"Search for resonant di-Higgs production in CMS and development of a model independent approach to look for new physics at the LHC"

François, Brieuc

Abstract

Despite the numerous experimental confirmations of the Standard Model (SM), this framework is believed not to be the complete theory of fundamental interactions. Many scenarios of physics beyond the SM (BSM) are postulated to solve some of the SM deficiencies. In this thesis we perform searches for new physics exploiting 13 TeV proton-proton collisions provided by the Large Hadron Collider and recorded by the Compact Muon Solenoid detector. We present the very first search for resonant Higgs pair production, X → hh, where one of the h decays into a b-quark pair and the other into two leptons plus two neutrinos via vector bosons. This process is predicted in several BSM scenarios such as Warped Extra Dimension or Two Higgs Doublet Model. The observations are found to be compatible with the SM expectations and limits are set on the signal production cross section for resonance masses from mX = 260 GeV to 900 GeV, for both spin-0 and spin-2 particles. The second part of the thesis dis...

Document type : *Thèse (Dissertation)*

Université catholique de Louvain
Secteur des Sciences et Technologies
Institut de Recherche en Mathtématique et Physique
Centre for Cosmology, Particle Physics and Phenomenology

Search for resonant di-Higgs production in CMS and development of a model independent approach to look for new physics at the LHC

Doctoral dissertation presented by

Brieuc FRANÇOIS

in fulfillment of the requirement for the degree of Doctor in Sciences

Thesis committee

Pr. Vincent LEMAÎTRE (*Advisor*)	UCL, Belgium
Pr. Jean-Marc GÉRARD (*Chairman*)	UCL, Belgium
Dr. Caroline COLLARD	IPHC, France
Pr. Laurent FAVART	ULB, Belgium
Dr. Andrea GIAMMANCO	UCL, Belgium

Remerciements

Les recherches menées à bien durant ces quatre années de thèse et résumées dans ce manuscript n'auraient pu voir le jour sans l'aide de mes collègues ni sans le soutient de mon entourage. Je souhaiterais, au travers de ces quelques lignes, les remercier chaleureusement pour avoir rendu ce travail possible.

Je remercie très sincèrement Vincent, mon promoteur, pour ces cinq années de collaborations fructueuses. Depuis le début de mon mémoire, et jusqu'à la fin de ma thèse, son enthousiasme contagieux fût le moteur qui m'a permis d'avancer dans mes recherches. Son esprit d'analyse aiguisé allié à une vision globale de la physique moderne furent d'une grande aide tout au long de mon parcours. Sur un plan plus personnel je souhaite également souligner le fait que, grâce à ses qualités humaines, ces années de travail furent un réel plaisir.

Bien entendu, il n'est pas le seul responsable de l'atmosphère positive qui m'a accompagnée durant cette thèse. Je remercie Alex et Michele pour m'avoir chaleureusement accueilli et aidé à démarrer. Je remercie Miguel pour sa joie de vivre et pour m'avoir permis de me sentir à l'aise dans mon travail. Merci aux Seb, mes voisins de bureau, pour leurs impressionantes compétences techniques et pour leur présence amicale. Merci à Olivier pour son aide précieuse et son humour. Merci aussi à Christophe[2], Jérôme, Martin, Pieter et ceux qui se reconnaitront pour avoir rendu les temps de midi si agréables.

Je souhaiterais également témoigner ma gratitude envers Jean-Marc, Caroline, Laurent et Andrea pour le soin avec lequel ils ont lu ce travail, pour leur nombreuses questions intéressantes et pour leurs suggestions qui ont significativement amélioré la qualité de ce manuscript.

Comme mentionné plus haut, ce travail n'aurait pas non plus été possible sans la présence de mes proches. Je souhaiterais remercier Marie, ma compagne et meilleure amie, pour m'avoir toujours soutenu, pour m'avoir supporté dans les moments plus difficiles et pour avoir enrichi ma vie. Merci à Béatrice, ma maman, pour ses encouragements qui m'ont poussés à persévérer dans la voie que j'avais choisie, pour ses bons conseils et pour son amour bienveillant.

Merci à mon papa, ma soeur, mon beau-père et tous les membres de ma famille pour leur présence.

Merci enfin à mes amis pour m'avoir également soutenu et pour m'avoir permis de décompresser quand j'en avais besoin ou tout simplement envie. Merci à la coloc' de Liernu, Yves, Nico, Jé et Jé, qui m'ont toujours accueilli chez eux comme si j'en faisais partie. Merci à Mehdi pour m'avoir si souvent fait rire, pour son coté chaleureux et sa générosité. Merci à Bru, Thom, Sim et Maick pour toutes les discussions intéressantes que nous avons. Je ne peux bien sûr pas citer tous les noms de ceux envers qui je suis reconnaissant c'est pourquoi je terminerai en remerciant tous mes proches; tous ceux qui, par leur brin de folie ou leur coté terre à terre, leur humour léger ou caustique, leur idéalisme ou leur réalisme, m'ont apporté la diversité nécessaire à l'équilibre.

Contents

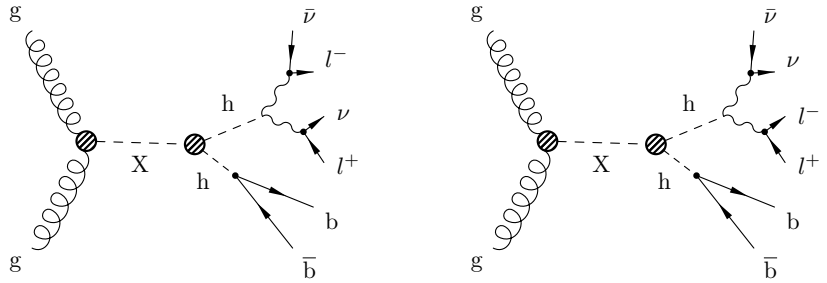
Introduction

Throughout the latter half of the 20th century, a theoretical framework describing three of the four known fundamental interactions of nature has been developed. This theory known as the *Standard Model* (SM) has met numerous successes at providing verified experimental predictions since then which made it the most used theory in elementary particle behavior characterization. Since the recent discovery of the Higgs boson in 2012, all the particles it predicted have been experimentally observed, adding further credence to this model.

Despite these experimental successes, this framework is believed not to be the complete theory of fundamental interactions due to observed phenomena it leaves unexplained. As one of many instances, it does not provide any viable dark matter candidate possessing all the required properties deduced from observational astrophysics and cosmology. Numerous scenarios of physics beyond the SM (BSM) have been and are still postulated to explain some of the SM deficiencies. BSM theories include various extensions of the SM such as *Supersymmetry* or propose entirely novel features such as *Extra Dimensions*. The question of which theory is the best step towards a deeper understanding of nature can only be settled via experiments and is one of the most active field of research in high energy physics.

The *Large Hadron Collider* (LHC) accelerator together with the *Compact Muon Solenoid* (CMS) detector are examples of experimental efforts trying to unravel this query. In 2015, this complex provided the first proton-proton collisions ever recorded with an energy in the center of mass of 13 TeV. This

thesis presents results from the confrontation of these experimental data with the SM predictions. In a first analysis, one tries to observe the production of a new particle that decays into two Higgs bosons in the final state with two b-quarks, two leptons and two neutrinos:



The presence of such a new particle is postulated in several BSM scenarios such as Warped Extra Dimension or Two Higgs Doublet Model.

As of this day, a very important number of experimental data analyses dedicated to specific BSM scenarios have been carried out but were unable to provide hints about where new physics could be lying. Given the plethora of BSM theories that have been and are still proposed, it is impossible to design an experimental analysis for each of them. In this context, the second part of the thesis is dedicated to the development of an innovative approach to search for new physics which would allow to probe any possible deviation from the SM prediction. While this is conceptually possible by looking at any kind of observable, emphasis is given to render the analysis as sensitive as possible to as many models as possible.

Chapter 1

Theoretical Background

The work presented in this manuscript pertains to high energy proton-proton (pp) collisions supplied by the *Large Hadron Collider* (LHC). The goal of this chapter is to provide the reader with the basics of the theoretical framework relevant to the studies reported in this thesis.

First we present the subset of events that have been studied together with the various processes that populates it. Second, we depict the physical effects taking place during a hadron collision and explain the procedure used to make predictions about these processes. After that, we go through the theoretical framework describing the behavior of elementary particles, the *Standard Model* (SM). We derive it based on few conditions and introduce interactions between particles by imposing symmetries to the theory. Next, we mention a few shortcomings of this theory which will motivates the introduction of the alternative models that are studied in this thesis. Finally, we present the Matrix Element Method which will be used to design the model independent search for physics beyond the SM, presented in Chap. 4.

1.1 Final state with two leptons and a $b\bar{b}$ quark pair

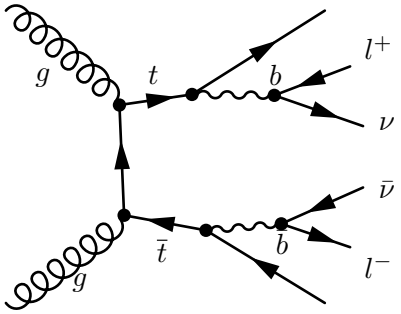
This thesis presents searches for new phenomena in pp collisions at the LHC. The rate of pp collisions provided by the LHC is so high ($\sim 40\,\text{MHz}$) that

current numerical resources do not allow to analyze all of them. Therefore a choice has to be made on which event subset one wants to study. The final state we analyze in this thesis consists of events where at least two leptons (electrons or muons) and a $b\bar{b}$ quark pair are produced (we will note this final state $\mathrm{llbb} + \mathrm{X}$ or simply llbb for convenience). It has the advantage of being easy to select with the trigger system thanks to the presence of two leptons and of being populated by SM processes with reasonably small cross-sections.

Main Standard Model backgrounds

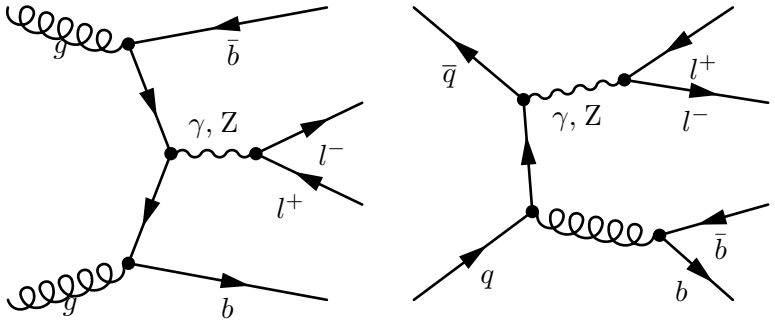
The dominant contributions from SM processes to this final state are (in decreasing order of expected yields):

- Pair production of two top quarks ($\mathrm{t\bar{t}}$) where the tops decay in a b-quarks a lepton and a neutrino. An example of Feynman diagram associated to this process is given by

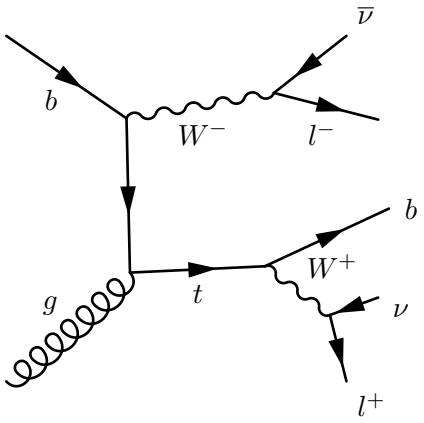


  This background dominates our final state and has a cross section of 87 pb which is nine orders of magnitude smaller than the inelastic pp cross-section at 13 TeV[1].

- Production of a Z/γ-boson (DY) in association with two b-quarks where the Z decays to two leptons:



- Production of a single top quark in association with a W boson (tW) both decaying in the leptonic channel. An example of Feynman diagram contributing to this process is given by:



- Production of two W or Z bosons (VV). Among these contributions, the dominant one is the ZZ process where one Z decays to two b-quarks and the other to two leptons.

- Production of a Z boson in association with a Higgs boson (Zh) decaying to two leptons and two b-quarks respectively.

- Production of a top quark pair plus a vector boson ($\mathrm{t\bar{t}}V$) or plus a Higgs boson ($\mathrm{t\bar{t}}H$).

Any process whose expected contribution to the llbb final state is smaller than the $\mathrm{t\bar{t}}H$ one is considered negligible. An example of such process is the production of three vector bosons or the SM production of two Higgs bosons.

Though the choice of final state to be studied is somehow arbitrary, the model-independent search approach developed in Chap. 4 is generic and could be applied to any topology.

The understanding of the processes populating the llbb final state is a key point of this thesis as one tries to observe something new on top of them. In the following sections one shall see how these processes can be described but also what are the fundamental mechanisms responsible for them.

1.2 Physics at the LHC

1.2.1 Cross section

The infinitesimal cross section between an initial state made of two particles A and B to a final state f, $A\ B\ \rightarrow\ f$, is given by [2]:

$$d\sigma = \frac{1}{2E_A 2E_B |\vec{v}_A - \vec{v}_B|} \left(\prod_f \frac{d^3 p_f}{(2\pi)^3} \frac{1}{2E_f} \right) \times |\mathcal{M}(p_A, p_B \rightarrow \{p_f\})|^2 (2\pi)^4 \delta^{(4)}(p_A + p_B - \sum p_f) \tag{1.1}$$

where $|\vec{v}_A - \vec{v}_B|$ is the relative velocity of the particles as viewed from the laboratory frame, $\prod_f \frac{d^3 p_f}{(2\pi)^3} \frac{1}{2E_f}$ represents the final state phase space, $\mathcal{M}$ is the quantum-mechanical probability amplitude of the process to occur (matrix element) and the Dirac delta function ensures the energy-momentum conservation. The matrix element holds all the dynamics of the considered process. Its exact expression is generally not known but one can express it as a perturbation series in the strength of the interaction and evaluate the first few terms in this series. The first order term in this series will be called "leading order" (LO), the second order term "next to leading order" (NLO), and so on. Each of these terms can be obtained from the theory Lagrangian by means of Feynman Diagrams [3], as one shall see in subsequent sections.

Hard process factorization

Protons are made of sub-constituents called partons (quarks and gluons). The energy of the LHC's protons is such that the hard interaction (high momentum transfer) during a collision can occur between these sub-constituents. Other processes described in the next two sections are characterized by lower momentum transfer (soft processes) and occur before/after the hard process. Given the difficulty to compute a matrix element amplitude describing the proton collision as a whole, one resorts to the *factorization theorem* which states that the soft processes can be factorized from the matrix element amplitude as they occur on a very different timescale [4].

Parton distribution function

As Eq. 1.1 shows, the cross section of a given process depends on the energy of the initial state. In the case of protons collision, there is no way to know *a priori* what is the energy of the interacting partons. The only available information is the probability density $f_i(x_i)$ to find a parton of type i with a fraction $0 < x_i < 1$ of the proton energy. The function describing this probability density is called the *parton distribution function* (pdf). Thanks to the theorem mentioned here-above, one can treat the cross section as a convolution of the pdf with the hard-process cross-section:

$$\begin{aligned} d\sigma =& dx_A f_A(x_A, \mu_F) dx_B f_B(x_B, \mu_F) \\ & \frac{1}{2E_A 2E_B |\vec{v}_A - \vec{v}_B|} \left(\prod_f \frac{d^3 p_f}{(2\pi)^3} \frac{1}{2E_f} \right) \\ & \times |\mathcal{M}(p_A, p_B \to \{p_f\})|^2 (2\pi)^4 \delta^{(4)}(p_A + p_B - \sum p_f) \end{aligned} \tag{1.2}$$

where $p_{A(B)}$ now depends on $x_{A(B)}$.

Due to theoretical difficulties related to non-perturbative QCD effects, no exact pdf calculation is available yet. However, the factorization theorem assumes that the pdf's are universal, these probability density are therefore fitted from a global dataset including HERA(-II), ATLAS, LHCb and CMS results [5]. The error on the obtained pdf is also treated as a systematic uncertainty impacting the final results of this thesis.

Radiation, confinement and hadronization

Now that we have seen how to deal with non elementary particle collisions, let us describe the way we treat soft processes occurring aside the hard process. For example, the scattering $q\bar{q} \to q\bar{q}$ seems very simple when looking only at the hard process but, to fully link the theoretical aspects discussed so far with what is experimentally observable, several other effects have to be taken into account:

- First the initial (final) state particles can radiate other particles before (after) undergoing the hard scattering process which will add more particles to the observed final state. These effects are respectively called *initial state radiation* (ISR) and *final state radiation* (FSR). Note that depending on the separation between hard and soft processes, some of these emissions may still be included in the hard process description.

- Another important effect makes the picture way more complicated: quarks (and gluons) are not observable as isolated particles due to the *color confinement*. This phenomenon implies that only colorless particles are observable as isolated. While so far no formal demonstration of this phenomenon exists, it has never been disproved by experiment. What is believed to happen is the following: when two quarks move away from each other, at some point it becomes energetically more favorable for a new quark–antiquark pair to spontaneously appear. What happens to our final state quarks is then the following: when produced, they carry a lot of energy and drift apart. While doing so, they can radiate gluons (and other particles) which can produce $q\bar{q}$ pairs and so on (*showering*). Confinement imposes that the produced colorful particles will form colorless bound states called hadrons (*hadronization*). Again based on the factorization theorem, one treats these effects separately from the hard-process cross section via the fragmentation function, $D_{f\to h}(z, \mu'_F)$, which encodes the probability of the final parton f to fragment into a hadron h with an energy fraction $z = \frac{E_h}{E_f}$. A consequence of the *showering* and *hadronization* is that instead of observing a single quark in the detector, one rather observes a spray of hadrons and other particles which is called

a *jet*. Note that there is an exception: the top quark, due to its very short lifetime, decays before undergoing hadronization.

While the factorization scales separating the pdf from the hard process and the hard process from the fragmentation can in principle be different, one generally assumes $\mu_F = \mu'_F$. The procedure described here-above allows to factorize out of the matrix element $\mathcal{M}$ all the effects but the hard process which is of most relevance for this thesis studies. Therefore, $\mathcal{M}$ encodes all the dynamics we are interested in.

In order to confront theoretical models with real data, one has to be able to predict the expected observations under the different hypotheses one wants to study. The simulation of the physical processes that have been mentioned so far are handled by the so-called Monte Carlo (MC) generators such as MADGRAPH [6], PYTHIA [7, 8] or POWHEG [9, 10, 11, 12, 13]. These MC generators provide a collection of four-momentum of "stable"[1] particles (colorless hadrons, leptons and photons) resulting from the requested process.

As mentioned earlier, one computes $\mathcal{M}$ as a perturbation series in the strength of the interactions. Divergences in the terms of this series are absorbed especially in the coupling constants at an arbitrary scale called *renormalization scale*, μ_R. When truncating the series, one introduces dependencies of the result on both the factorization and renormalizaton scale. The choice of these scales impacts thus our simulations. To take this into account, a systematic uncertainty on this thesis' results will be estimated by varying μ_F and μ_R.

So far we have seen the processes that populate our final state. We have seen how one can predict their behavior at hadron collider based on the factorization principle. We came to the observation that the matrix element is a central quantity to characterize the hard process' dynamics and mentioned that the terms contributing to this probability amplitude can be obtained from a theory by analyzing its Lagrangian density. It is now time to move on to the description of the theoretical framework we use to describe the processes mentioned in Sec. 1.1.

[1] The term "stable" is to be understood in the sense of having a mean path length sufficient to reach the detector e.g. ultra-relativistic muons can be considered stable due to their Lorentz dilated lifetime.

1.2.2 The Standard Model

Let us start with the conditions we impose to the SM Lagrangian density. First, we require it to be Poincaré invariant. Its general form is thus a function of the fields Ψ_i and their derivatives $\partial_\mu \Psi_i$:

$$\mathcal{L} = \mathcal{L}\left(\Psi_i(x), \partial_\mu \Psi_i(x)\right) \tag{1.3}$$

since Poincaré invariance forbids explicit space-time dependence. The action given by

$$S = \int d^4x \mathcal{L} \tag{1.4}$$

must be dimensionless in the natural unit system ($\hbar = c = 1$) which means that every Lagrangian term must have dimension E^{+4}. When performing amplitude calculations, we often hit infinities that have to be absorbed in order to get meaningful results. We want the Lagrangian to be *renormalizable* which requires that this absorption is possible and translates into the condition that any parameter in the Lagrangian must have a dimension in power of energy greater than or equal to zero [14].

Finally, we require the Lagrangian to be invariant under groups of local gauge transformations which elegantly introduces interactions between fields as it is illustrated here-under with the simple case of *Quantum Electrodynamics* (QED) based on the abelian group $U(1)$.

Gauge invariance: U(1)

Let us start with the Lagrangian of a free fermionic field, formulated by Paul Dirac

$$\mathcal{L}_{Dirac}^{Free} = \bar{\psi}(i\partial\!\!\!/ - m)\psi \tag{1.5}$$

and consider the following local U(1) transformation of the field ψ

$$\psi(x) \rightarrow e^{iq\alpha(x)}\psi(x). \tag{1.6}$$

Here q represents the strength of the phase transformation and $\alpha(x)$ is an arbitrary differentiable function of space-time coordinates.

At this stage, this transformation does not leave the Lagrangian in Eq. (1.5) invariant due to the term involving partial derivative of ψ. Let us now assume the existence of a spin-1 vector field $A_\mu(x)$ transforming as

$$A_\mu(x) \to A_\mu(x) + \partial_\mu \alpha(x) \tag{1.7}$$

and define the covariant derivative

$$D_\mu \equiv \partial_\mu - iqA_\mu(x). \tag{1.8}$$

One can easily see that rewriting the Lagrangian as:

$$\mathcal{L}^{Int}_{Dirac} = \bar{\psi}(i\not{D} - m)\psi \tag{1.9}$$

$$= \bar{\psi}(i\not{\partial} - m)\psi + q\bar{\psi}\gamma^\mu A_\mu \psi \tag{1.10}$$

makes it invariant under U(1). What is important to note now is that the second term in Eq. (1.10) implies an interaction between the spinor field ψ and the vector field A_μ.

Finally, the complete Lagrangian is obtained by adding the $U(1)$ gauge invariant term $F_{\mu\nu}F^{\mu\nu}$ with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ which corresponds to a kinetic term for spin-1 particles:

$$\mathcal{L}_{QED} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\not{\partial} - m)\psi + q\bar{\psi}\gamma^\mu A_\mu \psi. \tag{1.11}$$

Requiring $\mathcal{L}^{Free}_{Dirac}$ to be invariant under local U(1) gauge group boils down to the addition of a bosonic field A_μ interacting with the fermionic fields via the definition of a covariant derivative. Imposing to add to the Lagrangian all the renormalizable, gauge and Poincaré invariant terms leads to the complete theoretical QED description. It is worth stressing at this point that A_μ appears to be massless. Indeed, for spin-1 fields, the mass term is of the form $\frac{1}{2}m_A^2 A_\mu A^\mu$ which would violate gauge invariance. This approach of imposing gauge invariance to generate interactions between fields is a corner stone of the SM, described in the next section.

The Standard Model Lagrangian

This section is based on the reference [15]. The gauge group under which we require the SM Lagrangian density to be invariant is $SU(3)_C \otimes SU(2)_L \otimes$

$U(1)_Y$ where the subscript C refers to the *color*, L indicates that only left-handed[2] fermions transforms under $SU(2)_L$ reflecting the maximal parity violation of the weak interaction and Y represents the so-called *weak hypercharge*.

The group $SU(3)_C$ is associated to the strong interaction, its fundamental representation has eight generators $\frac{\lambda^a}{2}$ (with λ^a the Gell-Mann matrices) and the quanta of its gauge fields G^a_μ (a=1,...,8) are called the gluons. The associated gauge tensors are defined as :

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^b_\mu G^c_\nu \tag{1.12}$$

with g_s the coupling strength of the strong interaction and f^{abc} the structure constant of the group $SU(3)_C$.

The groups $SU(2)_L$ (whose fundamental representation generators are obtained from the three Pauli matrices: $\frac{\sigma^i}{2}$) and $U(1)_Y$ are associated to the electroweak interaction. They generate three and one spin-1 fields: W^i_μ (i=1,2,3) for $SU(2)_L$ and B_μ for $U(1)_Y$. Their associated tensors are

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon^{ijk} W^j_\mu W^k_\nu \tag{1.13}$$

and

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{1.14}$$

where g is the equivalent of g_s for $SU(2)_L$ and ϵ^{ijk} are the structure constants of the fundamental representation of $SU(2)$. Based on that, one can already write down the *gauge* part of the SM Lagrangian:

$$\mathcal{L}_{gauge} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a - \frac{1}{4} W^i_{\mu\nu} W^{\mu\nu}_i - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \tag{1.15}$$

which contains the kinetic and self-interaction terms of the gauge fields. The term involving three gluon fields is of particular importance as it contributes to the gluon induced $\mathrm{t\bar{t}}$ production, as shown below.

Moving on now to the fermionic matter content, one has to assign each field to a $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ representation according to the way they

[2] Here left-handed refers to chirality.

interact (based on experimental observations). It appears in nature that only left-handed chirality particles interacts *weakly*, therefore we define left- and right-handed fields with the following: $\psi_L = P_L\psi$ and $\psi_R = P_R\psi$ where

$$P_L = \frac{(1-\gamma^5)}{2} \tag{1.16}$$

and

$$P_R = \frac{(1+\gamma^5)}{2} \tag{1.17}$$

are the projection operators that extract left and right handed components from Dirac spinors. Finally we gather left-handed fermions from a same family into $SU(2)$ doublets

$$Q = \begin{pmatrix} u_L \\ d_L \end{pmatrix} \text{ and } L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \tag{1.18}$$

where $u(d)$ are up(down) type quarks, ν is the neutrino and e corresponds to the electron.

Matter field	Rep. dim.	Covariant derivative
$Q = \begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$(\mathbf{3},\mathbf{2},+\frac{1}{3})$	$D_\mu = \partial_\mu - ig_s\frac{\lambda^a}{2}G^a_\mu - ig\frac{\sigma^i}{2}W^i_\mu - ig'\frac{1}{6}B_\mu$
u_R	$(\mathbf{3},\mathbf{1},+\frac{4}{3})$	$D_\mu = \partial_\mu - ig_s\frac{\lambda^a}{2}G^a_\mu - ig'\frac{2}{3}B_\mu$
d_R	$(\mathbf{3},\mathbf{1},-\frac{2}{3})$	$D_\mu = \partial_\mu - ig_s\frac{\lambda^a}{2}G^a_\mu + ig'\frac{1}{3}B_\mu$
$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}$	$(\mathbf{1},\mathbf{2},-1)$	$D_\mu = \partial_\mu - ig\frac{\sigma^i}{2}W^i_\mu + ig'\frac{1}{2}B_\mu$
e_R	$(\mathbf{1},\mathbf{1},-2)$	$D_\mu = \partial_\mu + ig'B_\mu$

Table 1.1: Representation dimension of the SM matter fields and their corresponding covariant derivative. The three numbers in the second column are respectively the representation dimension under $SU(3)_C$, $SU(2)_L$ and the weak hypercharge.

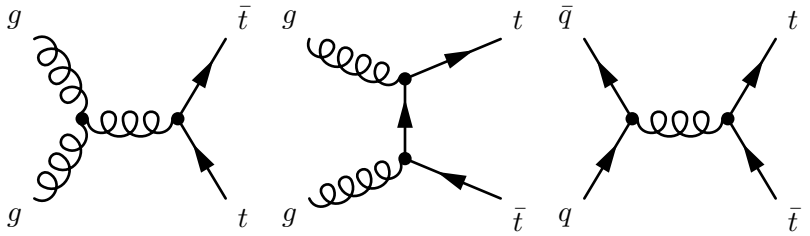
Based on the previous considerations, we can define the (massless) fermionic part of the SM Lagrangian density:

$$\mathcal{L}_{fermion} = \sum_{\psi,\, I} \bar{\psi}^I (i \not{D}) \psi^I \tag{1.19}$$

with the various fields[3] $\psi = Q,\ u_R,\ d_R,\ L,\ e_R$ and $I = 1,\ 2,\ 3$ the three families. The covariant derivatives (independent of the family) are define by

$$D_\mu = \partial_\mu - i g_s T^a G^a_\mu - i g T^i W^i_\mu - i g' \frac{Y}{2} B_\mu \tag{1.20}$$

where T^a and T^i are the generators of the $SU(3)_C$ and $SU(2)_L$ representation to which belong the fields and Y is its hypercharge. The representation dimension of SM matter fields and their corresponding covariant derivatives are collected in Tab. 1.1. As already mentioned, Eq. (1.15) implies in particular the triple gluon coupling. On the other hand, Eq. (1.19) and (1.20) introduce terms proportional to $\bar{\psi} \lambda^a G^a_\mu \psi$ for the quarks which imply couplings between them and the gluon fields. This already allows us to write the LO Feynman diagrams associated to the $\mathrm{t\bar{t}}$ production, the dominant SM process populating the llbb final state:



In the present situation, gauge invariance forbids mass terms both for gauge bosons and fermions. This theory which describes a world of massless particles has to be completed in order to meet experimental observations. The Brout-Englert-Higgs (BEH) mechanism [16, 17, 18] postulated in 1964 is an example of mechanism solving this issue.

[3] In the SM, neutrinos are assume to be massless and only left-handed neutrino are present. Experimental observations of neutrino oscillations suggest though that at least two of the three neutrino flavors are massive and would imposes us to include right-handed neutrinos to the theory but this is put aside because it goes out of the scope of this discussion.

BEH mechanism

This approach postulates the existence of one extra $SU(2)_L$ doublet of complex scalar fields

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{1.21}$$

which transforms under representation dimensions $(\mathbf{1}, \mathbf{2}, +1)$[4]and has therefore a covariant derivative $D_\mu = \partial_\mu - ig\frac{\sigma^i}{2}W_\mu^i - ig'\frac{1}{2}B_\mu$. The kinetic term of the scalar doublet $D_\mu H^\dagger D^\mu H$ will then generate mass terms for gauge bosons via the introduction of a potential V associated to the Higgs field. Gauge invariance under $SU(2)_L \otimes U(1)_Y$ requires terms involving $H^\dagger H$ [19] and, as mentioned at the beginning of Sec. 1.2.2, renormalizability forbids terms involving $H^\dagger H$ to a power higher than two. This leads to the potential

$$V(H^\dagger H) = -\mu^2 H^\dagger H + \lambda (H^\dagger H)^2. \tag{1.22}$$

Based on that, one can write down the scalar sector of the SM Lagrangian:

$$\mathcal{L}_{scalar} = D_\mu H^\dagger D^\mu H - V(H^\dagger H). \tag{1.23}$$

Vacuum stability demands λ to be greater than zero and to ensure the ground state to be different than zero, one requires $\mu > 0$. The ground state is degenerated and satisfies $H^\dagger H = \frac{\mu^2}{\lambda} \equiv v^2$ with v a real number called *vacuum expectation value* (vev). In order to preserve the electric charge conservation ϕ_1 and ϕ_2 must have a null vev while the remaining degree of freedom can be removed by choosing the ground state to be at $\phi_3 = v, \phi_4 = 0$ [15]:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{1.24}$$

The variation along the ϕ_3 component corresponding to the excited state of H is interpreted as the Higgs boson:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h \end{pmatrix}. \tag{1.25}$$

[4]Following convention form Tab. 1.1

By plugging Eq. (1.25) into $\mathcal{L}_{scalar}$ and expanding it we obtain:

$$\mathcal{L}_{scalar} \ni \frac{v^2}{8}(g^2 W_\mu^1 W^{1\mu} + g^2 W_\mu^2 W^{2\mu} + g^2 W_\mu^3 W^{3\mu} + g'^2 B_\mu B^\mu - 2g'g B_\mu W_3^\mu)\left(1 + \frac{h}{v}\right)^2. \tag{1.26}$$

To recover the charged fields $W^\pm$ and move to the mass eigenstates, we perform the following redefinitions [20]:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2) \tag{1.27}$$

$$A_\mu = \frac{1}{\sqrt{g^2 + g'^2}}(g B_\mu + g' W_\mu^3) \tag{1.28}$$

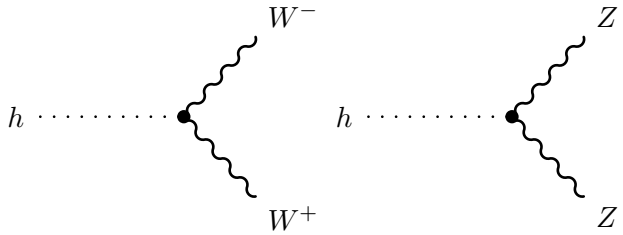
$$Z_\mu = \frac{1}{\sqrt{2}}(-g' B_\mu + g W_\mu^3). \tag{1.29}$$

These fields have now acquire a mass and correspond respectively to the W bosons with $m_W = \frac{1}{2}vg$, the photon with $m_A = 0$ and the Z boson with $m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$.

Using these definitions, Eq. (1.26) can be written as

$$\mathcal{L}_{scalar} \ni (m_W^2 W_\mu^+ W^{-,\mu} + \frac{1}{2} m_Z^2 Z_\mu Z^\mu)\left(1 + \frac{h}{v}\right)^2. \tag{1.30}$$

In this Lagrangian density one sees explicitly that the Higgs particle couples to the W and Z bosons. This coupling allows the Higgs to decay into a pair of W or Z bosons:

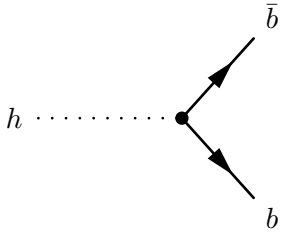


which is a characteristic of the di-Higgs signal we look for in Chap. 3.

Mass terms for fermions (except for neutrinos which are assumed to be massless in this discussion) arise from the so-called Yukawa sector of the SM:

$$\mathcal{L}_{Yukawa} = \sum_{I,J} -\frac{v}{\sqrt{2}} \left(\bar{u}_R^I (Y_u)^{IJ} u_L^J + \bar{d}_R^I (Y_d)^{IJ} d_L^J + \bar{e}_R^I (Y_e)^{IJ} e_L^J\right) \left(1 + \frac{h}{v}\right) + h.c. \tag{1.31}$$

where $Y_{u,d,e}$ are the three-by-three Yukawa matrices and I and J are indices over the generations. To complete the SM description of nature one should perform a rotation of the fields to their mass states but for the purpose of this discussion it is sufficient to note that $\mathcal{L}_{Yukawa}$ implies couplings between the Higgs boson and the fermions which are proportional to the fermion masses. On-shell Higgs bosons can not decay into a pair of top quark since its mass is too small which implies that its dominant fermionic decay channel is into a b quark pair:



This coupling is present in the Zh background mentioned in Sec. 1.1 and in the di-Higgs signal studied in Chap. 3.

1.3 Beyond Standard Model physics

1.3.1 Current questioning and unexplained phenomena

Once the free parameters of the SM have been measured, all sorts of reactions among particles or particle properties can be predicted and confronted with experiment. Numerous such test of the SM have been carried out so far and none of them lead to compelling evidence of SM hypothesis exclusion. The SM is thus a very successful theory however it is likely not the end of the road as suggested by various hints related to the free parameters of the theory or to

presently unexplained phenomena. A non-exhaustive list of these hints (chosen mainly to introduce the various signals studied in this thesis) is presented hereunder.

- **Strong θ puzzle:** When building the SM Lagrangian based on Poincaré plus gauge invariance and renormalizability, a term of the form $\theta \frac{g^2}{32\pi^2} G_a^{\mu\nu} \tilde{G}_{a,\mu\nu}$ with $\tilde{G}_{a,\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} G_a^{\rho\sigma}$ and θ a free parameter should be added. On the one hand, the measured value of θ receives contributions from the Yukawa and QCD sectors, on the other hand, constraints from the non-observation of neutron electric dipole moment impose $\theta \lesssim 10^{-9}$ [21]. Having two contributions coming from two sectors seemingly blind to each other but which result in a (almost) perfect cancellation requires an important fine tuning. Even though the theory works as such, this seems very unnatural and claims for a deeper explanation.

- **Flavor puzzle:** The way masses are generated for fermions via Eq. (1.31) introduces one parameter per massive particle. The Yukawa sector does not say anything about the value or relation between these masses. However, there is a regular pattern in the generation masses: except in the case of neutrinos for which the following statement awaits experimental verification, each particle from a family is lighter than its counterpart from the next one. It is thus natural to expect that a more general mechanism, with fewer free parameters, could be responsible for generating this pattern. The SM lacks of such explanation.

- **Hierarchy puzzle:** The free parameters of the SM scalar potential have to be renormalized as explained at the beginning of Sec. 1.2.2. The bare Higgs mass has corrections that diverge quadratically. If one regularizes the divergences with a high energy cut off such as the Planck Mass, these corrections are thus proportional to M^2_{Planck}. Given that the physical value of the Higgs boson mass is of the order of the electroweak scale a very delicate fine tuning of the bare parameters is required to meet experimental observation [15]. This puzzle is related to the following question for which we have no explanation so far: why is the gravity interaction so weak at energies of the order of the electroweak scale?

- **Dark matter:** Various astrophysical and cosmological observations such as the galactic rotation curves, bullet galaxy cluster or cosmic microwave background anisotropies [22] (coming thus from very different scales) tell us that, if one assumes that the way we describe gravity at large scale is correct, the SM only explains a small part of the total matter present in the universe. The remaining matter should be made of additional particles for which the SM does not provide any candidate [23].

- **Matter/anti-matter asymmetry:** The observed universe seems very unbalanced since matter heavily dominates over anti-matter whereas the Big Bang is expected to have produced equal amounts of matter and antimatter. The SM does not provide any viable mechanism to explain such an asymmetry. Therefore, except if another mechanism is responsible for this observation (e.g. it may be that we happen to live in a region of the universe where matter dominates while antimatter dominates elsewhere), the SM has to be extended to explain why only matter survived.

In light of the points mentioned here-above, it looks like the SM could be amended towards a deeper understanding of nature. The next section is dedicated to various theoretical framework proposals going in this direction.

1.3.2 Beyond Standard Model Scenarios

As mentioned in the previous section, the SM leaves room for improvement in various aspects. Numerous models called *Beyond Standard Model* (BSM) have been (and are still) proposed to tackle one or more "problem(s)" of the SM. The difficulty being to extend/modify the SM while consistently describing all the experimentally established results.

Given the amount of BSM scenarios available in the literature and the free parameters accompanying them, it is impossible to design an experimental search for every possible scenarios. Therefore, the community is making a lot of effort to allow experimental searches being sensitive to large numbers of BSM models or parameters. Within this context, a large part of this thesis is dedicated to study the possibility of designing a method able to probe as many of

the BSM scenarios as possible while still being reasonably sensitive. In order to evaluate the general sensitivity of the method, one has to provide results for various BSM benchmarks. The choice of which BSM scenarios we provide results for was driven by the availability of an equivalent "model-dependent" (dedicated) search performed in reasonably similar conditions. The various BSM families they belong to is described here-under.

- **Scalar sector extensions:** Two-Higgs-Doublet Models (2HDM) [24] are simple extensions of the SM where one more $SU(2)$ doublet is added to the scalar sector. Enlarged Higgs sectors can be associated with larger symmetry group and can provide additional sources of *CP* violation addressing thus the matter/antimatter asymmetry problem. In addition to that, they could also provide Dark Matter candidates [25].

  With two complex scalar $SU(2)$ doublets, eight fields are generated. As in the SM scenario, three of them get "eaten" to give mass to the W and Z bosons and the remaining five correspond to particles that could potentially be created at the LHC. There are two charged scalars ($H^{\pm}$), two neutral scalars (H and h) and one pseudo-scalar (A). An example of process that would be visible in the llbb final state is $\mathrm{pp} \rightarrow H \rightarrow ZA \rightarrow l^{+}l^{-}b\bar{b}$ [26].

  These extensions of the SM are not complete theories per se, they are rather seen as part of more complete theory. An example of theory implementing scalar sector extensions is *supersymmetry*. Supersymmetry models are based on the introduction of a new symmetry between bosons and fermions which implies that every SM particle has a *superpartner* with the same quantum numbers but differing by one half-unit of spin. This theory solves the hierarchy puzzle mentioned in previous section by introducing new radiative corrections canceling the quadratic divergences in the scalar sectors. Phenomenologically speaking, one can test these models by searching for superpartners like the stop quark ($\tilde{t}$) which is the bosonic equivalent of the top quark. The pair production of stop quark leaves a signature in the $\mathrm{llbb} + \mathrm{X}$ final state via the decay to a top quark plus an invisible particle (the top quark decaying subsequently to a b-quark a lepton and a neutrino).

- **Dark matter models:** There is an important amount of models trying to provide dark matter candidate(s). As a consequence, an interesting strategy is to try to constrain viable dark matter scenarios in a model independent way via simplified models. In these models, one simply consider a single dark matter candidate with arbitrary spin that couples to visible matter. The simplified model used to assess the power of this thesis' model independent search belongs to the family of 'top-philic dark matter scenarios' where the dark matter candidate dominantly couples to top quark via an s-channel scalar mediator [27]. Again, this BSM scenario leaves signature in the llbb final state thanks to the presence of a top quark pair.

- **Wrapped Extra Dimension:** This family of theories (Randall-Sundrum models [28]) propose an explanation on why gravity is so small compared to the weak interaction by introducing an additional compactified dimension. The universe would be a five dimensional space with two branes, one where the gravity is a relatively strong force and the other corresponding to our usual (3+1)-dimensional space. Phenomenologoically, these models can imply new massive resonances that could be probed at the LHC [29].An example of signature populating the llbb final state is the decay of the new massive resonance to two Higgs bosons which subsequently decay to two b-quarks and two W bosons going to a lepton and a neutrino. This signal will be used to optimize this thesis' dedicated search and will provide one additional benchmark to study the power of the model-independent search.

1.4 Matrix Element Method

As mentioned in the previous sections, the llbb final state is populated by various SM processes but is also potentially sensitive to several BSM signatures. In order to maximize the sensitivity of a search to BSM signals, one usually tries to extract regions where the ratio "SM contribution over BSM signal contribution" is small. In most analysis, this region is defined based on kinematic quantity (discriminant) which partially characterize the processes one wants to separate such as e.g. invariant masses. This thesis' model independent search

has the peculiarity of not having a well identified signal which renders the analysis optimization challenging. Since one can only rely on the backgrounds, one wants to have discriminants that maximally encodes the processes dynamics. The *Matrix Element Method* (MEM) described here-under is an example of tool providing such discriminants.

1.4.1 Definition

The MEM weight under a process hypothesis[5] α for a given experimental event x[6], $W(x|\alpha)$, gives the probability density to observe x in the detector provided that the process α occurred in the hard interaction.

Its mathematical definition at hadron collider relies on the cross section formula defined in Sec. 1.2.1 and is given by

$$W(\mathbf{x}|\alpha) \equiv \sum_{i,j} \int d\Phi(\mathbf{y}) dq_i dq_j f_i(q_i) f_j(q_j) |M_\alpha(q_i, q_j, \mathbf{y})|^2 T(\mathbf{x}|\mathbf{y}). \quad (1.32)$$

The sum over i and j takes into account the fact that the process α may be induced by different initial state ($q\bar{q}$, gg, etc.). The integration is performed over the initial state partons energy and over the partonic final state phase space $\Phi(\mathbf{y})$. The matrix element M_α encloses the dynamic of the hard process α as explained in Sec. 1.2.1 and the *transfer function* $T(\mathbf{x}|\mathbf{y})$ models the evolution of the hard process partonic final state $\mathbf{y}$ into the detector level reconstructed event $\mathbf{x}$, following the factorization theorem described earlier. The transfer function takes into account the physics described in Sec. 1.2.1 (showering, hadronization, etc.) together with the experimental reconstruction effects such as the detector resolution. The physical quantity described by the transfer function is the probability density to reconstruct the event x in the detector provided that the hard process led to the partonic configuration y. It is therefore normalized as follows:

$$\int T(x|y) dx = 1 \quad (1.33)$$

[5] This method can also be used to extract information about theoretical parameters of a model but this possibility is of no relevance in our context.

[6] Here x represents the measured momenta of final state particles.

in the hypothetical scenario of a detector with 100% acceptance and selection efficiency. More details on the transfer functions will be given in the next chapter.

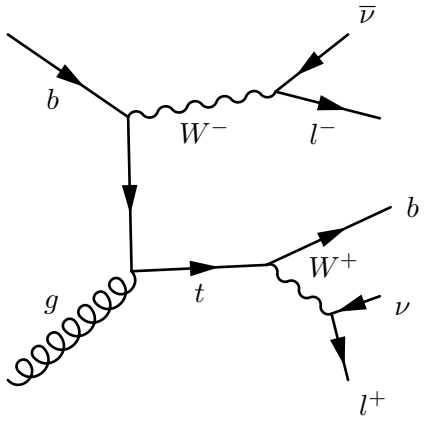
The MEM weight thus quantifies the "agreement" between the theoretical process α and the experimental event x which makes it a very useful tool to characterize the reconstructed events.

1.4.2 Integration

The main difficulty when computing MEM weights resides in performing the numerical integration of a multi-dimensional function which exhibits in general a highly non trivial behavior. Indeed, the integrand in Eq. (1.32) can have several localized peaks e.g. in presence of resonances in the matrix element or in the transfer function when the partonic configuration $\mathbf{y}$ gets kinematically close to the reconstructed event $\mathbf{x}$.

The MEM weights computation is performed with the software package MoMEMta [30] which is the successor of MadWeight [31] and is characterized by a great modularity together with a coverage for the needs of an experimental analysis work flow. The integration is performed with the CUBA library [32] which interfaces the VEGAS algorithm [33]. Conceptually, this algorithm tries to optimize the integration grid towards regions where the integrand makes the greatest contributions to the final integral. Narrow peaks in the multi-dimensional domain may thus be overlooked if one does not take additional precautions.

To illustrate the way we perform such integrals, let us take the example of MEM weight computation under tW^- hypothesis whose implementation was performed within this thesis' work. Considering that the initial b-quark is present in the proton (i.e. factorizing the gluon splitting it comes from inside the pdf), an example of Feynman diagram contributing to this process in the llbb final state is given by



One can see that this process involves three resonances associated to the two W bosons and the top quark. As a consequence, kinematic configurations where at least one of the resonances is away from its on-shell mass will bring small contribution to the total integral. In order to ease the integration, one performs changes of variable to integrate over the resonance masses instead of over the final state particles four vectors.

To evaluate the matrix element, one has to reconstruct the full partonic final state. This is not immediate due to the presence of two neutrinos since their individual momentum is inaccessible based on experimental information. As it will be further described in Sec. 2.3.5, the measurement of missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$) gives two constraints while the neutrinos momenta represent six degrees of freedom, leaving the system under-constrained. Let us explain in details how one technically performs this integral.

First we randomly generate the azimuthal angle of the neutrino ν associated to the W^+. Next, since one knows the four momenta of the b and l^+ from measurements, one can obtain p_ν based on the top and W^+ masses, solving the system of equations

$$m_{W^+}^2 = (p^\nu + p^{l^+})^2 \tag{1.34}$$

$$m_t^2 = (p^\nu + p^{l^+} + p^b)^2. \tag{1.35}$$

This adds two integration dimensions because the W^+ and t masses are randomly generated according to Breit-Wigner distributions. Finally, one derives

$p_{\bar{\nu}}$ based on the W mass and on the $E_{\mathrm{T}}^{\mathrm{miss}}$ by solving the system of equations

$$m_{W^-}^2 = (p^{\bar{\nu}} + p^{l^-})^2 \tag{1.36}$$

$$p_x^{E_{\mathrm{T}}^{\mathrm{miss}}} = (p^{\bar{\nu}} + p^{\nu})_x \tag{1.37}$$

$$p_y^{E_{\mathrm{T}}^{\mathrm{miss}}} = (p^{\bar{\nu}} + p^{\nu})_y. \tag{1.38}$$

$p_{E_{\mathrm{T}}^{\mathrm{miss}}}^{x(y)}$ are known from the reconstructed event x while the W^- mass is randomly generated according to its Breit-Wigner distribution. Since we have two reconstructed b-quarks in the final state while the matrix element has only one, the weight is computed by averaging over the permutation of the reconstructed b-jets. To summarize, this technique uses the kinematic information of the process to constrain the partonic final state momenta and part of the integration over the neutrino momenta is traded for an integration over the resonance masses. The Jacobians associated to these change of variables are automatically evaluated by the MoMEMta software for each integration point.

Chapter 2

Experimental setup and event reconstruction

So far we have seen how to make the link between a theoretical framework and a well identified observable. We have covered the most relevant theoretical aspects of the current understanding in high energy physics and presented some extension proposals together with their typical signatures. We have described the final state which will be studied and the tool used to perform the model independent search analysis. One central piece is still missing to apprehend all the aspects of this thesis: the experimental set-up providing data allowing to test our current particle physics knowledge. This chapter is entirely dedicated to this matter. First we describe the LHC accelerator. Second we present the CMS detector. Next we cover the reconstruction of physics objects relevant for the $\mathrm{llbb} + \mathrm{X}$ final state and finally we derive the transfer functions used to compute the MEM weights described in Sec. 1.4.

2.1 The Large Hadron Collider

The *Large Hadron Collider* (LHC) [34, 35, 36] is, as of today, the most powerful accelerator ever built. This project was officially endorsed by the *Organisation européenne pour la recherche nucléaire* (CERN) in 1994 to take over from the *Large Electron Positron* (LEP). It consisted of replacing the existing electron-positron accelerator facility by a more powerful apparatus able

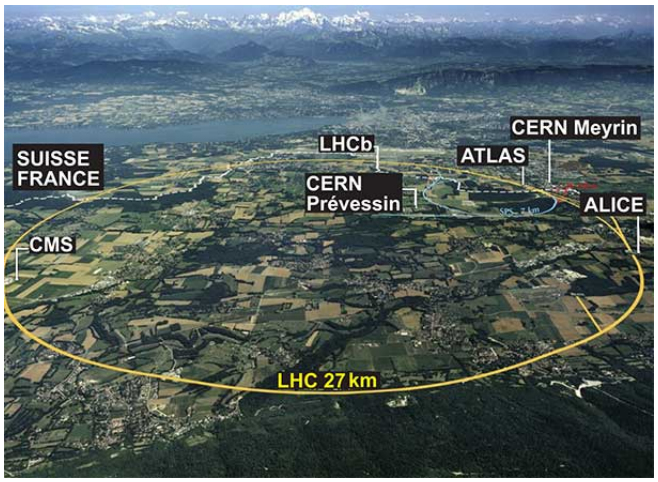


Figure 2.1: Aerial view of the LHC ring geographic location. The picture also shows the LHC four biggest experiments: ALICE, ATLAS, LHCb and CMS.

to accelerate protons and heavy ions up to respectively 7 TeV and 1150 TeV (namely 2.75 GeV per nucleon). The accelerator forms a 27 km circumference ring buried at approximately 100 m underground and is situated near Geneva beneath the France-Switzerland border as shown on Fig. 2.1.

2.1.1 Luminosity

The total luminosity delivered by an accelerator is given by

$$L = \int \mathscr{L}(t)dt. \tag{2.1}$$

If identical bunches containing each N particles collide head-on with frequency f, the instantaneous luminosity, $\mathscr{L}$, is given by

$$\mathscr{L} = f\frac{N^2}{4\pi\sigma_x\sigma_y} \tag{2.2}$$

where σ_x and σ_y are the standard deviations of the supposedly gaussian distributed particle density in the bunch. We also assumed that the profile of the particle distribution in the beam is independent of the position along the

bunch. This simple formula reveals the accelerator parameters that most impact the event rate: the number of bunch crossings per second, the number of particles per bunch and their spatial spreading in the transverse plane.

At the LHC, protons are accelerated from 450 GeV to 6.5 TeV using radiofrequency cavity which are metallic chambers containing an oscillating electromagnetic field. In order to minimize the scattering of protons by gas molecules, the beam pipes must be kept at ultra-high vacuum ($\sim 10^{-13}$ atm). Protons travel at velocities close to the speed of light in the accelerator ring and must follow a circular trajectory. This is achieved by means of superconducting dipole electromagnets. Quadrupole magnets are used to keep the beam focused, increasing thus the instantaneous luminosity. The LHC was designed to have bunches crossing every 25 ns (40 MHz) with 1.15×10^{11} protons each and a "transverse dispersion" of $\sigma_x = \sigma_y = 17\,\mu\mathrm{m}$. Putting all these parameters into the formula given in Eq. (2.2) leads to an instantaneous luminosity of the order of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ which is the design luminosity of the LHC. Note that in this discussion we neglected smaller effects due to, for instance, the fact that the beams do not collide head-on.

The integrated luminosity is a very important quantity when performing a measurement or a search since it is directly related (together with the cross sections) to the number of expected event. This quantity has to be measured and is never perfectly known. The error on the total luminosity is treated as a systematic uncertainty in this work.

2.1.2 Data taking eras

The LHC provided its first 7 TeV pp collisions in 2010 as shown on Fig. 2.2 followed by another data taking period in 2011 at the same energy providing 6.1 fb^{-1} of integrated luminosity. In 2012, the energy in the center of mass was raised to 8 TeV, energy at which 23.3 fb^{-1} were delivered. This marked the end of the so-called run I. After a two years shut down planned for maintenance, consolidation and preparation for running at higher energy, the LHC resumed data taking in 2015 providing 4.2 fb^{-1} pp collisions at 13 TeV, triggering the start of run II. The work presented in this thesis is based on these very first 13 TeV pp collisions. Since then, LHC continued providing data at this energy,

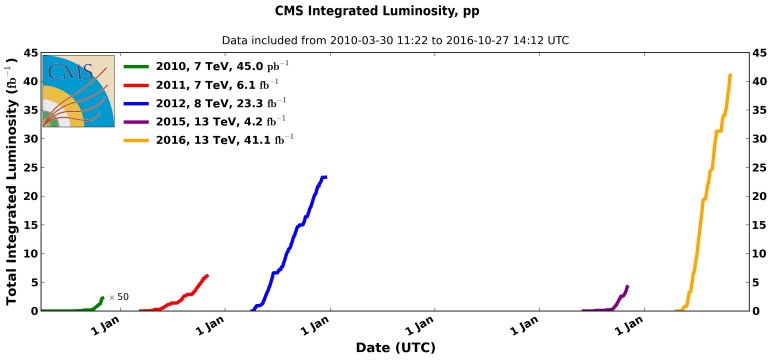


Figure 2.2: Luminosity delivered by the LHC to CMS versus time for 2010, 2011, 2012, 2015 and 2016 considering proton-proton collisions only.

with a total of 41.1 fb^{-1} in 2016. Note that some of the apparently empty period visible on Fig. 2.2 correspond to shorter maintenance or to heavy ion runs not discussed in this thesis.

2.2 The CMS detector

The data analyzed in this thesis are provided by CMS [37, 38] which is a general-purpose detector allowing for a wide physics program such as the study of the Higgs sector, the search for new particles, the precise measurement of SM processes, etc.

The CMS detector is 21.5 m long with a diameter of 15 m and weights 12 500 tonnes. The total inelastic cross section for proton-proton collisions at a center of mass energy of 13 TeV is about 70 mb [1] which corresponds, at the LHC design luminosity, to approximately 7×10^8 inelastic collisions per second. The latter means that a bunch crossing (every 25 ns) leads to an average of $\sim$17 concomitant collisions. These additional collisions happening at the same time than the collision of interest are called pile-up (PU). Each of these collisions produces several particles flying through the detector, especially in the direction close to the beam pipe. These extreme conditions imply very

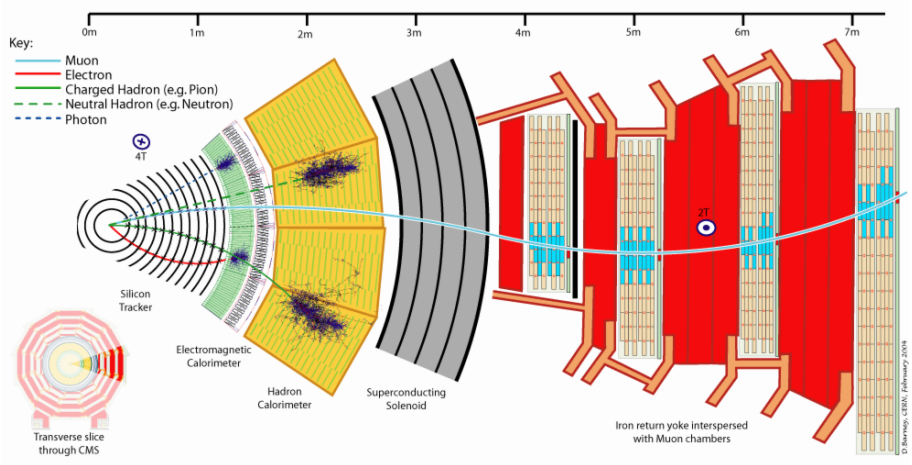


Figure 2.3: Transverse slice of the CMS detector, showing the individual detector subsystems and the particle signatures in each of them. The particle type can be inferred by combining the detector response in the different subcomponents.

strict requirements about resistance to radiation and timing response on the CMS design.

The layout of the detector shown on Fig. 2.3 is driven by the type of long lived particles one can detect. These particles are hadrons, electrons, muons and photons. The first layer of the detector encountered by the collisions products is a highly granular tracker which allows to reconstruct the charged particles trajectory without significantly affecting their momenta. A high magnetic field permeates the volume in order to bend the particle paths and infer their momentum by studying their trajectory. The second layer is the electromagnetic calorimeter which measures the energy of electrons and photons. The next layer is the hadron calorimeter which is designed to determine the energy of particles made of quarks. Finally, among the above-metioned stable particles, the only one that generally survives the passage through the calorimeters is the muon to which is dedicated the last layer of the detector called muon chambers. Before going through a more detailed description of each detector layer, let us establish our coordinate system.

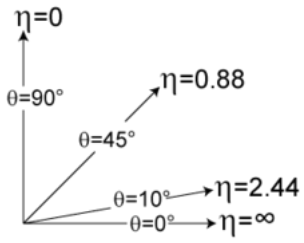


Figure 2.4: Visualization of the relation between the polar angle θ and the pseudo-rapidity $\eta \equiv -\ln(\frac{\theta}{2})$.

2.2.1 Coordinate conventions

The coordinate system adopted in CMS has the origin at the nominal collision point inside the detector. The x-axis points radially towards the LHC center, the y-axis points vertically upward while the z-axis is defined pointing along the beam direction towards the Jura mountains forming a right-handed coordinate system. The azimuth angle ϕ is measured from the x-axis in the xy-plane and the polar angle θ is relative to the z-axis. We define the convenient pseudo-rapidity quantity $\eta \equiv -\ln(\frac{\theta}{2})$, and provide the reader with a visualization of the relation between θ and η on Fig. 2.4. Finally, any quantity measured in the transverse plane (xy-plane) will be labeled with a subscript T e.g. the transverse momentum is noted p_T and the missing transverse energy will be noted $E_{\mathrm{T}}^{\mathrm{miss}}$.

2.2.2 Magnet

The magnet is a key element of the CMS detector as it allows, through the bending of the trajectories, to measure the energy of muons which are invisible to the calorimeters. This bending also improves other particles momentum resolution by making available a measurement complementary to the one realized with the calorimeters.

The magnet consists of a solenoid in NbTi material cooled down to 4.5 K, a temperature at which the material is in superconducting state. It is located between the calorimeters and the muon chambers. This location allows to have

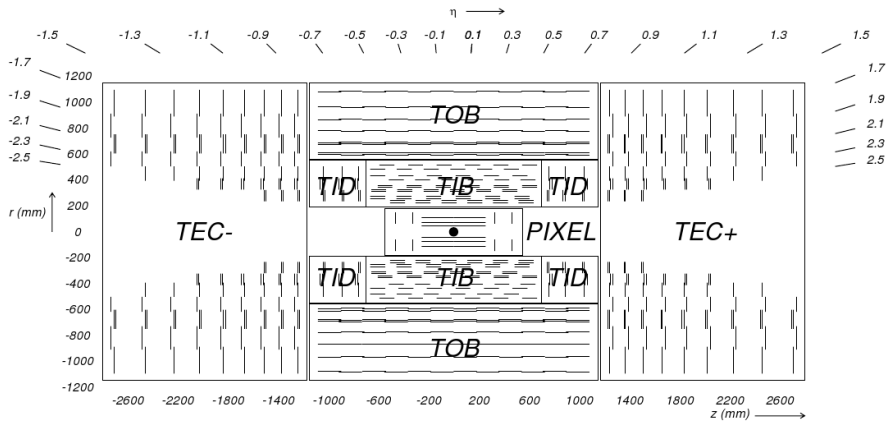


Figure 2.5: Longitudinal cross section of the CMS tracker showing the geometric arrangement of its sub-components. The black dot represents the interaction point.

a strong (3.8 T) and constant magnetic field inside the tracker and calorimeter volumes. The field in the muon chambers is about 2 T and has an inverted direction. A more thorough description of the CMS magnet can be found elsewhere [39].

2.2.3 Tracking system

The CMS tracker [40] is the sub-detector that lies closest to the beam line. It is an all-silicon detector with a sensitive area of about 200 m^2 covering a pseudo-rapidity up to 2.5 as shown on Fig. 2.5. The purpose of this detector is to measure the trajectories of charged particles allowing precise momentum measurement and determination of interaction vertices as well as secondary-vertices originating from long-lived particle decay such as B-hadrons.

It is made of two independent entities, the Strip detector and the Pixel detector. The basic components are made of a bulk of an n-type semi-conductor on top of which is placed an array of p-type pixels or strips. When a charged particle passes through the n-bulk it creates electron-hole pairs which migrate towards the pixels/strips leading to a signal in the electronic readout called a hit.

The Pixel detector is made of three barrel layers (BPix) distant of 4.4, 7.3 and 10.2 cm from the beam pipe and of four forward disks (FPix) at coordinates $\pm$ 34.5 and $\pm$ 46.5 cm on the z-axis. The BPix together with FPix are made of 1440 modules sharing 66 millions of pixels with a size of 100×150 µm. It has a very good spatial resolution (10 µm in the $r \times \phi$ plane and 20 µm in the z-direction) which makes it the key player in vertex position determination.

The Strip detector, surrounding the Pixel, is made of four inner barrels (TIB), six inner disks (TID), six outer barrels (TOB) and nine end-cap disks (TEC) at each side of the TOB. This sub-detector is made of 15 148 modules with 9.6 million strips. The charges produced by crossing particles are collected on several strips, increasing thus the resolution. The overall achieved resolution is $\sim$ 30-40 µm in the $r \times \phi$ plane and $\sim$ 230-300 µm in the z-direction.

From these hits one reconstructs tracks representing the particle trajectory as discussed in Sec. 2.3.1.

2.2.4 Electromagnetic calorimeter

The electromagnetic calorimeter [41] (ECAL) is the detector responsible for the energy measurement of photons and electrons. Its design had to face the strict LHC conditions with a high magnetic field, high level of radiation and only 25 ns between each collision.

The functioning principle is the following: at high energy (>100 MeV), electrons mainly loose their energy via *bremsstrahlung* emission of photons. In matter, high energy photons convert into electron-positron pairs which will in turn radiate photons, etc. This phenomenon called electromagnetic shower is triggered when a photon or an electron enters the calorimeter. The material used is a lead tungstate crystal ($PbWO_4$) which has the interesting properties of having small radiation length (0.89 cm) and Molière radius ($\sim$2 cm)[1], being transparent and to scintillate when electrons and photons pass through it. Thanks to the scintillating property, light is emitted consequently to the electromagnetic shower. The crystals have been design in order to contain the whole

[1] These two quantities are characteristic constants of materials quantifying respectively the mean longitudinal distance over which a high energy electron looses all but $\frac{1}{e}$ of its energy and the the radius of a cylinder containing on average 90% of the shower's energy.

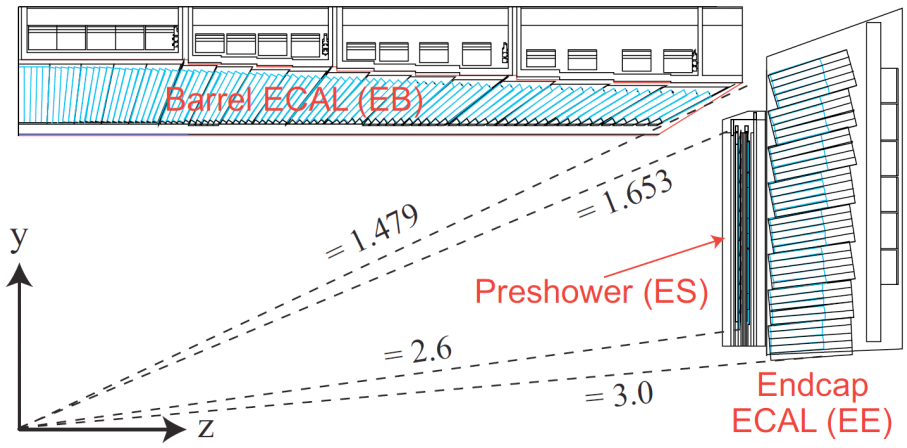


Figure 2.6: Longitudinal cross section of one quarter of the CMS electromagnetic calorimeter.

electromagnetic showers and photo-detectors have been glued on their back to collect the resulting light emission. With a proper calibration, the collected light can be related to the energy of the incoming particle.

The ECAL is made of two sub-components: the barrel which is made of 61 200 $PbWO_4$ $22 \times 22 \times 230$ mm crystals covering pseudo-rapidity $|\eta| < 1.479$ and the end-caps with ~15 000 further $30 \times 30 \times 220$ mm crystals covering pseudo-rapidity up to 3 (see Fig. 2.6.) A module with a much higher granularity (2 mm wide cells) called Preshower is placed in front of the end-caps to help differentiating π^0 decaying into two co-linear photons from prompt photons.

At energies $\mathcal{O}$(GeV), relevant at the LHC, the resolution can be parametrized as:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{a}{E}\right)^2 + \left(\frac{b}{\sqrt{E}}\right)^2 + c^2 \tag{2.3}$$

with E in GeV. This formula shows that the higher the energy, the better the relative resolution.

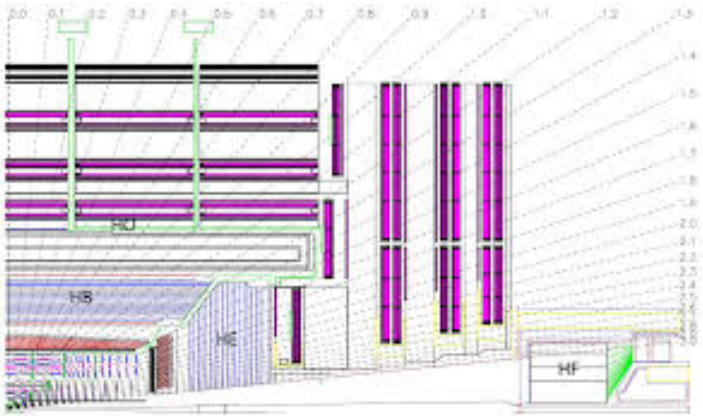


Figure 2.7: Transverse view of one quarter of the CMS HCAL with pseudo-rapidity benchmarks shown as dashed lines.

2.2.5 Hadron calorimeter

The hadron calorimeter [42] (HCAL) measures the energy of long lived hadrons, particles made up of quarks and gluons. When hadrons enter the HCAL material, they interact with the nuclei, exciting them and/or producing new particles which will either decay or also interact with medium nuclei. This process called hadronic shower differs from electromagnetic shower by the variety of processes that are in play. In general, a hadronic shower contains an electromagnetic component as e.g. π^0 can be produced and will very often decay into two photons.

The HCAL mostly consists of alternating layers of 5 cm brass "absorbers" which trigger the hadronic showers followed by 5 mm plastic scintillators which measure the energy. It is organized into four sub-components: the inner and outer barrels (HB and HO), the end-cap (HE) and the forward section (HF) (see Fig. 2.7). The HB and HO goes to pseudo-rapidity up to 1.3, HE covers $1.3 \leq |\eta| \leq 3$ and HF extends the pseudo-rapidity coverage to 5.2.

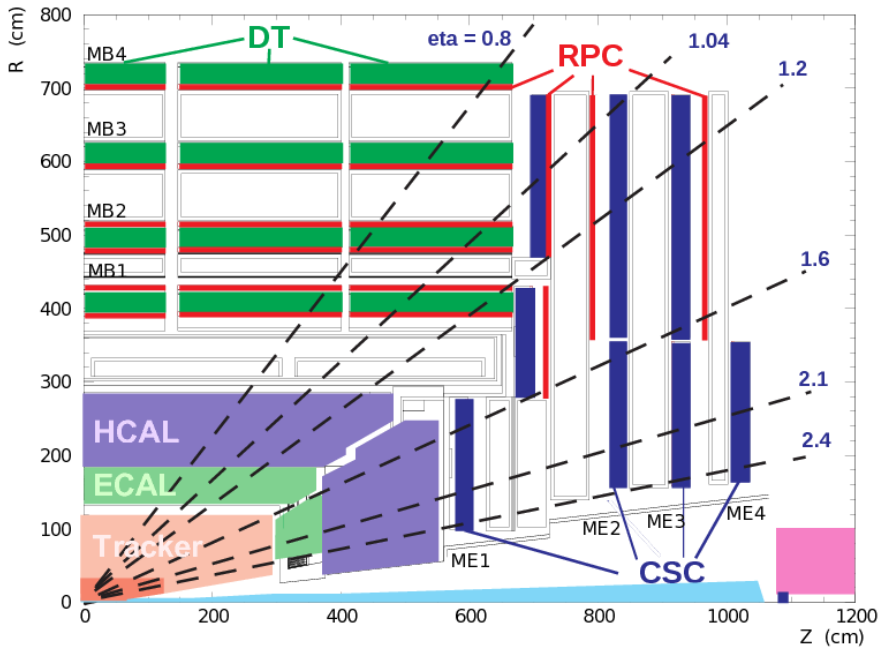


Figure 2.8: Transverse view of one quarter of the CMS muon chambers with pseudo-rapidity benchmarks shown as dashed lines. The place of the different chamber technologies (DT in green, RPC in red and CSC in blue) is also shown.

2.2.6 Muon chambers

Muons are, with the neutrinos, the only prompt particle likely to reach the last layer of the CMS detector: the muon chambers [43]. These chambers are made of three different technologies and have a total pseudo-rapidity coverage corresponding to $|\eta| < 2.4$ (see Fig. 2.8). Their role is to assess the presence of muons in the event and to complement the tracker in the reconstruction of their trajectories.

The Drift Tube (DT) system measures the muon position in the barrel region of the detector ($|\eta| < 1.2$). Each tube is 4 cm-wide and contains a positively charged wire within a gas volume. When a muon passes through the tube, it knocks electrons off the atoms of the gas. These electrons follow the electric

field until the wire and leave an electric signal when reaching it. Two coordinates for the muon position can thus be extracted: one is given by the position of the hit along the wire and the other is obtained from the time taken by the electrons to drift until the wire. The DT chambers are arranged such that the three muons coordinates can be extracted by crossing information from the different chambers.

Cathode Strip Chambers (CSC) are used in the end-cap disks where particle rate is higher $(1 \leq |\eta| \leq 2.4)$. They consists of arrays of positively charged wires crossed with negatively charged copper strips within a gas volume. The functioning principle is the same as for DT except that positive ions are also detected.

Finally, Resistive Plate Chambers (RPC) are placed in the whole muon system to complement information from DT and CSC with fast measure of the muon momentum very helpful for the trigger system. RPC's consist of two parallel plates, a positively-charged anode and a negatively-charged cathode, both made of plastic material and separated by a gas volume.

2.2.7 Simulation

As mentioned in the previous chapter, one resorts to simulation in order to confront theoretical predictions to experimental data. We have seen in this section all the hardware that is used to detect the product of the pp collisions which, obviously, also has to be simulated to predict what will be observed under a given hypothesis.

We have seen at the end of Sec. 1.2.1 that the hard process, showering and hadronization were handled by the MC generators which provide the four-momenta of the "stable" particles resulting from the process we simulate. We will call these particles the "gen-level particles", in opposition to the "reco-level objects" that are the result of the detector reconstruction. To predict what is actually observed, the outcome of the event generation is exploited by a program that simulates the passage of such particles through the CMS detector. This tasks is handled by the GEANT4 [44] package. Based on their four-momenta, particles are traced through a simulated version of the CMS detector, modeling the passage of particles through matter and the generation

of secondary particles resulting from interactions with the detector material. The hits left by these particles are provided by simulating the response of the front-end electronics. After this step, the event reconstruction chain in simulation is very similar to the one for real data. The next section is dedicated to the description of this event reconstruction.

2.3 Particle reconstruction

We have seen in the previous section that each sub-detector of CMS focuses on a specific task. We have also seen that each particle leaves a specific signature across this variety of sub-detectors. For instance, electrons leaves in general signatures only in the Tracker and in the ECAL while muons are visible in the Tracker and in the Muon Chambers. One can thus try to identify the particle species by crossing information from various sub-detectors. The reconstruction of the particle four momentum is also performed using a thorough combination of all relevant CMS sub-detectors. This reconstruction technique called *Particle Flow* [45] is applied to reconstruct all the objects used in this thesis. In the coming sections, one shall use *PF particle* to refer to all the physical objects resulting from this reconstruction technique.

In Sec. 2.2, we have always referred to the *typical* behavior of particles in the detector while in reality we are dealing with probabilistic processes. Therefore, translating a detector signature into particle label is absolutely not straightforward since the signature of one type of particle can significantly vary from one case to another. Moreover the *typical* signature of a specific particle can be mimicked by other type of particles. As a consequence, the object definition based on detector information is always a trade-off between purity and efficiency.

In this section one will present the reconstructed object definition relevant for detecting the llbb final state. We will start from low level objects such as tracks and vertices to more complicated objects such as b-jets.

2.3.1 Tracks and vertices

Track reconstruction [46, 47] refers to the process of using the hits described in previous section to obtain the trajectories of charged particles allowing to infer their momentum. It can be decomposed in four main logical parts:

- **Seed generation** which provides initial track candidates and parameters using only 2 or 3 hits from the innermost tracker layers.
- **Trajectory building** where the track seeds are extrapolated outwards taking into account the magnetic field and the possible interactions with the material. Accounting for the uncertainties on this extrapolation, one looks for hits compatible with the extrapolated trajectory. If such a hit is found, one adds it to the trajectory and update the uncertainties on the track parameters. This process continues until no valid hit is found or the tracker material has ended. The procedure has parameters that can be tuned depending on the purpose such as the minimum track p_T, the allowed number of consecutive invalid hits or the minimum number of hits required to form a track.
- We define a track candidate with five parameters: d_0 which is the distance in the transverse plane between the origin and the impact point (point of closest approach between the track and the beam axis), z_0 which is the longitudinal coordinate of the impact point, ϕ is the azimuth angle of the track at the impact point, θ is the corresponding polar angle and p_T, the transverse momentum. **Track fitting** is the step where one extracts these five parameters based on the associated hits and their uncertainties.
- **Ambiguity resolution** is necessary as the trajectory building step may reconstruct a given track starting from different seeds or may reconstruct several track candidates starting from the same seed. To mitigate this, one applies a criteria on the number of shared hits :

$$f_{shared} = \frac{N_{shared}^{hits}}{min(N_1^{hits}, N_2^{hits})} \tag{2.4}$$

  where $N_{1(2)}^{hits}$ are the number of hits in the first (second) track candidate. If f_{shared} exceeds 0.5 the track with the least number of hits is discarded.

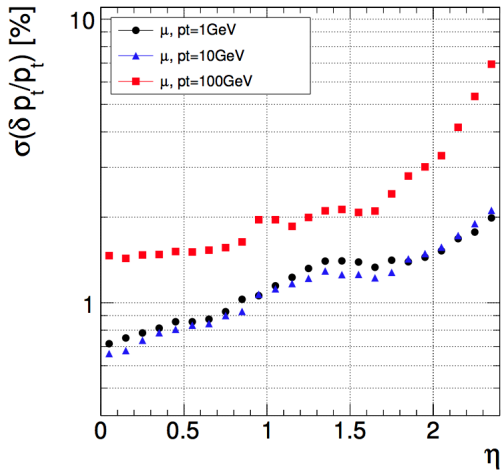


Figure 2.9: Resolution on transverse momentum as a function of speudo-rapidity (η) for single muon simulations with transverse momentum of 1 (black dots), 10 (blue triangles) and 100 GeV (red squares).

> If both track candidates possess the same number of hits, the one with the highest χ^2 per degree of freedom in track fitting step is discarded.

For illustration, Fig. 2.9 shows the resolution on the quantity of most relevance for this analysis, the transverse momentum p_T, as a function of pseudo-rapidity and for different p_T magnitudes. One can see that the resolution is typically of the order of the percent and worsen towards higher pseudo-rapidity or towards too high transverse momentum.

The reconstructed tracks can be grouped together to build a higher level quantity called vertex [48]. Indeed, when several particles are produced at the same place their tracks should cross at the point where they originate (within uncertainties). The resolution on the vertex position is typically of the order of 10 to 100 µm depending on the number of tracks associated to it. We distinguish two types of vertices: *primary vertices* which are believed to originate from the interaction point of a pp collision and *secondary vertices* which arise when particles with sufficient lifetime decay in the detector volume. Since, in general, multiple pp interactions occur per bunch crossing several primary vertices are reconstructed per event.

In order to make truthful predictions, the event simulation mentioned in Sec. 1.2.1 and 2.2.7 is superimposed with low momentum transfer pp collisions. The number of such collisions to be superimposed is randomly chosen according to a Poisson distribution of mean equal to the number of expected interactions. The latter is computed based on the total inelastic pp cross-section together with the instantaneous luminosity. We face here an unavoidable difficulty: at the moment of the MC event generation one generally does not know exactly what will be the instantaneous luminosity delivered by the LHC (except for post data taking MC generation). Indeed, as one can see on the left hand side distribution from Fig. 2.10, the number of reconstructed primary vertices in the event subset containing two leptons and two b-jets shows tensions between data and predictions.

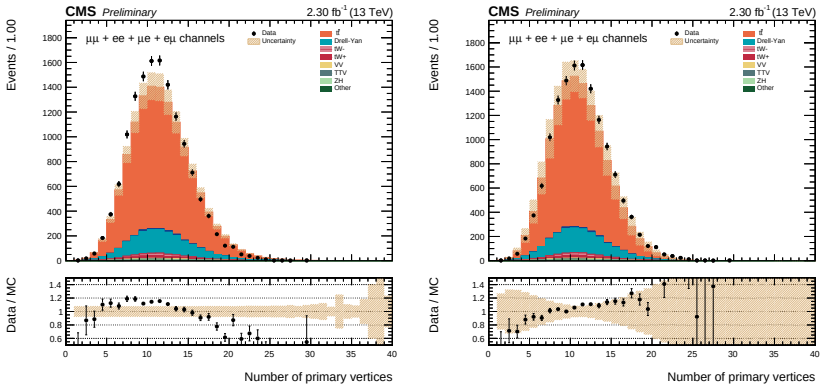


Figure 2.10: Number of reconstructed primary vertices for data (black dots) and simulation at $\sqrt{s}$ =13 TeV in the event subset containing at least two leptons and two b-jets. MC simulations (see Sec. 1.2.1 and 2.2.7) are represented by solid histograms and are stacked together. The $\mathrm{t\bar{t}}$ process corresponds to orange, Drell-Yan to light(dark) blue for $m_{\mathrm{ll}} > (<) 50\,\mathrm{GeV}$, tW to burgundy, di-boson (VV) to beige, $\mathrm{t\bar{t}}$ V to grey, Zh to light green and the other SM processes with very small contributions such as $\mathrm{t\bar{t}}$ h or Wh are in dark green. Dashed brown bands represent uncertainty on the SM MC expectations. The left(right) hand side plot shows the distribution before(after) applying the pile-up event reweighting.

To cope with this, one applies an event-by-event reweighting based on the instantaneous luminosity that was actually delivered and the total pp cross section (69 mb with an uncertainty of 5% [49]). The number of primary vertices distribution resulting from this event reweighting is shown on the right hand side of Fig. 2.10. One sees that, though the data to MC ratio is still not flat, the differences are covered by the uncertainty on the total inelastic pp cross-section.

2.3.2 Electrons

As mentioned in the previous section, electrons leave a signature both in the tracker and in the electromagnetic calorimeter. Due to the material budget present between the beam spot and the ECAL, electrons are likely to emit bremsstrahlung photons before to reach the latter. These photons in turn can create e^+/e^- pairs and so on. The produced shower generally follows the electron trajectory and is detected in the ECAL as several adjacent crystal hits (cluster). Therefore electron candidates are obtained by associating a track to a cluster of energy deposit in the ECAL [50].

Once all electron candidates have been reconstructed, one has to apply further identification (ID) and isolation (ISO) criteria in order to discriminate prompt isolated electrons (the one of interest for the llbb final state) from those that result from photon conversion, B-hadron decays, misidentified jets, etc. We can split these discriminating variables into three categories:

- Calorimetric observables such as the ratio of the energy deposit in the HCAL over the one in the ECAL or the shape of the ECAL deposit are exploited to reject jets with large electromagnetic shower component.
- Genuine electrons within jets coming from e.g. semileptonic decay of B-hadrons will in general have significant energy flow near their trajectories. To greatly reduce these unwanted electrons, we apply an upper threshold on the relative amount of energy contained inside a cone of size R=0.3 around the electron direction.
- Secondary electrons produced from photon conversions inside the Tracker material are rejected by applying criteria on the track pattern. Indeed,

their tracks usually have big impact parameters since they start further away from the beam pipe and are characterized by missing hits in the innermost layers of the Tracker.

A set of cuts on the various quantities defined here-above is called a *working point* (WP). The one we work with in this thesis is the so-called *Tight* WP characterized by an efficiency of 70% for prompt electrons [51].

In order to account for the different efficiency of ID and ISO selection between data and MC simulations, a reweighting of MC events is applied. The scale factors are applied electron by electron and are derived with the *tag-and-probe* method by the *Higgs to WW* group [51]. The uncertainty on these scale factors leads to a sub-dominant systematic effect in this thesis.

2.3.3 Muons

As muons produce hits both in the Tracker and Muon Chambers, they are reconstructed making use of both sub-detectors. Muons are defined as tracks in the Muon Chambers which match tracks in the tracking system or as tracks in the tracking system matching muon segments in the chambers. The extrapolation from one subsystem to the other is done taking into account the expected energy loss and the uncertainty due to multiple scattering. If one of the mentioned matching criteria is fulfilled, a global track fit is performed using hits both from Tracker and Muon Chambers.

The main sources of backgrounds for prompt muons are coming from heavy flavor particle decays producing real muons and the so-called *punch-through* fake muons which are particles passing through the calorimeter and producing hits in the muon chambers. As for the electrons, we apply several selection criteria to reduce these backgrounds. The most interesting quantity are the relative amount of energy around the muon direction, the goodness of the global fit mentioned above, the compatibility between Tracker and Muon Chambers track segments and the impact parameter of the global track.

As for electrons, scale factors provided by *Higgs to WW* group [51] are applied to reweight MC events in order to absorb the efficiency difference between data and MC. The uncertainty on these scale factors is treated as a systematic error.

2.3.4 Jets

The goal of the jet reconstruction is to identify quarks and gluons corresponding to external legs of the hard scattering process. As mentioned is Sec. 1.2.1 partons produced during the hard interaction do not propagate freely to reach the CMS detector, they instead radiate other partons and forms a myriad of hadrons. Some of these hadrons decay into lighter hadrons, leptons or photons. A jet corresponds therefore to a spray of collimated particles. In practice, we use jet reconstruction algorithms which define a set of rule for grouping these particles together inside a cone of predefined size. The algorithm applied for jet reconstruction in this thesis is the anti-k_T [52] with a cone size of 0.4.

The clustering algorithm is applied event by event and takes as input all the PF particles but the one associated to charged hadrons identified as coming from PU interactions. Purity criteria based in particular on the hadronic/electromagnetic energy deposit fractions are applied in order to reject fake jets while keeping an efficiency of $\sim$99 % for real jets.

The obtained jet energy is scaled (JES) to subtract the PU contribution in the jet cone and to take into account the non-uniform detector response in pseudo-rapidity and transverse momentum. This last correction is derived on MC simulation by comparing the gen-level jet energy to the reconstructed jet energy. The remaining scale difference between data and MC jet energy response is taken into account by a residual correction [53]. One more correction is applied to MC jets to cope for the fact that the jet energy resolution (JER) is different for the real and simulated detector. The derivation of these corrections is impacted by uncertainties which are treated as sources of systematic errors on this thesis results.

b-tagging

In this thesis we are mostly interested in jets coming from the hadronization of b-quarks as they are part of the llbb final state. We want thus to be able to differentiate b-jets and jets coming from the hadronization of lighter quarks. This technique known as b-tagging exploits the particularity of B-hadrons with respect to other hadrons such as their longer lifetime (leading to a secondary vertex as explained in Sec. 2.3.1) or their mass.

The b-tagging algorithm used in this thesis is called *CSVv2* and corresponds to an artificial neural network which is fed with quantities such as the impact parameter significance[2] of the tracks associated to the jet or the secondary vertex mass. These two quantities are shown on Fig. 2.11 for events passing a trigger selection requiring the presence of at least one jet with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$. These plots were produced in the context of the commissioning of b-tagging algorithms [54] which represents an important part of the service work achieved during this thesis. One sees that, as suggested in previous paragraph, the jets coming from the hadronization of b-quarks have tracks with a higher impact parameter significance and secondary vertices which are characterized by a higher invariant mass compared to other jets.

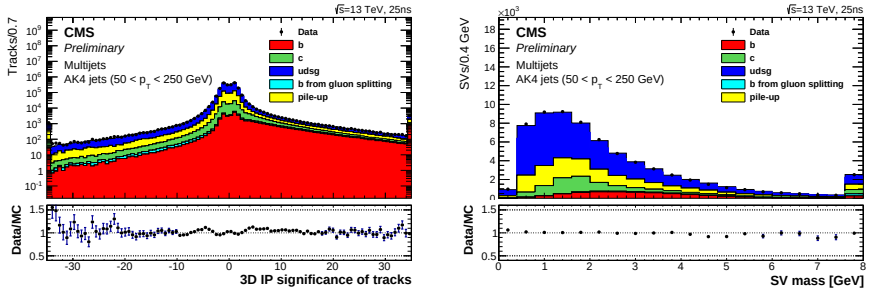


Figure 2.11: Example of variables used as input to train the CSVv2 b-tagging algorithm for events passing a trigger selection requiring the presence of at least one jet with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$. The distributions are made with jets verifying $50 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ and show the contributions from different jet flavor separately. On the left hand side one shows the three-dimensional impact parameter significance of all the tracks associated to the jet while the right-hand side distribution shows the mass of the jet secondary vertices. The total number of entries in the simulation is normalized to the observed number of entries in data. Underflow and overflow are added to the first and last bins, respectively.

The distribution of the CSVv2 discriminant in events containing two leptons and two jets (without requiring them to come from a b-quark) is shown on Fig. 2.12 for the leading jet (defined as the jet having the highest CSVv2 discriminant). One sees that $\mathrm{t\bar{t}}$ events lie mainly at high CSVv2 values which

[2] Significance refers here to the value divided by its error.

reflects the fact that they contain two b-quarks in their final state. The jets present in DY events entering the selection mentioned here-above are of any kind which explains why this process is characterized by lower CSVv2 scores. DY events lying at high CSVv2 scores mainly correspond to $Z + b\bar{b}$ events.

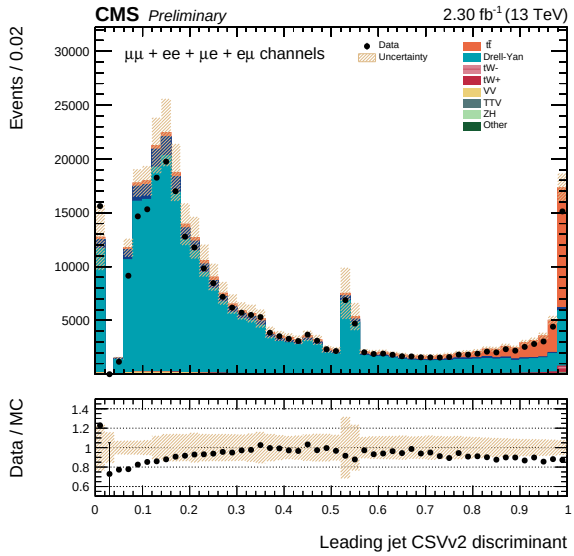


Figure 2.12: CSVv2 discriminant for the jet with highest CSVv2 in events containing at least two leptons and two jets. Jets without a selected track and secondary vertex are arbitrarily assigned a discriminator value of -10. The rightmost bin includes all events from the overflow, while the underflow is added to the first bin. No b-tagging MC correction is applied on this distribution.

Once the b-tagger discriminant is defined, one has to choose a working point to define what we call a b-tagged jet. Three distinct working points are centrally provided by the CMS collaboration: Tight (T), Medium (M) and Loose (L) which are respectively defined as leading to 0.1, 1 and 10 % of mis-tagging rate while keeping a typical efficiency between about 50% and 80% for b-jets [54]. The Medium working point, corresponding to CSVv2 discriminant value higher than 0.8, has been chosen to define b-tagged jets in this thesis.

As one can see on Fig. 2.12, the predictions do not perfectly match the observations which induces a difference in b-tag selection efficiency between data and MC simulations. In order to account for this difference, jet by jet scale factors are applied to reweight MC events. These scale factors are derived in bins of p_{T} and η and are provided separately for jets coming from light-quarks (g, u, d, s) and for jets coming from c or b-quarks. The typical value of these scale factors is around 0.95 for the chosen working point. The uncertainty on these scale factors is treated as a source of systematic error affecting this thesis' results.

2.3.5 Missing transverse energy

Thanks to the almost hermetic coverage of the CMS detector, one can assume in first approximation that any detectable particle created during a pp collision will leave a trace in the detector. Since the initial state protons have no transverse momentum when they interact, the outcome of the collision (seen has a whole) can not have any transverse momentum by energy-momentum conservation law[3]:

$$\vec{p}_T^{\,tot} = \left(\sum_{i \in \{\text{all particles}\}} \vec{p_i} \right)_T = \vec{0}. \tag{2.5}$$

where $\vec{p}_T^{\,tot}$ has to be understood as a 2D vector in the xy plane. On the other hand, if a particle created during the collision does not interact with the detector, the total visible interaction products will have a non-zero transverse momentum:

$$\vec{p}_T^{\,visible} = \left(\sum_{i \in \{\text{visible particles}\}} \vec{p_i} \right)_T \neq \vec{0}. \tag{2.6}$$

The only SM particles behaving as such are the neutrinos but many BSM models, such as the ones introducing dark matter candidates, predict more particles escaping the detector without interacting. Based on the above observations, the

[3]The same statement does not hold for the longitudinal momentum p_Z^{tot} since the interacting partons inside the proton can carry different energy fractions.

invisible decay product(s) transverse momentum ($\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$) can be inferred from the visible final state particles via

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \equiv -\vec{p}_{T}^{\,visible}. \tag{2.7}$$

Of course, when several invisible particles are created during the collision, their individual transverse momentum is inaccessible via this method. Finally, one can introduce the notion of *missing transverse energy*:

$$E_{\mathrm{T}}^{\mathrm{miss}} \equiv ||\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}||. \tag{2.8}$$

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is one of the most complicated object to be experimentally reconstructed as it directly depends on all the other objects. In practice, we define it as the negative vectorial sum over the transverse momenta of all PF particles [55]. An event with no invisible particle in the final state would thus never lead to zero $E_{\mathrm{T}}^{\mathrm{miss}}$ due to the finite precision of the detector and to particles outside of acceptance. Note that the jet energy corrections defined in Sec. 2.3.4 as well as their associated uncertainties are propagated to the $E_{\mathrm{T}}^{\mathrm{miss}}$. The distribution of this quantity for our $\mathrm{llbb} + \mathrm{X}$ final state is well reproduced by MC simulations as shown on Fig. 2.13. As expected, processes with neutrinos in the partonic final state such as $\mathrm{t\bar{t}}$ lie in general at high $E_{\mathrm{T}}^{\mathrm{miss}}$ values. The DY process has essentially no "real" source of $E_{\mathrm{T}}^{\mathrm{miss}}$ (except for the neutrinos inside the jets or coming from tau lepton decays), its distribution lies thus at lower $E_{\mathrm{T}}^{\mathrm{miss}}$ value and illustrates the $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution.

2.3.6 Typical llbb event

Now that we have covered the reconstruction of all the objects that are relevant to detect the $\mathrm{llbb} + \mathrm{X}$ final state, let us provide the reader with a visualization of a complete event reconstruction. To this end, one shows on Fig. 2.14 the display of en event that has been recorded during the 2015 data taking period and that has been analyzed within this work. One sees on the top right corner a track segment in the muon chamber that is matched to a track in the inner volume of the detector and which is therefore associated to the presence of a muon, shown as a red line. On the middle right, one sees an important electromagnetic energy deposit which is associated to a track and which is free

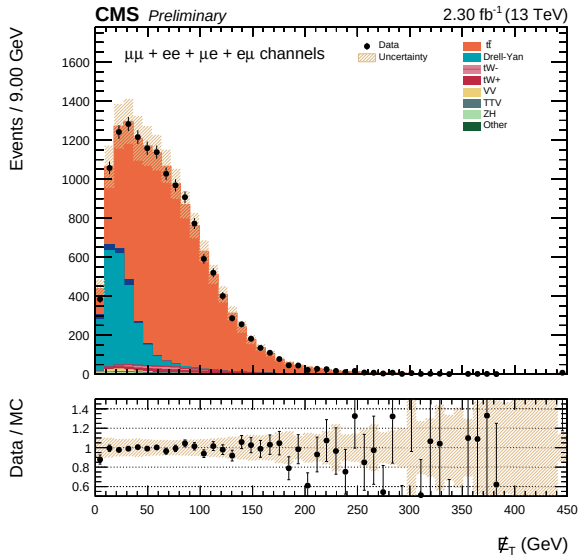


Figure 2.13: Distribution of the missing transverse energy in event containing at least two leptons and two b-jets for data (black dots) and MC simulations (colored stacked histogram). All the MC reweigthing corrections are applied on this distribution.

of hadronic activity. This is the typical signature of an electron. The length of the purple arrow gives us information about the $E_{\mathrm{T}}^{\mathrm{miss}}$: this particular event is characterized by a $E_{\mathrm{T}}^{\mathrm{miss}}$ of 85 GeV indicating the presence of neutrino(s). Finally, two jets have been reconstructed as shown by the two yellow cones encompassing tracks and pointing towards both hadronic and electromagnetic deposits. The event shown here has been reconstructed with 12 additional primary vertices which are likely to come from PU interactions and lead to extra tracks or low energy calorimeter deposits. The reconstructed event exhibits thus a muon, an electron, two jets and an important amount of $E_{\mathrm{T}}^{\mathrm{miss}}$, which is the typical signature of a $\mathrm{t\bar{t}}$ event.

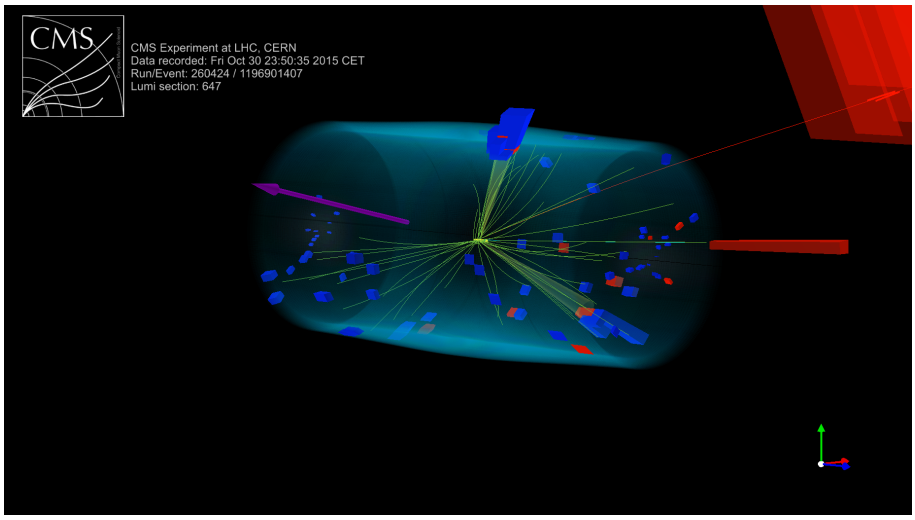


Figure 2.14: Display of a typical llbb event recorded by the CMS experiment during the 2015 data taking era. Tracks are represented by yellow lines, electromagnetic(hadronic) calorimeter deposit by red(blue) blocks and the missing transverse energy is shown as a purple arrow. The event shown belongs to the $e\mu$ category and has been analyzed in this work.

2.4 Trigger System

As mentioned previously, the LHC delivers events at a rate of 40 MHz in nominal conditions. With the current technologies, there is no way to store the information of every event occurring in CMS. However, most of these events correspond to soft QCD interactions which are unlikely to reveal the new phenomena we are looking for in this thesis. We can therefore afford to drop most of them and keep only potentially interesting events. The trigger system is responsible of this sorting.

The CMS trigger system is made of two parts: the so-called Level 1 Trigger (L1) which reduces the event rate from 40 MeV to 100 kHz and the High Level Trigger (HLT) which cut it further down to 100 Hz. Since the decision whether to keep an event or not has to be taken in a short amount of time, the object reconstruction at the trigger level (online) uses less information than the

one which is used to provide the final datasets (offline), resulting in a worse resolution.

The **L1 trigger** [56] is composed of custom hardware processors and uses only information from the calorimeters and muon detectors.

The **HLT** [57] receives only events passing the L1 trigger requirements which allow for a more advanced event reconstruction. It consists of a farm of CPU's performing algorithmic operations to take the final decision on whether the event will be kept for physics analysis or not.

The analyses presented in this thesis study the llbb final state. Given that the jet rate at LHC is much higher than the electron or muon rate, events are collected based on a set of triggers requiring the presence of two leptons (electron or muon). We use the most inclusive unprescaled[4] trigger which require transverse momentum $p_{\mathrm{T}} > 17\,\mathrm{GeV}$ for the first lepton – called leg 1 – and $p_{\mathrm{T}} > 12(8)\,\mathrm{GeV}$ for the second electron(muon) – called leg 2.

Due to the coarser precision of the trigger level object reconstruction, offline events that do not match these requirements may still have fired it. On the other hand, offline events that would have fulfilled these criteria are not always selected. To take these effects into account in our simulations, one reweights the MC events by the efficiency of a given lepton configuration to fire the required trigger.

The trigger efficiencies are calculated leg per leg, from data, based on the *tag-and-probe* method [58] and are centrally provided by the "$h \to WW$" CMS group [51] as a function of p_{T} and η. To illustrate the importance of applying this prescription to the simulations, one shows on Fig. 2.15 the efficiency of the electron trigger leg 1 ($p_{\mathrm{T}} > 17\,\mathrm{GeV}$) as a function of the offline electron transverse momentum. The plateau ($p_{\mathrm{T}} > 30\,\mathrm{GeV}$) reaches approximately 98% and the turn on ($p_{\mathrm{T}} \sim 17\,\mathrm{GeV}$) rapidly moves from 20 to 80% efficiency while not reweighting the MC means assuming 100% efficiency over the whole p_{T} range (introducing thus kinematic differences between data and MC for the variables correlated to the trigger efficiency). Note that these trigger efficien-

[4]Triggers with lower p_{T} thresholds exist but due to their too high rate, some events firing them must randomly be discarded which would artificially lower the analyzed luminosity.

cies are provided with an uncertainty which is treated as a source of systematic error in this work.

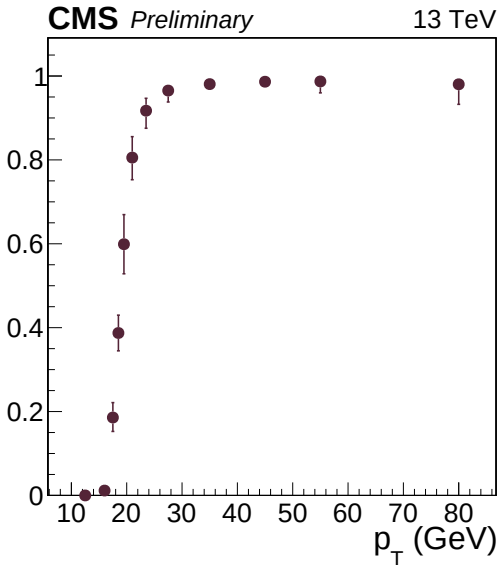


Figure 2.15: Efficiency as a function of transverse momentum for an electron to fire the trigger leg requiring $p_{\mathrm{T}} > 17\,\mathrm{GeV}$. This efficiency is measured from data based on the *tag-and-probe* method [58].

2.5 Determination of the transfer function

When discussing the MEM in Sec. 1.4, one introduced the transfer function, $T(\mathbf{x}|\mathbf{y})$, whose role is to bridge hard process partonic final $\mathbf{y}$ state to what is actually observed in the detector $\mathbf{x}$. This section discusses how one derives T in our final state.

As we have seen in previous sections, several effects take place between the hard process final state $\mathbf{y}$ and the detector level reconstructed event $\mathbf{x}$. Taking the example of a final state quark, T has to include showering, hadronization, clustering and detector resolution effects to properly build the probability den-

sity function of reconstructing a jet with kinematic $\mathbf{x}$ provided that it came from a hard process particle $\mathbf{y}$.

In practice, we derive the transfer function with simulated events by comparing the reconstructed objects to the *matched* hard process final state particles, assuming that T does not depend on the type of hard process considered. The matching between reconstructed and gen-level particles is performed based on the type of the particle (reco and gen-level particles are required to be of the same type) and on the angular distance $\Delta R < 0.2$ for b-quarks and $\Delta R < 0.1$ for leptons. To perform the matching, we use the collection of gen-level particles that have already undergone FSR radiations: a ΔR matching with gen-particles before FSR would often fail for particles emitting a radiation affecting the direction, biasing thus the transfer function. If we find a match, the hard process parton (before FSR) associated to it is unambiguously found by going back in the MC simulation history. This gen-level particle is the one compared to the reco-level object.

The llbb final state consists of three different types of objects: b-jets, electrons and muons. We work under the approximation that the *event* transfer function can be factorized into a product of *object* transfer functions

$$T(\mathbf{x}|\mathbf{y}) = T_b(\mathbf{x}_{b1}|\mathbf{y}_{b1})T_b(\mathbf{x}_{b2}|\mathbf{y}_{b2})T_l(\mathbf{x}_{l1}|\mathbf{y}_{l1})T_{l'}(\mathbf{x}_{l2}|\mathbf{y}_{l2}) \tag{2.9}$$

where T_b is the transfer function for b-jets and T_l the transfer function for leptons. Since the evolution from partonic final state to reconstructed object is different for electrons and muons, we derive separate transfer functions for the two objects; hence the $T_{l'}$ in Eq. (2.9).

We factorize further the object transfer function by separating the angular and energy components:

$$T_i(\mathbf{x}_i|\mathbf{y}_i) = T_i^E(E_i^{\text{reco}}|E_i^{\text{parton}})T_i^\eta(\eta_i^{\text{reco}}|\eta_i^{\text{parton}})T_i^\phi(\phi_i^{\text{reco}}|\phi_i^{\text{parton}}). \tag{2.10}$$

Considering the high granularity of the CMS detector, the angular transfer functions is approximated by Dirac delta function which leads to:

$$T_i(\mathbf{x}_i|\mathbf{y}_i) = T_i^E(E_i^{\text{reco}}|E_i^{\text{parton}})\delta(\eta_i^{\text{reco}}, \eta_i^{\text{parton}})\delta(\phi_i^{\text{reco}}, \phi_i^{\text{parton}}). \tag{2.11}$$

The probability density function $T_i^E(E_i^{\text{reco}}|E_i^{\text{parton}})$ is derived as a smoothened step function obtained from a $\mathrm{t\bar{t}}$ generated sample by filling a histogram with

the quantity: $\Delta E = E^{\text{reco}} - E^{\text{parton}}$. Since the relative resolution on the energy depends on the energy itself, we derive several step functions for different E^{parton} ranges. The binning in E^{parton} was chosen to be dynamic to allow for a good trade off between statistics and granularity in E^{parton}. Note that E^{parton} has been considered until 2000 GeV and that a prescription in MoMEMta ensures to take the transfer function corresponding to this extreme value in case an integration phase space point goes out of boundary.

As mentioned in Sec. 1.4, we normalize these step functions so that

$$\int T_i^E(E_i^{\text{reco}}|E_i^{\text{parton}})dE_i^{\text{reco}} = 1 \tag{2.12}$$

separately for each E^{parton} range. By doing so, we partially take into account the object selection efficiency coming from the cut applied on reconstructed quantities at the analysis level because we build the transfer functions without applying any of these cuts[5]. Let us illustrate this statement with the following. If one requires, at the analysis level, reconstructed electrons with an energy above 20 GeV, generated electrons at an energy of e.g. 15 GeV will have a certain probability, $\epsilon < 1$, to pass the selection cut. The transfer function for electrons with $E^{\text{parton}} = 15$ GeV will only be probed, at the analysis level, for $\Delta E > +5$ GeV. When building the transfer function without applying the cut on reconstructed electrons and normalizing it to unity as shown in Eq.(2.12), the integral of the part which is actually probed amounts to ϵ.

Figure 2.16 shows examples of T for various particles and E^{parton} ranges. By comparing the b-quark transfer functions for $45 < E^{\text{parton}} < 50$ GeV (top histogram) and for $145 < E^{\text{parton}} < 150$ GeV (middle), one sees how T_b evolves with E^{parton}. The relative energy resolution improves when going to higher energy. One also notices that the peak of the b-quark transfer function is not at zero. This is explained by the fact that the jet reconstruction technique omits the neutrinos inside the cone which are frequent for b-jets. By comparing the top histogram to the bottom one (transfer function for electrons with $45 < E^{\text{parton}} < 50$ GeV), one sees that the electron energy resolution is about a factor 10 better than the one for b-jets in this range of E^{parton}. The muon energy resolution has a behavior close to the electron one. The complete transfer

[5]This statement is not totally exact as CMS provides datasets including cuts on the various objects but they are in general much looser than the analysis cuts.

functions (including the full E^{parton} ranges considered) are shown on Fig. 5.1 from App. 5.1.1 separately for b-jets, electrons and muons.

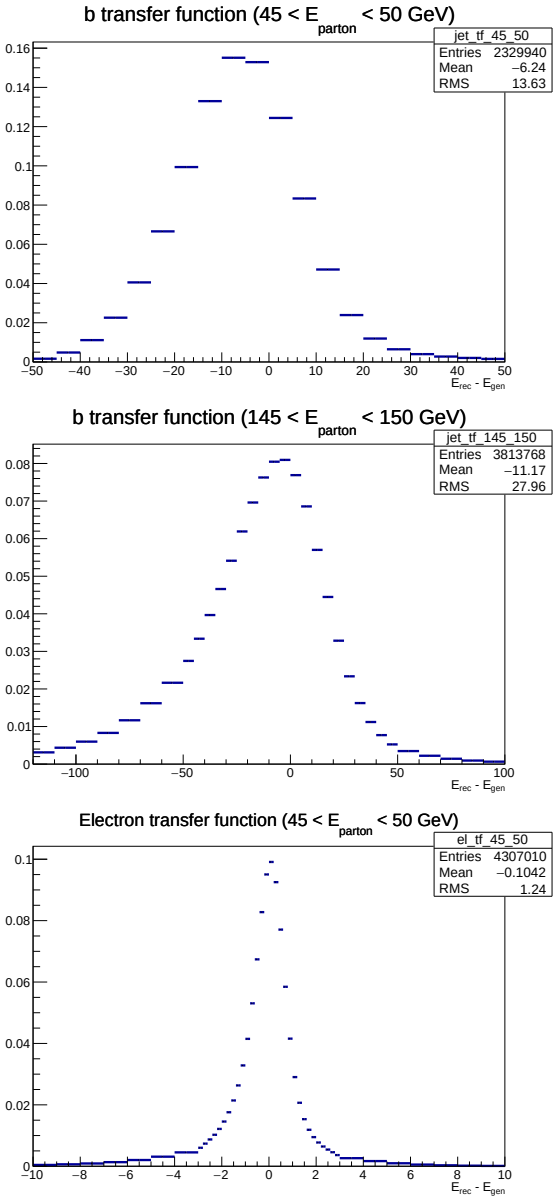


Figure 2.16: Energy transfer functions for b-quarks of energy $45 < E^{\text{parton}} < 50$ GeV (top) and $145 < E^{\text{parton}} < 150$ GeV (middle). The bottom plot shows the transfer function for electrons with $45 < E^{\text{parton}} < 50$ GeV. The x-axis shows $\Delta E = E^{\text{reco}} - E^{\text{parton}}$ and the y-axis is the value of the probability density function to reconstruct the object with an energy away of ΔE from the hard-process particle energy. Each histogram is normalized to unity to allow for probability interpretation.

Chapter 3

Search for resonant di-Higgs production decaying into $\mathrm{b\bar{b}}\ \mathrm{l}^+\nu_\mathrm{l}\ \mathrm{l}^-\bar{\nu}_\mathrm{l}$

The recent discovery of the Higgs boson by CMS [59] and ATLAS [60] opened numerous windows to search for physics beyond the SM. One of these new windows is the simultaneous production of two Higgs bosons. This process exists in the SM but is indeed enhanced in various BSM scenarios. CMS already published several 8 TeV analysis exploring this new window via the $\mathrm{b\bar{b}}\ \mathrm{b\bar{b}}$ [61], $\mathrm{b\bar{b}}\ \tau^+\tau^-$ [62] and $\mathrm{b\bar{b}}\ \gamma\gamma$ [63] decay channels.

The final state $\mathrm{b\bar{b}}\ \mathrm{l}^+\nu_\mathrm{l}\ \mathrm{l}^-\bar{\nu}_\mathrm{l}$ (that we will note $\mathrm{bb}\ \mathrm{l}\nu\mathrm{l}\nu$ for convenience) has a reasonable branching ratio (see next section for a more thorough discussion) but is expected to be challenging due to the impossibility of reconstructing the mass of the Higgs decaying to $l\nu l\nu$. However, in case new physics pops up in di-Higgs production, it should be consistently observed in all decay channels to draw reliable conclusions. Moreover, if the signal is expected to be small, one may need all the non-negligible decay channels in order to reach a sufficient significance.

This chapter describes the CMS public analysis [64] which corresponds to an important part of this thesis work. It consists in a search for resonant di-Higgs production in the $\mathrm{bb}\ \mathrm{l}\nu\mathrm{l}\nu$ final state. Though the analysis design is performed using a well identified signal, the only important assumption made is that the

two Higgs result from the decay of a resonance with a given spin. As such, the results may be interpreted in several BSM scenarios which is a highly desired feature as explained in Sec. 1.3.2. The interest of performing this analysis is twofold. First, the $\mathrm{bb\,l\nu l\nu}$ final state has never been studied in any of the di-Higgs searches so far. Second, it provides an additional benchmark to assess the power of the thoroughly model-independent search described in the next chapter.

We first present an overview of di-Higgs production. Second we go through analysis technical details and event selection. Finally we describe the analysis optimization before extracting the final results.

3.1 Di-Higgs production

Looking at the Lagrangian presented in Sec. 1.2.2, on sees that the SM predicts the existence of double Higgs production, in particular via $\mathcal{L}_{Yukawa}$ thanks to the top-Higgs coupling and via the term involving three Higgs fields in Eq. (1.22) – the other couplings allowing di-Higgs production such as hhVV from Eq. (1.26) lead to subdominant contributions [65]. The dominant Feynman diagrams for the SM gluon fusion di-Higgs production are given in Fig. 3.1. The 13 TeV SM cross section is predicted to be very small, $\sigma^{\mathrm{hh}}_{\mathrm{NNLO}} = 37^{+5.3}_{-6.7}$ fb [66], due to the destructive interference between these two dominant diagrams. We expect thus $\sim$80 hh events with an integrated luminosity of 2.3 fb^{-1} without considering the Higgs decays branching ratios. As a comparison, we expect $\sim$ 200 000 $\mathrm{t\bar{t}}$ event in the fully leptonic channel which makes the SM di-Higgs process impossible to observe with the luminosity available for this work.

However, many BSM scenarios predict an enhancement of the di-Higgs production cross section and can thus be already probed with the available data. A first example is based on the assumption that the scale where new physics lives is far beyond the electroweak scale probed at the LHC but that it has visible impacts at lower energies. Under this assumption, indirect effects at the electroweak scale, due to these BSM phenomena living at higher scale, can be parametrized in an effective field theory (EFT) [67, 68]. In this framework, one can test these BSM effects at the LHC because they imply both the appear-

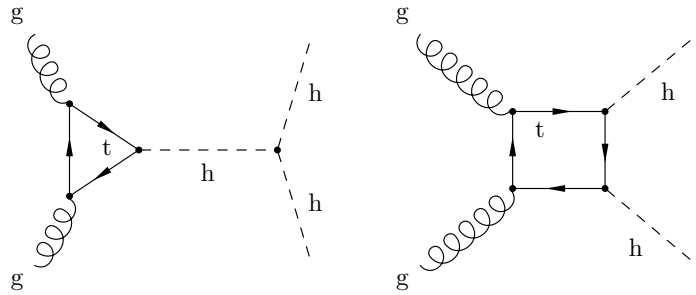


Figure 3.1: Dominant Feynman diagrams for the SM hh production. The left-hand side diagram is coming both from $\mathcal{L}_{Yukawa}$ and $\lambda v h^3$ while the right-hand side contribution is possible thanks to $\mathcal{L}_{Yukawa}$ only. Any massive colored particle can potentially circulate in the loops but the top quark largely dominates due to the magnitude of its Yukawa coupling.

ance of new contact interactions such as tthh and modifications of the Higgs couplings. In particular, modifying the Higgs self coupling, $\lambda_h hh$, which appears in the left hand side diagram from Fig. 3.1 removes the destructive interference and implies important modification of the cross section. A second example is based on the assumption that new physics can already be probed directly at LHC energies. It gathers essentially all the BSM scenarios predicting the existence of new particles that can be produced at the LHC and that couples to the Higgs boson. These new resonances decay thus into two Higgs boson and can be experimentally tracked down as they lead to a higher hh production cross section. We find such new resonances for instance in Higgs singlet models [69, 70], in the 2HDM [24] family or in models inspired by warped extra dimensions [28, 29] as discussed in Sec. 1.3.2. This chapter focuses on the second scenarios: it presents a search for narrow width ($\sim 1\,\mathrm{MeV}$ i.e. well below the experimental resolution) resonance, scanning masses from 260 GeV to 900 GeV and deriving results for spin 0 or spin 2 particle.

The table on the left hand side of Fig. 3.2 shows the theoretical prediction of the Higgs branching ratios ($\mathcal{B}_{\mathrm{h}\to\mathrm{XX}}$) and the chart on the right shows the respective $\mathcal{B}_{\mathrm{hh}\to\mathrm{XXYY}}$. One can see that the dominant Higgs decay is to a pair of $\mathrm{b\bar{b}}$ which explains why most di-Higgs analyses focus on a final state where at least one of the Higgs decays into this channel. The choice of the other channel is a trade off between the signal purity and the size of the branching ratio. For

Final state	$\mathcal{B}_{\mathrm{h}\to\mathrm{XX}}$ (%)
$\mathrm{b\bar{b}}$	57.7
$\mathrm{W^+W^-}$	21.5
gg	8.6
$\tau^+\tau^-$	6.3
$\mathrm{c\bar{c}}$	2.9
ZZ	2.6
$\gamma\gamma$	0.2
$\mathrm{Z}\gamma$	0.15

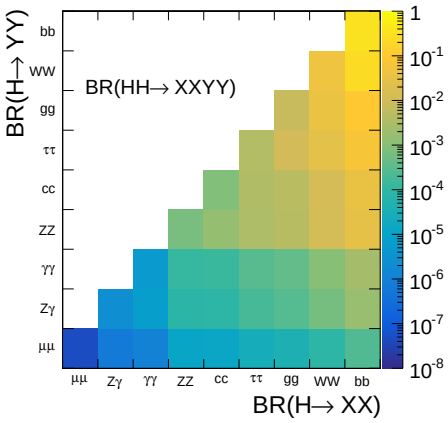
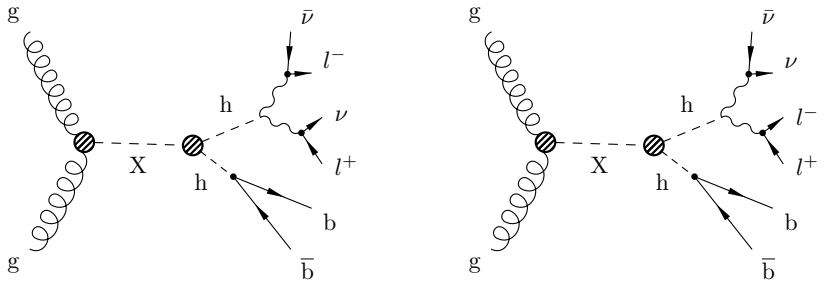


Figure 3.2: Left hand side table shows the SM theoretical Higgs branching ratios for a Higgs mass of 125 GeV [71]. Right hand side plot shows the corresponding di-Higgs branching ratios.

instance, $\mathrm{b\bar{b}}\,\mathrm{b\bar{b}}$ final state has the highest branching ratio but is affected by a very large QCD background whereas $\mathrm{b\bar{b}}\,\gamma\gamma$ is little contaminated by SM processes but has a low branching ratio. The final state we chose to study, $bb\,VV$, is in between these two extremes, with a reasonable $\mathrm{t\bar{t}}$ background and a moderate branching ratio. We further require the leptonic decay of the V's, leading to the diagrams:



These diagrams interfere with the SM $\mathrm{t\bar{t}}$ and hh processes but the impact of the inference can be neglected due to the very narrow width of the X resonance.

3.2 Samples

3.2.1 Data

As mentioned previously, we analyze 2.3 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13\,\text{TeV}$ collected in 2015. The different datasets considered for the analysis are listed in Table 3.1 for each leptonic channel respectively. The presence of two datasets for the same run period is due to changes in the reconstruction that took place during the 2015D era: PromptReco-v4 corresponds to the most recent reconstruction chain and therefore the data taken before the modifications had to be re-reconstructed to ensure an homogeneous dataset.

Channel	Dataset	lumi [fb^{-1}]
ee	DoubleEG/Run2015D-05Oct2015-v1	0.59
	DoubleEG/Run2015D-PromptReco-v4	1.71
$\mu\mu$	DoubleMuon/Run2015D-05Oct2015-v1	0.59
	DoubleMuon/Run2015D-PromptReco-v4	1.71
$e\mu$	MuonEG/Run2015D-05Oct2015-v1	0.59
	MuonEG/Run2015D-PromptReco-v4	1.71

Table 3.1: Data samples used in the analysis.

3.2.2 Monte Carlo Simulation

In order to establish the expectations under the different tested hypotheses, one resorts to MC simulations as explained in Sec. 1.2.1 and 2.2.7. Parton distribution functions are modeled with NNPDF30_nlo at NLO and NNPDF30_lo at LO [5]. Background samples have been generated using MADGRAPH 5 [6], POWHEG 2 [9, 10, 11, 12, 13] and PYTHIA 8 [7, 8]. The signal samples have been generated using MADGRAPH 5 (or spin-2) object decaying into two SM Higgs bosons with a mass of 125 GeV. One of the Higgs bosons is required to decay into a pair of b-quarks, while the second one is required to decay to a pair of weak bosons and subsequently to final states containing two leptons and two neutrinos assuming SM branching ratios. This implies that the signal

samples contain both $\mathrm{h} \to \mathrm{Z(ll)Z}(\nu\nu)$ and $\mathrm{h} \to \mathrm{W(l\nu)W(l\nu)}$ decay legs. The resonance has been generated using a narrow width (1 MeV).

Table 3.2 shows details on the main background MC simulations used in the analysis. The effective luminosity generated for the dominant $\mathrm{t\bar{t}}$ background is largely sufficient to describe a dataset of 2.3 fb^{-1}. Top pair production and tW backgrounds have been generated with NLO precision while the DY samples used are at LO accuracy. Note that MADGRAPH Drell-Yann samples at NLO were available but could not be used here because the effective statistics (taking into account events with negative weights present in this sample) was too small. The complete list of background MC samples used in the analysis are listed on Tab. (5.1) and Tab. (5.2) in App. 5.2.1 together with their cross section.

Process	Generator	Precision	σ [pb]	$\mathcal{L}$ [pb^{-1}]
$\mathrm{t\bar{t}}$ (inclusive)	POWHEG	NLO	831.76	116421
tW (inclusive)	POWHEG	NLO	71.2	27866
DY $\to l^{+}l^{-}$				
$m_{ll} \in [5, 50]$	MADGRAPH	LO	71310	22793
$m_{ll} > 50$	MADGRAPH	LO	6025.2	77031

Table 3.2: Summary table describing the main background processes, the generator and precision used for their simulation, the corresponding cross section [72] and effective luminosity that was actually generated.

3.3 Event selection

We select events with two oppositely charged leptons ($e^{+}e^{-}, \mu^{+}\mu^{-}, e^{\pm}\mu^{\mp}$) and two b-tagged jets. The electrons (muons) are required to have a p_{T} greater than 20 GeV and 15(10) GeV, for the higher and lower p_{T} lepton, respectively. This choice ensures to be above the trigger requirements described in Sec. 2.4. Muons (electrons) with pseudo-rapidity $|\eta| < 2.4$ ($|\eta| < 2.5$) are considered. A di-lepton mass requirement of $m_{\mathrm{ll}} > 12\,\mathrm{GeV}$ is applied and a matching is

performed between offline and online leptons by asking $\Delta R < 0.1$ and $\frac{\Delta p_T}{p_T} < 0.5$. Jets are required to have $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$, be separated from identified leptons by a distance of $\Delta R > 0.3$ and pass the medium CSVv2 b-tagging working point (described in Sec. 2.3.4).

In events which have more than two jets passing these requirements, a criteria is needed to choose among them. Figure 3.3 shows the m_{jj} distributions for the different jet pairing criteria investigated in $t\bar{t}$ and signal events. The two jets with the highest CSVv2 output is the selected condition because its m_{jj} distribution for signal events is more peaked around the Higgs than the one of other conditions not presenting a resonant behavior for $t\bar{t}$ events (the "mh" condition seems to be the best regarding signal events but presents an undesired resonant behavior around the Higgs mass for the $\mathrm{t\bar{t}}$ background).

Before to go further, let us describe the legend and color convention applied to the plots shown in this chapter:

- MC simulations are represented by solid histograms and are stacked together. Drell-Yan process corresponds to light blue, $\mathrm{t\bar{t}}$ to orange, single-top to burgundy, di-boson (VV) to beige, $\mathrm{t\bar{t}}$ V to grey and SM higgs processes (including $\mathrm{t\bar{t}}$ h) are in light green. Other SM contributions such as W+jets were processed but lead to negligible contributions and are thus disregarded.
- Signal processes are represented by single lines and are arbitrarily renormalized to a total cross section of 1 pb. To avoid overloading the plots, only the spin-0 $\mathrm{m_X}$ = 400, 650 and 900 GeV benchmarks are shown.
- Data are represented by black dots with statistical uncertainty shown as vertical black lines.
- The dashed brown band represents uncertainty on the SM MC expectations associated to the various reweighting described in previous chapter, the jet energy corrections, the pdf's, the choice of renormalization and factorisation scales and the limited MC statistics.

On top of the object selection described here-above, further cuts are applied in order to ease the analysis optimization described in the next section. As

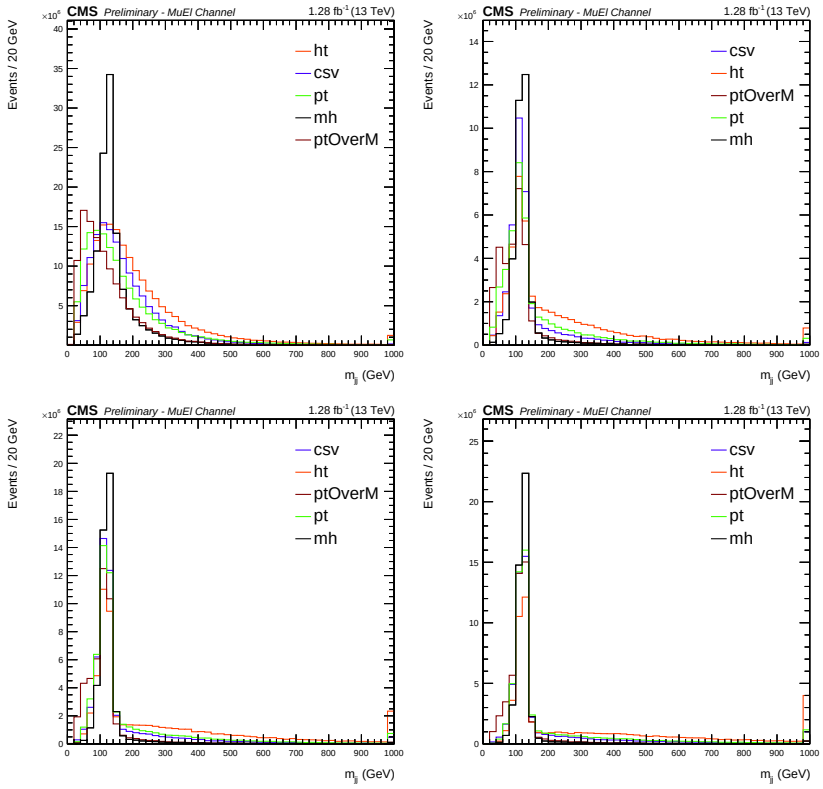


Figure 3.3: m_{jj} distribution for different di-jet selections in $t\bar{t}$ events (upper left), and signal events with masses m_X =400 (upper right), 650 (lower left), and 900 GeV (lower right). The various jet pairing considered are: the two highest p_T jets ("ht"), the jets with the highest $(\vec{p}_1 + \vec{p}_2)_T$ ("pt"), the two jets with the highest $(\vec{p}_1 + \vec{p}_2)_T / m_{jj}$ ("ptOverM"), the di-jet system with the closest mass to the Higgs boson ("mh"), and the two jets with the highest CSVv2 discriminant ("csv").

mentioned earlier the signal generation includes both $h \to ZZ$ and $h \to WW$ decays. However, the $h \to Z(\to ll)Z(\to \nu\nu)$ component of the signal is drowned under the DY SM process and would require a dedicated analysis strategy. We focus here on the $h \to WW$ component which lies at lower dilepton invariant mass. Therefore we select events with $\mathrm{m_Z} - \mathrm{m_{ll}} > 15\,\mathrm{GeV}$ which heavily suppresses the DY background as shown on Fig. 3.4. Though

the DY contribution in the different flavor channel ($e^{\pm}\mu^{\mp}$) is negligible, we also apply the m_{ll} cut for this lepton flavor configuration since it suppresses a background dominated region as shown on Fig. 3.5. One can see on this distribution that the data and MC shapes are in good agreement while an overall excess of MC is observed. The latter will be absorbed via the uncertainty on the background normalization (not shown on the plot uncertainty bands) when extracting the final result as described later.

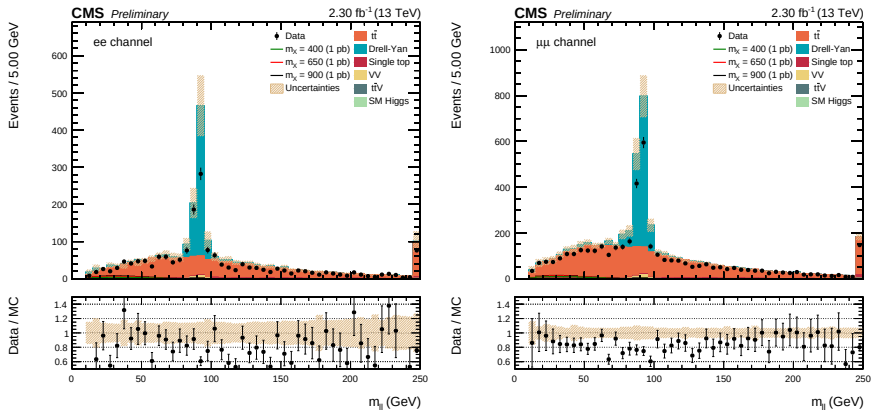


Figure 3.4: The m_{ll} distribution for data and simulated events after requiring two leptons, and two b-tagged jets. The e^+e^- lepton flavor combinations is shown on the left while the $\mu^+\mu^-$ channel is shown on the right. The last bin includes the overflow.

We further suppress the following background dominated regions: $\Delta R_{ll} < 2.2$, $\Delta R_{jj} < 3.1$, and $\Delta\phi_{ll,jj} > 1.5$ whose distributions are shown on Fig. 3.6.

A summary of the background yields after all the selection requirements described in this section is shown on Table 3.3. One can notice that at this stage the dominant background is by far $\mathrm{t\bar{t}}$. The signal to $\mathrm{t\bar{t}}$ ratio is expected to be the same in the three lepton channels. Furthermore, the signal and $\mathrm{t\bar{t}}$ kinematics are also expected to be similar in the three lepton channels. For these

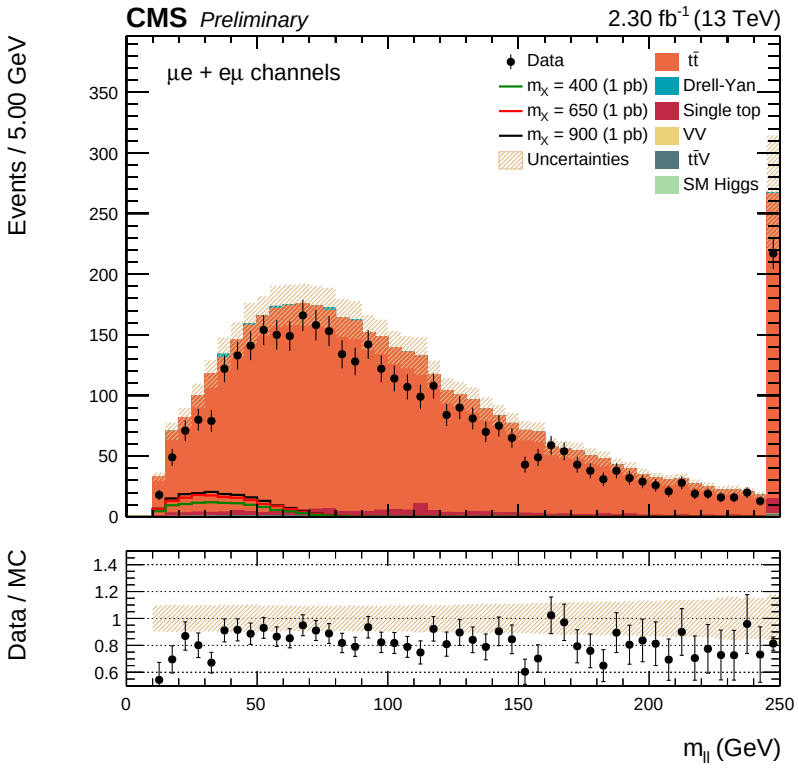


Figure 3.5: The m_{ll} distribution in the different flavor channel ($e^{\pm}\mu^{\mp}$) for data and simulated events after requiring. The last bin includes the overflow.

reasons, we chose to perform the analysis optimization considering all the lepton channels together.

3.4 Analysis optimization

We further optimize the event selection towards signal dominated phase spaces using a boosted decision tree (BDT) discriminant and the m_{jj} distribution shown on Fig. 3.7.

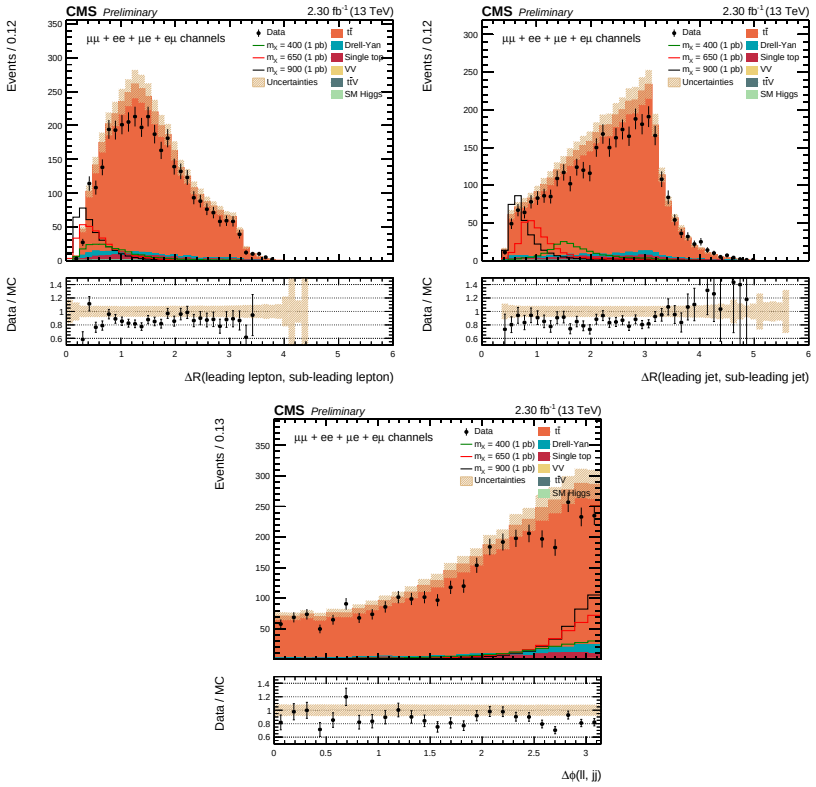


Figure 3.6: ΔR_{ll}, ΔR_{jj}, and $\Delta\phi_{ll,jj}$ distributions for data and simulated events after requiring two leptons, two b-tagged jets, and $\mathrm{m_Z} - \mathrm{m_{ll}} > 15$ GeV. All lepton flavour combinations shown together.

The analysis relies on a cut and count method based on four "regions" in the 2D plane of the BDT discriminants and $\mathrm{m_{jj}}$ distribution. We define two regions from the $\mathrm{m_{jj}}$ distribution: $\mathrm{m_{jj}}$-peak ($\mathrm{m_{jj}}$-P) corresponds to the signal like region around the Higgs mass and $\mathrm{m_{jj}}$-sidebands ($\mathrm{m_{jj}}$-SB) corresponds to the background like region away from the Higgs mass. Two regions are also defined from the BDT discriminant: *low-BDT-scores* region corresponds to the background like phase space while the *high-BDT-scores* region is the signal like region. This defines the four analysis regions shown on Fig. 3.8: high-BDT-scores & $\mathrm{m_{jj}}$-P, high-BDT-scores & $\mathrm{m_{jj}}$-SB, low-BDT-scores & $\mathrm{m_{jj}}$-P and low-BDT-scores & $\mathrm{m_{jj}}$-SB. The cyan area (high-BDT-scores & $\mathrm{m_{jj}}$-P)

	2l + 2b-jets + selection cuts
$t\bar{t}$	1913.1
Single Top	56.1
DY	53.9
$t\bar{t}V$	4.6
SM Higgs	3.3
VV	2.1
Total bkg	2033.8

Table 3.3: Background yields requiring two leptons, two b-tagged jets, $\mathrm{m_Z} - \mathrm{m_{ll}} > 15$ GeV, $\Delta R_{ll} < 2.2$, $\Delta R_{jj} < 3.1$, and $\Delta\phi_{ll,jj} > 1.5$.

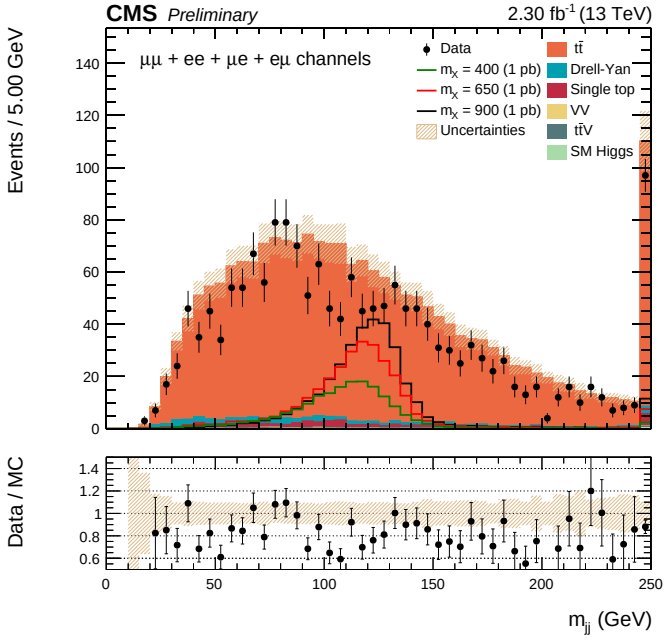


Figure 3.7: The $\mathrm{m_{jj}}$ distribution for data and simulated events after requiring all selection cuts described in Sec. 3.3. The last bin includes the overflow.

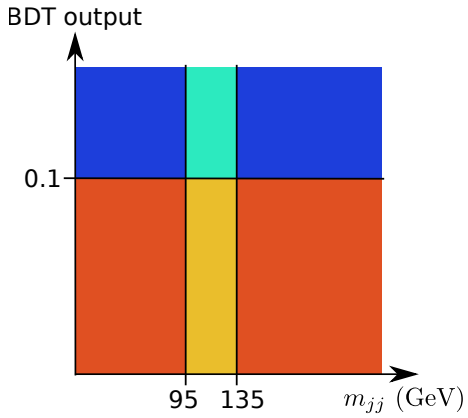


Figure 3.8: Analysis categories represented in the BDT output vs $\mathrm{m_{jj}}$ plane. The cyan area corresponds to the high-BDT-scores & $\mathrm{m_{jj}}$-P, the blue area corresponds to the high-BDT-scores & $\mathrm{m_{jj}}$-SB, the orange area corresponds to the low-BDT-scores & $\mathrm{m_{jj}}$-SB, and the yellow area corresponds to the low-BDT-scores & $\mathrm{m_{jj}}$-P.

corresponds to the most signal-like region while the other regions are useful to normalize the SM backgrounds to data before extracting the limits.

First we optimize the cuts on $\mathrm{m_{jj}}$ distribution by maximizing the expected analysis sensitivity based on a cut and count method using only the $\mathrm{m_{jj}}$-P and $\mathrm{m_{jj}}$-SB regions. The signal mass points considered for this optimization are $\mathrm{m_X}$ = 400, $\mathrm{m_X}$ = 650 and $\mathrm{m_X}$ = 900 GeV. One shows on Fig. 3.9 the expected 95% CLs limits on the production cross section obtained for the signal $\mathrm{m_X}$ = 650 GeV as a function of the lower and upper $\mathrm{m_{jj}}$ cuts. On sees that the best limit (minimal value on this figure) is reached for "$\mathrm{m_{jj}}$ low cut" at 95 GeV and "$\mathrm{m_{jj}}$ high cut" at 140 GeV.

The equivalent plots for $\mathrm{m_X}$ = 400 and $\mathrm{m_X}$ = 900 GeV are shown on Fig. 5.2 in App. 5.2.2). The best $\mathrm{m_{jj}}$ cuts are very close for the three considered benchmarks which allows us to chose one common region definition without loosing too much sensitivity. The chosen trade-off is: $\mathrm{m_{jj}}$-P $\equiv m_{\mathrm{jj}} \in [95, 135]$ GeV and $\mathrm{m_{jj}}$-SB $\equiv m_{\mathrm{jj}} < 95$ or $m_{\mathrm{jj}} > 135$ GeV.

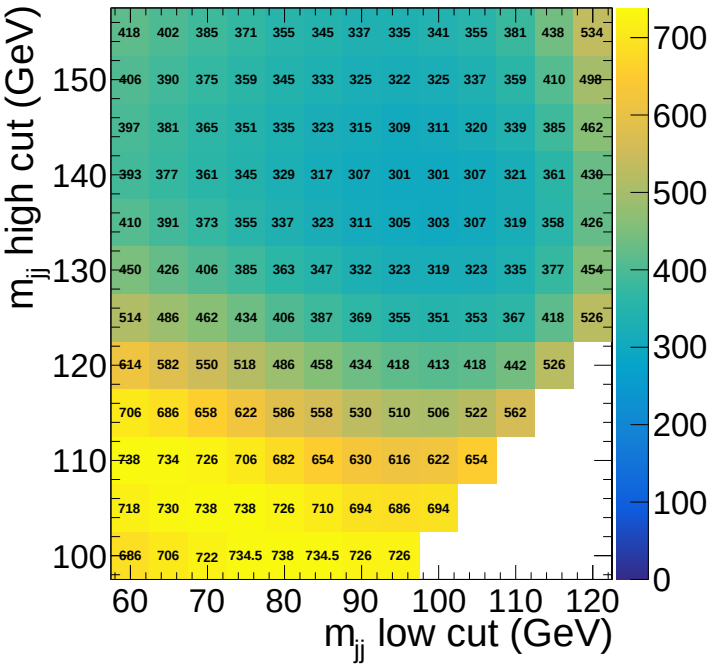


Figure 3.9: Expected 95% CLs limit on the signal production cross section (fb) based on a cut-and-count method using the $\mathrm{m_{jj}}$-P and $\mathrm{m_{jj}}$-SB regions, as a function of the lower and upper $\mathrm{m_{jj}}$ cuts, for signal with $\mathrm{m_X}$ =650 GeV.

The following set of kinematic variables were used as input of the BDT training: $\mathrm{m_{ll}}$, ΔR_{ll}, ΔR_{jj}, $\Delta\phi_{ll,jj}$, p_T^{ll}, p_T^{jj}, $\min\Delta R_{j,l}$, and M_T. Where $\min\Delta R_{j,l}$ is the minimal angular distance between the lepton/jet pairs among the four possible combinations and $M_T \equiv \sqrt{2p_T^{ll}E_T^{miss}(1-\cos(\Delta\phi(ll,E_T^{miss})))}$. Figure 3.10 shows the ΔR_{jj} and $\min\Delta R_{j,l}$ distributions. Since the two jets come from the same object in the signal, the more the Higgs boson is boosted, the lower ΔR_{jj} distribution peaks (as can be seen by comparing this distribution for the three resonance masses). For the dominant $\mathrm{t\bar{t}}$ background the two jets are coming from different objects which explains why these events lie in general at higher ΔR_{jj} values. Since the top quarks decay into a lepton, a jet and a neutrino, $\min\Delta R_{j,l}$ is small when at least one of the top is boosted. For the signal, the leptons come from one Higgs and the jets from the other. Therefore

if the X resonance is not boosted (which is generally the case for high m_X) the two Higgs are back to back and the $\min \Delta R_{j,l}$ is large, providing thus a good discriminating power between the signal and $t\bar{t}$ background. The other BDT input variables are shown on Fig. 5.3 in App. 5.2.3.

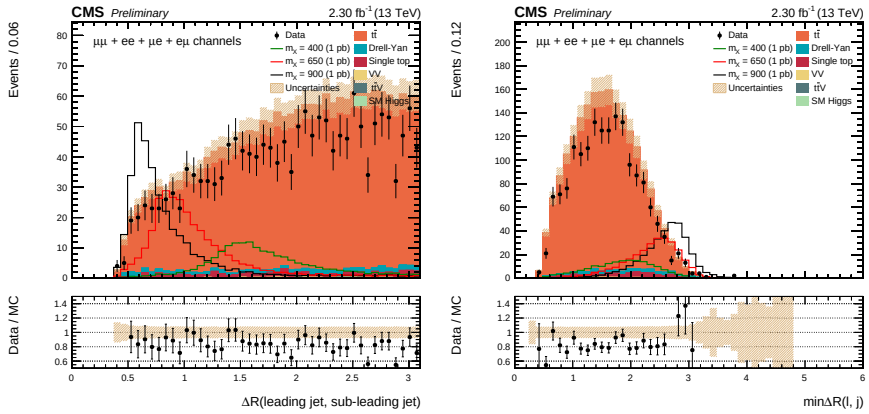


Figure 3.10: The ΔR_{jj} (left) and min$\Delta R_{j,l}$ (right) distributions for data and simulated events after requiring all selection cuts described in Sec. 3.3. Both are used as input to train the BDT discriminant. All lepton flavor combinations are shown together.

The training of the BDT is performed using spin-0 samples as signals and the three dominant SM processes ($t\bar{t}$, tW and DY) as background. The MC reweighting scale factors are taken into account for the training and each background is given in proportion relative to their expected yield. Half of the samples statistics is used for the training and the other half is used to ensure the BDT did not undergo over-training. Two BDT's trained with signals m_X = 400 and 650 GeV, whose distributions are shown on Fig. 3.11, are used in the analysis. Two additional trainings with m_X = 500 and 900 GeV have been considered but were discarded because they brought no significant gain to the sensitivity while making the analysis flow heavier.

The two signal benchmarks chosen to optimize the analysis are the one used for the two BDT trainings: m_X = 400 and m_X = 650 GeV. The optimization

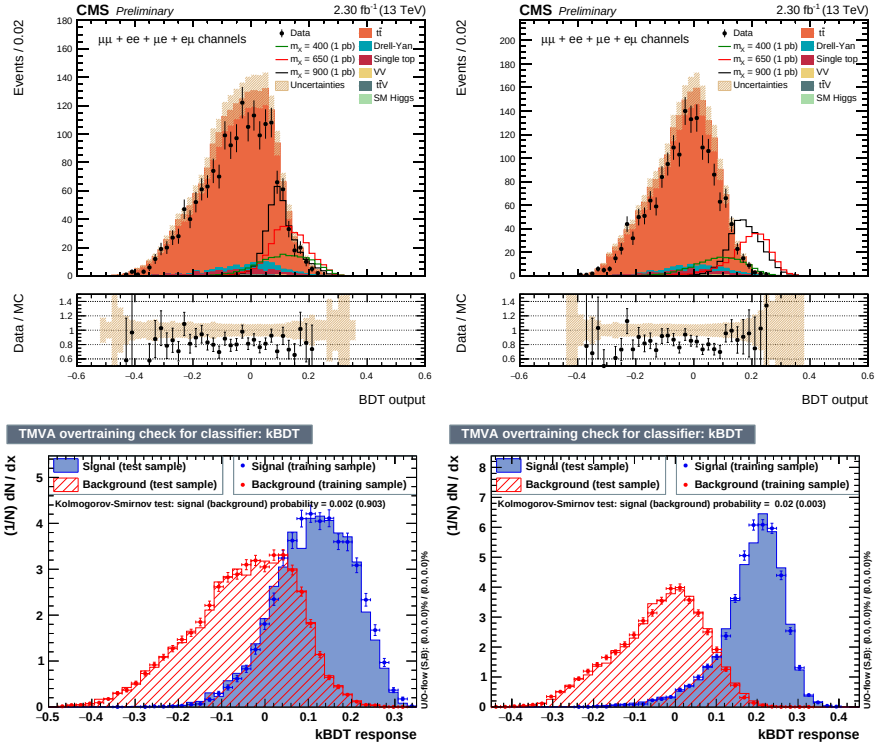


Figure 3.11: BDT output distributions for the training with $m_X = 400$ GeV (left) and $m_X = 650$ GeV (right), after requiring all selections cuts described in Sec. 3.3. The bottom plots show the BDT output distribution in the test (plain histogram) and training (dotted histogram) samples for the background (red histogram) and the signal (blue histogram). The training with m_X = 400 GeV is shown on the left, the one with m_X = 650 GeV on the right.

of the BDT region definition is performed separately for each benchmark and is based on the maximization of the analysis sensitivity using the four regions mentioned above (the m_{jj} regions are already defined when optimizing the BDT regions). Figure 3.12 shows the expected 95% CLs limit on the signal cross section as a function of the BDT cut defining the regions. As expected, the nominal BDT's (the one trained with the corresponding signal) are the one performing best. One sees that the optimal region definitions are close for the two signal benchmarks m_X= 400 GeV and m_X= 650 GeV. We choose thus the

same cut for BDT-m_X = 400 and BDT-m_X = 650 GeV : $\mathrm{BDT_{out}} < 0.1$ which corresponds to the low-BDT-scores and $\mathrm{BDT_{out}} \geq 0.1$ corresponding to the high-BDT-scores.

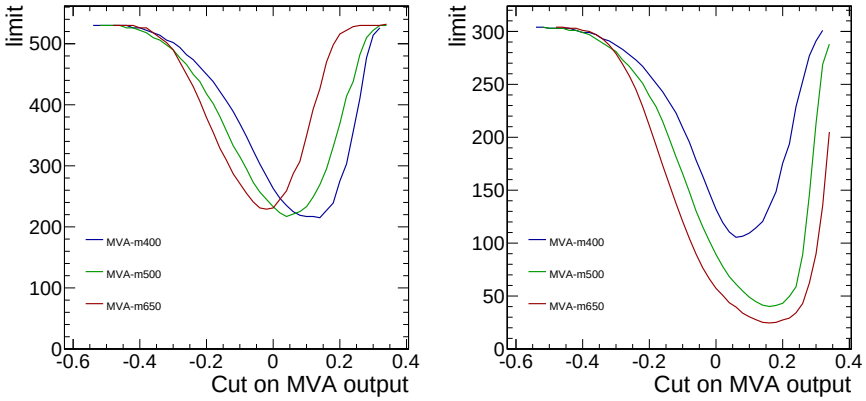


Figure 3.12: Expected 95% CLs limit on the signal cross section (fb) using the four regions in the two dimensional "m_{jj} - BDT output" plane (see Fig. 3.8) as a function of the BDT cut. The two m_{jj} windows are already defined as m_{jj}-P $\equiv m_{jj} \in [95, 135]$ GeV and m_{jj}-SB $\equiv m_{jj} < 95$ or $m_{jj} > 135$ GeV. The signal with m_X= 400(650) GeV is shown on the left(right). Three BDT's are considered with different signal samples used for the training: m_X = 400 (blue), 500 (green) and 650 (red).

The mass range of the search is defined from m_X = 260 GeV to m_X = 900 GeV. The choice of the BDT training to apply for a given mass is based on the expected analysis sensitivity obtained with each BDT as a function of the X mass, as shown in Fig. 3.13. The BDT trained with m_X = 400 GeV is applied to signals with $m_X \leq 450$ GeV, and the BDT trained with m_X = 650 GeV is applied to signals with $m_X \geq 450$ GeV.

While the BDT's are trained using spin-0 signals, their performances in terms of signal versus background efficiency are comparable with those trained using

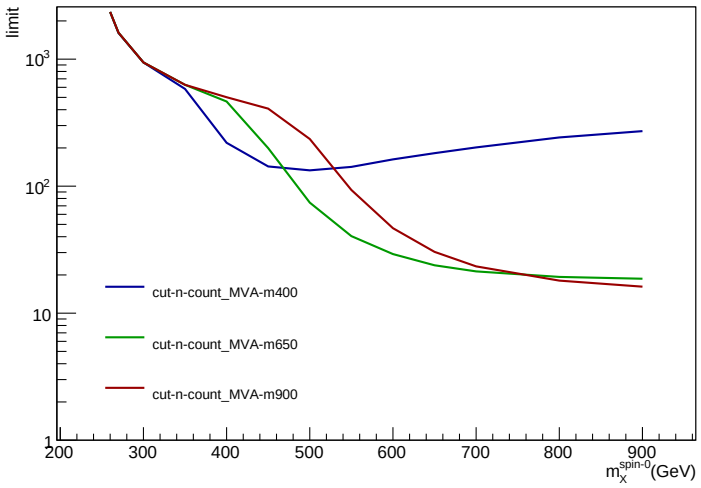


Figure 3.13: Expected 95% CLs limit on the signal cross section (fb) using the four regions in the two dimensional "m_{jj} - BDT output" plane, as a function of the X mass and for different BDT trainings. Three BDT's trained with different signals are shown: $m_X = 400$ GeV (blue curve), $m_X = 650$ GeV (green curve) and $m_X = 900$ GeV (red curve).

spin-2 as signals as shown on Fig. 3.14. This allows us to apply the spin-0 BDT's to the spin-2 samples without important loss of sensitivity.

3.5 Systematic uncertainties

Any source of uncertainties affecting the normalization and/or the shape of the background and signal expectations will affect the final results of this analysis and are considered as systematic uncertainties.

We consider the following as experimental sources of systematic uncertainties. As explained in Sec. 2.3.2, 2.3.3, 2.3.4 and 2.4, the leptons selection (identification and isolation), b-tagging scale factors and trigger efficiencies, are not known perfectly. The associated uncertainties are estimated by vary-

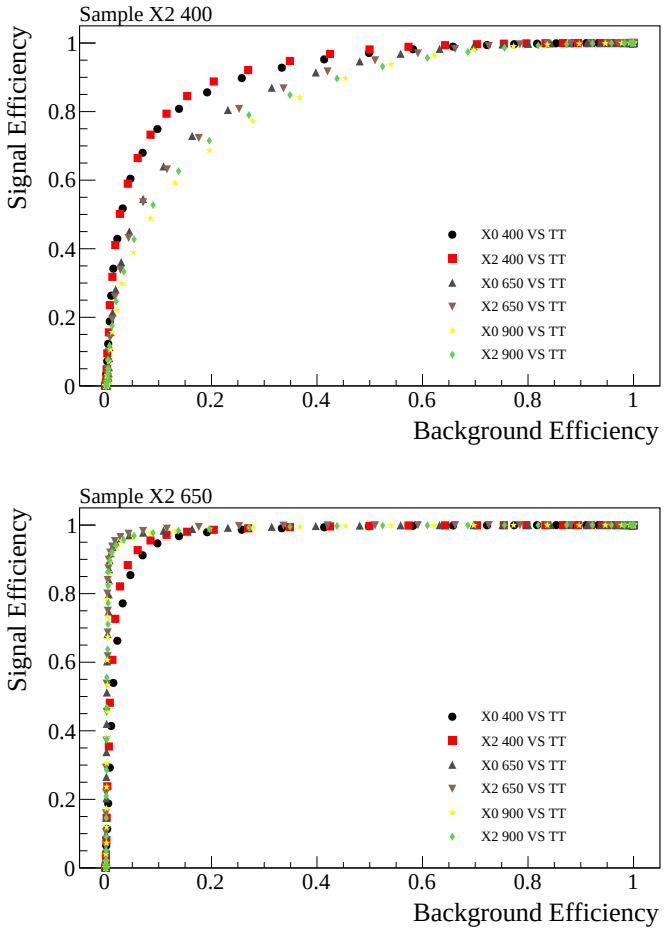


Figure 3.14: Spin-2 signal versus background efficiencies for BDTs trained with spin-0 ("X0") and spin-2 ("X2") signal samples. Spin-2 signal samples with m_X= 400 GeV (left) and m_X= 650 GeV (right) are shown.

ing them by one standard deviation. The impact of the JES/JER uncertainty (see Sec. 2.3.4) is evaluated by shifting the jet energy correction/resolution factors for each jet up and down by one standard deviation, and the variations are propagated to the $E_{\mathrm{T}}^{\mathrm{miss}}$. The magnitude of the uncertainties related to the pdf for each simulated background process is obtained by using different pdf sets than the nominal one (NNPDF 3.0). As explained in Sec. 2.3.1, the PU

reweighting uncertainty is also taken into account. To estimate it, we apply the PU reweighting procedure by using shifted total pp interaction cross section. Finally, the global uncertainty coming from the LHC luminosity measurement amounts to 2.7% [73].

Theoretical uncertainties on the cross sections used to predict the yield estimates are taken into account process per process. Their impact on the final yields is estimated by varying the process cross section by one standard deviation (obtained from Ref. [72]). The uncertainties on the theory predictions due to the arbitrary choice of renormalization and factorization scales as explained in Sec. 1.2.1 are obtained by taking the envelope of the shapes obtained with modified scales. The various configurations considered to derive the envelope are: $(\frac{\mu_R}{2}, \mu_F)$, $(\mu_R, \frac{\mu_F}{2})$, $(\frac{\mu_R}{2}, \frac{\mu_F}{2})$, $(2\mu_R, \mu_F)$, $(\mu_R, 2\mu_F)$ and $(2\mu_R, 2\mu_F)$.

In addition, we consider global scaling uncertainties on the $t\bar{t}$ (10%), Drell-Yan (30%) and single-top (20%) processes to account for the global normalization discrepancy between data and MC.

Finally, systematic uncertainties due to the limited statistics of MC samples are also taken into account.

The effect of the various systematic uncertainties on the total yields in the fours final regions are summarized in Tab. 3.4.

3.6 Results

Selected events are classified into 4 categories, as described in Sec. 3.4. The most signal-like region is the high-BDT scores & $\mathrm{m_{jj}}$-P while the three other regions help constraining the background normalizations within uncertainties. We will refer in the following to the so-called post-fit and pre-fit uncertainties to specify whether they are constrained by a fit to data or not.

Pre-fit yields in final regions are shown in Tab. 3.5 for the BDT 650 GeV training. The equivalent table for BDT 400 GeV training is given in Tab. 5.3 from App. 5.2.4. Quoted uncertainties include both statistical and systematics un-

Source	Sig. ($m_X = 400$)	Sig. ($m_X = 650$)	Background
Trigger efficiency	5.1 - 6.0%	6.7 - 7.4%	4.5 - 5.3%
Jet b-tagging	4.9 - 6.5%	5.7 - 7.3%	5.1 - 6.0%
Jet energy scale	1.6 - 3.0%	0.6 - 3.9%	1.0 - 3.6%
Jet energy resolution	0.5 - 4.1%	1.8 - 3.5%	0.1 - 2.4%
Electon ID & ISO	1.3 - 1.6%	1.3 - 1.7%	1.4 - 1.5%
Muon ID & ISO	0.9 - 1.4%	1.0 - 1.1%	1.2 - 1.5%
Pileup	0.4 - 1.8%	0.1 - 0.6%	0.5 - 2.2%
Parton distributions	0.4 - 0.5%	0.2 - 0.5%	0.5 - 0.6%
QCD scale	0.3 - 0.4%	0.2 - 0.4%	0.8 - 2.4%
Luminosity		2.7%	
Signal MC stat.	1.4 - 2.4%	0.9 - 3.2%	-
Affecting only $t\bar{t}$ (87.0 - 95.3% of the total bkg)			
$t\bar{t}$ cross section	-	-	6.5%
$t\bar{t}$ normalization	-	-	10%
$t\bar{t}$ MC stat.	-	-	0.6 - 2.3%
Affecting only Drell-Yan (1.8 - 7.1% of the total bkg)			
Drell-Yan normalization	-	-	30%
Drell-Yan MC stat.	-	-	4.4 - 22.7%
Affecting only single top (2.5 - 4.6% of the total bkg)			
Single top normalization	-	-	20%
Single top MC stat.	-	-	6.6 - 24.4%

Table 3.4: Summary of the systematic uncertainties and their individual impact range on total yields, for signal $m_X = 400$ GeV, signal $m_X = 650$ GeV, and background. The first(second) number corresponds to the smallest(biggest) impact among the four final regions.

Signal samples	high-BDT 650, m_{jj}-P	high-BDT 650, m_{jj}-SB	low-BDT 650, m_{jj}-P	low-BDT 650, m_{jj}-SB
m_X = 650 (1 pb)	185.0 ± 18.9	69.3 ± 8.9	14.5 ± 1.7	21.0 ± 2.6
SM samples				
$t\bar{t}V$	0.2 ± 0.1	0.7 ± 0.2	0.8 ± 0.2	2.9 ± 0.4
SM Higgs	0.2 ± 0.0	0.5 ± 0.1	0.6 ± 0.1	2.0 ± 0.2
VV	0.2 ± 0.1	0.5 ± 0.1	0.3 ± 0.1	1.6 ± 0.2
Single top	2.7 ± 1.3	5.0 ± 1.3	11.4 ± 1.9	37.0 ± 4.4
Drell-Yan	3.0 ± 0.7	8.2 ± 1.6	8.4 ± 1.6	35.0 ± 5.0
$t\bar{t}$	53.0 ± 4.7	99.5 ± 9.1	433.4 ± 33.7	1327.2 ± 106.5
Total ± (stat.) ± (syst.)	59.2 ± 1.7 ± 5.7	114.4 ± 2.4 ± 10.9	454.9 ± 4.4 ± 35.7	1405.4 ± 7.9 ± 114.2
Data ± (stat.)	53 ± 7.3	110 ± 10.5	349 ± 18.7	1172 ± 34.2

Table 3.5: Pre-fit yields in final regions for the BDT-m_X = 650 GeV training. Quoted uncertainties include both statistical and systematic uncertainties, as detailed in Tab. 3.4, except the normalization and cross-section uncertainties.

certainties, as detailed in Tab. 3.4, except normalization and cross-section uncertainties.

Signal efficiency as a function of the X mass hypothesis, taking into account all the data/MC correction factors, in the four final regions defined in the analysis, is shown in Fig. 3.15. The efficiency is interpolated in between the fully-simulated mass points.

3.6.1 Maximum likelihood fit

We perform a maximum likelihood fit in the four categories of the analysis, letting all the nuisance parameters floating in order to extract the best fit cross-section for each mass hypothesis.

The post-fit pull of the different nuisances are shown in Fig. 3.16 for the BDT 650 GeV training and in Fig. 5.4 in App. 5.2.4 for the BDT 400 GeV training. As expected from the excess of MC, the $\mathrm{t\bar{t}}$ normalization and cross-section are the most pulled nuisance parameters by the fit, in order to restore a good agreement between the simulation and the data. Nevertheless, the likelihood has enough degrees of freedom to describe the data, none of the nuisance param-

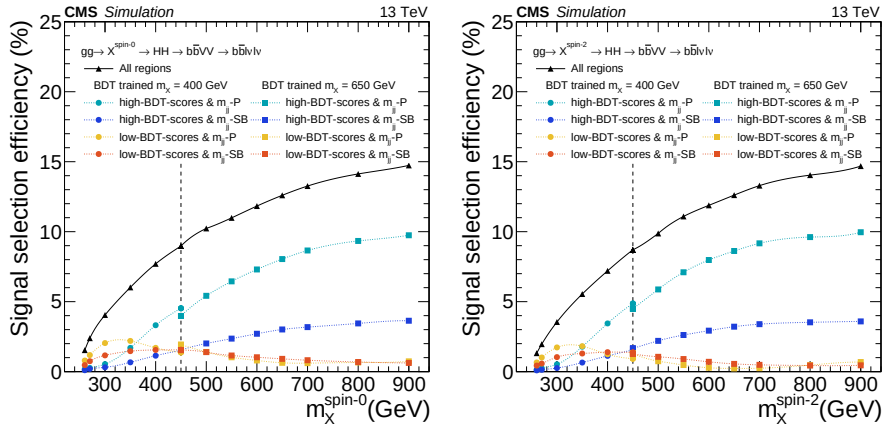


Figure 3.15: Signal efficiency as a function of the X mass hypothesis for spin-0 (left) and spin-2 (right) hypotheses, taking into account all the data/MC correction factors, in the four final regions defined in the analysis. The markers correspond to efficiencies evaluated on fully-simulated signal samples, while in between fully-simulated samples the efficiencies have been interpolated. The color code is the same as in Fig. 3.8: the cyan corresponds to the high-BDT-scores & $\mathrm{m_{jj}}$-P, the blue corresponds to the high-BDT-scores & $\mathrm{m_{jj}}$-SB, the orange corresponds to the low-BDT-scores & $\mathrm{m_{jj}}$-SB, and the yellow corresponds to the low-BDT-scores & $\mathrm{m_{jj}}$-P. The black line, being the sum of all (mutually exclusive) regions, correspond to the full signal acceptance after selection.

eters being pulled by more than 1 standard deviation. Best fit cross-sections are extracted from the same fit, as a function of the X mass hypothesis and are shown in figure 3.17. No sign of new physics is observed, all the cross-sections are compatible with zero and limits on the resonant Higgs production cross-sections are thus set.

We also extract post-fit normalization scale factors for each background sample, as well as associated uncertainties. For this, we do not consider the most signal-like category (high-BDT-scores & $\mathrm{m_{jj}}$-P), and we perform a background-only fit (the signal strength is fixed to 0) to the remaining 3 categories. The

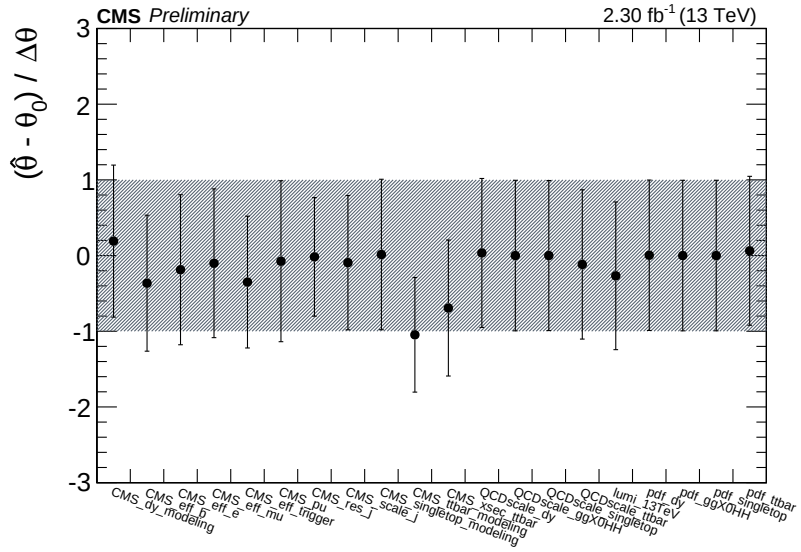


Figure 3.16: Post-fit pull distribution of the different nuisances for the BDT 650. Note that the MC stat uncertainties, included in the fit, are not pulled nor their uncertainties reduced, so they are not shown in the plot for readability.

obtained scale factors can be found in Tab. 3.6, for the BDT-$\mathrm{m_X}$ =650 GeV training. As one can see on the resulting post-fit yields for the BDT $\mathrm{m_X}$ = 650 GeV, shown on Tab. 3.7, the predicted yields are compatible with the observations within one standard deviation also in the high-BDT-scores & $\mathrm{m_{jj}}$-P region, which gives another evidence for the absence of significant fluctuation and motivates the limit setting described in the next section. The equivalent tables for the BDT $\mathrm{m_X}$ = 400 GeV are shown in App.5.2.5 on Tab. 5.4 and 5.5.

3.6.2 Limits

As demonstrated in previous section, no excess is visible in the whole mass range of the analysis. Limits on the resonant Higgs pair production cross section times branching fraction, $gg \to X \to hh \to bbVV \to bbl\nu l\nu$, are thus set.

Background process	Scale factors	Post-fit uncertainties
$t\bar{t}$	0.82	2.22%
Single top	0.94	14.17%
Drell-Yan	0.98	25.32%
SM Higgs	0.95	5.70%
VV	0.94	8.61%
$t\bar{t}V$	0.94	7.65%

Table 3.6: Post-fit normalization scale-factors for each background sample obtained from the background only fit in the three less signal like regions defined according to the BDT trained with m_X = 650 GeV.

Signal samples	high-BDT 650, m_{jj}-P	high-BDT 650, m_{jj}-SB	low-BDT 650, m_{jj}-P	low-BDT 650, m_{jj}-SB
m_X = 650 (1 pb)	185.0 ± 1.7	69.3 ± 1.0	14.5 ± 0.5	21.0 ± 0.6
SM samples				
$t\bar{t}$	43.6 ± 1.5	81.9 ± 2.5	356.6 ± 8.6	1092.0 ± 25.0
Single top	2.5 ± 0.8	4.7 ± 1.1	10.8 ± 2.0	34.9 ± 5.5
Drell-Yan	2.9 ± 0.8	8.0 ± 2.1	8.2 ± 2.2	34.4 ± 8.9
SM Higgs	0.2 ± 0.0	0.5 ± 0.0	0.6 ± 0.0	1.9 ± 0.1
VV	0.2 ± 0.0	0.5 ± 0.1	0.3 ± 0.0	1.5 ± 0.2
$t\bar{t}V$	0.2 ± 0.0	0.7 ± 0.1	0.8 ± 0.1	2.7 ± 0.3
Total ± (stat.) ± (syst.)	49.6 ± 1.4 ± 1.3	96.3 ± 2.0 ± 2.8	377.3 ± 3.7 ± 8.3	1167.5 ± 6.7 ± 26.2
Data ± (stat.)	53 ± 7.3	110 ± 10.5	349 ± 18.7	1172 ± 34.2

Table 3.7: Post-fit yields in final regions, high-BDT & m_{jj}-P, high-BDT & m_{jj}-SB, low-BDT & m_{jj}-P, and low-BDT & m_{jj}-SB resulting from the background only fit excluding the most signal like region, for the BDT 650 GeV training.

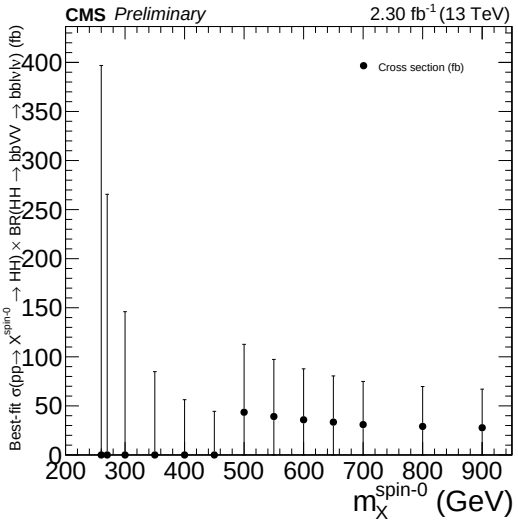


Figure 3.17: Best-fit cross-section times branching ratio as a function of the mass hypothesis, obtained from the maximum likelihood fit based on the four regions defined in Sec. 3.4. At the transition point, $m_{\mathrm{X}} = 450\,\mathrm{GeV}$,

Upper limits at 95% confidence level with the *asymptotic* CL_s [74] method are computed as function of the X mass hypothesis, using the data, background, and signal yields in the four final regions (see Sec. 3.4) taking the uncertainties mentioned in Sec.3.5 into account. The results are shown on Fig. 3.18 for spin-0 and spin-2 resonances.

The observed limits are compatible within 2 standard deviations to the expected ones. The change of trend in the observed limits at 450 GeV corresponds to the transition point of the analysis between the two BDT's, one optimized for low mass resonances and the other for high mass resonances.

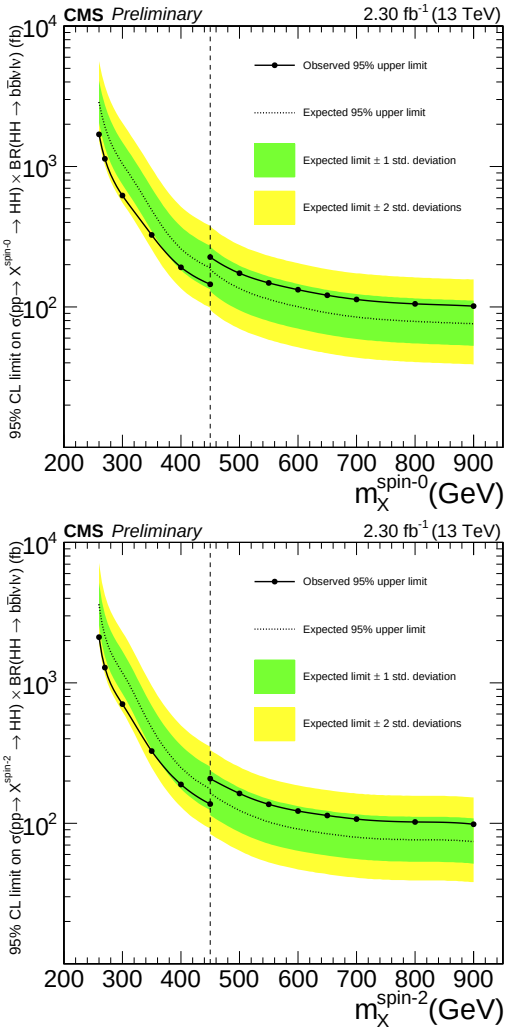


Figure 3.18: Expected and observed 95% CL_s upper limits on the the resonant Higgs pair production cross section times branching fraction for $hh \rightarrow bbVV \rightarrow bbl\nu l\nu$, computed using the asymptotic CL_s method. Spin-0 results are shown on the top plot and spin-2 limits are visible on the bottom plot. The markers correspond to limits evaluated on fully-simulated signal samples, while in between fully-simulated samples the limits have been set on interpolated signal yields and systematics.

3.7 Conclusion

We have presented the very first search for resonant Higgs pair production, $X \to hh$, where one of the h decays as $h \to \mathrm{b\bar{b}}$, and the other as $h \to \mathrm{WW} \to \ell\nu\ell\nu$. Resonance masses were considered in the range from $\mathrm{m_X}$ = 260 GeV to 900 GeV and 2.3 fb^{-1} of LHC proton-proton collision data at $\sqrt{s}$ = 13 TeV were analyzed.

Data and predictions from the SM are in agreement within uncertainties. For mass hypotheses from $m_{\mathrm{X}} = 500$ GeV to $m_{\mathrm{X}} = 900$ GeV, the data are observed (expected) to exclude a production cross-section times branching ratio from 174 to 101 (135 to 75.8) fb.

The comparison of the results obtained in the $\mathrm{bbl}\nu\mathrm{l}\nu$ final state to the results obtained in other final states is shown on Fig. 3.19. While the $\mathrm{bbl}\nu\mathrm{l}\nu$ final state does not have the best expected sensitivity in any of the resonance mass regime, one sees that for low mass resonances we expect to exclude cross sections less than one order of magnitude bigger than e.g. the $\mathrm{bb}\tau\tau$ final state which exploited 12.9 fb^{-1} of luminosity.

Considering the fact that this analysis was the first of its kind and that a lot of room is left for improvement, we can state that the $\mathrm{bbl}\nu\mathrm{l}\nu$ final state is worth being investigated further. Even as such, we expect this final state to bring non-negligible contribution when combining all the results from the different final state together.

To close this chapter, let us provide a list of the very promising analysis modification that would improve its contribution to the other channels:

- Consider separately the different lepton flavor channels. The gain is expected from the absence of DY event in the different flavor channel.
- Exploits the full differential shape of $\mathrm{m_{jj}}$ and BDT discriminant instead of considering four regions.
- Derive background expectations from data instead of MC. For instance, the DY expectations could be derived from the observations in the two

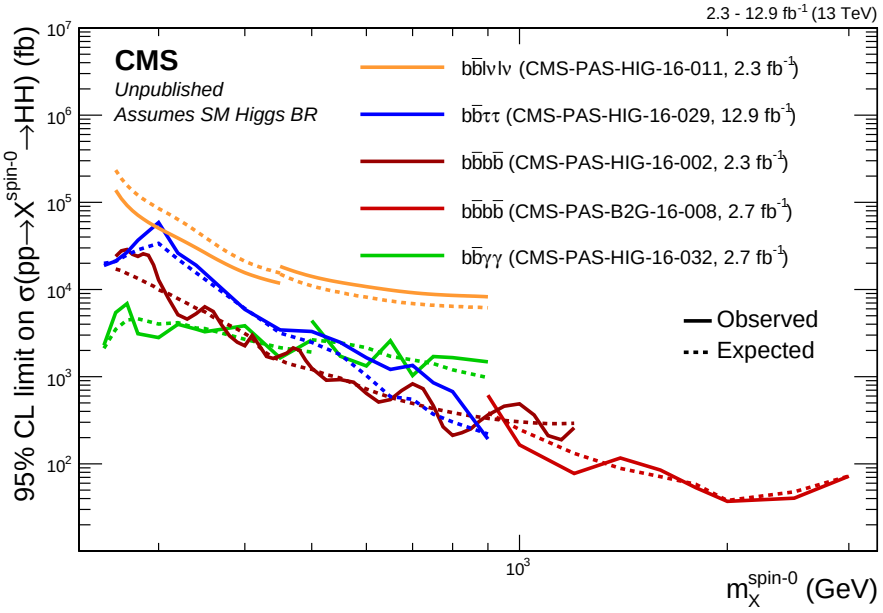


Figure 3.19: Observed and expected 95% CLs upper limits on the production cross section of $\sigma(gg \rightarrow X \rightarrow hh)$ obtained by different run II analyses assuming spin-0 hypothesis.

leptons and two jets (without b-tagging) by applying a reweighting procedure mimicking the b-tagging effects.

- Have a dedicated analysis optimization for each simulated mass point instead of using two BDT discriminant for the full mass range.

Chapter 4

Model independent search for new physics at the LHC

In previous chapter we have presented a dedicated search for new resonances decaying into two Higgs bosons. The observations of this search were in agreement with the SM predictions. We set thus limits on the resonance production cross section. A very large number of other searches dedicated to specific models have been carried out with LHC data and, so far, no striking evidence supporting any of the tested BSM hypotheses have been found.

On the one hand, this situation suggests that, if new physics is visible with current LHC data, it might have been missed because we have been looking at wrong places. Given the amount of possible BSM physics scenarios, it is difficult to develop dedicated searches for every single one of them. Moreover, there is also the possibility that new physics lies somewhere nobody ever thought of yet. Designing searches for specific BSM hypotheses would thus always lead us to look at wrong places. These two observations motivate the development of alternative approaches to look for new physics while relying the least possible on particular BSM hypotheses.

A lot of effort have been deployed to enlarge the scope of single searches. For example, the framework of *Effective Field Theory* (EFT) allows to parametrize the effect of high scale new physics on the scales probed at the LHC while being almost agnostic about the details of the new phenomena happening at higher scale. This EFT approach allows to probe new physics in a more model-

independent way but also makes the assumption that there is an important gap between the scales probed at the LHC and the new physics scale. Another model-independent approach relaxing this assumption is to perform systematic scans of various distributions a priori sensitive to hard scale phenomena. Such analyses have been carried out with 2015 data by scanning e.g. the di-jet [75], di-photon [76], di-lepton [77] or di-boson [78] invariant mass spectra. No significant deviation could be spotted. A more thorough approach based on 8 TeV data has been developed to study a very large number of final state topologies by considering events with arbitrary number of jet(s), photon(s), electron(s) and muon(s) [79]. The search consists in a scan of three kinematic quantities : the sum of all scalar transverse momenta, the combined invariant mass of all objects and the missing transverse energy of the final state. Again, the SM hypothesis could not be excluded.

In this chapter, we investigate the possibility of designing an innovative method to search for new physics without any bias towards a specific BSM scenarios and by studying highly non-trivial phase space regions. This approach is complementary to the one mentioned in the previous paragraph in the sense that the latter probes hard scales using various final states while the former probes a large phase space of a more restricted final state. The event sub-set containing two leptons and two b-jets has been chosen because it is easy to select by the trigger system, populated by SM processes with reasonably small cross sections and potentially sensitive to a large number of BSM signatures, as discussed in Sec.1.1 and 1.3.2. The phase space regions we propose to study are defined based on the MEM weights defined in Sec. 1.4 and 2.5 which allow to quantify the agreement of a given event with the various SM background processes.

4.1 Analysis set-up

This analysis is performed under the same conditions than the hh dedicated analysis, technical aspects such as object selection or MC datasets used can thus be found in Chap. 3. The only difference with respect to the dedicated search is that a reprocessing of data and MC reconstruction algorithm was performed with an updated CMSSW version including, in particular, a better

detector alignment knowledge. During this new MC campaign, more DY NLO events were generated which allowed us to use this sample instead of the LO one used in the previous chapter's analysis.

4.1.1 Signal samples

Even though we aim at an analysis design without any bias towards a specific BSM, we need signal benchmarks to assess the power of the developed model independent search (MIS) and to allow comparison with dedicated searches. The choice of BSM benchmarks has been driven by the availability of such dedicated analyses in reasonably similar conditions. In total, signals from five different scenarios have been considered:

- **Two-Higgs-Doublet models:** These signal samples are generated in the 2HDM framework described in Sec. 1.3.2. They correspond to a specific scenario called *alignment limit* where the h boson behaves as the SM Higgs boson [26]. The studied signature is $\mathrm{pp} \to H \to ZA \to l^+l^-b\bar{b}$ with various combination of H and A masses. The corresponding dedicated search is described in Ref. [80]. For our purpose, a complete mass scan is of no relevance, the chosen probed masses (in GeV) are: $(m_H = 500, m_A = 300)$ and $(m_H = 800, m_A = 700)$.

- **New spin-0 resonances X coupled to the Higgs boson:** Signal samples corresponding to the process $gg \to X \to hh \to \mathrm{bbl}\nu\mathrm{l}\nu$ with $m_X = 400, 650$ and 900 GeV have been considered. The dedicated search is one of the main results in this thesis and is presented in Chap. 3.

- **SM production of a Higgs boson pair:** this sample corresponds to the process $gg \to \mathrm{hh} \to \mathrm{bbl}\nu\mathrm{l}\nu$ as predicted by the SM. As mentioned in Sec. 3.1, this process exists in the SM but the predicted cross section is very small which renders its observation very challenging. The dedicated search is described in Ref. [81].

- **Dark Matter:** signal samples corresponding to dark matter production in association with a top quark pair have been considered. These samples are generated with a spin-0 scalar mediator Φ which couples to the top

quark and decays into dark matter particle χ as described in Sec. 1.3.2 and in the dedicated search document [82]. The considered masses are : $(m_\Phi = 10,\ m_\chi = 1)$, $(m_\Phi = 100, m_\chi = 1)$ and $(m_\Phi = 500, m_\chi = 1)$ in GeV.

- **Supersymmetry:** finally, the power of the method is tested on signal samples belonging to the supersymmetric model family described in Sec. 1.3.2. This sample corresponds to stop ($\tilde{t}$) pair production where both stops decay into a SM top quark and the *lightest supersymmetric particle* χ (LSP). Two samples are considered: $(m_{\tilde{t}} = 500, m_\chi = 325)$ and $(m_{\tilde{t}} = 850, m_\chi = 100)$ in GeV. The dedicated search [83] with the event selection closest to ours has been performed requiring two leptons plus at least one b-jet and exploits 12.9 fb^{-1} which complicates the comparison. The sensitivity of the dedicated search is extrapolated to 2.3 fb^{-1} of integrated luminosity (L) assuming that it is proportional to $\sqrt{L}$. For these reasons the comparison between the MIS and the dedicated search is approximate and should be interpreted with caution.

For the sake of readability, only one benchmark per BSM family will be shown on the forthcoming plots. The total cross sections times branching ratio of all these BSM processes are arbitrarily set to 1 pb. The color convention for the MC backgrounds is as described in Fig. 2.10. Note also that, unless otherwise stated, the uncertainty band on the plots includes all the systematic uncertainties mentioned in Sec. 3.5 except the normalization uncertainties.

4.1.2 Background reweighting

One of the major difficulty of a MIS comes from the absence of a well identified signal region. This indeed implies the absence of control regions to normalize the backgrounds and constrain the systematic uncertainties. As we have already seen in the previous analysis, there is a normalization discrepancy between data and MC that can easily be absorbed by applying process dependent global scale factors (SF).

We make thus the assumption that new physics is not visible at the inclusive llbb stage and reweight the two dominant backgrounds ($\mathrm{t\bar{t}}$ and DY) based on

	$\mu\mu$	ee	μe
$\mathrm{t\bar{t}}$	0.98 ± 0.02	0.89 ± 0.02	0.95 ± 0.001
DY	0.79 ± 0.03	0.73 ± 0.03	-

Table 4.1: $\mathrm{t\bar{t}}$ and DY reweighting scale factors together with their statistical uncertainties as obtained by the *Maximum Likelihood Estimation* method on the di-lepton invariant mass distributions.

the m_{ll} distribution. This assumption is legitimated by the fact that if new physics was already visible at such an inclusive stage, it would have been already spotted by other analyses. The $\mathrm{m_{ll}}$ distribution has been chosen for its good $\mathrm{t\bar{t}}$ to DY discrimination, as illustrated on Fig. 4.1.

To absorb potential trigger or reconstruction discrepancies across the different lepton channels, one derives SF separately for the $\mu\mu$, ee and μe categories (only $\mathrm{t\bar{t}}$ SF are derived in the μe category as the DY contribution is very small in this channel). The SF extracted via the *Maximum Likelihood Estimation* technique are shown on Tab. 4.1 together with their statistical uncertainties. The Fig. 4.1 also shows the di-lepton distributions without and with the SF applied on.

4.2 Method description

As explained at the beginning of this chapter, one wants to investigate the possibility of developing a MIS which probes highly differential phase spaces in the llbb topology. One also wants to maximize the sensitivity of the search to any potential deviation from SM prediction without involving any BSM hypotheses. The only handle we have are thus the SM backgrounds populating the llbb topology. In this respect, the idea pursued when designing the method algorithm is to maximize the separation among the different SM backgrounds. This requires observables best characterizing them. We have thus naturally

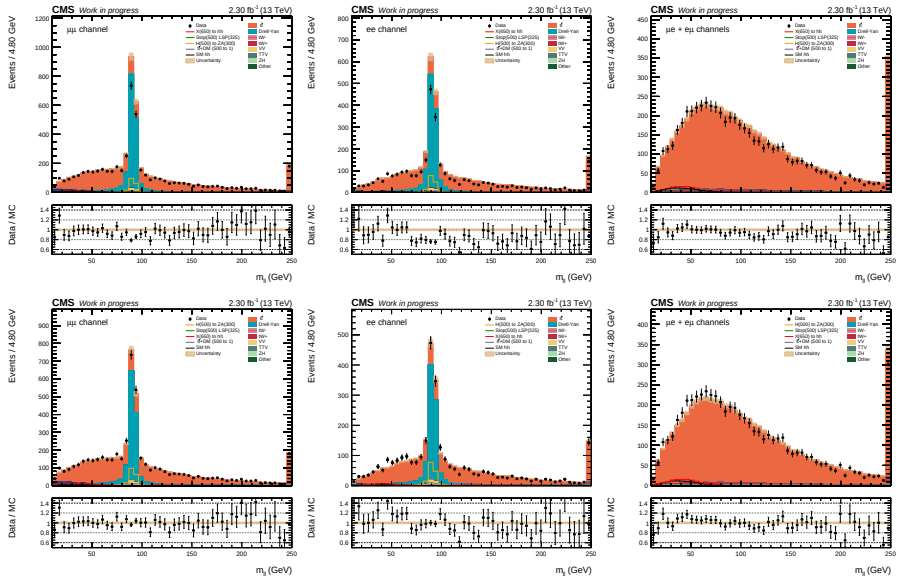


Figure 4.1: Di-lepton invariant mass distribution used to derive $\mathrm{t\bar{t}}$ and DY reweighting scale factors for the $\mu\mu$ (left), ee (middle) and μe (right) channels. Top and bottom plots show the distributions before and after applying the reweighting, respectively. The uncertainty bands represent the MC statistical error and the luminosity uncertainty.

chosen the MEM weights under the different background hypotheses to build our final discriminant.

4.2.1 Matrix Element Method weights

For technical reasons, it is difficult to compute the MEM weights under all the hypotheses of the backgrounds populating the llbb final state. Indeed, the MEM weight computation is heavily time consuming and some backgrounds such as $\mathrm{t\bar{t}}V$ have more final state objects than the one we reconstruct. We chose thus to compute the MEM weights under the hypotheses of the dominant SM backgrounds corresponding to the llbb final state. Table 4.2 shows the expected yields for the different SM contributions after requiring two leptons and two b-jets. The VV contribution is heavily dominated by the $ZZ \rightarrow l^+l^-q\bar{q}$ process which accounts for more than 85% of the quoted yield while the dominant Zh contribution is due to $Z(l^+l^-)h(b\bar{b})$. We compute therefore the MEM weights under the following hypotheses: $t\bar{t}$, DY, $\bar{t}W^+$, tW^-, ZZ and Zh hypotheses. The interest of separating $\bar{t}W^+$ and tW^- hypotheses is that it avoids wasting sensitivity to new physics scenarios having different impact on the two processes such as departure from the SM prediction of the CKM element $|V_{td}|$ [84].

We compute thus six MEM weights for each reconstructed event. To mitigate the fact that the typical order of magnitude of these weights differs from one hypothesis to another, each weight is divided by a constant defined as its average value over the full MC samples:

$$W_i^{\text{Rescaled}} = \frac{W_i}{< W_i >}. \tag{4.1}$$

This constant $< W_i >$ is computed a posteriori (i.e. after having computed the non-rescaled weights on all the MC samples), for each hypothesis i separately and by taking into account the different MC corrections. Figures 4.2 and 4.3 show the opposite of the base-10 logarithm of the various W_i^{Rescaled} for data and MC, assessing that these complicated quantities are well modeled by simulations. Note that the events lying on the left-hand side of these distributions are the most compatible with the tested hypothesis.

SM samples	
$t\bar{t}$	11123.8 ± 301.0
Drell-Yan	2325.9 ± 62.4
$\bar{t}W^+$	194.4 ± 6.5
tW^-	190.6 ± 6.4
$\mathrm{t\bar{t}}V$	59.9 ± 1.7
VV	56.8 ± 1.6
Zh	28.6 ± 0.8
Other	24.1 ± 1.4
Total $\pm$ (stat.) $\pm$ (syst.)	$14004.1 \pm 28.4 \pm 378.1$

Table 4.2: SM processes yields after requiring two leptons and two b-jets, applying the background scale factors reweighting. The quoted uncertainty includes to the MC statistical error and the luminosity uncertainty.

The weights corresponding to processes where two initial state particles lead to four final state particles (llbb) with well reconstructed resonances (i.e. without neutrino in the final state) are shown on Fig. 4.2 while Fig. 4.3 shows the weights under hypotheses with two neutrinos in the final state. Looking for instance at the W_{DY}^{Rescaled} distribution from Fig. 4.2, one sees that this quantity brings a very important discriminating power with $\mathrm{t\bar{t}}$ and tW processes while Zh and ZZ events tend to be more compatible with the DY hypothesis, as expected from the presence of a Z boson decaying to two leptons. As shown in next section, a great discriminating power between SM processes can be achieved by combining these weights together.

A thorough analysis has been performed on the data events present in the tail of the distributions shown on Fig. 4.2, in a phase space region characterized by a low compatibility with the tested hypotheses. Three events have been found to lie simultaneously in the tails of the DY, ZZ and Zh distributions. A detailed description of their kinematics is shown on Tab. 4.3. One sees that all of them are characterized by a di-jet invariant mass above 200 GeV. Two

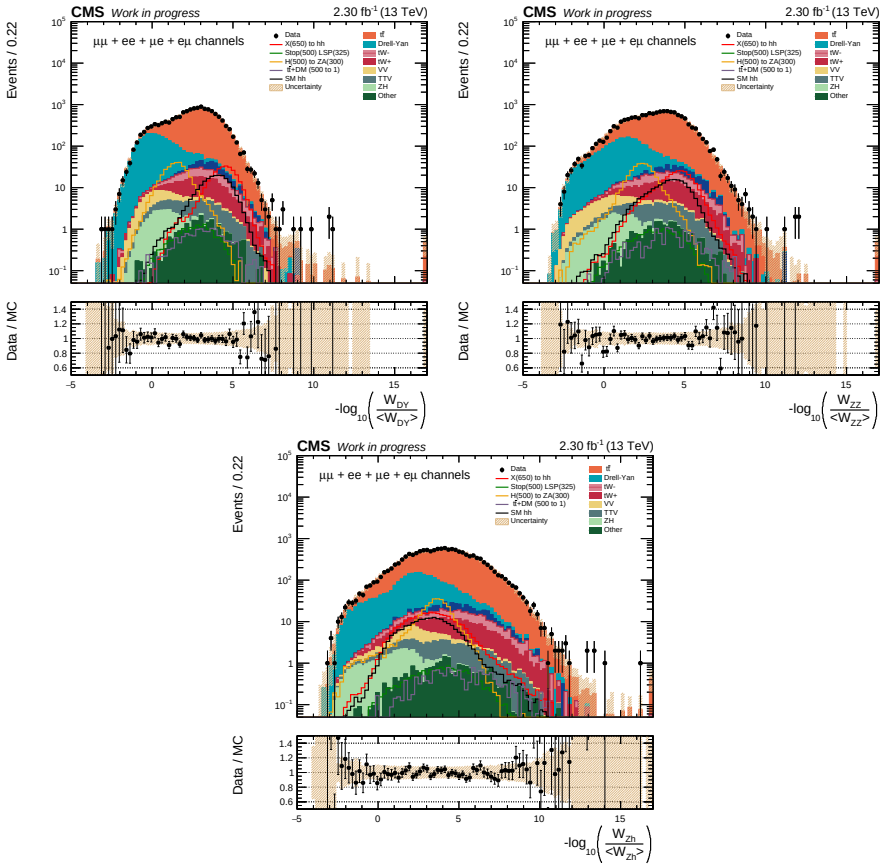


Figure 4.2: Opposite of the base-10 logarithm of the rescaled MEM weights after requiring two leptons and two b-jets, applying the background scale factors reweighting. From top to bottom and left to right, one shows the weights under the hypotheses Drell-Yan, ZZ and Zh. The last bin includes the overflow, populated by events with a low compatibility with the tested hypothesis.

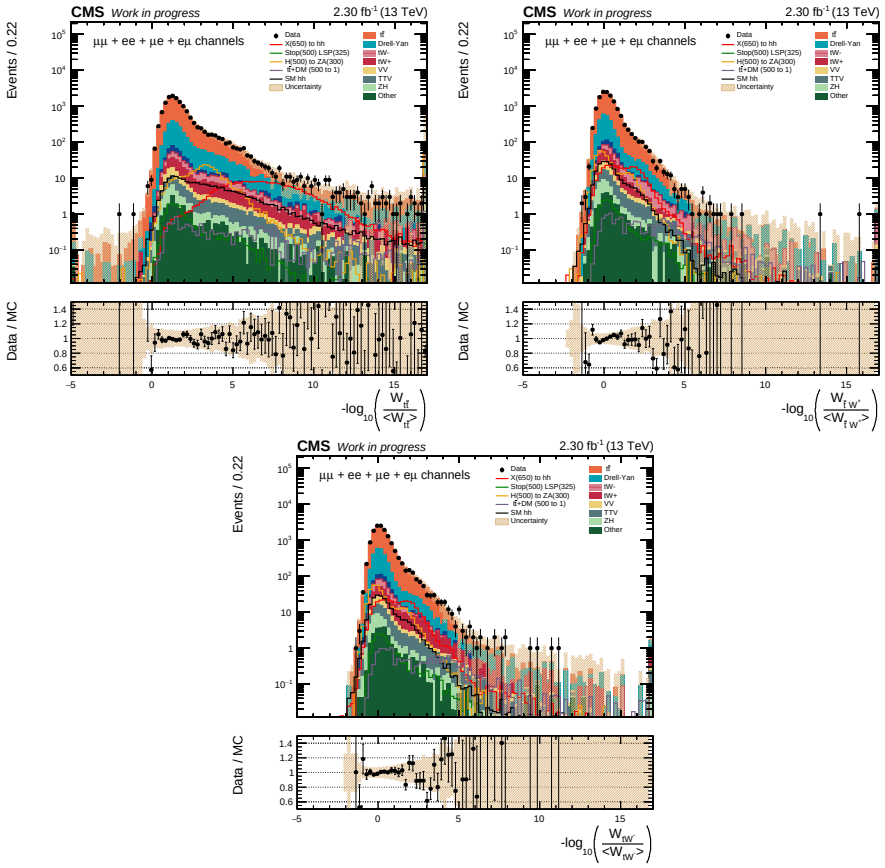


Figure 4.3: Opposite of the base-10 logarithm of the rescaled MEM weights after requiring two leptons and two b-jets, applying the background scale factors reweighting. From top to bottom and left to right, one shows the weights under the hypotheses $t\bar{t}$, tW^{+} and tW^{-}. The last bin includes the overflow, populated by events with a low compatibility with the tested hypothesis.

p_T^{lep1}	p_T^{lep2}	p_T^{jet1}	p_T^{jet2}	$E_{\mathrm{T}}^{\mathrm{miss}}$	m_{ll}	ΔR_{ll}	m_{jj}	ΔR_{jj}	Flavor
60	23	146	56	93	60	1.8	277	3.7	$\mu\mu$
103	18	158	57	49	90	3.1	218	3.3	$\mu\mu$
68	30	178	65	9	117	2.7	219	3.1	μe

Table 4.3: Characteristics of the three data events lying in the tail of the DY, ZZ and Zh weight distributions. From left to right one shows the leading lepton p_{T}, sub-leading lepton p_{T}, leading jet p_{T}, sub-leading jet p_{T}, missing transverse energy, di-lepton invariant mass, angular distance between the two leptons, di-jet invariant mass, angular distance between the two jets and the lepton flavor category. Energies are expressed in GeV.

have a di-lepton invariant mass away from the Z mass by about 30 GeV and one is, in addition, characterized by a $E_{\mathrm{T}}^{\mathrm{miss}}$ of about 100 GeV. To quantify a possible local excess, one derives the observed and expected yields requiring, arbitrarily, $W_{DY}^{\mathrm{Rescaled}} > 8.37$ & $W_{ZZ}^{\mathrm{Rescaled}} > 10.32$ & $W_{Zh}^{\mathrm{Rescaled}} > 13.13$. After these requirements, we observe 3 ± 1.6 data events while expecting 0.9 ± 0.2 events (without considering the JEC/JER and cross-section uncertainties) which does not allow to draw significant conclusion.

4.3 Method implementation

As stated at the beginning of this section, one wants to build a discriminant which reveals phase space regions potentially never probed by other dedicated analyses and which separate the different backgrounds among each other. One way to achieve this is to recursively separate one of the SM background against the others, using the Matrix Element Method weights under the respective process hypotheses. The recursive splitting (based on MC) starts with the whole llbb topology and builds a tree structure where, at each step, two daughter boxes are defined: one corresponding to the typical phase space of the process one wants to separate and the other corresponding to the phase space characterizing all the other SM processes considered. Note that the tree is built using

only events whose two b-jets pass the Tight b-tagging working point to place ourselves in a phase space region as pure as possible in b-quark content and to moderate the needs in computing resources (around a factor five less events have to be processed at each step of the splitting).

To actually perform this recursive splitting one needs discriminants which reasonably separate *each* of the considered SM backgrounds against *all* the others. The discriminant allowing to separate the background i from the others is based on the ratio of the rescaled weight under the hypothesis i over the sum of the rescaled weights under the other hypotheses:

$$\frac{W_i^{\text{Rescaled}}}{\sum\limits_{j\neq i} W_j^{\text{Rescaled}}}. \tag{4.2}$$

For technical reasons one takes the opposite logarithm of the ratio in Eq. (4.2) and normalizes it between 0 and 1 with the following function:

$$\text{Norm}(x) \equiv \frac{\arctan(x)}{\pi} + \frac{1}{2}. \tag{4.3}$$

Putting all this together one defines the discriminant used to separate the SM process i from the others as

$$D_i \equiv \text{Norm}\left(-\log_{10} \frac{W_i^{\text{Rescaled}}}{\sum\limits_{j\neq i} W_j^{\text{Rescaled}}} \right). \tag{4.4}$$

The six discriminants used to build the tree are shown on Fig. 4.4. One sees that the process i lies at high D_i values whereas the other processes tend to be at lower D_i values, as desired. Looking for instance at D_{tW^+} and D_{tW^-} (middle plots), one sees that the built discriminant is, in particular, able to separate the two $tW^{\pm}$ processes.

We have now defined, based on SM information only, the discriminant used to recursively split the phase space. When building the tree described earlier, one has to choose at each step which background will be separated from the others and where to cut on the associated discriminant. Both choices are based on

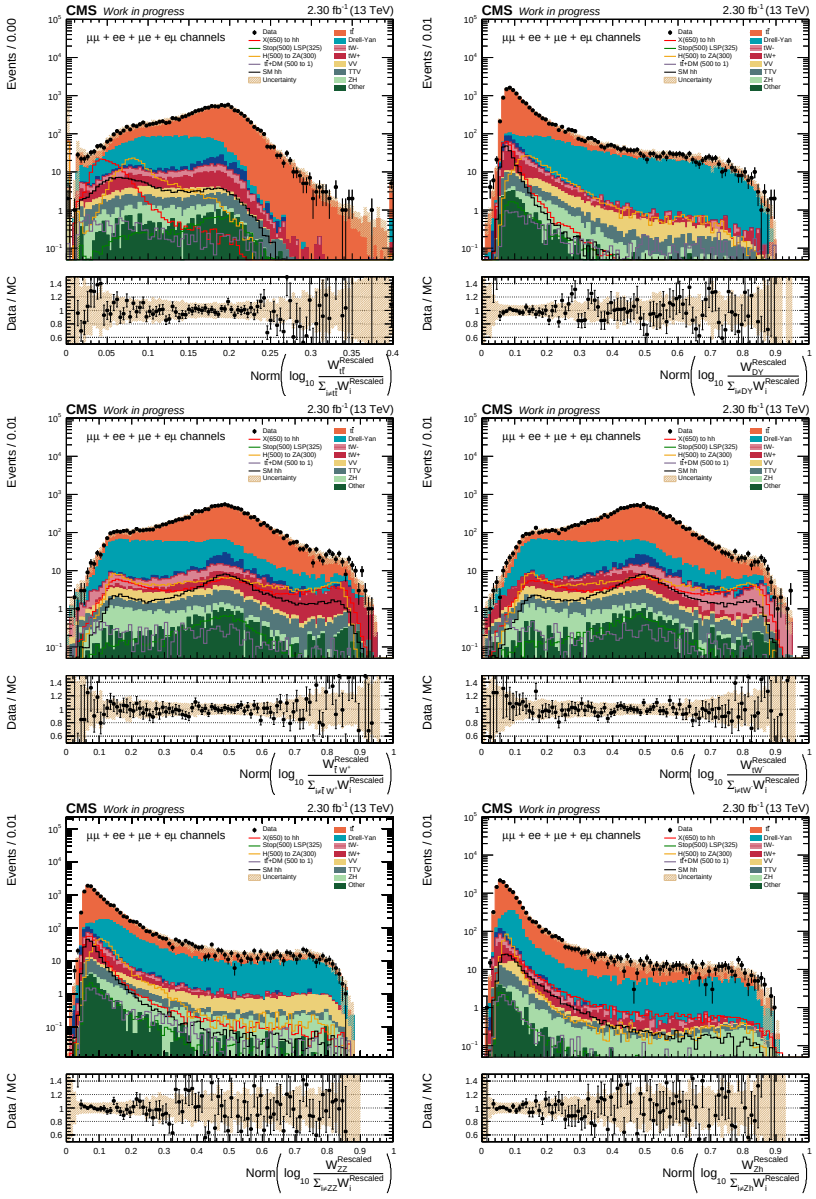


Figure 4.4: D_i distributions after requiring two leptons and two b-jets, applying the background scale factors reweighting. From top to bottom and left to right one shows: $D_{t\bar{t}}$, D_{DY}, D_{tW^+}, D_{tW^-}, D_{ZZ} and D_{Zh}. The last bin in $D_{t\bar{t}}$ includes the overflow, populated by events with a high(low) compatibility with the $\mathrm{t\bar{t}}$(other SM processes) hypotheses.

so-called *Gini* index broadly used in *Boosted Decision Trees*. Let us define the *purity* of the process i, P_i, as $\frac{N_i}{N_{\text{tot}}}$ where N_i is the number of expected events of type i and N_{tot} is the number of expected events considering all the SM processes. The *Gini* index is defined as:

$$Gini \equiv N_{\text{tot}} P(1-P). \tag{4.5}$$

The discriminant (hence the background to be separated from the others) and the value of the cut defining the two daughter boxes from the mother box are chosen to maximize the quantity

$$G \equiv Gini_{\text{mother}} - Gini_{\text{daughter}_1} - Gini_{\text{daughter}_2}. \tag{4.6}$$

We compare thus, at each step, six different possibilities to separate the phase space and the splitting is chosen to maximize the G index.

To prevent the tree from growing indefinitely one needs criteria to stop its construction: a given box will not be split anymore if one of the two following conditions is fulfilled.

- A phase space region where no background can efficiently be separated anymore from the others is reached. This stopping criteria is fulfilled when the selection efficiency of the background i for the chosen cut is smaller than ϵ_{cut} times the other background selection efficiency:

$$\frac{\epsilon_i}{\epsilon_{\text{All bkg but } i}} < \epsilon_{\text{cut}}.$$

- The error from finite MC statistics in at least one of the two daughter boxes exceeds $Stat_{\text{cut}}$ of the total yield in this box.

The method has thus two arbitrary parameters, ϵ_{cut} and $Stat_{\text{cut}}$, whose values will determine the deepness of the tree. One of the goal of this work is to give insights about how the sensitivity of the method evolves when modifying these parameters. Let us describe in the next section a first example of tree built with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$.

Process	Cut Value	$\epsilon_{\text{process}}$	ϵ_{other}	G
$t\bar{t}$	0.15	0.72	0.31	24.86
Drell-Yan	0.20	0.74	0.15	45.44
$\bar{t}W^+$	0.77	0.18	0.01	0.80
tW^-	0.76	0.19	0.01	0.63
ZZ	0.24	0.58	0.1	0.06
Zh	0.53	0.17	0.02	0.01

Table 4.4: Parameters of the cuts leading to the best G for each background at the first step of the tree building with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$ (see Sec. 4.3). The first row is the process we try to separate from the other, the second row gives the value of the chosen cut on the corresponding discriminant while the last rows show the efficiency for the considered process, the efficiency for the other SM processes and the G parameter corresponding to this cut, respectively.

4.4 Tree example

The tree building starts with the six discriminants shown on Fig. 4.4. Each of them is scanned to define the set of cuts verifying the criteria on the statistical error with $Stat_{\text{cut}} = 0.1$. The one leading to the best G is chosen for each of the six distributions. Table 4.4 gives details about the chosen cuts for the six different SM processes one tries to separate at the first step of the splitting. One sees that each of these cuts verifies the criteria $\frac{\epsilon_i}{\epsilon_{\text{All bkg but } i}} < 2$. We choose thus the one maximizing G, which is the cut on D_{DY}, to perform the first phase space splitting. Two daughter boxes are thus defined according to $D_{DY} \leq 0.2$ and $D_{DY} > 0.2$. The procedure is repeated in each of these boxes and so on until the stopping criteria do not allow to split further. Let us detail one of the paths leading to a final box.

Figure 4.5 shows the beginning of the tree resulting from the parameter choice $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$ (called nominal tree in next section). One sees that, as described in the previous paragraph, the first discriminant used to split

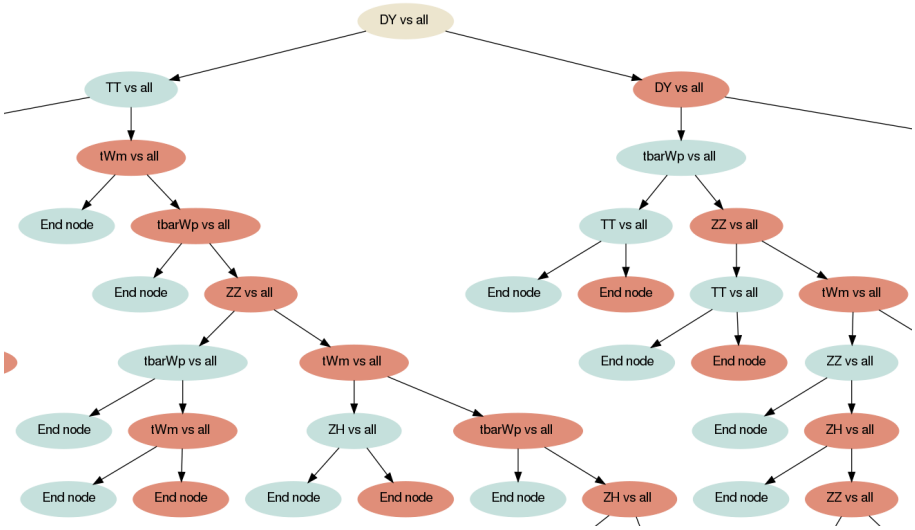


Figure 4.5: Tree visualization. We adopt the following convention: the name of the mother box writes *i vs all* which implies that the daughter box on the left hand side (blue node) corresponds to the 'i-like' phase space region ($D_i > x$) while the right hand side daughter box (red node) corresponds to the phase space region populated by the other backgrounds ($D_i \leq x$). We show here the beginning of the tree resulting from the recursive phase space splitting with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$.

Process	Cut Value	$\epsilon_{\text{process}}$	ϵ_{other}	G
$t\bar{t}$	0.17	0.41	0.12	10.1260
Drell-Yan	0.38	0.53	0.29	6.1572
$\bar{t}W^+$	0.67	0.13	0.01	0.0307
tW^-	0.66	0.18	0.02	0.0371
ZZ	0.23	0.63	0.28	0.0258
Zh	0.11	0.74	0.39	0.0042

Table 4.5: Parameters of the cuts leading to the best G for each background at the second level of the tree building with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$, in the "DY-like" daughter box. The first row is the process we try to separate from the other, the second row gives the value of the chosen cut on the corresponding discriminant while the last rows show the efficiency of the considered process, the efficiency of the other SM processes and the G parameter corresponding to this cut, respectively.

the phase space is D_{DY}. In the left hand side daughter box (the one populated with events verifying $D_{DY} > 0.2$), the cut on the $\mathrm{t\bar{t}}$ discriminant is the one leading to the best G as shown on Tab. 4.5. In the resulting right hand side box (populated by events verifying $D_{DY} > 0.20$ & $D_{\mathrm{t\bar{t}}} \leq 0.17$), the cuts on the DY and $\mathrm{t\bar{t}}$ discriminants lead to the two best G but do not verify $\frac{\epsilon_i}{\epsilon_{\text{All bkg but } i}} < 2$ (see Tab.4.6). The discriminant leading to the third best G, D_{tW^-}, is thus used to further split the phase space. The tree growing stops in the resulting left hand side box because no further phase space splitting verifies the two criteria mentioned in previous section. One can notice that even at the early stage of the tree building shown on Fig. 4.5, each of the considered SM backgrounds intervenes the phase space splitting several times. The full tree, shown on Fig. 5.5 in App. 5.3.1, defines 148 final boxes and spans over 18 levels of depth i.e. some of the final boxes are the result of 18 consecutive phase space splittings.

The figure of interest i.e. the discriminant we use to test the SM and look for BSM physics is built from the yields in each of the final boxes resulting from

Process	Cut Value	$\epsilon_{\text{process}}$	ϵ_{other}	G
$t\bar{t}$	0.14	0.41	0.30	0.9451
Drell-Yan	0.39	0.57	0.38	2.742
$\bar{t}W^+$	0.67	0.16	0.02	0.0310
tW^-	0.66	0.23	0.02	0.0373
ZZ	0.24	0.64	0.31	0.0219
Zh	0.11	0.76	0.42	0.0036

Table 4.6: Parameters of the cuts leading to the best G for each background at the third level of the tree building with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$, in the box corresponding to the path "DY-like & Non-$t\bar{t}$-like". The first row is the process we try to separate from the other, the second row gives the value of the chosen cut on the corresponding discriminant while the last rows show the efficiency of the considered process, the efficiency of the other SM processes and the G parameter corresponding to this cut, respectively.

the recursive background separation. This discriminant is shown on Fig. 4.6 with all the SM processes shown separately and on Fig. 4.7 with all the SM processes merged together. Looking at the latter, one sees that this discriminant has the peculiarity to reveal phase space regions poorly populated by SM events (at the end of the histogram) but where some signal contributions remain significant. This means that it is possible to improve an analysis sensitivity to some signals by performing an optimization totally agnostic of the BSM model they belong to. Looking for instance at the resonant di-Higgs production with m_X = 650 GeV (red histogram), one has several bins with a signal to background ratio improved with respect to the one we have in inclusive distributions such as the one shown on Fig. 4.1. On the other hand, signals such as stop pair production (green histogram) populate more the first bins (where the SM contribution is large) due to a kinematic behavior closer to the $\mathrm{t\bar{t}}$ one. The built discriminant reveal in some sense the "bar code" of the various signals. In the next section one derives results using this first phase space splitting.

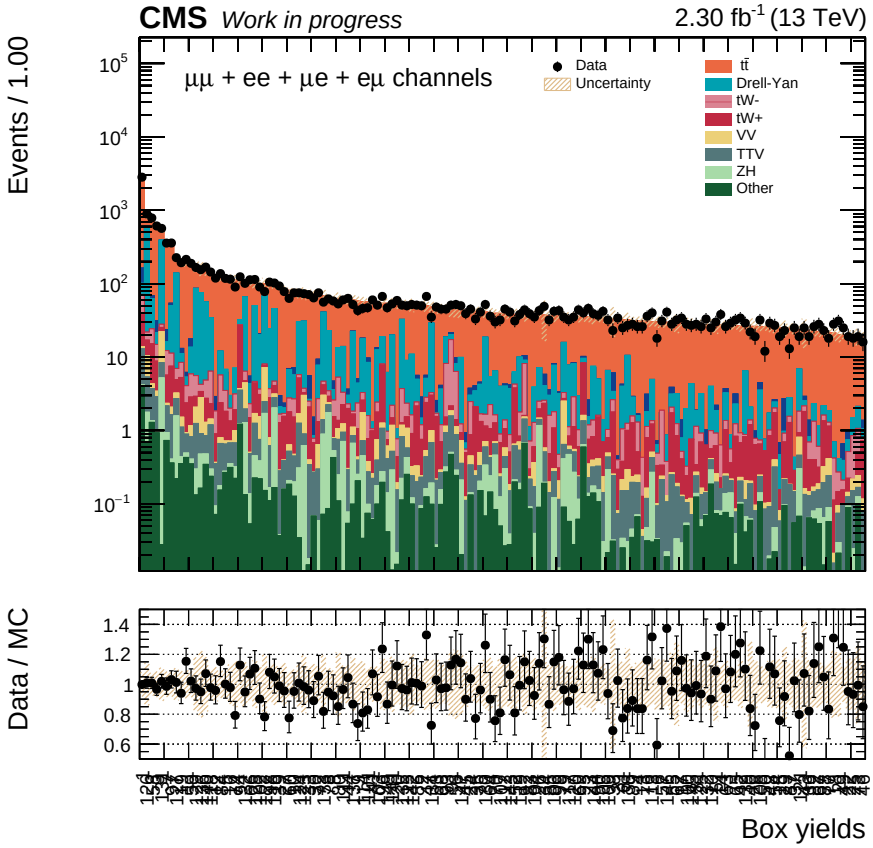


Figure 4.6: Yields in each of the final boxes resulting from the recursive phase space splitting with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$. The histogram bins have been sorted by decreasing SM expected yield.

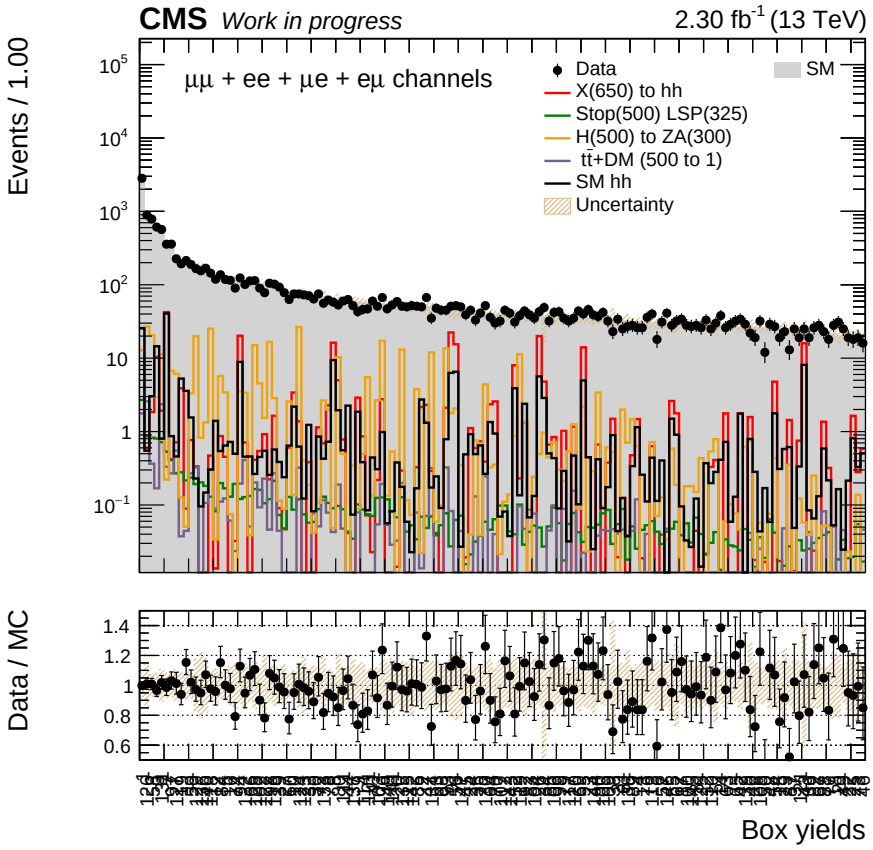


Figure 4.7: Yields in each of the final boxes resulting from the recursive phase space splitting with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$. The histogram bins have been sorted by decreasing SM expected yield and all the SM contributions are merged together.

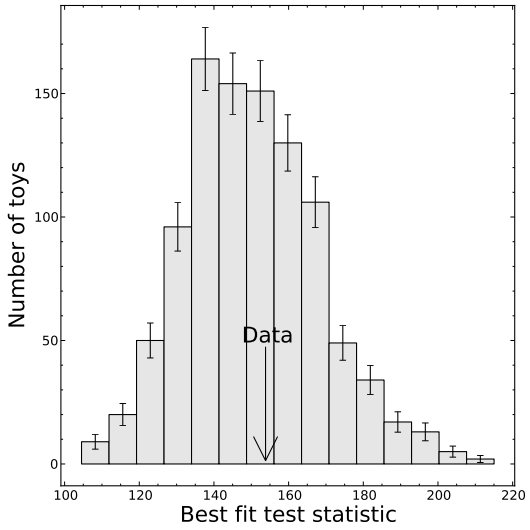


Figure 4.8: Best fit statistics as obtained from the goodness of fit with the saturated algorithm. The gray histogram represents the toys best fit statistics (1000 toys have been considered) and the black arrow is the equivalent quantity for the data distribution.

4.5 Results

The first thing we can do with this highly differential distribution is to check whether the SM hypothesis is compatible with the observed data. To this end, one performs a *Goodness of Fit* with the so-called *saturated* algorithm [85] both on data and on MC toys whose template is the predicted SM distribution. The systematic uncertainties considered for the fit are the same as the one described in Sec. 3.5 except for the additional global scaling uncertainties which are irrelevant since the DY and $\mathrm{t\bar{t}}$ background have already been renormalized to data. The values of the best fit statistics for 1000 toys and for data are shown on Fig. 4.8. One sees that the data are characterized by a best fit statistic sitting at a central place compared to the toys best fit statistics assessing thus the compatibility of the observations with the SM expectations.

Signal (masses in GeV)	CLs (Dedicated)	CLs (MIS)	Ratio
$t\bar{t}$ + DM (m_χ=1, m_ϕ=10)	25467	24469	0.96
$t\bar{t}$ + DM (m_χ=1, m_ϕ=100)	9412	20719	2.20
$t\bar{t}$ + DM (m_χ=1, m_ϕ=500)	2051	17578	8.57
$X^0_{400} \to hh$	259	545	2.10
$X^0_{650} \to hh$	91	300	3.31
$X^0_{900} \to hh$	76	264	3.49
SM hh	93	624	6.72
$H(500) \to Z\ A(300)$	237	365	1.54
$H(800) \to Z\ A(700)$	117	379	3.25
$\tilde{t}\tilde{t}\ (m_{\tilde{t}} = 500, m_\chi = 325)$	≈ 47000	42188	≈ 1
$\tilde{t}\tilde{t}\ (m_{\tilde{t}} = 850, m_\chi = 100)$	≈ 190	13078	≈ 70

Table 4.7: 95% expected CLs limits in fb obtained by the dedicated searches (second raw) and by the MIS method (third row) using the yields in each of the final boxes resulting from the tree with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$. All the systematic uncertainties described in Sec. 3.5 are taken into account except for the additional global scaling uncertainties. The cross section uncertainties on $\mathrm{t\bar{t}}$ and DY are replaced by the uncertainty on the background normalization fit. Each line in the table corresponds to the limit on a different BSM scenarios/parameter set.

Given the compatibility of the observed data with the SM prediction, one derives 95% CLs exclusion limits on the various BSM hypotheses considered. Comparing the expected limits obtained with the MIS method to the one delivered by the dedicated searches allows to assess the power of the discriminant built without any optimization towards a specific BSM. As already mentioned, we are limited by the availability of dedicated searches realized in reasonably similar condition. Though it implies that one can only do the exercise for some specific benchmarks, this comparison is a valuable information allowing to state whether or not this way of designing a BSM agnostic search method leads to reasonable sensitivity. The obtained limits are given on Tab. 4.7.

One sees that several limits from the MIS are close (about a factor 2 worse) to the one from the corresponding dedicated search which is encouraging provided that no optimization towards specific BSM's has been considered. The lowest value (0.96) corresponding to the signal "$t\bar{t}$+DM (m_χ=1 GeV, m_ϕ=10 GeV)" is explained by the fact that the dedicated search focused on signals with higher m_ϕ. The event selection in this analysis indeed requires $E_{\mathrm{T}}^{\mathrm{miss}} > 50\,\mathrm{GeV}$ which suppresses about 25% of the signal at such low ϕ mass. Looking carefully at Tab. 4.7 allows to characterize our method: MIS limits seem more competitive when the signal considered is kinematically close to a SM process. Indeed, the processes where MIS method is most competitive are

- "$t\bar{t}$ + scalar DM (m_χ=1 GeV, m_ϕ=10 GeV)" which is close to $\mathrm{t\bar{t}}$ since the amount of $E_{\mathrm{T}}^{\mathrm{miss}}$ due to the presence of DM candidate is reasonably small in such a low mass scenarios.
- "$X \to hh$ with $m_X^{spin-0} = 400$ GeV" which is the most difficult benchmark (i.e. closest to $\mathrm{t\bar{t}}$) in resonant di-Higgs search (see Sec. 3.4).
- "$H(500) \to Z\ A(300)$" which is close to the DY process due to the Z resonance and the relatively low H/A masses.
- "$\tilde{t}\tilde{t}$ $(m_{\tilde{t}} = 500, m_\chi = 325)$" which has a kinematics closer to $\mathrm{t\bar{t}}$ than the benchmark with $(m_{\tilde{t}} = 850, m_\chi = 100)$.

Inside a given model family, the MIS method tends to be less competitive in the cases where the kinematic strongly differs from the SM backgrounds such as "$t\bar{t}$ + scalar DM (m_χ=1 GeV, m_ϕ=500 GeV)"; "$X \to hh$ with $m_X^{spin-0} = 900$ GeV", "$H(800) \to Z\ A(700)$" and $\tilde{t}\tilde{t}$ $(m_{\tilde{t}} = 850, m_\chi = 100)$. For these extreme kinematic behavior it is in general easy to define a signal region free of SM background in the context of a dedicated analysis while the MIS method lacks of such signal regions by construction.

Tree	ϵ_{cut}	$Stat_{\text{cut}}$	# of final boxes
Nominal	2	0.1	148
Middle-deep	1.5	0.1	296
Deep	2	0.5	579

Table 4.8: Characteristics of the three trees built by varying the two arbitrary parameters of the method, ϵ_{cut} and $Stat_{\text{cut}}$, described in Sec. 4.3.

4.6 Study of the tree parameters

4.6.1 Sensitivity to BSM scenarios

Other trees have been built by choosing different ϵ_{cut} and $Stat_{\text{cut}}$ to characterize how the sensitivity of the method evolves with the granularity of the final discriminant. In total, three trees have been studied. The parameters of these trees are shown on Tab. 4.8 together with the deepness and number of final boxes they lead to.

The data and MC yields in each of the final boxes for the *middle-deep* and for the *deep* trees are shown on Figs. 5.6 and 5.7 in App. 5.3.1, respectively. The 95% expected CLs limits on signal production cross sections obtained from these distributions are shown on Tab. 4.9 for the three trees separately. The results for the deeper trees have to be taken with some care for the following reason: when building deep trees, the assumptions of the *asymptotic* method used to derive limits are not fulfilled everywhere (bins with very few expected events are present in the discriminant). In the case of the nominal tree, one could cross check some of the limits obtained with the asymptotic method using a more robust approach called *modified frequentist CLs* [86]. Unfortunately, this cross check could not be performed for the deeper trees due to numerical limitations.

Keeping this caution in mind, one still notices that for almost all signals considered, building deeper trees improves the analysis sensitivity. The last row in Tab. 4.9 shows the ratio of the limit obtained from the *deep* tree to the limit ob-

Signal (masses in GeV)	Nominal	Middle-Deep	Deep	$\frac{\text{Deep}}{\text{Dedicated}}$
$t\bar{t}$ + DM (m_χ=1, m_ϕ=10)	24469	18140	14484	0.57
$t\bar{t}$ + DM (m_χ=1, m_ϕ=100)	20719	15984	12891	1.37
$t\bar{t}$ + DM (m_χ=1, m_ϕ=500)	17578	13828	10641	5.19
$X^0_{400} \to hh$	545	456	467	1.80
$X^0_{650} \to hh$	300	191	142	1.57
$X^0_{900} \to hh$	264	177	76	1.68
SM hh	624	432	400	4.31
$H(500) \to Z\ A(300)$	365	312	291	1.23
$H(800) \to Z\ A(700)$	379	321	263	2.25
$\tilde{t}\tilde{t}\ (m_{\tilde{t}} = 500, m_\chi = 325)$	42187	36656	29531	≈ 0.6
$\tilde{t}\tilde{t}\ (m_{\tilde{t}} = 850, m_\chi = 100)$	13078	9422	11016	≈ 55

Table 4.9: 95% expected CLs limits on signal production cross sections (fb) using the yields in each of the final boxes resulting from three different recursive phase space splitting with the following parameter choices: $\epsilon_{\text{cut}} = 2$ & $Stat_{\text{cut}} = 0.1$ (*nominal*), $\epsilon_{\text{cut}} = 1.5$ & $Stat_{\text{cut}} = 0.1$ (*middle-deep*) and $\epsilon_{\text{cut}} = 2$ & $Stat_{\text{cut}} = 0.5$ (*deep*). The last row shows the ratio of the expected limit obtained from the *deep* tree to the limit obtained in the dedicated search. All the systematic uncertainties described in Sec. 3.5 are taken into account except for the additional global scaling uncertainties. The cross section uncertainty on $\mathrm{t\bar{t}}$ and DY are replaced by the uncertainty on the background normalization fit. Each line in the table corresponds to the limit on a different BSM scenarios/parameter set.

tained in dedicated searches. Except for the "$t\bar{t}$+DM m_χ=1 GeV m_ϕ=10 GeV" and "$\tilde{t}\tilde{t}$ $(m_{\tilde{t}} = 500, m_\chi = 325)$" signals, the MIS method is never characterized by a better sensitivity than the dedicated searches, as expected regarding our analysis strategy. However, the gap between the two approaches tends to decrease when building discriminants with higher granularity. While we had a MIS-to-dedicated limit ratio between 0.96 and 70 with 148 final boxes, it shrinks between 0.57 and 55 with 579 final boxes and most of the obtained limits start to be competitive with the dedicated one.

4.6.2 Sensitivity to SM background normalization

So far we have exploited our discriminant to check the consistency of the observed data with the SM prediction and to assess the sensitivity of the method in constraining various BSM physics scenario. Another example of application would be to use the obtained distribution to extract all the SM background processes normalization simultaneously in a given topology. Indeed, such a measurement benefits from the good separation between all the SM processes, inherent to the developed methodology.

To illustrate this use case one derives, based on a maximum likelihood fit to pseudo-data, scale factors for the six SM processes considered during the tree building. The distribution used is the yield in each of the final boxes. The pseudo data are generated using as template the sum of the six distributions corresponding to the six aforementioned processes and considering a Poisson law for the expected yields. In this respect, only the statistical error is considered.

The obtained scale factors resulting from 1000 toy experiments thrown with the nominal tree yields as template are shown on Fig. 4.9, separately for the six processes. Considering only the statistical error, the method allows to derive the background normalization of tW processes at 30% precision level in the llbb final state. The equivalent distributions for the *middle-deep* and *deep* trees are shown on Figs.5.8 and 5.9 in App. 5.3.2.

To quantify the evolution of this figure of merit as a function of the tree deepness, one provides on Tab. 4.10 the standard deviation of the scale factors ob-

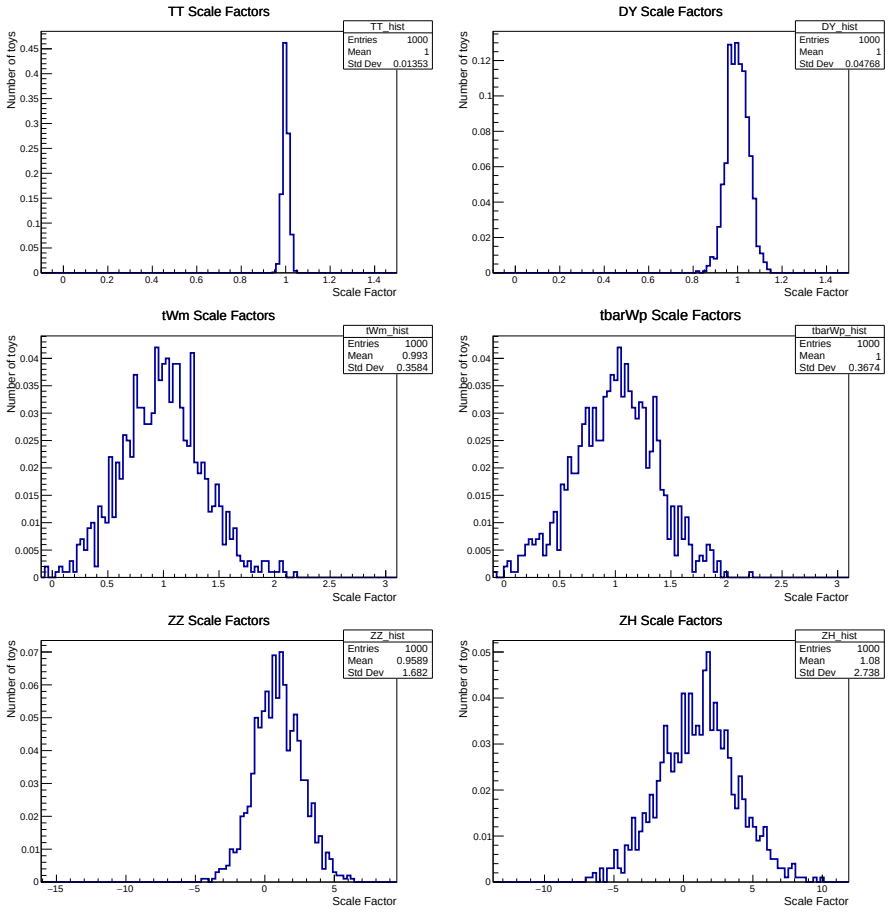


Figure 4.9: Scale factors for the six SM processes obtained from a maximum likelihood fit to pseudo-data. The distribution used to perform the fit is the yields in each of the final box resulting from the nominal splitting. A thousand pseudo-data experiments have been generated according to Poisson distribution with a mean corresponding to the expected yields considering the six SM processes.

Tree	$t\bar{t}$	DY	tW^-	$\bar{t}W^+$	ZZ	Zh
Nominal	0.013	0.048	0.36	0.37	1.7	2.7
Middle-deep	0.012	0.047	0.33	0.33	1.7	2.7
Deep	0.012	0.046	0.31	0.30	1.5	2.2

Table 4.10: Standard deviation of the scale factors for the six SM processes obtained from a maximum likelihood fit to the yields in each of the final box over a thousand toy experiments. The results from the three trees are shown separately.

tained over the 1000 pseudo-experiments, for the different trees. One sees that, the rarest the process, the better the relative improvement on the precision of its normalization when building more granular discriminant. This indicates that building deeper trees improves the sensitivity of the analysis to rare processes.

4.7 Conclusion

We have proposed an alternative methodology to search for new physics which has the peculiarity of not relying on any BSM hypothesis when designing the final discriminant. This innovative way of analyzing experimental data relies only on the known backgrounds and has the advantage of being potentially sensitive to a large spectrum of deviations from the SM predictions, including the one that would not be foreseen by any of the available BSM models. It also brings complementarity to the other model independent searches developed so far as it probes a larger phase space in a more restricted final state.

The designed methodology has been applied to the LHC 2015 dataset with events containing at least two leptons and two b-jets. No evidence for new physics has been observed. The sensitivity of the method has thus been evaluated by studying its capacity to constrain various BSM signals. Expected 95% CLs limits have been derived on the production cross section of eleven signals belonging to five different BSM model families. The limits obtained via the MIS approach have been found to be, in most cases, competitive with the cor-

responding limits obtained via dedicated searches. We also observed that the MIS methodology tends to be more competitive with dedicated searches for signals having a kinematic behavior close to one of the SM background populating the studied final state, confirming its complementarity with the searches that are already in place.

The discovery of a significant deviation from SM expectation by the MIS method would be the first step in discovering a new physics phenomenon. Indeed, the strength of this approach is to probe a large spectrum of BSM scenarios. It is however insufficient to properly characterize this deviation. A dedicated search would have to be carried out in the spotted region to, first, ensure the source of the deviation is not coming from an incomplete understanding of the detector or of the SM dynamics and second, to identify the new physics process responsible for this deviation. Hints about the latter could however be obtained with this method by looking whether the deviations in the final box yields are consistent with the introduction of a given new process.

Let us close this chapter by providing a few examples of possible extensions and modifications that could be performed to the presented analysis:

- The current implementation of the method was performed by computing the MEM weights only under six SM processes hypotheses. One could benefit from having the MEM weights from other processes such as $\mathrm{t\bar{t}}\,V$.
- Another interesting extension of this work would be to develop an automated tool that would scan a large number of trees built by varying the free parameters of the method or by varying the processes one takes into account when building the tree. Studying the ones with deviation from the SM predictions would provide hints about where new physics could be lying.
- In order to have a really thorough exploration of the available experimental data, one should not only perform this analysis in the llbb topology but extend it to all the different final states. Though this is conceptually possible, it would require very important numerical resources.

Conclusion

In 2015, the LHC delivered the very first 13 TeV proton-proton collisions. These experimental data allowed us to probe an entirely new phase space and to test our current understanding of nature at energies never reached before. In this work, we have analyzed a subset of these new data and confronted the *Standard Model* (SM) predictions to the observations. A dataset corresponding to an integrated luminosity of 2.3 fb^{-1} has been analyzed and events with two leptons and two b-jets have been considered.

In the first part of this work, we have studied the resonant production of two SM Higgs bosons. This process arise in all the *Beyond Standard Model* (BSM) scenarios postulating the existence of a new particle that couples to the Higgs boson. If the new resonance is heavy enough, it will indeed decay into two Higgs boson which results in an increased di-Higgs event yield with respect to the SM prediction. Both spin-0 and spin-2 resonances have been considered and masses were scanned between 260 GeV and 900 GeV. The observations have been found to be in agreement with the SM predictions and limits on the resonance production cross section times branching ratio have been set. For masses from $m_{\mathrm{X}} = 500$ GeV to $m_{\mathrm{X}} = 900$ GeV, the data are observed (expected) to exclude a production cross-section times branching ratio from 174 to 101 (135 to 75.8) fb. This analysis was the first one studying the $\mathrm{bb}\,\mathrm{l}\nu\mathrm{l}\nu$ di-Higgs decay mode. Comparing the obtained limits with the one from other decay channels allowed to state that the $\mathrm{bb}\,\mathrm{l}\nu\mathrm{l}\nu$ final state brings non negligible contributions to searches for di-Higgs production. At low mass resonance (below 400 GeV) the expected limit is about a factor 3 worse than the one

from $b\bar{b}\,\tau^{+}\tau-$ final state which analyzed data corresponding to an integrated luminosity of 12.9 fb^{-1} while at high mass resonance (above 500 GeV), the expected limits are between a factor 5 and 10 worse than the one from $b\bar{b}\,\gamma\gamma$ analysis which exploited a comparable integrated luminosity. This result triggered two new analyses exploiting the $\mathrm{bb\,l\nu l\nu}$ channel: one looking for non-resonant di-Higgs production with the 2015 dataset and the other looking for both scenarios (resonant and non-resonant) using the 2016 dataset.

In the second part of this work, we have investigated the possibility of designing a new kind of analysis whose optimization does not rely on any BSM scenario. While such an analysis is by construction sub-optimal for a given BSM model, it offers the advantage of being potentially sensitive to any deviation from the SM predictions. In order to increase the sensitivity of the analysis, a highly differential discriminant has been built based on the *Matrix Element Method* weights under hypotheses corresponding to the SM processes populating the studied final state. This discriminant corresponds to the yields in each of the final nodes of a tree resulting from a recursive separation of the SM processes among each other. While this new method has been implemented in the llbb final state, it is generic and could be applied to any event topology.

Yet again, the SM predictions have been found to be compatible with the observations and the power of this new approach has thus been studied by deriving expected limits for several signal benchmarks belonging to five different BSM scenarios. For most of the considered signals, the obtained limit has been found to be competitive with the one set by other searches that use the information on the signal kinematic when optimizing the final observable. For analyses realized in similar conditions, the model independent search limits are between a factor 1 and 8 worse than the one from dedicated searches. This newly developed methodology has been observed to be more competitive for signals with kinematics close to a SM process considered for the phase space splitting. This result assesses that this kind of model independent search is complementary to the other one that are already in place. The approach has two free parameters which have an impact on the structure of the tree resulting from the recursive phase space splitting. Three different trees have been studied by varying these parameters and the analysis sensitivity has been found to improve with the deepness of the tree.

As of today, no experimental analysis has been able to provide strong evidence in favor of a specific BSM scenario. Furthermore, the very large number of proposed BSM models prevents us from developing an analysis for each of them. In this context, the model independent search appears to be a promising tool to complement the other searches. Indeed, it helps ensuring that we did not fail at discovering new physics because we have not looked at the right place.

The results presented in this thesis are of course not only the product of my own work. The CMS collaboration provided the basic blocks of both analyses including the MC simulations, the data taking to which I participated via online shifts in the Data Acquisition and Tracker teams and via offline b-tagging commissioning, the scale factors derivations and trigger efficiencies measurements. I realized the resonant di-Higgs production analysis in close collaboration with two post-doctoral researchers. While every one of us took part in all the different aspects and decisions of the analysis development we could coarsely summarize the work sharing as follows. One of the post-doctoral researcher was mainly in charge of developing a general framework used also by other analyses. I developed mainly the actual implementation of the hh object selection, the MC corrections, the trigger efficiencies, the analysis strategy (event selection, BDT trainings, validations and evaluations, ...) and the treatment of the systematic uncertainties. The other post-doctoral researcher mainly developed the code related to the limit setting and to the optimization of the four regions based on the expected sensitivity. I realized alone the implementation of the model independent search.

Chapter 5

Appendix

5.1 Experimental setup and event reconstruction: Extra Material

5.1.1 Transfer function

Complete 2-dimensional energy transfer functions for the three objects that are found in the llbb final state are shown on Fig. 5.1.

5.2 Search for resonant di-Higgs production decaying into $b\bar{b}\ l^+\nu_l\ l^-\bar{\nu}_l$: Extra Material

5.2.1 Samples

The complete list of MC samples used in the analyses is shown on Tab. 5.1 and 5.2.

Process	Dataset	σ [pb]
$t\bar{t}$	TTTo2L2Nu_13TeV-powheg	87.31
	TT_TuneCUETP8M1_13TeV-powheg-pythia8	831.76
Drell-Yan	DYJetsToLL_M-10to50_TuneCUETP8M1_13TeV-amcatnloFXFX	18610
	DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX	6025.2
	DYJetsToLL_M-5to50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	71310
	DYJetsToLL_M-5to50_HT-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	224.2
	DYJetsToLL_M-5to50_HT-200to400_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	37.2
	DYJetsToLL_M-5to50_HT-400to600_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	3.581
	DYJetsToLL_M-5to50_HT-600toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	1.124
	DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	6025.2
	DYJetsToLL_M-50_HT-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	147.4
	DYJetsToLL_M-50_HT-200to400_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	40.99
	DYJetsToLL_M-50_HT-400to600_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	5.678
	DYJetsToLL_M-50_HT-600toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	2.198
VV	VVTo2L2Nu_13TeV_amcatnloFXFX_madspin_pythia8	12.05
	ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	3.22
	WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	5.595
	WZTo1L3Nu_13TeV_amcatnloFXFX_madspin_pythia8	3.033
	WWToLNuQQ_13TeV-powheg	49.997
	WZTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8	10.71
	WZTo3LNu_TuneCUETP8M1_13TeV-powheg-pythia8	4.42965
	ZZTo4L_13TeV_powheg_pythia8	1.256
single-top	ST_tW_(anti)top_5f_inclusiveDecays_13TeV-powheg	35.6
	ST_s-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8_TuneCUETP8M1	10.38
	ST_t-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8_TuneCUETP8M1	70.69
$W+$ jets	WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	61526.7
$t\bar{t}+V$	TTWJetsToQQ_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8	0.4062
	TTZToLLNuNu_M-10_TuneCUETP8M1_13TeV-amcatnlo-pythia8	0.2529
	TTZToQQ_TuneCUETP8M1_13TeV-amcatnlo-pythia8	0.5297
	TTWJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8	0.2043
$h \to WW$ (GGF)	GluGluHToWWTo2L2Nu_M125_13TeV_powheg_JHUgen_pythia8	2.05
$h \to WW$ (VBF)	VBFHToWWTo2L2Nu_M125_13TeV_powheg_JHUgen_pythia8	0.175
$Wh(h \to WW)$	HWplusJ_HToWW_M125_13TeV_powheg_pythia8	0.0393
$Wh(h \to WW)$	HWminusJ_HToWW_M125_13TeV_powheg_pythia8	0.0252
$Zh(h \to WW)$	HZJ_HToWW_M125_13TeV_powheg_pythia8	0.0406
$tth(h \to \text{non}bb)$	ttHToNonbb_M125_13TeV_powheg_pythia8	0.2151
$h \to bb$ (GGF)	GluGluHToBB_M125_13TeV_powheg_pythia8	25.34
$h \to bb$ (VBF)	VBFHToBB_M-125_13TeV_powheg_pythia8_weightfix	2.1626
$Wh(h \to bb)$	WH_HToBB_WToLNu_M125_13TeV_amcatnloFXFX_madspin_pythia8	0.173
$Zh(h \to bb)$	ZH_HToBB_ZToLL_M125_13TeV_amcatnloFXFX_madspin_pythia8	0.173
$Zh(h \to bb)$	ZH_HToBB_ZToNuNu_M125_13TeV_amcatnloFXFX_madspin_pythia8	0.159
$Zh(h \to bb)$	ggZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8(_ext1)	0.00695
$Zh(h \to bb)$	ggZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8(_ext1)	0.00695
$tth(h \to bb)$	ttHTobb_M125_13TeV_powheg_pythia8	0.2934

Table 5.1: Background Monte Carlo samples used in the analysis and their cross sections in pb.

Process	Dataset
Spin-0	/GluGluToRadionToHHTo2B2VTo2L2Nu_M-*_narrow_13TeV-madgraph/ RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM
Spin-2	/GluGluToBulkGravitonToHHTo2B2VTo2L2Nu_M-*_narrow_13TeV-madgraph/ RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

Table 5.2: Signal Monte Carlo samples used in the analysis.

5.2.2 Region definitions on m_{jj}

The signal extraction process described in section 3.4 shows two regions definition from the BDT output and the m_{jj} distribution. These regions define the final categories in which we extract the limits. The m_{jj} window is optimized by computing the expected sensitivity of the analysis using a cut-and-count in the two m_{jj} regions as a function of the m_{jj} window. Fig. 5.2 shows the expected sensitivity as a function of the lower and upper m_{jj} cuts, for signal masses of 400 and 900 GeV.

5.2.3 Boosted Decision Tree studies

BDT training input variables that are not shown in Sec. 3.4 are shown on Fig. 5.3.

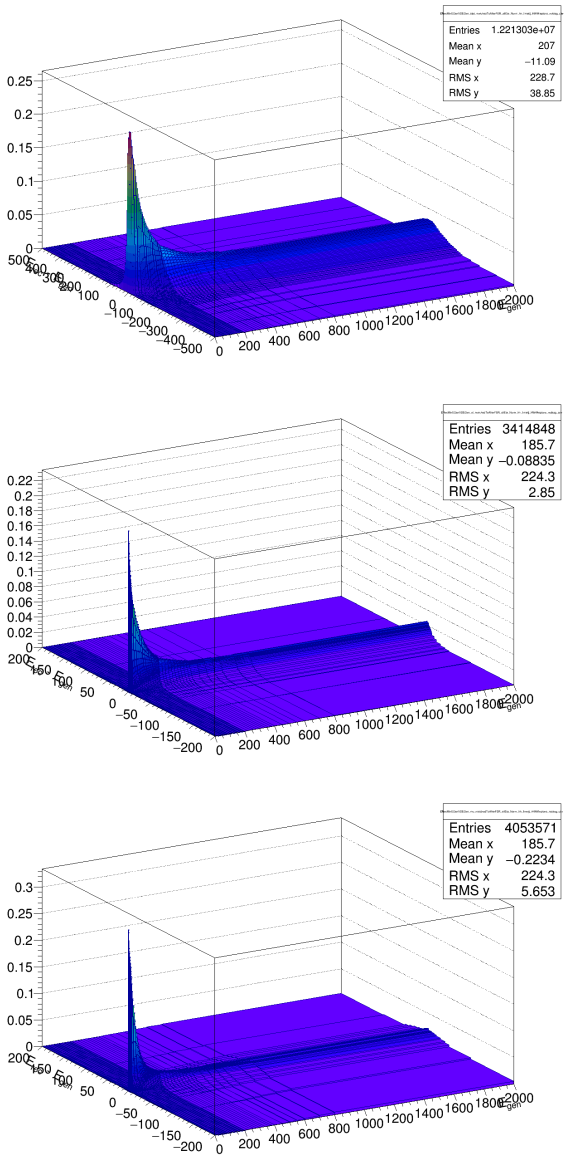


Figure 5.1: Energy transfer function for b-jets (top), electrons (middle) and muons (bottom). The x-axis shows E^{parton} from 0 to 2000 GeV, the y-axis shows $E^{\text{reco}} - E^{\text{parton}}$ with various range depending on the object (from -500 to 500 GeV for b-jets and from -200 to 200 GeV for leptons) and the z-axis is the value of the probability density function. Each "slice" in E^{parton} is normalized to unity to allow for probability interpretation.

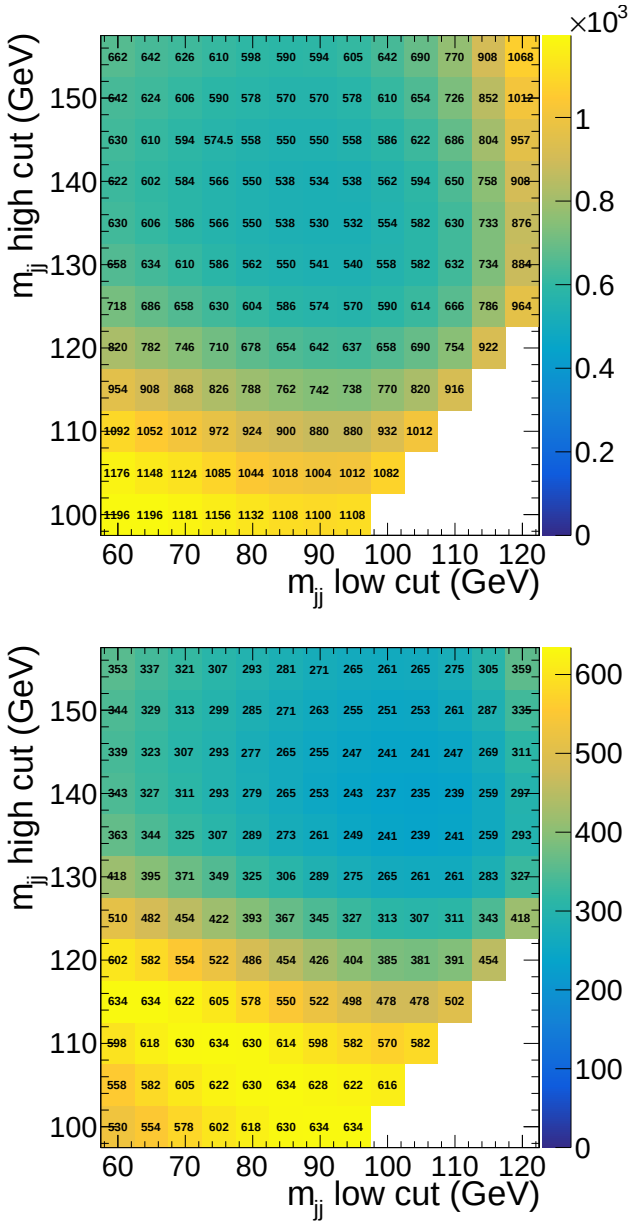


Figure 5.2: Expected CLs limit based on cut-and-count limits, as a function of the lower and upper m_{jj} cuts, for signal with m_X = 400 GeV (top) and for signal with m_X =900 GeV (bottom).

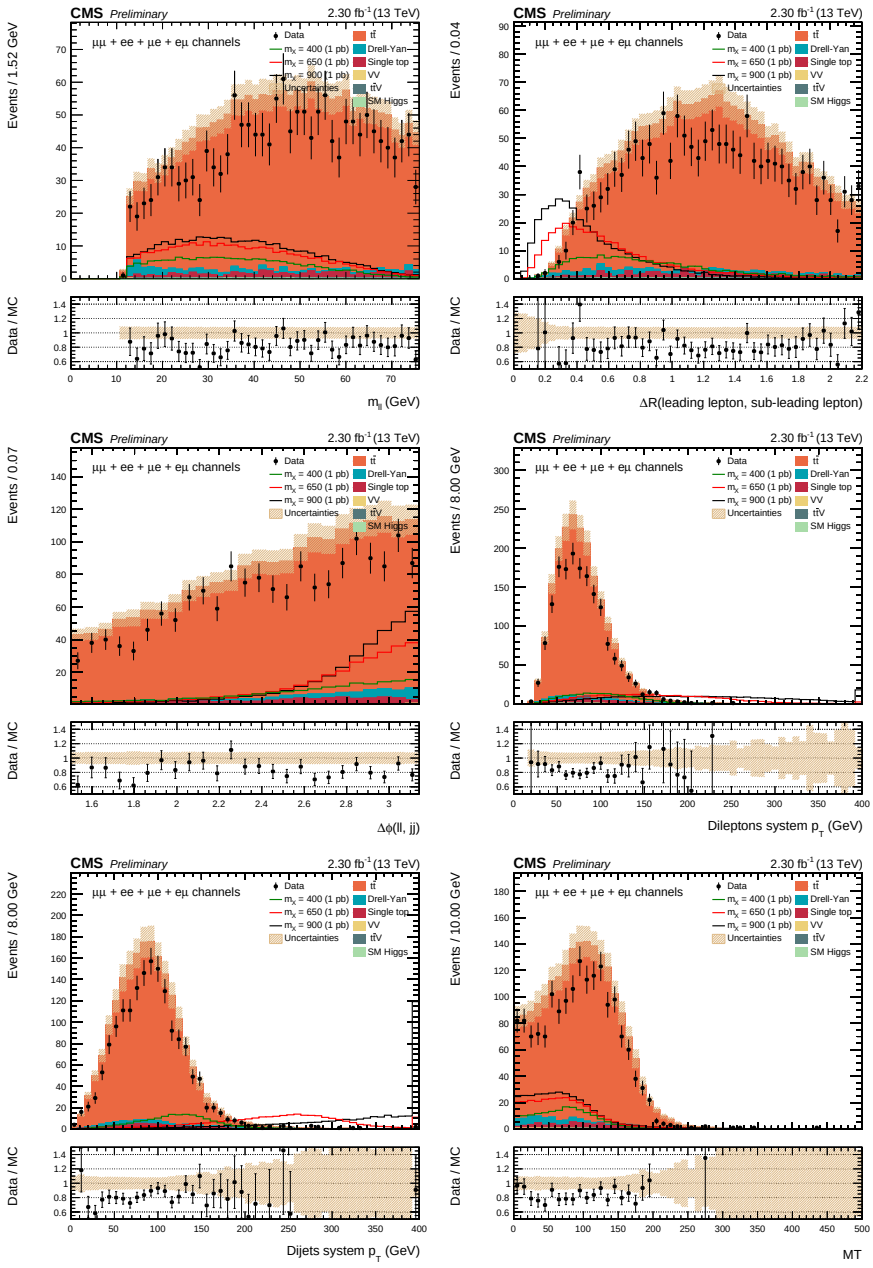


Figure 5.3: BDT training input variables that are not shown in the text body. From top left to bottom right: $\mathrm{m_{ll}}$, ΔR_{ll}, $\Delta\phi_{ll,jj}$, p_T^{ll}, p_T^{jj}, and MT distributions for data and simulated events after requiring all selection cuts described in Sec. 3.3. All lepton flavour combinations are shown together.

5.2.4 Maximum likelihood fit

The pre-fit yields in the four final regions defined in Sec. 3.4 are shown on Tab. 5.3 for the BDT 400 GeV training.

Signal samples	high-BDT 400, m_{jj}-P	high-BDT 400, m_{jj}-SB	low-BDT 400, m_{jj}-P	low-BDT 400, m_{jj}-SB
m_X = 400 (1 pb)	76.3 ± 6.7	26.2 ± 2.7	39.1 ± 3.3	35.6 ± 3.6
SM samples				
$t\bar{t}V$	0.2 ± 0.1	0.5 ± 0.1	0.8 ± 0.2	3.0 ± 0.4
SM Higgs	0.3 ± 0.1	0.4 ± 0.1	0.6 ± 0.1	2.1 ± 0.2
VV	0.2 ± 0.1	0.4 ± 0.1	0.3 ± 0.1	1.7 ± 0.3
Single top	3.2 ± 1.1	3.7 ± 1.1	10.9 ± 2.0	38.3 ± 4.6
Drell-Yan	2.9 ± 0.8	6.2 ± 1.3	8.4 ± 1.7	37.0 ± 5.3
$t\bar{t}$	75.6 ± 6.2	92.0 ± 8.1	410.8 ± 32.5	1334.7 ± 107.5
Total ± (stat.) ± (syst.)	82.4 ± 1.9 ± 6.9	103.2 ± 2.2 ± 9.5	431.7 ± 4.3 ± 34.7	1416.5 ± 8.0 ± 115.6
Data ± (stat.)	64 ± 8.0	85 ± 9.2	338 ± 18.4	1197 ± 34.6

Table 5.3: Pre-fit yields in final regions for the BDT 400 GeV training. Quoted uncertainties include both statistical and systematic uncertainties, as detailed in Tab. 3.4, except normalization and cross-section uncertainties.

We perform a maximum likelihood fit using the 4 categories of the analysis. Figure 5.4 summarizes the pull for each nuisance parameters used in the analysis, for the BDT 400.

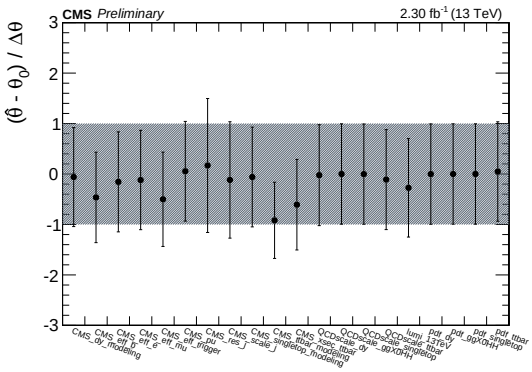


Figure 5.4: Post-fit pull distribution of the different nuisances for the BDT 400. Note that the MC stat uncertainties, included in the fit, are not pulled nor their uncertainties reduced, so they are not shown in the plot for readability.

5.2.5 Post-fit data / MC comparisons

The set of scale factors obtained from the fit described in Sec. 3.6.1 are shown on Tab. 5.4 for the BDT 400 training.

Background process	Scale factors	Post-fit uncertainties
$t\bar{t}$	0.83	2.14%
Single top	0.92	14.77 %
Drell-Yan	0.91	25.83%
SM Higgs	0.93	5.96%
VV	0.93	9.29%
$t\bar{t}V$	0.93	7.78%

Table 5.4: Post-fit normalization scale-factors for each background sample obtained from the background only fit in the three less signal like regions defined according to the BDT trained with m_X = 400 GeV.

Yields in the four final regions using the post-fit distributions can be found in Tab. 5.5 for the BDT 400 training.

5.3 Model independent search for new physics at the LHC: Extra Material

5.3.1 Tree visualization

Figure 5.5 shows the full tree visualization resulting from the phase space splitting described in Sec. 4.4. This gives intuition about how deep goes the tree and shows that some branches stop way earlier than others, depending on the path they took at earlier stages.

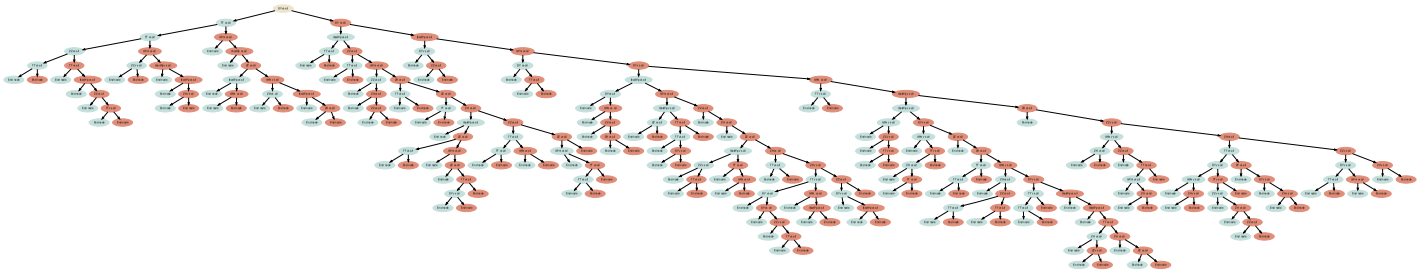


Figure 5.5: Tree visualization for the splitting realized with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.1$. We adopt the following convention: the name of the father box writes *i vs all* which implies that the daughter box on the left hand side (blue node) corresponds to the 'i-like' phase space region while the right hand side daughter box (red node) corresponds to the phase space region populated by the other backgrounds.

Signal samples	high-BDT 400, m_{jj}-P	high-BDT 400, m_{jj}-SB	low-BDT 400, m_{jj}-P	low-BDT 400, m_{jj}-SB
m_X = 400 (1 pb)	76.3 ± 1.1	26.2 ± 0.6	39.1 ± 0.8	35.6 ± 0.7
SM samples				
$t\bar{t}V$	0.2 ± 0.1	0.5 ± 0.1	0.8 ± 0.1	2.8 ± 0.3
SM Higgs	0.3 ± 0.0	0.4 ± 0.0	0.5 ± 0.0	2.0 ± 0.1
VV	0.2 ± 0.1	0.6 ± 0.1	0.4 ± 0.1	2.5 ± 0.3
Single top	2.9 ± 0.8	3.4 ± 0.9	10.1 ± 1.9	35.3 ± 5.7
Drell-Yan	2.7 ± 0.8	5.7 ± 1.6	7.6 ± 2.1	33.7 ± 8.8
$t\bar{t}$	62.8 ± 2.0	76.3 ± 2.3	340.8 ± 8.0	1107.5 ± 24.5
Total ± (stat.) ± (syst.)	69.0 ± 1.6 ± 1.6	86.9 ± 1.8 ± 2.3	360.2 ± 3.7 ± 7.7	1183.8 ± 6.7 ± 25.8
Data ± (stat.)	64 ± 8.0	85 ± 9.2	338 ± 18.4	1197 ± 34.6

Table 5.5: Post-fit yields in final regions, high-BDT & m_{jj}-P, high-BDT & m_{jj}-SB, low-BDT & m_{jj}-P, and low-BDT & m_{jj}-SB for the BDT 400 GeV training.

Figures 5.6 and 5.7 show the final discriminant resulting from the *middle-deep* and for the *deep* tree, respectively, described in Sec. 4.6.1.

5.3.2 SM background scale factors

Figure 5.8 and 5.9 show the scale factors for the six SM processes considered during the tree building obtained based on a maximum likelihood fit to pseudo-data, as explained in Sec. 4.6.2. These figures correspond to the scale factors obtained using the yields in each of the final boxes resulting from the *middle-deep* and *deep* splitting, respectively.

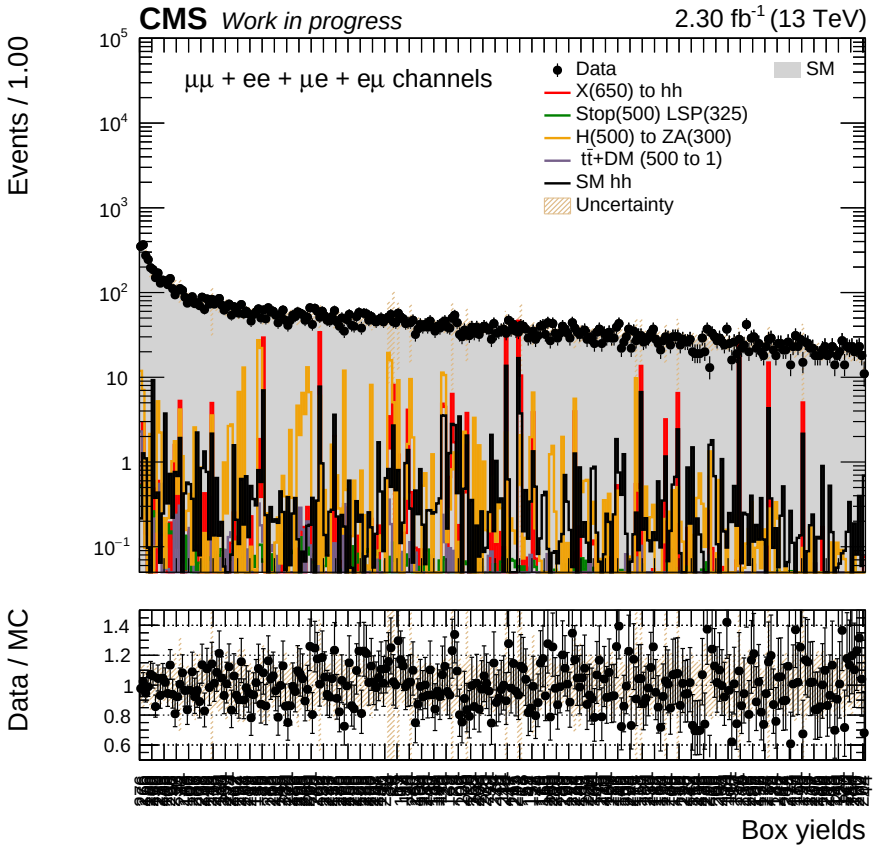


Figure 5.6: Yields in each of the final boxes resulting from the recursive phase space splitting with $\epsilon_{\text{cut}} = 1.5$ and $Stat_{\text{cut}} = 0.1$. The histogram bins have been sorted by decreasing SM expected yield and the uncertainty band contains all the uncertainties described in Sec. 3.5 except for the normalization uncertainties. The histogram is shown with all the SM processes merged.

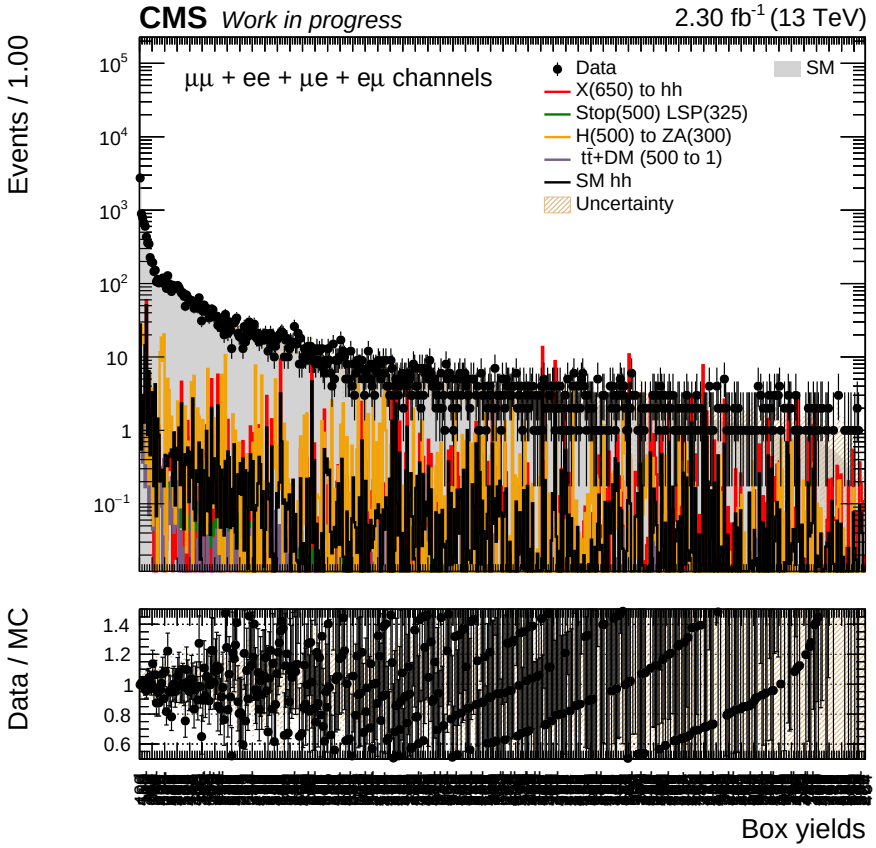


Figure 5.7: Yields in each of the final boxes resulting from the recursive phase space splitting with $\epsilon_{\text{cut}} = 2$ and $Stat_{\text{cut}} = 0.5$. The histogram bins have been sorted by decreasing SM expected yield and the uncertainty band contains all the uncertainties described in Sec. 3.5 except for the normalization uncertainties. The histogram is shown with all the SM processes merged.

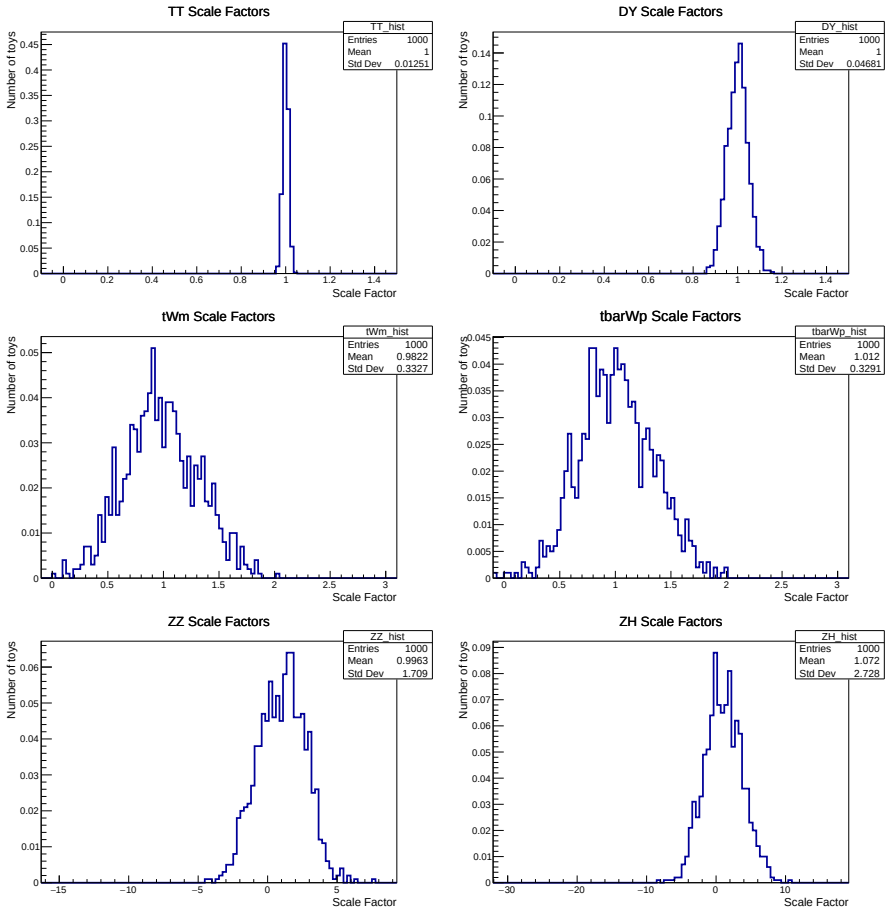


Figure 5.8: Scale factors for the six SM processes obtained from a maximum likelihood fit to pseudo-data. The distribution used to perform the fit is the yields in each of the final box resulting from the *middle-deep* splitting. A thousand pseudo-data experiments have been generated according to Poisson distribution with a mean corresponding to the expected yields considering the six SM processes.

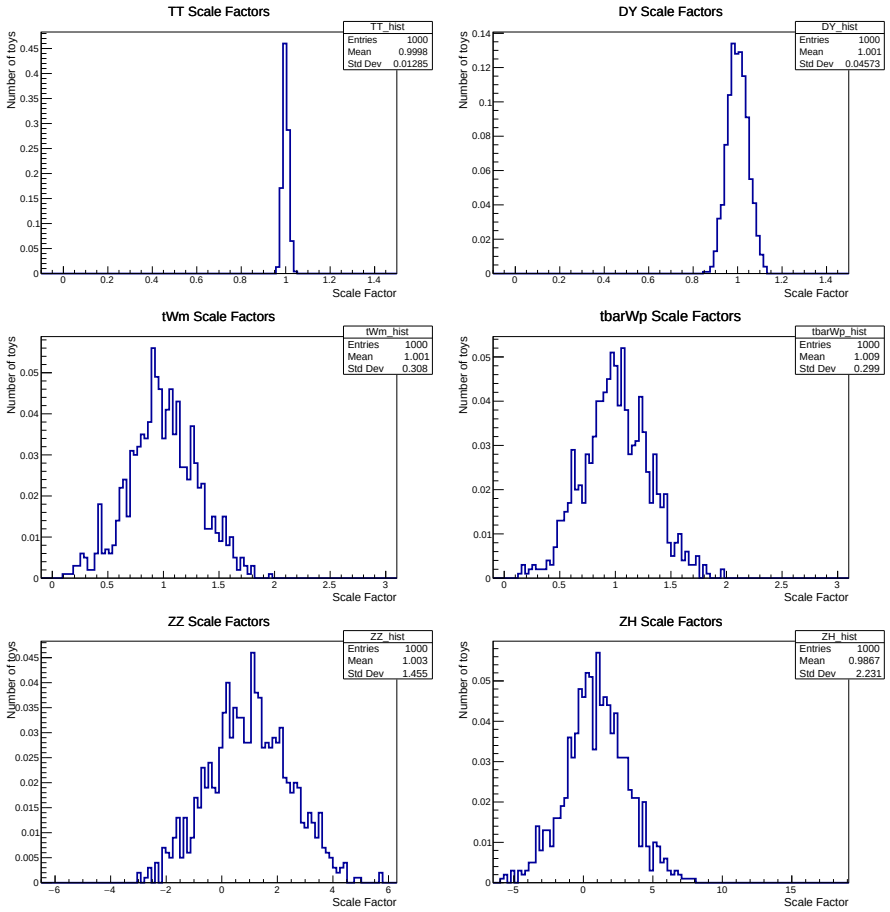


Figure 5.9: Scale factors for the six SM processes obtained from a maximum likelihood fit to pseudo-data. The distribution used to perform the fit is the yields in each of the final box resulting from the *deep* splitting. A thousand pseudo-data experiments have been generated according to Poisson distribution with a mean corresponding to the expected yields considering the six SM processes.

Bibliography

[1] CMS Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV", Tech. Rep. CMS-PAS-FSQ-15-005, CERN, Geneva, 2016.

[2] Daniel V. Schroeder Michael E. Peskin, *An Introduction to Quantum Field Theory*, chapter 4.5, ABP, 1995.

[3] Daniel V. Schroeder Michael E. Peskin, *An Introduction to Quantum Field Theory*, chapter 1, ABP, 1995.

[4] George Sterman John C. Collins, Davison E. Soper, "Factorization of hard process in QCD", *Adv.Ser.Direct.High Energy Phys.5:1-91*, 2004.

[5] Richard D. Ball et al., "Parton distributions for the LHC Run II", *JHEP*, vol. 04, pp. 040, 2015.

[6] 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*, vol. 07, pp. 079, 2014.

[7] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP*, vol. 05, pp. 026, 2006.

[8] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.*, vol. 178, pp. 852–867, 2008.

[9] Paolo Nason, "A New method for combining NLO QCD with shower Monte Carlo algorithms", *JHEP*, vol. 11, pp. 040, 2004.

[10] Stefano Frixione, Paolo Nason, and Carlo Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", *JHEP*, vol. 11, pp. 070, 2007.

[11] 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*, vol. 06, pp. 043, 2010.

[12] Simone Alioli, S.O. Moch, and P. Uwer, "Hadronic top-quark pair-production with one jet and parton showering", *JHEP*, vol. 1201, pp. 137, 2012.

[13] Simone Alioli, P. Nason, C. Oleari, and E. Re, "Single-top production in the s- and t-channel", *JHEP*, vol. 0909, pp. 111, 2009.

[14] Mark Srednicki, *Quantum Field Theory*, chapter 18, University of California, Santa Barbara, 2006.

[15] Christopher Smith, *Introduction to the Standard Model*, chapter 2, Lecture Notes, 2013.

[16] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett. 13, 321*, 1964.

[17] Peter W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett. 13, 508*, 1964.

[18] Guralnik G. S., Hagen C. R., and Kibble T. W., "Global conservation laws and massless particles", *Phys. Rev. Lett. 13, 585*, 1964.

[19] Kien Nguyen, "The Higgs mechanism", `https://www.theorie.physik.uni-muenchen.de/lsfrey/teaching/archiv/sose_09/rng/higgs_mechanism.pdf`, Lecture notes.

[20] Daniel V. Schroeder Michael E. Peskin, *An Introduction to Quantum Field Theory*, chapter 20.2, ABP, 1995.

[21] Roberto D. Peccei, *The Strong CP Problem and Axions*, pp. 3–17, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[22] Gianfranco Bertone, Dan Hooper, and Joseph Silk, "Particle dark matter: evidence, candidates and constraints", *Physics Reports*, vol. 405, no. 5, pp. 279 – 390, 2005.

[23] K. A. Olive et al., "Review of Particle Physics", *Chin. Phys.*, vol. C38, pp. 090001, 2014.

[24] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, Marc Sher, and Joao P. Silva, "Theory and phenomenology of two-Higgs-doublet models", *Phys. Rept.*, vol. 516, pp. 1–102, 2012.

[25] Laura Lopez Honorez, Emmanuel Nezri, Josep F Oliver, and Michel H G Tytgat, "The inert doublet model: an archetype for dark matter", *Journal of Cosmology and Astroparticle Physics*, vol. 2007, no. 02, pp. 028, 2007.

[26] Simon de Visscher, Jean-Marc Gérard, Michel Herquet, Vincent Lemaître, and Fabio Maltoni, "Unconventional phenomenology of a minimal two-Higgs-doublet model", *JHEP*, vol. 08, pp. 042, 2009.

[27] Chiara Arina, Mihailo Backović, Eric Conte, Benjamin Fuks, Jun Guo, Jan Heisig, Benoît Hespel, Michael Krämer, Fabio Maltoni, Antony Martini, Kentarou Mawatari, Mathieu Pellen, and Eleni Vryonidou, "A comprehensive approach to dark matter studies: exploration of simplified top-philic models", *Journal of High Energy Physics*, vol. 2016, no. 11, pp. 111, 2016.

[28] L. Randall and R. Sundrum, "A large mass hierarchy from a small extra dimension", *Phys.Rev.Lett.*, vol. 83, pp. 3370–3373, 1999.

[29] Alexandra Oliveira, "Gravity particles from Warped Extra Dimensions, predictions for LHC", *hep-ph*, 2014.

[30] Sébastien Wertz, "The Matrix Element Method in the LHC era", *EPJ Web Conf.*, vol. 137, pp. 11010, 2017.

[31] Pierre Artoisenet, Vincent Lemaître, Fabio Maltoni, and Olivier Mattelaer, "Automation of the matrix element reweighting method", *Journal of High Energy Physics*, vol. 2010, no. 12, pp. 68, Dec 2010.

[32] T. Hahn, "Cuba—a library for multidimensional numerical integration", *Computer Physics Communications*, vol. 168, no. 2, pp. 78 – 95, 2005.

[33] T. Ohl, "Vegas revisited: Adaptive monte carlo integration beyond factorization", *Computer Physics Communications*, vol. 120, no. 1, pp. 13 – 19, 1999.

[34] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock, *LHC Design Report Volume 1*, CERN, Geneva, 2004.

[35] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock, *LHC Design Report Volume 2*, CERN, Geneva, 2004.

[36] Michael Benedikt, Paul Collier, V Mertens, John Poole, and Karlheinz Schindl, *LHC Design Report Volume 3*, CERN, Geneva, 2004.

[37] CMS Collaboration, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*, Technical Design Report CMS. CERN, Geneva, 2006.

[38] CMS Collaboration, "CMS Physics: Technical Design Report Volume 2: Physics Performance", *J. Phys. G*, vol. 34, no. CERN-LHCC-2006-021. CMS-TDR-8-2, pp. 995–1579. 669 p, 2007, revised version submitted on 2006-09-22 17:44:47.

[39] CMS Collaboration, "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays", *Journal of Instrumentation*, vol. 5, no. 03, pp. T03021, 2010.

[40] V Karimäki, M Mannelli, P Siegrist, H Breuker, A Caner, R Castaldi, K Freudenreich, G Hall, R Horisberger, M Huhtinen, and A Cattai, *The CMS tracker system project: Technical Design Report*, Technical Design Report CMS. CERN, Geneva, 1997.

[41] CMS Collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*, Technical Design Report CMS. CERN, Geneva, 1997.

[42] CMS Collaboration, *The CMS hadron calorimeter project: Technical Design Report*, Technical Design Report CMS. CERN, Geneva, 1997.

[43] CMS Collaboration, *The CMS muon project: Technical Design Report*, Technical Design Report CMS. CERN, Geneva, 1997.

[44] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J.J. Gómez Cadenas, I. González, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F.W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampén, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M.G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J.P. Wellisch, T. Wenaus, D.C. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschiesche, "Geant4—a simulation toolkit", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250 – 303, 2003.

[45] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", Tech. Rep. CMS-PRF-14-001. CMS-PRF-14-001-004, CERN, Geneva, Jun 2017.

[46] Wolfgang Adam, Boris Mangano, Thomas Speer, and Teddy Todorov, "Track Reconstruction in the CMS tracker", Tech. Rep. CMS-NOTE-2006-041, CERN, Geneva, Dec 2006.

[47] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *Journal of Instrumentation*, vol. 9, no. 10, pp. P10009, 2014.

[48] CMS Collaboration, "Measurement of $\mathrm{B\bar{B}}$ angular correlations based on secondary vertex reconstruction at $\sqrt{s} = 7$ TeV", *Journal of High Energy Physics*, vol. 2011, no. 3, pp. 136, 2011.

[49] CMS Collaboration, "Cms pile-up reweighting recommandations", `https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJSONFileforData`, Private access.

[50] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV", *Journal of Instrumentation*, vol. 10, no. 06, pp. P06005, 2015.

[51] CMS Collaboration, "Common analysis object definitions and triggers efficiencies for the h $\rightarrow$ WW run-2 analysis", CMS AN-2015/299. Private access.

[52] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, "The anti- k t jet clustering algorithm", *Journal of High Energy Physics*, vol. 2008, no. 04, pp. 063, 2008.

[53] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at $\sqrt{s} = 8$ TeV", *Journal of Instrumentation*, vol. 12, no. 02, pp. P02014, 2017.

[54] CMS Collaboration, "Identification of b quark jets at the CMS Experiment in the LHC Run 2", Tech. Rep. CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[55] CMS Collaboration, "Missing transverse energy performance of the CMS detector", *Journal of Instrumentation*, vol. 6, no. 09, pp. P09001, 2011.

[56] CMS Collaboration, *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*, Technical Design Report CMS. CERN, 2000.

[57] CMS Collaboration, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project*, Technical Design Report CMS. CERN, Geneva, 2002.

[58] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *Journal of Instrumentation*, vol. 7, no. 10, pp. P10002, 2012.

[59] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B*, vol. 716, no. 1, pp. 30 – 61, 2012.

[60] ATLAS Collaboration, "Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC", *Physics Letters B*, vol. 716, no. 1, pp. 1 – 29, 2012.

[61] CMS Collaboration, "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*, vol. 749, no. arXiv:1503.04114. CMS-HIG-14-013. CERN-PH-EP-2015-042, pp. 560. 29 p, Mar 2015.

[62] CMS Collaboration, "Searches for a heavy scalar boson h decaying to a pair of 125 GeV Higgs bosons hh or for a heavy pseudoscalar boson A decaying to Zh, in the final states with h $\rightarrow \tau\tau$", *Physics Letters B*, vol. 755, pp. 217 – 244, 2016.

[63] CMS Collaboration, "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*, vol. 94, no. CMS-HIG-13-032. CMS-HIG-13-032. CERN-EP-2016-050, pp. 052012. 46 p, Mar 2016.

[64] CMS Collaboration, "Search for resonant Higgs boson pair production in the $\mathrm{b\bar{b}l\nu l\nu}$ final state at $\sqrt{s} = 13$ TeV", Tech. Rep. CMS-PAS-HIG-16-011, CERN, Geneva, 2016.

[65] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, P. Torrielli, E. Vryonidou, and M. Zaro, "Higgs pair production at the LHC with NLO and parton-shower effects", *Phys. Lett.*, vol. B732, pp. 142–149, 2014.

[66] Bruce Mellado Garcia, Pasquale Musella, Massimiliano Grazzini, and Robert Harlander, "CERN Report 4: Part I Standard Model Predictions", LHCHXSWG-DRAFT-INT- 2016-008, CERN, 2016.

[67] Aleksandr Azatov, Roberto Contino, Giuliano Panico, and Minho Son, "Effective field theory analysis of double Higgs boson production via gluon fusion", *Phys. Rev.*, vol. D92, no. 3, pp. 035001, 2015.

[68] Florian Goertz, Andreas Papaefstathiou, Li Lin Yang, and José Zurita, "Higgs boson pair production in the D=6 extension of the SM", *JHEP*, vol. 04, pp. 167, 2015.

[69] T. Binoth and J.J. van der Bij, "Influence of strongly coupled, hidden scalars on higgs signals", *Zeitschrift für Physik C Particles and Fields*, vol. 75, no. 1, pp. 17–25, 1997.

[70] Robert Schabinger and James D. Wells, "Minimal spontaneously broken hidden sector and its impact on higgs boson physics at the cern large hadron collider", *Phys. Rev. D*, vol. 72, pp. 093007, Nov 2005.

[71] A. Denner, S. Heinemeyer, I. Puljak, D. Rebuzzi, and M. Spira, "Standard model higgs-boson branching ratios with uncertainties", *The European Physical Journal C*, vol. 71, no. 9, pp. 1753, 2011.

[72] CMS Collaboration, "Summary table of samples produced for the 1 billion campaign, with 25ns bunch-crossing", `https://twiki.cern.ch/twiki/bin/view/CMS/SummaryTable1G25ns` `https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat13{{{TeV}}}`, Private access.

[73] CMS Collaboration, "CMS luminosity measurement for the 2015 data-taking period", Tech. Rep. CMS-PAS-LUM-15-001, CERN, Geneva, 2017.

[74] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *The European Physical Journal C*, vol. 71, no. 2, pp. 1554, Feb 2011.

[75] CMS Collaboration, "Search for narrow resonances decaying to dijets in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. Lett.*, vol. 116, pp. 071801, Feb 2016.

[76] CMS Collaboration, "Search for new physics in high mass diphoton events in proton-proton collisions at $\sqrt{s} = 13$ TeV", Tech. Rep. CMS-PAS-EXO-15-004, CERN, Geneva, 2015.

[77] CMS Collaboration, "Search for a Narrow Resonance Produced in 13 TeV pp Collisions Decaying to Electron Pair or Muon Pair Final States", Tech. Rep. CMS-PAS-EXO-15-005, CERN, Geneva, 2015.

[78] CMS Collaboration, "Search for massive resonances decaying into pairs of boosted W and Z bosons at $\sqrt{s}$ = 13 TeV", Tech. Rep. CMS-PAS-EXO-15-002, CERN, Geneva, 2015.

[79] CMS Collaboration, "MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8\,\mathrm{TeV}$", Tech. Rep. CMS-PAS-EXO-14-016, CERN, Geneva, 2017.

[80] CMS Collaboration, "Search for H to Z(ll)+A(bb) with 2015 data", Tech. Rep. CMS-PAS-HIG-16-010, CERN, Geneva, 2016.

[81] CMS Collaboration, "Search for Higgs boson pair production in the $\mathrm{b\bar{b}l\nu l\nu}$ final state at $\sqrt{s} = 13\,\mathrm{TeV}$", Tech. Rep. CMS-PAS-HIG-16-024, CERN, Geneva, 2016.

[82] CMS Collaboration, "Search for dark matter in association with a top quark pair at $\sqrt{s} = 13$ TeV in the dilepton channel", Tech. Rep. CMS-PAS-EXO-16-028, CERN, Geneva, 2016.

[83] CMS Collaboration, "Search for direct top squark pair production in the dilepton final state at $\sqrt{s} = 13$ TeV", Tech. Rep. CMS-PAS-SUS-16-027, CERN, Geneva, 2016.

[84] Ezequiel Alvarez, Leandro Da Rold, Mariel Estevez, and Jernej F. Kamenik, "Measuring $|V_{td}|$ at LHC", 2017.

[85] Robert D. Cousins, "Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms", `http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf`, 2013.

[86] Thomas Junk, "Confidence level computation for combining searches with small statistics", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 434, no. 2, pp. 435 – 443, 1999.