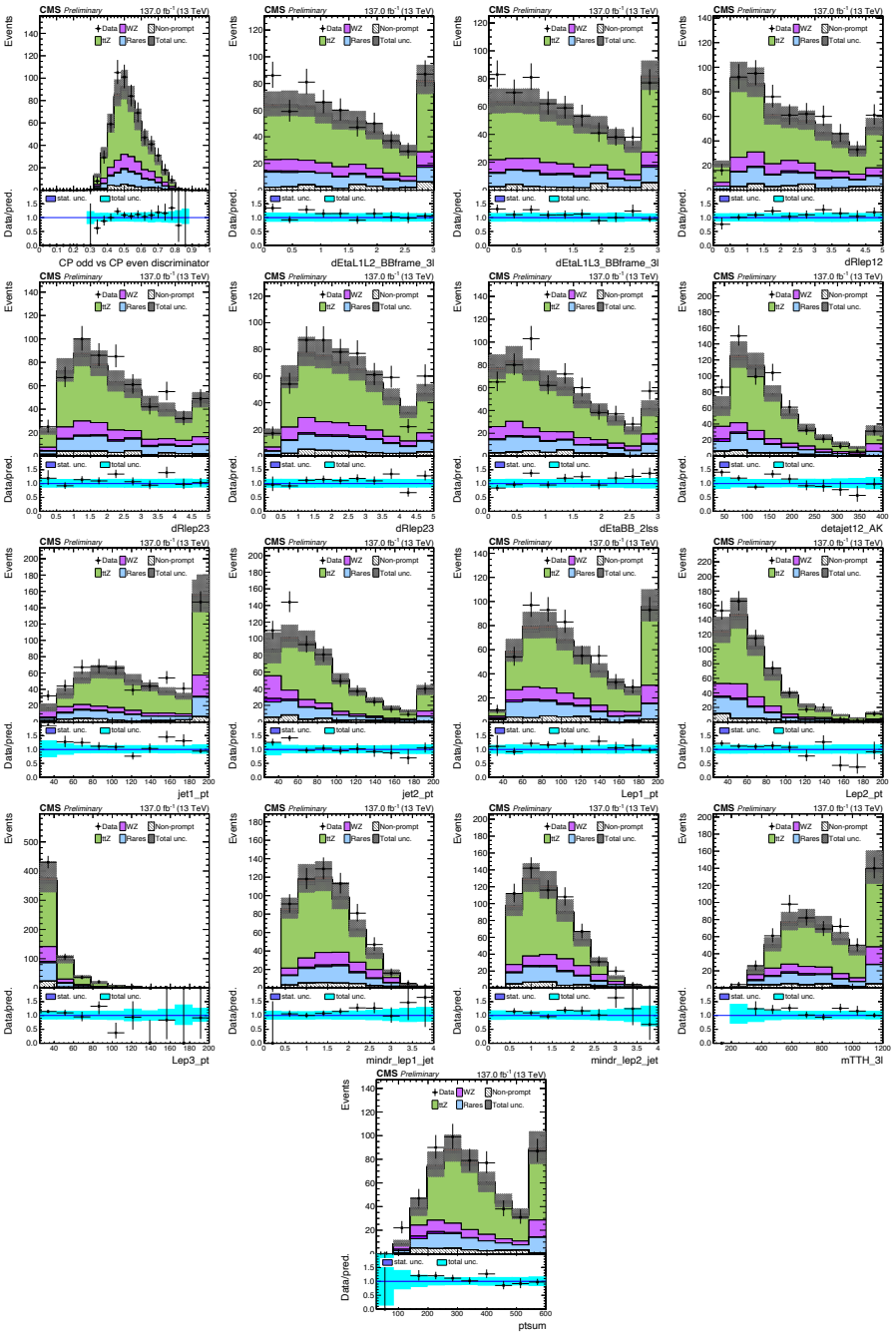


Figure C.2: Control region plots for the ttZ control region

C.2.3 WZ control region

The WZ control region region is constructed starting from the ttZ control region and inverting the b tag jet multiplicity requirement in that region. Such a region is enriched in WZ events with additional jets.

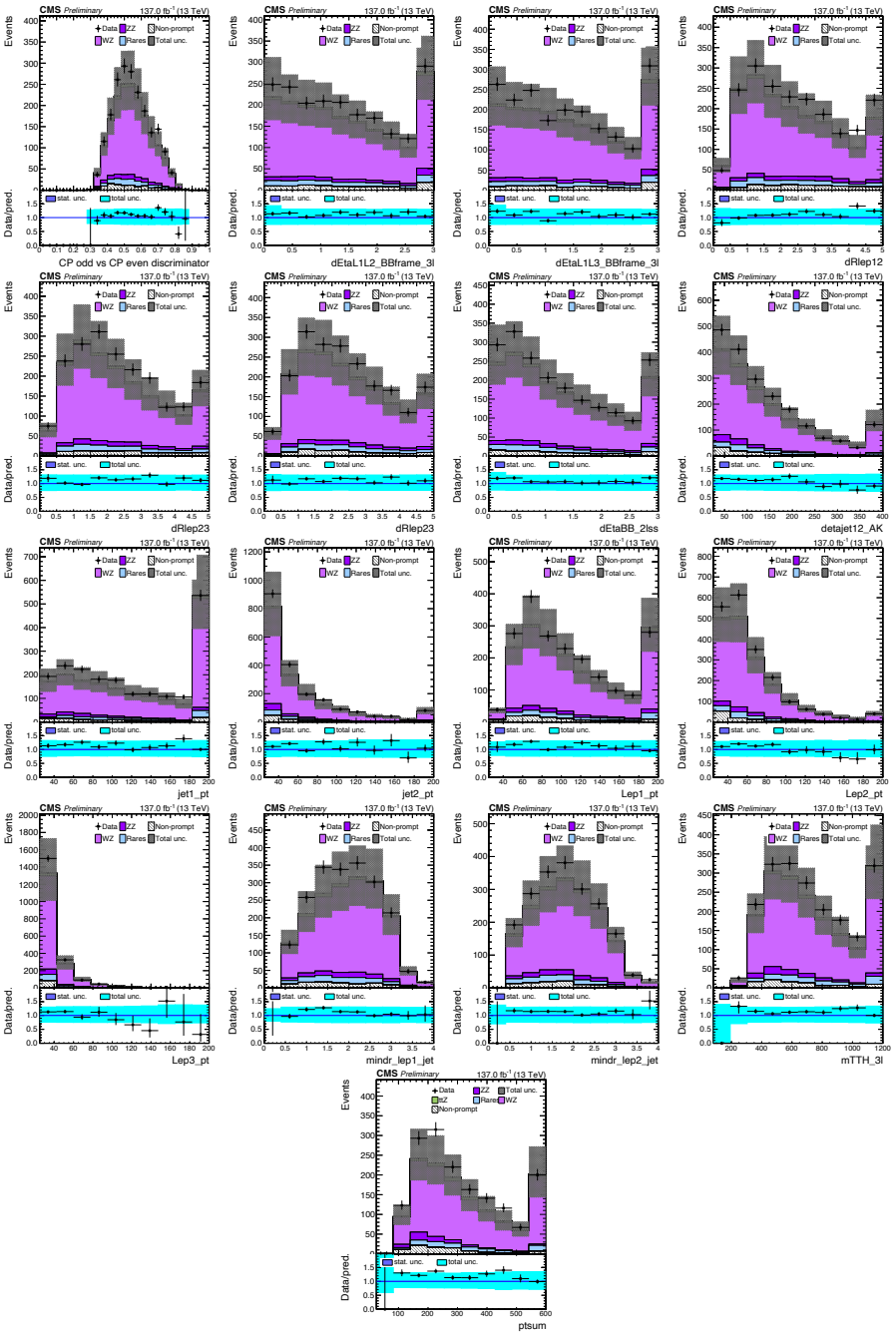


Figure C.3: Control region plots WZ for the control region

C.3 Systematic Uncertainties

Various imprecisely-known or simulated effects may alter the event yield of the ttH and tH signals (tH is considered a signal in the BSM analysis and a background in the SM analysis). It and of background processes, as well as the shape of the distributions in the discriminating observables that are used for the signal extraction. In this section these effects, usually called systematic uncertainties, are described and their impact on the yields in the various final states is discussed. The systematic uncertainties are modelled as nuisance parameters in the maximum-likelihood fits described in Sections 6.8. The nuisance parameters are allowed to change the event yield, accounting also for the migration of events among regions and among different bins in the distributions fitted in each region. We consider broadly two categories of nuisance parameters: those which purely affect the yield in a category (*rate uncertainties*) are assigned a log-normal probability density function, whereas those which affect also the shape of the distributions (*shape uncertainties*) are modelled via a polynomial interpolation with a Gaussian constraint and are also allowed to change the event yields in a category. The correlations between the various uncertainty sources across the three years of data-taking are detailed in the text and summarized in Table C.2.

- **Trigger efficiency**
  The impact on the final discriminants due to the trigger efficiency is estimated. In the categories used in this analysis purely leptonic triggers are used with uncertainties between 1 and 2%. For 2lss0τ and 2lss1τ categories, the uncertainty is taken as shape systematics and is correlated across categories. For the 3l category, the uncertainty accounts for 1% and is treated as correlated across categories. Trigger uncertainties are treated as uncorrelated across years for all categories.

- **Identification and isolation efficiency for e and μ**
  The impact on the final discriminants due to the identification and isolation efficiency of the leptons is treated separately:

  - The loose electron and muon identification criteria are affected by uncertainties which are estimate according to the CMS recommendations for muons and electrons. For muons a total shape uncertainty is assigned to the muon identification and isolation scale factors (see 4.3.4), while for the electrons the uncertainties correspond

to the electron identification and GSF tracking efficiency. In both cases, the uncertainties are derived as a function of the lepton p_T and η.

– The tight electron and muon identification criteria is estimated in closure tests performed separately for the three years and as a function of the lepton p_T and η. Uncertainties due to lepton selection efficiency determination uncertainty constitute one of the main sources of uncertainty in many analyses. These uncertainties can be factorized in two groups: the contribution arising from the measurement of the efficiency in Z events in data, and the contribution of the potential topology differences between the events used as reference (Z events) and signal events ($t\bar{t}$ events).

The total uncertainty in the lepton identification and isolation efficiency is taken as shape uncertainty and is correlated across years.

- **Identification efficiency for τ_h**
The uncertainty associated to the τ_h identification efficiency is estimated as function of the p_T and decay mode of the τ leptons (recommendations of CMS for Tau [172]). This uncertainty is dominated by statistical effects and is treated as uncorrelated across years, p_T bins and decay modes. The total uncertainty is taken as shape uncertainty and is treated as uncorrelated across years.

- **Energy scale of e, μ and τ_h**
The energy scales of electrons and muons are known with an uncertainty of less than 1% and are neglected in the analysis. The uncertainties associated to the energy scale of τ_h are obtained by varying the corresponding scale factor by its uncertainties which are lower than 1.1%. The total uncertainty in the τ_h energy scale is taken as shape uncertainty and is treated as uncorrelated across years.

- **Jet energy corrections**
The jet energy scale (JES) systematic uncertainties are evaluated by shifting the JES applied to the reconstructed jets up and down by one standard deviation, following the recommendations of the CMS collaboration [64]. The events are then re-analyzed, including the re-application of the jet-based selection and the computation of all relevant kinematic quantities, to derive the varied kinematic distributions, which enter as

shape uncertainties (including rate effects) in the final fit. This procedure results in a set of 11 JEC uncertainty sources. Some of them are correlated across the years and some are not, while all of them are treated as uncorrelated among themselves.

- **b-tag efficiency and mistag rate**
  Uncertainties in the b-tagging efficiencies and mistag rates as function of the jet p_T and η are evaluated by CMS collaboration [173]. The effect of these uncertainties on the yields is evaluated by varying the data-to-MC correction factors within their uncertainties and reanalyzing the events.

- MET **resolution and response**
  Uncertainties in the MET resolution and response are taken into account by varying the jet energy scale and resolution within their respective uncertainties and recomputing MET and all MET related observables after each variation. The uncertainty on the unclustered MET is taken into account and found to have a marginal impact in the analysis. The total uncertainty is taken as shape uncertainty and is treated as correlated across years.

- **Signal rate**
  The signal rate is measured in units of the SM ttH and tH production rates; the measurement is therefore affected by uncertainties in the ttH and tH cross sections. The uncertainties associated to these processes are the following:

  – The uncertainty in the SM ttH cross section, computed at NLO accuracy, amounts to $^{+6.8\%}_{-10.0\%}$, of which $^{+5.8\%}_{-9.3\%}$ are due to missing higher orders and 3.6% arises from uncertainties in the PDF and α_s [174].

  – The uncertainty in the SM tH cross section, computed at NLO accuracy, amounts to $^{+4.2\%}_{-6.8\%}$, of which $^{+4.1\%}_{-6.7\%}$ are due to missing higher orders and 1.0% arises from uncertainties in the PDF and α_s [174].

  – The uncertainty in the ITC tH cross section, computed at NLO accuracy, amounts to $^{+2.8\%}_{-4.0\%}$, of which $^{+2.7\%}_{-3.9\%}$ are due to missing higher orders and 0.9% arises from uncertainties in the PDF and α_s [174].

  – The uncertainty in the branching fraction for the Higgs boson to decay into WW ($\tau\tau$) amounts to 1.54% (1.65%).

  - Uncertainties in the acceptance that are due to missing higher orders are treated as shape systematics, as they affect the shape of the distribution in the observable used for signal extraction. They are estimated by varying the renormalization (μ_{R}) and factorization (μ_{F}) scales between 0.5 and 2 times their default values, with the constraint that $0.5 \leq \mu_{\mathrm{F}}/\mu_{\mathrm{R}} < 2$.

- **NLO vs. LO differences**
  ttH is simulated at LO in this analysis. Hence, to cover for discrepancies between NLO and LO simulation this uncertainty is considered. This uncertainty is constructed as the symmetrized ratio of yields with LO and NLO samples.

  All the theoretical uncertainties associated to the signal rate are correlated across years.

- **Background rates**
  The uncertainties associated to each of the relevant background contributions in this analysis are the listed below:

  - The contribution of the WZ+jets background is known with an uncertainty of 30%. The uncertainty represents the uncertainty in the extrapolation to the SR from the WZ+jets dominated control region. This control region is defined on the tHreeLeptonZeroTau category and is described in Section 6.4. The uncertainty includes the statistical uncertainty in the data yields in the control region; the uncertainty in the b-tag efficiency and mistag rate, and the uncertainty in the relative composition of light quark and gluon, c quark and b quark jets in WZ+jets events contributing to the control region and to the signal region.
  - A 30% normalization uncertainty is applied to the WZ+jets and ZZ+jets background processes which contain 3 or more jets, to account for the mismodelling observed in the MC in these regions.
  - The uncertainty in the background arising from production of top quark pairs in association with real or virtual photons amounts to 30%.
  - An uncertainty of 50% is assigned to other rare backgrounds which are modeled using MC simulation.

- No assumption is made on the rates of the irreducible ttW, $ttWW$, and ttZ backgrounds. The normalization of these backgrounds is determined simultaneously with the rate of the ttH and tH signals by the maximum-likelihood fit used for the signal extraction, described in Sections 6.8.
- The uncertainties on the fake background estimation are related to shape variations which arise from statistical uncertainties in the measurement and application regions, from the subtraction of the prompt lepton contamination in the measurement region and from the differences in the background composition between measurement region (dominated by multijet background) and application region (dominated by tt+jets background).
- The uncertainty on the yield of the charge flip background in the $2lss + 0\tau_h$ and $2lss + 1\tau_h$ categories amounts to 30%.

The uncertainties associated to the fake background estimation techniques are treated as uncorrelated across the years, while the ones associated to the charge flips and conversions are treated as correlated. The uncertainties associated to electroweak and rare processes are treated as correlated across the years.

- **Luminosity**
  The uncertainty in the integrated luminosity is split among correlated and uncorrelated effects across the years. The uncorrelated effects amount to 2.2%, 2.0% and 1.5% in 2016, 2017 and 2018, respectively [175]. The correlated effects amount to 1.4%, 1.3% and 2.1% in 2016, 2017 and 2018, respectively. These effects are applied as rate uncertainties.

- **Pileup**
  Uncertainties in the instantaneous luminosity and on the inelastic cross section may affect the event yield of the ttH and and tH signals and of backgrounds obtained from the MC simulation, as different PU conditions may modify the efficiency to pass the event selection criteria. We vary the product of instantaneous luminosity and inelastic cross section, the number of PU interactions, by 5%, reweight simulated events to the different PU conditions, and rerun the analysis. We find that the effect on the signal and of background yields amounts to less than 1%.

- **L1 ECAL prefiring**
  The effects due to the uncertainty in the L1 prefiring correction are taken

into account by varying the scale factors provided by the Level-1 group by their uncertainty. The uncertainty, which is taken into account only in 2016 and 2017 data-taking eras, is taken as shape uncertainty and is uncorrelated across years.

Uncertainty source	Type	Correlation 2016-2018
Trigger efficiency	Norm. / Shape	Uncorrelated
Identification and isolation efficiency for and	Shape	Correlated
Identification efficiency for τ_h	Shape	Uncorrelated
Energy scale of , and τ_h	Shape	Uncorrelated
Jet energy scale	Shape	Uncorr. / Corr.
b-tag efficiency and mistag rate	Shape	Uncorr. (stat.) / Corr. (exp.)
E_T resolution and response	Shape	Correlated
Signal rates	Norm. / Shape	Correlated
Fake background rate	Shape	Uncorrelated
Flip background rate	Norm.	Correlated
EWK and rare background rates	Norm.	Correlated
Luminosity	Norm.	Uncorr. / Corr.
Level-1 ECAL prefiring	Shape	Uncorrelated

Table C.2: Summary of the main uncertainty sources, their type and the correlations across the three data-taking years.

Bibliography

[1] Angela Lahee Nachtmann, Otto and W Wetzel. Elementary particle physics,concepts and phenomena. URL: http://lib.ugent.be/catalog/rug01:000295749.

[2] Xianhao Xin. Glashow-weinberg-salam model: An example of electroweak symmetry breaking. 2007.

[3] S. Chatrchyan, V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, and 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. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021, doi:10.1016/j.physletb.2012.08.021.

[4] CMS Collaboration. Measurements of production cross sections of the higgs boson in the four-lepton final state in proton–proton collisions at 13 tev. *Eur. Phys. J. C 81 (2021) 488*. doi:10.1140/epjc/s10052-021-09200-x.

[5] A.M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic, J. Erö, A. Escalante Del Valle, M. Flechl, R. Frühwirth, and et al. A measurement of the higgs boson mass in the diphoton decay channel. *Physics Letters B*, 805:135425, Jun 2020. URL: http://dx.doi.org/10.1016/j.physletb.2020.135425, doi:10.1016/j.physletb.2020.135425.

[6] CMS collaboration. Observation of higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121:121801, Sep 2018. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.121801, doi:10.1103/PhysRevLett.121.121801.

[7] A M Sirunyan, A Tumasyan, W Adam, T Bergauer, M Dragicevic, J Erö, A Escalante Del Valle, R Frühwirth, M Jeitler, N Krammer,

L Lechner, D Liko, I Mikulec, F M Pitters, N Rad, J Schieck, R Schöfbeck, M Spanring, S Templ, and W Waltenberger. Evidence for higgs boson decay to a pair of muons. *Journal of high energy physics : JHEP.*, 2021(1), 2021-01.

[8] LHC Higgs XSeciton WG. Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector. 2, 2017. doi:10.23731/CYRM-2017-002.

[9] Biagio Di Micco, Maxime Gouzevitch, Javier Mazzitelli, and Caterina Vernieri. Higgs boson potential at colliders: Status and perspectives. *Reviews in Physics*, 5:100045, 2020. URL: https://www.sciencedirect.com/science/article/pii/S2405428320300083, doi:https://doi.org/10.1016/j.revip.2020.100045.

[10] Combined Higgs boson production and decay measurements with up to 137 fb-1 of proton-proton collision data at sqrts = 13 TeV. Technical report, CERN, Geneva, 2020. URL: https://cds.cern.ch/record/2706103.

[11] Fabio Maltoni, Eleni Vryonidou, and Cen Zhang. Higgs production in association with a top-antitop pair in the Standard Model Effective Field Theory at NLO in QCD. *JHEP*, 10:123, 2016. arXiv:1607.05330, doi:10.1007/JHEP10(2016)123.

[12] Kim C. W. Giunti, C. *Fundamentals of neutrino physics and astrophysics*. 2007.

[13] P. Hut and K.A. Olive. A cosmological upper limit on the mass of heavy neutrinos. *Phys. Lett. B*, 87(1):144 – 146, 1979. doi:https://doi.org/10.1016/0370-2693(79)90039-X.

[14] K. A. Olive et al. Review of Particle Physics. *Chin. Phys. C*, 38:090001, 2014. See the review section on Neutrino Mass, Mixing, and Oscillations, and references therein. doi:10.1088/1674-1137/38/9/090001.

[15] Michael Gronau, Chung Ngoc Leung, and Jonathan L. Rosner. Extending Limits on Neutral Heavy Leptons. *Phys. Rev. D*, 29:2539, 1984. doi:10.1103/PhysRevD.29.2539.

[16] Frank F. Deppisch, P. S. Bhupal Dev, and Apostolos Pilaftsis. Neutrinos and Collider Physics. *New J. Phys.*, 17:075019, 2015. arXiv:1502.06541, doi:10.1088/1367-2630/17/7/075019.

[17] Albert M Sirunyan et al. Search for heavy Majorana neutrinos in same-sign dilepton channels in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 01:122, 2019. arXiv:1806.10905, doi:10.1007/JHEP01(2019)122.

[18] Yves Baconnier, Giorgio Brianti, P Lebrun, A G Mathewson, R Perin, and Yves Baconnier. *LHC: the Large Hadron Collider accelerator project*. CERN, Geneva, 1993. URL: https://cds.cern.ch/record/257706.

[19] CMS collaboration. The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment. *JINST*, 3:S08004. 361 p, 2008. Also published by CERN Geneva in 2010. URL: https://cds.cern.ch/record/1129810, doi:10.1088/1748-0221/3/08/S08004.

[20] A Tricomi. Upgrade of the CMS tracker. *Journal of Instrumentation*, 9(03):C03041–C03041, mar 2014. doi:10.1088/1748-0221/9/03/c03041.

[21] Simon van der Meer. Calibration of the effective beam height in the isr. Technical report, 1968.

[22] CMS Collaboration et al. Precision luminosity measurement in proton-proton collisions at $sqrt\{s\}$=13 tev in 2015 and 2016 at cms. *arXiv preprint arXiv:2104.01927*, 2021. doi:https://doi.org/10.1140/epjc/s10052-021-09538-2.

[23] O. Aberle, Béjar Alonso, et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2020. URL: https://cds.cern.ch/record/2749422, doi:10.23731/CYRM-2020-0010.

[24] Juliette Alimena et al. Searching for long-lived particles beyond the Standard Model at the Large Hadron Collider. *J. Phys. G*, 47(9):090501, 2020. arXiv:1903.04497, doi:10.1088/1361-6471/ab4574.

[25] The Phase-2 Upgrade of the CMS Tracker. Technical report, CERN, Geneva, Jun 2017. URL: https://cds.cern.ch/record/2272264, doi:10.17181/CERN.QZ28.FLHW.

[26] The Phase-2 Upgrade of the CMS Barrel Calorimeters. Technical report, CERN, Geneva, Sep 2017. This is the final version, approved by the LHCC. URL: https://cds.cern.ch/record/2283187.

[27] The Phase-2 Upgrade of the CMS Muon Detectors. Technical report, CERN, Geneva, Sep 2017. This is the final version, approved by the LHCC. URL: https://cds.cern.ch/record/2283189.

[28] Technical proposal for a MIP timing detector in the CMS experiment Phase 2 upgrade. Technical report, CERN, Geneva, Dec 2017. URL: https://cds.cern.ch/record/2296612, doi:10.17181/CERN.2RSJ.UE8W.

[29] The ATLAS collaboration. The Phase-2 Upgrade of the CMS L1 Trigger Interim Technical Design Report. Technical report, CERN, Geneva, Sep 2017. This is the CMS Interim TDR devoted to the upgrade of the CMS L1 trigger in view of the HL-LHC running, as approved by the LHCC. URL: https://cds.cern.ch/record/2283192, doi:10.17181/CERN.UUWZ.FMIS.

[30] A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, M. Flechl, and et al. Particle-flow reconstruction and global event description with the cms detector. *Journal of Instrumentation*, 12(10):P10003–P10003, Oct 2017. URL: http://dx.doi.org/10.1088/1748-0221/12/10/P10003, doi:10.1088/1748-0221/12/10/p10003.

[31] The CMS collaboration. The performance of the cms muon detector in proton-proton collisions at $\sqrt{s}$ = 7 tev at the lhc. *Journal of Instrumentation*, 8(11):P11002–P11002, Nov 2013. URL: http://dx.doi.org/10.1088/1748-0221/8/11/P11002, doi:10.1088/1748-0221/8/11/p11002.

[32] CMS collaboration. Performance of electron reconstruction and selection with the cms detector in proton-proton collisions at 8 tev. *Journal of Instrumentation*, 10(06):P06005–P06005, Jun

2015. URL: http://dx.doi.org/10.1088/1748-0221/10/06/P06005, doi:10.1088/1748-0221/10/06/p06005.

[33] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-ktjet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008. URL: http://dx.doi.org/10.1088/1126-6708/2008/04/063, doi:10.1088/1126-6708/2008/04/063.

[34] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12(02):P02014–P02014, Feb 2017. URL: http://dx.doi.org/10.1088/1748-0221/12/02/P02014, doi:10.1088/1748-0221/12/02/p02014.

[35] Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector. Technical report, CERN, Geneva, 2016. URL: https://cds.cern.ch/record/2205284.

[36] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30, 2012. arXiv:1207.7235, doi:10.1016/j.physletb.2012.08.021.

[37] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1, 2012. arXiv:1207.7214, doi:10.1016/j.physletb.2012.08.020.

[38] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321, 1964. doi:10.1103/PhysRevLett.13.321.

[39] Peter W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys. Lett.*, 12:132, 1964. doi:10.1016/0031-9163(64)91136-9.

[40] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508, 1964. doi:10.1103/PhysRevLett.13.508.

[41] G.S. Guralnik, C.R. Hagen, and T.W.B. Kibble. Global Conservation Laws and Massless Particles. *Phys. Rev. Lett.*, 13:585, 1964. doi:10.1103/PhysRevLett.13.585.

[42] Peter W. Higgs. Spontaneous Symmetry Breakdown without Massless Bosons. *Phys. Rev.*, 145:1156, 1966. doi:10.1103/PhysRev.145.1156.

[43] T.W.B. Kibble. Symmetry breaking in nonAbelian gauge theories. *Phys. Rev.*, 155:1554, 1967. doi:10.1103/PhysRev.155.1554.

[44] J. Baglio, A. Djouadi, R. Gröber, M. M. Mühlleitner, J. Quevillon, and M. Spira. The measurement of the Higgs self-coupling at the LHC: theoretical status. *JHEP*, 04:151, 2013. arXiv:1212.5581, doi:10.1007/JHEP04(2013)151.

[45] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, U. Schubert, and T. Zirke. Higgs Boson Pair Production in Gluon Fusion at Next-to-Leading Order with Full Top-Quark Mass Dependence. *Phys. Rev. Lett.*, 117(1):012001, 2016. [Erratum: Phys. Rev. Lett.117,no.7,079901(2016)]. arXiv:1604.06447, doi:10.1103/PhysRevLett.117.079901,10.1103/PhysRevLett.117.012001.

[46] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, and T. Zirke. Full top quark mass dependence in Higgs boson pair production at NLO. *JHEP*, 10:107, 2016. arXiv:1608.04798, doi:10.1007/JHEP10(2016)107.

[47] Daniel de Florian and Javier Mazzitelli. Higgs pair production at next-to-next-to-leading logarithmic accuracy at the LHC. *JHEP*, 09:053, 2015. arXiv:1505.07122, doi:10.1007/JHEP09(2015)053.

[48] Giuseppe Degrassi, Pier Paolo Giardino, and Ramona Gröber. On the two-loop virtual QCD corrections to Higgs boson pair production in the Standard Model. *Eur. Phys. J.*, C76(7):411, 2016. arXiv:1603.00385, doi:10.1140/epjc/s10052-016-4256-9.

[49] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015. arXiv:1410.8849, doi:10.1007/JHEP04(2015)040.

[50] S. Alioli, P. Nason, C. Oleari, and E. Re. NLO vector-boson production matched with shower in POWHEG. *JHEP*, 07:060, 2008. arXiv:0805.4802, doi:10.1088/1126-6708/2008/07/060.

[51] Paolo Nason. A new method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004. arXiv:hep-ph/0409146, doi:10.1088/1126-6708/2004/11/040.

[52] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *JHEP*, 11:070, 2007. arXiv:0709.2092, doi:10.1088/1126-6708/2007/11/070.

[53] E. Bagnaschi, G. Degrassi, P. Slavich, and A. Vicini. Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM. *JHEP*, 02:088, 2012. arXiv:1111.2854, doi:10.1007/JHEP02(2012)088.

[54] Paolo Nason and Carlo Oleari. NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG. *JHEP*, 02:037, 2010. arXiv:0911.5299, doi:10.1007/JHEP02(2010)037.

[55] Heribertus B. Hartanto, Barbara Jager, Laura Reina, and Doreen Wackeroth. Higgs boson production in association with top quarks in the POWHEG BOX. *Phys. Rev.*, D91(9):094003, 2015. arXiv:1501.04498, doi:10.1103/PhysRevD.91.094003.

[56] Gionata Luisoni, Paolo Nason, Carlo Oleari, and Francesco Tramontano. $HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO. *JHEP*, 10:083, 2013. arXiv:1306.2542, doi:10.1007/JHEP10(2013)083.

[57] Yanyan Gao, Andrei V. Gritsan, Zijin Guo, Kirill Melnikov, Markus Schulze, and Nhan V. Tran. Spin determination of single-produced resonances at hadron colliders. *Phys. Rev. D*, 81:075022, 2010. [Erratum: 10.1103/PhysRevD.81.079905]. arXiv:1001.3396, doi:10.1103/PhysRevD.81.075022.

[58] Paolo Nason and Giulia Zanderighi. W^+W^-, WZ and ZZ production in the POWHEG-BOX-V2. *Eur. Phys. J.*, C74(1):2702, 2014. arXiv:1311.1365, doi:10.1140/epjc/s10052-013-2702-5.

[59] John M. Campbell and R. K. Ellis. MCFM for the Tevatron and the LHC. *Nucl. Phys. Proc. Suppl.*, 205:10, 2010. arXiv:1007.3492, doi:10.1016/j.nuclphysbps.2010.08.011.

[60] John M. Campbell, R. Keith Ellis, and Ciaran Williams. Bounding the Higgs width at the LHC using full analytic results for $gg- > e^-e^+\mu^-\mu^+$. *JHEP*, 04:060, 2014. arXiv:1311.3589, doi:10.1007/JHEP04(2014)060.

[61] Vardan Khachatryan et al. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. *JINST*, 10:P06005, 2015. doi:10.1088/1748-0221/10/06/P06005.

[62] Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-HIG-18-001, CERN, Geneva, 2018. URL: https://cds.cern.ch/record/2621419.

[63] PF Jet Performances at High Level Trigger using Patatrack pixel tracks. Apr 2021. URL: https://cds.cern.ch/record/2765489.

[64] Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV. Sep 2021. URL: http://cds.cern.ch/record/2792322.

[65] A. M. Sirunyan et al. Identification of heavy-flavour jets with the CMS detector in collisions at 13 TeV. *Submitted to JINST*, 2017. arXiv:1712.07158.

[66] Massimiliano Grazzini, Stefan Kallweit, and Dirk Rathlev. ZZ production at the LHC: fiducial cross sections and distributions in NNLO QCD. *Phys. Lett.*, B750:407–410, 2015. arXiv:1507.06257, doi:10.1016/j.physletb.2015.09.055.

[67] John M. Campbell, R. Keith Ellis, and Ciaran Williams. Vector boson pair production at the LHC. *JHEP*, 07:018, 2011. arXiv:1105.0020, doi:10.1007/JHEP07(2011)018.

[68] Marco Bonvini, Fabrizio Caola, Stefano Forte, Kirill Melnikov, and Giovanni Ridolfi. Signal-background interference effects in $gg \rightarrow$

$H \to WW$ beyond leading order. *Phys. Rev. D*, 88:034032, 2013. arXiv:1304.3053, doi:10.1103/PhysRevD.88.034032.

[69] Stefano Catani and Massimiliano Grazzini. An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC. *Phys. Rev. Lett.*, 98:222002, 2007. arXiv:hep-ph/0703012, doi:10.1103/PhysRevLett.98.222002.

[70] Massimiliano Grazzini. NNLO predictions for the Higgs boson signal in the H —> WW —> lnu lnu and H —> ZZ —> 4l decay channels. *JHEP*, 02:043, 2008. arXiv:0801.3232, doi:10.1088/1126-6708/2008/02/043.

[71] Massimiliano Grazzini and Hayk Sargsyan. Heavy-quark mass effects in Higgs boson production at the LHC. *JHEP*, 09:129, 2013. arXiv:1306.4581, doi:10.1007/JHEP09(2013)129.

[72] J. Baglio, F. Campanario, S. Glaus, M. Mühlleitner, J. Ronca, and M. Spira. $gg \to HH$: Combined uncertainties. *Phys. Rev. D*, 103(5):056002, 2021. arXiv:2008.11626, doi:10.1103/PhysRevD.103.056002.

[73] LHC Higgs Cross Section Working Group, S. Heinemeyer, C. Mariotti, G. Passarino, and R. Tanaka (Eds.). Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. *CERN-2013-004*, CERN, Geneva, 2013. arXiv:1307.1347, doi:10.5170/CERN-2013-004.

[74] Georges Aad et al. Search for Higgs boson pair production in the $\gamma\gamma$ final state using pp collision data at $\sqrt{s} = 8$ TeV from the ATLAS detector. *Phys. Rev. Lett.*, 114:081802, 2015. arXiv:1406.5053, doi:10.1103/PhysRevLett.114.081802.

[75] Georges Aad et al. Search for Higgs boson pair production in the final state from pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 75:412, 2015. arXiv:1506.00285, doi:10.1140/epjc/s10052-015-3628-x.

[76] Georges Aad et al. Searches for Higgs boson pair production in the $HH \to \tau\tau, \gamma\gamma^*, \gamma\gamma$, channels with the ATLAS detector. *Phys. Rev. D*, 92:092004, 2015. arXiv:1509.04670, doi:10.1103/PhysRevD.92.092004.

[77] ATLAS Collaboration. Search for pair production of Higgs bosons in the final state using proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector. *Phys. Rev. D*, 94:052002, 2016. arXiv:1606.04782, doi:10.1103/PhysRevD.94.052002.

[78] Vardan Khachatryan et al. Search for two Higgs bosons in final states containing two photons and two bottom quarks in proton-proton collisions at 8 TeV. *Phys. Rev. D*, 94:052012, 2016. arXiv:1603.06896, doi:10.1103/PhysRevD.94.052012.

[79] Vardan Khachatryan et al. Search for resonant pair production of Higgs bosons decaying to two bottom quark-antiquark pairs in proton-proton collisions at 8 TeV. *Phys. Lett. B*, 749:560, 2015. arXiv:1503.04114, doi:10.1016/j.physletb.2015.08.047.

[80] CMS Collaboration. Search for a massive resonance decaying to a pair of Higgs bosons in the four b quark final state in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. *Phys. Lett. B*, 781:244, 2018. arXiv:1710.04960, doi:10.1016/j.physletb.2018.03.084.

[81] CMS Collaboration. Search for higgs boson pair production in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$. *Phys. Rev. D*, 96:072004, 2017. arXiv:1707.00350, doi:10.1103/PhysRevD.96.072004.

[82] Albert M Sirunyan et al. Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton–proton collisions at $\sqrt{s}$ = 13 TeV. *Phys. Lett. B*, 778:101, 2018. arXiv:1707.02909, doi:10.1016/j.physletb.2018.01.001.

[83] CMS Collaboration. Search for resonant and nonresonant Higgs boson pair production in the $\ell\nu\ell\nu$ final state in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. *JHEP*, 01:054, 2018. arXiv:1708.04188, doi:10.1007/JHEP01(2018)054.

[84] Combined Higgs boson production and decay measurements with up to 137 of proton-proton collision data at $\sqrt{s}$ = 13 TeV. CMS Physics Analysis Summary CMS-PAS-HIG-19-005, 2020. URL: http://cds.cern.ch/record/2706103.

[85] Georges Aad et al. Combination of searches for Higgs boson pairs in collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett.*

B, 800:135103, 2020. arXiv:1906.02025, doi:10.1016/j.physletb.2019.135103.

[86] Albert M Sirunyan et al. Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 122:121803, 2019. arXiv:1811.09689, doi:10.1103/PhysRevLett.122.121803.

[87] Prospects for HH measurements at the HL-LHC. Technical report, CERN, Geneva, 2018. URL: http://cds.cern.ch/record/2652549.

[88] Search for long-lived heavy neutral leptons with displaced vertices in pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ with the CMS detector. 2021.

[89] P. Abreu et al. Search for neutral heavy leptons produced in Z decays. *Z. Phys. C*, 74:57–71, 1997. [Erratum: Z.Phys.C 75, 580 (1997)]. doi:10.1007/s002880050370.

[90] Albert M Sirunyan et al. Search for heavy Majorana neutrinos in same-sign dilepton channels in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 01:122, 2019. arXiv:1806.10905, doi:10.1007/JHEP01(2019)122.

[91] A. M. Sirunyan, Tumasyan, et al. Search for heavy neutral leptons in events with three charged leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 120:221801, May 2018. URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.221801, doi:10.1103/PhysRevLett.120.221801.

[92] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. arXiv:1405.0301, doi:10.1007/JHEP07(2014)079.

[93] Anupama Atre, Tao Han, Silvia Pascoli, and Bin Zhang. The Search for Heavy Majorana Neutrinos. *JHEP*, 05:030, 2009. arXiv:0901.3589, doi:10.1088/1126-6708/2009/05/030.

[94] Daniel Alva, Tao Han, and Richard Ruiz. Heavy majorana neutrinos from w fusion at hadron colliders. *Journal of High Energy Physics*, 2015(2), feb 2015. URL: https://doi.org/10.1007%2Fjhep02%282015%29072, doi:10.1007/jhep02(2015)072.

[95] Céline Degrande, Olivier Mattelaer, Richard Ruiz, and Jessica Turner. Fully automated precision predictions for heavy neutrino production mechanisms at hadron colliders. *Physical Review D*, 94(5), sep 2016. URL: https://doi.org/10.1103%2Fphysrevd.94.053002, doi:10.1103/physrevd.94.053002.

[96]

[97] Vardan Khachatryan et al. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. *JINST*, 10:P06005, 2015. arXiv:1502.02701, doi:10.1088/1748-0221/10/06/P06005.

[98] Serguei Chatrchyan et al. Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV. *JINST*, 7:P10002, 2012. arXiv:1206.4071, doi:10.1088/1748-0221/7/10/P10002.

[99] A. M. Sirunyan et al. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV. *JINST*, 13(06):P06015, 2018. arXiv:1804.04528, doi:10.1088/1748-0221/13/06/P06015.

[100] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. arXiv:0802.1189, doi:10.1088/1126-6708/2008/04/063.

[101] Wolfgang Waltenberger. Adaptive Vertex Reconstruction. Technical report, CERN, Geneva, Jul 2008. URL: https://cds.cern.ch/record/1166320.

[102] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 121, 2019. arXiv:1810.05165, doi:10.1007/JHEP01(2019)121.

[103] M. Zaheer, S. Kottur, S. Ravanbakhsh, Póczos B., Salakhutdinov R. R., and A. J. Smola. Deep sets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 3394–3404, december 2017. arXiv:1703.06114.

[104] The LHC Higgs Combination Group Collaboration. Procedure for the LHC Higgs boson search combination in Summer 2011, 2012. CMS NOTE-2011/005.

[105] G Cowan, K Cranmer, E Gross, and O Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J.*, C71:1554, 2011. arXiv:1007.1727, doi:10.1140/epjc/s10052-011-1554-0.

[106] Marco Drewes and Jan Hajer. Heavy Neutrinos in displaced vertex searches at the LHC and HL-LHC. *JHEP*, 02:070, 2020. arXiv:1903.06100, doi:10.1007/JHEP02(2020)070.

[107] Alain Blondel, E. Graverini, N. Serra, and M. Shaposhnikov. Search for Heavy Right Handed Neutrinos at the FCC-ee. *Nucl. Part. Phys. Proc.*, 273-275:1883–1890, 2016. arXiv:1411.5230, doi:10.1016/j.nuclphysbps.2015.09.304.

[108] Albert M Sirunyan et al. Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at 13 TeV. 11 2020. arXiv:2011.03652.

[109] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B 716:1, 2012. arXiv:1207.7214, doi:10.1016/j.physletb.2012.08.020.

[110] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett.*, B 716:30, 2012. arXiv:1207.7235, doi:10.1016/j.physletb.2012.08.021.

[111] CMS D0 ATLAS, CDF. First combination of tevatron and lhc measurements of the top-quark mass. 2014. arXiv:1403.4427.

[112] Bogdan A. Dobrescu and Christopher T. Hill. Electroweak symmetry breaking via top condensation seesaw. *Phys. Rev. Lett.*, 81:2634–2637, 1998. arXiv:hep-ph/9712319, doi:10.1103/PhysRevLett.81.2634.

[113] R. Sekhar Chivukula, Bogdan A. Dobrescu, Howard Georgi, and Christopher T. Hill. Top Quark Seesaw Theory of Electroweak Symmetry Breaking. *Phys. Rev.*, D59:075003, 1999. arXiv:hep-ph/9809470, doi:10.1103/PhysRevD.59.075003.

[114] D. Delepine, J. M. Gerard, and R. Gonzalez Felipe. Is the standard Higgs scalar elementary? *Phys. Lett.*, B372:271–274, 1996. arXiv:hep-ph/9512339, doi:10.1016/0370-2693(96)00048-2.

[115] Albert M Sirunyan, Armen Tumasyan, Adam, et al. Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with leptonic $\mathrm{t\bar{t}}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 03:026. 59 p, Apr 2018. URL: https://cds.cern.ch/record/2312382, arXiv:1804.03682, doi:10.1007/JHEP03(2019)026.

[116] Albert M Sirunyan et al. Measurements of $\mathrm{t\bar{t}}H$ Production and the CP Structure of the Yukawa Interaction between the Higgs Boson and Top Quark in the Diphoton Decay Channel. *Phys. Rev. Lett.*, 125(6):061801, 2020. arXiv:2003.10866, doi:10.1103/PhysRevLett.125.061801.

[117] Higgs boson production in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV. Technical report, CERN, Geneva, 2020. URL: http://cds.cern.ch/record/2725523.

[118] K. Kondo. Dynamical likelihood method for reconstruction of events with missing momentum. 1: method and toy models. *J. Phys. Soc. Jap.*, 57:4126–4140, 1988. doi:10.1143/JPSJ.57.4126.

[119] K. Kondo. Dynamical likelihood method for reconstruction of events with missing momentum. 2: mass spectra for $2 \to 2$ processes. *J. Phys. Soc. Jap.*, 60:836–844, 1991. doi:10.1143/JPSJ.60.836.

[120] Search for CP violation in $\mathrm{t\bar{t}H}$ and tH production in multilepton channels at $\sqrt{s} = 13$ TeV. Technical report, CERN, Geneva, 2022. URL: http://cds.cern.ch/record/2803420.

[121] Massironi, A. and Boldrini, G. EFT model for SMP measurements. *CMS Analysis Note*, 2020/204, 2020.

[122] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. arXiv:1405.0301, doi:10.1007/JHEP07(2014)079.

[123] Paolo Nason. A new method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004. arXiv:hep-ph/0409146, doi:10.1088/1126-6708/2004/11/040.

[124] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *JHEP*, 11:070, 2007. arXiv:0709.2092, doi:10.1088/1126-6708/2007/11/070.

[125] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010. arXiv:1002.2581, doi:10.1007/JHEP06(2010)043.

[126] Peter Skands, Stefano Carrazza, and Juan Rojo. Tuning PYTHIA 8.1: the Monash 2013 tune. *Eur. Phys. J.*, C 74(8):3024, 2014. arXiv:1404.5630, doi:10.1140/epjc/s10052-014-3024-y.

[127] R. D. Ball, V. Bertone, S. Carrazza, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, and J. Rojo. Parton distributions with QED corrections. *Nucl. Phys.*, B 877:290, 2013. arXiv:1308.0598, doi:10.1016/j.nuclphysb.2013.10.010.

[128] R. D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015. arXiv:1410.8849, doi:10.1007/JHEP04(2015)040.

[129] Richard D. Ball et al. Parton distributions from high-precision collider data. *Eur. Phys. J.*, C77(10):663, 2017. arXiv:1706.00428, doi:10.1140/epjc/s10052-017-5199-5.

[130] Kirill Melnikov and Frank Petriello. Electroweak gauge boson production at hadron colliders through O(α_s^2). *Phys. Rev.*, D 74:114017, 2006. arXiv:hep-ph/0609070, doi:10.1103/PhysRevD.74.114017.

[131] Michal Czakon and Alexander Mitov. Top++: A program for the calculation of the top-pair cross section at hadron colliders. *Comput. Phys. Commun.*, 185:2930, 2014. arXiv:1112.5675, doi:10.1016/j.cpc.2014.06.021.

[132] T. Seva et al. Measurement of the differential cross section of bosons produced in association with jets in collisions at $\sqrt{s}$ = 8 TeV. *CMS Analysis Note*, 2013/049, 2013.

[133] Kant, P. and Kind, O. M. and Kintscher, T. and Lohse, T. and Martini, T. and Mölbitz, S. and Rieck, P. and Uwer, P. HATHOR for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions. *Comput. Phys. Commun.*, 191:74, 2015. arXiv:1406.4403, doi:10.1016/j.cpc.2015.02.001.

[134] M. Aliev, H. Lacker, U. Langenfeld, S. Moch, P. Uwer, and M. Wiedermann. HATHOR: HAdronic Top and Heavy quarks crOss section calculatoR. *Comput. Phys. Commun.*, 182:1034–1046, 2011. arXiv:1007.1327, doi:10.1016/j.cpc.2010.12.040.

[135] Nikolaos Kidonakis. Two-loop soft anomalous dimensions for single top quark associated production with a $^-$ or H^-. *Phys. Rev.*, D 82:054018, 2010. arXiv:1005.4451, doi:10.1103/PhysRevD.82.054018.

[136] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016. arXiv:1610.07922, doi:10.23731/CYRM-2017-002.

[137] Rikkert Frederix and Ioannis Tsinikos. Subleading ew corrections and spin-correlation effects in $t\bar{t}w$ multi-lepton signatures, 2020. arXiv:2004.09552.

[138] S. Agostinelli et al. GEANT4—a simulation toolkit. *Nucl. Instrum. Meth.*, A 506:250, 2003. doi:10.1016/S0168-9002(03)01368-8.

[139] CMS collaboration. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{}$is/i=13 TeV, jun 2018. doi:10.1088/1748-0221/13/06/p06015.

[140] Performance of reconstruction and identification of tau leptons in their decays to hadrons and tau neutrino in LHC Run-2. Technical report, CERN, Geneva, 2016. URL: https://cds.cern.ch/record/2196972.

[141] Performance of the DeepTau algorithm for the discrimination of taus against jets, electron, and muons. Oct 2019. URL: http://cds.cern.ch/record/2694158.

[142] Serguei Chatrchyan et al. Determination of jet energy calibration and transverse momentum resolution in CMS. *JINST*, 6:P11002, 2011. arXiv:1107.4277, doi:10.1088/1748-0221/6/11/P11002.

[143] Serguei Chatrchyan et al. Identification of quark jets with the CMS experiment. *JINST*, 8:P04013, 2013. arXiv:1211.4462, doi:10.1088/1748-0221/8/04/P04013.

[144] K. A. Olive et al. Review of Particle Physics. *Chin. Phys. C*, 38:090001, 2014. See the review section on Neutrino Mass, Mixing, and Oscillations, and references therein. doi:10.1088/1674-1137/38/9/090001.

[145] A. M. Sirunyan et al. Measurements of properties of the Higgs boson decaying into the four-lepton final state in collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:047, 2017. arXiv:1706.09936, doi:10.1007/JHEP11(2017)047.

[146] Rikkert Frederix and Ioannis Tsinikos. Subleading EW corrections and spin-correlation effects in $t\bar{t}W$ multi-lepton signatures. *Eur. Phys. J. C*, 80:803, 2020. arXiv:2004.09552, doi:10.1140/epjc/s10052-020-8388-6.

[147] Jeff Asaf Dror, Marco Farina, Ennio Salvioni, and Javi Serra. Strong tW Scattering at the LHC. *JHEP*, 01:071, 2016. arXiv:1511.03674, doi:10.1007/JHEP01(2016)071.

[148] CMS collaboration. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{}$is/i=13 TeV. *Journal of Instrumentation*, 13(06):P06015–P06015, jun 2018. doi:10.1088/1748-0221/13/06/p06015.

[149] Fabio Maltoni, Davide Pagani, Ambresh Shivaji, and Xiaoran Zhao. Trilinear Higgs coupling determination via single-Higgs differential measurements at the LHC. *Eur. Phys. J.*, C77(12):887, 2017. arXiv:1709.08649, doi:10.1140/epjc/s10052-017-5410-8.

[150] François Chollet et al. Keras, 2015. URL: https://keras.io.

[151] Martín Abadi et al. TensorFlow: large-scale machine learning on heterogeneous systems, 2015. URL: https://www.tensorflow.org.

[152] W. Beenakker, S. Dittmaier, M. Kramer, B. Plumper, M. Spira, and P. M. Zerwas. Higgs radiation off top quarks at the Tevatron and the LHC. *Phys. Rev. Lett.*, 87:201805, 2001. arXiv:hep-ph/0107081, doi:10.1103/PhysRevLett.87.201805.

[153] W. Beenakker, S. Dittmaier, M. Kramer, B. Plumper, M. Spira, and P. M. Zerwas. NLO QCD corrections to t anti-t H production in hadron collisions. *Nucl. Phys. B*, 653:151–203, 2003. arXiv:hep-ph/0211352, doi:10.1016/S0550-3213(03)00044-0.

[154] L. Reina and S. Dawson. Next-to-leading order results for t anti-t h production at the Tevatron. *Phys. Rev. Lett.*, 87:201804, 2001. arXiv:hep-ph/0107101, doi:10.1103/PhysRevLett.87.201804.

[155] L. Reina, S. Dawson, and D. Wackeroth. QCD corrections to associated t anti-t h production at the Tevatron. *Phys. Rev. D*, 65:053017, 2002. arXiv:hep-ph/0109066, doi:10.1103/PhysRevD.65.053017.

[156] S. Dawson, L. H. Orr, L. Reina, and D. Wackeroth. Associated top quark Higgs boson production at the LHC. *Phys. Rev. D*, 67:071503, 2003. arXiv:hep-ph/0211438, doi:10.1103/PhysRevD.67.071503.

[157] S. Dawson, C. Jackson, L. H. Orr, L. Reina, and D. Wackeroth. Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections. *Phys. Rev. D*, 68:034022, 2003. arXiv:hep-ph/0305087, doi:10.1103/PhysRevD.68.034022.

[158] Rikkert Frederix, Stefano Frixione, Valentin Hirschi, Fabio Maltoni, Roberto Pittau, and Paolo Torrielli. Scalar and pseudoscalar Higgs production in association with a top–antitop pair. *Phys. Lett. B*, 701:427–433, 2011. arXiv:1104.5613, doi:10.1016/j.physletb.2011.06.012.

[159] M. V. Garzelli, A. Kardos, C. G. Papadopoulos, and Z. Trocsanyi. Standard Model Higgs boson production in association with a top anti-top pair at NLO with parton showering. *EPL*, 96(1):11001, 2011. arXiv:1108.0387, doi:10.1209/0295-5075/96/11001.

[160] Steven Weinberg. Phenomenological Lagrangians. *Physica A*, 96(1-2):327–340, 1979. doi:10.1016/0378-4371(79)90223-1.

[161] W. Buchmuller and D. Wyler. Effective Lagrangian Analysis of New Interactions and Flavor Conservation. *Nucl. Phys. B*, 268:621–653, 1986. doi:10.1016/0550-3213(86)90262-2.

[162] Chung Ngoc Leung, S. T. Love, and S. Rao. Low-Energy Manifestations of a New Interaction Scale: Operator Analysis. *Z. Phys. C*, 31:433, 1986. doi:10.1007/BF01588041.

[163] Andrew Kobach. Baryon Number, Lepton Number, and Operator Dimension in the Standard Model. *Phys. Lett. B*, 758:455–457, 2016. arXiv:1604.05726, doi:10.1016/j.physletb.2016.05.050.

[164] D. Barducci et al. Interpreting top-quark LHC measurements in the standard-model effective field theory. 2 2018. arXiv:1802.07237.

[165] Nathan P. Hartland, Fabio Maltoni, Emanuele R. Nocera, Juan Rojo, Emma Slade, Eleni Vryonidou, and Cen Zhang. A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector. *JHEP*, 04:100, 2019. arXiv:1901.05965, doi:10.1007/JHEP04(2019)100.

[166] Andrea Dainese, Michelangelo Mangano, Andreas B Meyer, Aleandro Nisati, Gavin Salam, and Mika Anton Vesterinen. Report on the Physics at the HL-LHC, and Perspectives for the HE-LHC. Technical report, Geneva, Switzerland, 2019. URL: https://cds.cern.ch/record/2703572, doi:10.23731/CYRM-2019-007.

[167] A. Sirunyan et al. Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS. *Eur. Phys. J. C*, 81:800, 2021. arXiv:2104.01927, doi:10.1140/epjc/s10052-021-09538-2.

[168] CMS Collaboration. CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-17-004, 2018. URL: http://cds.cern.ch/record/2621960.

[169] CMS Collaboration. CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-18-002, 2019. URL: http://cds.cern.ch/record/2676164.

[170] CMS Collaboration. Search for long-lived heavy neutral leptons with displaced vertices in proton-proton collisions at $\sqrt{s}$ =13 tev. 2022. URL: https://arxiv.org/abs/2201.05578, doi:10.48550/ARXIV.2201.05578.

[171] Luka Lambrecht et. al,. Study of displaced vertex reconstruction using light neutral hadrons in Drell-Yan events. *CMS Physics Analysis Note*, AN-2020/111, 2020. URL: http://cms.cern.ch/iCMS/jsp/openfile.jsp?tp=draft&files=AN2020_111_v2.pdf.

[172] CMS Collaboration. Performance of reconstruction and identification of leptons in their decays to hadrons and in LHC Run 2. CMS Physics Analysis Summary CMS-PAS-TAU-16-002, 2016. URL: http://cds.cern.ch/record/2196972.

[173] CMS Collaboration. Performance of b-tagging algorithms at 13 tev. *CMS AN*, 2016/036, 2016.

[174] S Heinemeyer, Mariotti, et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Sec-

tion Working Group. Technical Report arXiv:1307.1347. CERN-2013-004, Geneva, 2013. Comments: 404 pages, 139 figures, to be submitted to CERN Report. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections. URL: http://cds.cern.ch/record/1559921.

[175] CMS Collaboration. CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-17-004, 2017. URL: http://cds.cern.ch/record/2621960.