UCL
Université
catholique
de Louvain

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

The final state with two *b* jets and two leptons at the LHC as a probe of the scalar sector

Doctoral dissertation presented by

Adrien CAUDRON

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

Jury de thèse

Pr. Christophe DELAERE (*Advisor*)	UCL, Belgium
Pr. Vincent LEMAÎTRE (*Chairman*)	UCL, Belgium
Pr. Jean-Marc GÉRARD	UCL, Belgium
Pr. Barbara CLERBAUX	ULB, Belgium
Dr. Gaëlle BOUDOUL	IPNL, France

October 2016

L'expérience est une lanterne attachée dans notre dos, qui n'éclaire que le chemin parcouru.

Confucius

Remerciements

Durant ces cinq années de thèse, nombreuses ont été les personnes qui ont croisé mon chemin, auxquelles s'ajoutent bien entendu toutes celles rencontrées en amont. Il est difficile de remercier toutes ces personnes qui m'ont donné envie de suivre ce parcours ou qui m'ont accompagné tout au long de celui-ci. Je voudrais donc commencer par un remerciement général à tous ceux et toutes celles avec qui j'ai pu partager des moments, des discussions, échanger lors des diverses écoles doctorales et conférences, les séjours au CERN et bien sûr ici en Belgique, en particulier à Louvain-la-Neuve et Bruxelles.

Je voudrais en premier lieu remercier mon promoteur, Christophe Delaere, pour m'avoir permis de réaliser cette thèse et pour m'avoir accorder sa confiance. Cela a été une expérience très enrichissante. Je voudrais par la même occasion remercier Vincent Lemaître, Jean-Marc Gérard, Barbara Clerbaux et Gaëlle Boudoul pour avoir accepter d'être membre de mon jury. Il est certain que vos commentaires ont apporté un plus significatif à cette thèse.

J'aimerais aussi remercier Andrea Giammanco pour m'avoir orienté sur la possibilité de continuer mon parcours universitaire en Belgique après m'avoir motiver à poursuivre sur la voie de mon stage de master. Dans la même optique, je voudrais remercier Adriana qui m'a pousser à tenter ma chance et qui a une grande responsabilité dans l'accomplissement de ce travail. Et merci aussi pour tous les commentaires stylistiques des graphiques et de leur interprétations très imagées qui m'ont donné une vision disons originale de mes résultats.

Ensuite il y à tous les collègues du CP3. La liste est assez longue mais je dois forcément commencer par Tristan. Je voudrais le remercier pour avoir partager notre bureau pendant l'essentiel de ma thèse. Cela a été pour moi une collaboration très productive. J'ai pu apprendre énormément et j'ai pu apprécier les critiques constructives sur mes présentations et mes résultats. Et bien sûr, je voudrais le remercier pour m'avoir pousser pour obtenir de la visibilité et prendre des responsabilités au sein de la collaboration CMS et en particulier dans le groupe BTV. Et tout cela sans compter sur toutes les anecdotes footballistiques sans lesquels cette thèse serait certainement moins bien réussi (Dank u well dhr. Voetbal!). Je tiens aussi à remercier chaleureusement (on est bientôt en hiver...) Ludivine pour avoir partagé une grande partie de nos thèses respectives ensemble. Je voudrais remercier tant bien la collègue que l'amie pour les discussions de physique ou non au CP3, les voyages plus ou moins réussis de Lisbonne à Hambourg en passant par Vietri-sul-Mare mais aussi et surtout pour toutes ces soirées sur Bruxelles (à parler travail, bien sûr...). Merci aussi à Roberto, pour sa disponibilité malgré son emploie du temps très chargé et pour tous ces échanges, sur l'alignement en particulier, qui m'ont été très bénéfiques (Grazie mille cumpà!). Un grand merci à Alex avec qui j'ai beaucoup apprécié collaborer sur l'analyse BSM, bientôt ton tour... Je souhaiterais aussi remercier Arnaud pour la collaboration sur l'analyse Higgs mais aussi pour les verres après le travail.

Enfin merci à tous les autres membres du groupe *llbb* : Jesus, Miguel, Brieuc, Olivier, les Sébastiens, Camille, Martin... Et merci à tous les autres collègues du CP3 : Christophe R., Michele G., Suzan, Gauthier, Lucia, Federico, David L., Michele S., Elisa, Marcello, Nicolas, Antoine... et tous les autres que je n'ai pas cité ici. Et je n'oublie pas, bien évidemment, Laurent et Bob qui en plus d'être de bons collègues ont été de très bons amis : merci pour tous les bons moments sur Bruxelles !

Je souhaite aussi remercier Jérôme et Pavel pour le soutien informatique qui a été crucial pour la réalisation et l'écriture de cette thèse dans les meilleures conditions. Merci aussi à Ginette et tout le staff administrative pour la bonne humeur et le soutien logistique.

Je souhaite remercier toutes les personnes qui ont collaboré avec moi dans le groupe BTV de CMS en particulier les conveners successifs qui m'ont donné la chance de trouver une place dans la collaboration CMS et ont cru en moi pour me donner la responsabilité de devenir co-convener software du groupe.

Un grand merci à tous les membres du Butokukan pour l'esprit positif et combatif et tout particulièrement à Jacques pour les corrections de l'anglais.

Enfin je tiens à remercier mes amis : Adrien T. (ou Adri.2 car malgré mon retard je vais quand même finir ma thèse en 1er !), Alisson (Mlle 0 déchet, merci pour les réflexions utiles autour de comment faire mieux !), Anna-Maria (grazie per la buon umore e il sostegno nei momenti giusti) et merci à tous les autres pour les moments partagés. Merci à mes collocs passés et présents ! Et à tous ceux qui ont eu la patience d'essayer de comprendre le sujet de ma thèse ! Je voudrais aussi remercier Béatrice pour son accueil et pour toutes les informations et les conseils sur la vie en Belgique. Et pour finir merci à ma famille. Merci à ma mère, Lætitia, mon père, mes grands-parents et tous les autres membres de ma famille. Merci pour le soutien et les encouragements !

Abstract

The data from the LHC produced in 2012 at 8 TeV and collected by the CMS experiment are used to probe the scalar sector of the Standard Model (SM) and beyond. Two analyses have been conducted based on an integrated luminosity of 19.5-19.8 fb^{-1} in the final state with two *b* jets and two leptons.

The decay of the SM Higgs boson to a pair of *b* quarks is challenging at the LHC. In this thesis we propose the use of a Matrix Element (ME) technique to search for the process *Z*(*ll*)*h*(*bb*). The ME weights for the background and signal hypotheses are combined thanks to neural networks to search for the presence of the signal. The sensitivity of the analysis does not allow to conclude on the presence or absence of the signal and an upper limit on $\mu = \sigma/\sigma_{SM}$ of 1.6 is derived compatible at 2 standard deviation (s.d.) with the presence of the signal. A search to the well-known *ZZ* process in the same final state is also conducted giving a result compatible from the expectation from other CMS measurements. The impact of the pile-up on the results of this analysis is also discussed.

Several observations are still unexplained by the SM. In this context a search in the frame of two-Higgs-doublets models (2HDMs) is presented. Such models can arise in several beyond standard model theories. The $H/A \longrightarrow Z(ll)A/H(bb)$ processes have been chosen to potentially discover two new resonances. It constitutes the first search at the LHC for these processes. A generic analysis was conducted in order to allow possible recasting of the results in a wide range of models. One interesting excess with a local (global) significance of 2.6 (1.5) s.d. was observed, compatible both in shape and in amplitude with the probed signal with $m_A = 104$ GeV and $m_H = 270$ GeV. Generic limits as a function of m_A and m_H are derived and used to test a type II 2HDM. Limits for a specific model are then obtained as a function of m_A, m_H, *cos*$(\beta - \alpha)$ and *tan*(β). Perspectives to explore low m_A in the boosted regime are also discussed.

Contents

Introduction

Interactions of elementary particles are well described by the Standard Model (SM) which corresponds to the state of the art of our knowledge in high energy physics.

The recent discovery of the Higgs boson [1, 2] at the Large Hadron Collider (LHC) confirms the Brout-Englert-Higgs (BEH) mechanism [3–5] and the SM. Until now, no significant deviation from the SM expectation has been found [6–8]. The decay of the Higgs boson to a pair of b quarks is one of the most challenging decays to observe at the LHC. Indeed, despite the fact that the branching ratio (BR) for this decay is the largest BR for a Higgs boson with a mass of 125 GeV, the chance of observing this decay suffers from a huge background. This background arises from prompt quark and gluon productions from strongly interacting processes. It is however possible to search for it using the vector boson fusion production mode or the associated production modes with a Z or a W boson. However these production modes have lower production rates at hadron colliders. The search for this decay also suffers from the lower experimental resolution on the reconstructed Higgs mass due to the less precise resolution on the b-quark energy compared to muon or photon. In such a challenging context, this search can benefit from the use of advanced techniques in data analysis such as Multi-Variate Analysis methods (e.g. Neural Networks, Boosted Decision Trees) but also Matrix Element techniques. Developing and testing such techniques plays an important role in the search of rare and challenging processes or in the search of yet undiscovered particles. The analysis presented in Chapter 3 was performed in this context.

Despite the global good description of Nature by the SM, several known issues are still unexplained (e.g. mass hierarchy, baryogenesis, dark matter). In the last decades, different theories Beyond the Standard Model (BSM) have been developed in order to address these issues. Amongst these, the SUSY (SUperSYmmetry) models [9] are between the main BSM models which was searched for in the past decades. One of the properties of these models is the prediction of at least a second Higgs doublet. A large variety of other BSM models also predict a second doublet and are conventionally called 2HDMs for Two-Higgs-Doublets-Models [10]. In these models, in addition to the SM Higgs boson, four other particles are predicted: one other scalar boson H, one pseudoscalar boson A and two charged scalar bosons H^+ and H^-. The analysis presented in Chapter 4 intends to look for such new potential particles.

The Higgs boson was discovered thanks to the data delivered by the LHC and collected by the CMS and ATLAS experiments during 2011 and 2012. The LHC is a circular particle accelerator which provides proton-proton collision data. The centre of mass energy was of 7 and 8 TeV in 2011 and 2012, respectively and raised to 13 TeV in 2015. The results, which will be presented here, are mostly based on the data collected by the CMS experiment in 2012.

In Chapter 1 the physics processes and the theoretical context of the analyses are introduced. In Chapter 2 the experimental setup is presented as well as some relevant information on the data reconstruction and the machine performance at the time of the data taking in 2012. In Chapter 3 a Matrix Element technique is explored in the context of a search for the SM Higgs boson produced in association with a Z boson and decaying to a pair of b quarks. The results based on the 2012 data are also combined with the results based on the 2011 data. In Chapter 4 a search for new particles in the context of a generic 2HDM will be reviewed. This analysis looked for an H or an A boson decaying to a Z boson and an A or an H boson decaying to a pair of b quarks. For this analysis only the 2012 data were used. The interpretation of the results in a specific 2HDM model is also considered.

Chapter 1

Physics at the LHC

1.1 Standard Model processes at hadron colliders

Physicists express the current understanding of Nature at small scales in terms of elementary particles. The Standard Model (SM) describes these elementary particles and their interactions. This theory has worked well over the past decades allowing physicists to make accurate predictions. Nonetheless, some phenomena remain unexplained which suggests that it might be incomplete. Its detailed description can be found elsewhere [11–13].

In the past years, the search for the so called 'Higgs' boson was one of the most lively fields in particle physics. The dynamism of this field was stimulated by the experimental results from the LHC. It led, in 2012 [1, 2], to the discovery of a new boson with a mass of about 125 GeV by the CMS and ATLAS collaborations. So far, the properties of this newly found boson matches the expectations from the SM Higgs boson [14–16].

This achievement was possible thanks to the rediscovery and the detailed studies of all the dominant SM processes with a great precision by these collaborations. Indeed, hadron collider physics covers a wide range of physics processes predicted by the SM. In this thesis the discussion will focus on the case of proton-proton (*p*-*p*) collisions produced at the LHC because the results which will be presented in Chapters 3 and 4 are based on these collisions. Due to the high energy at which such collisions are produced, the interactions are occurring between the subconstituents of the protons, meaning the quarks and gluons (commonly called 'partons'). The presence of these partons inside the proton and the fraction of the proton energy they are bringing are defined by the parton distribution functions (PDFs). These allow us to define the interaction rate between the different partons of two protons. The amplitude of a given interaction leading to a given final state is defined by the matrix element (ME) of the process. These two elements, the PDFs and the ME, allow the computation of the cross section of a given process produced by *p*-*p* collisions.

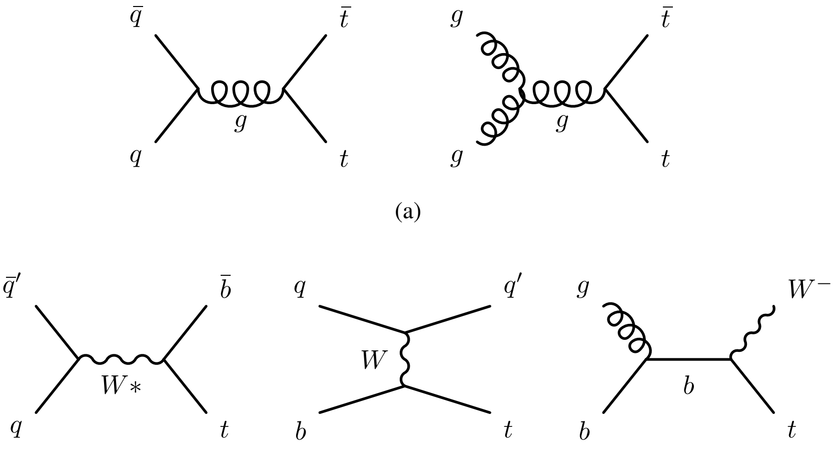
Because most of the particles of interest are experimentally detectable only through their decay products (leptons/quarks/gluons/photons/neutrinos), several interactions can lead to the same final state particles. This makes difficult the identification of the processes which actually happened in the data. If some observables like the mass of the intermediate resonances can help in discriminating different processes, it is also important to take into consideration all the processes which can resemble the process of interest. For the studies presented in Chapters 3 and 4 the most relevant processes are the productions of quarks and gluons, the top quark productions and the single and di-boson productions. They will be briefly introduced in what follows.

In hadron collisions strongly interacting processes are dominant. They lead to the presence of several quarks and gluons in the final state. Events in which such final state are produced are commonly called **multi-jets** events from **QCD** (for Quantum Chromodynamics) processes. 'Jets' refer to the high multiplicity of particles produced in the direction of the final states quarks and gluons by the hadronisation process following their production. QCD processes, despite the fact that they can be interesting for probing the SM description, are mostly considered as a background for less frequent processes. Multi-jets events can however be easily rejected by requiring the presence of leptons in the final state.

The **top quark** productions which happen primarily through strong interactions, can be considered as independent processes with respect to the other strongly interacting processes. The reasons for this can be summarised in this way:

- The top quark is the only SM quark not hadronising due to its large mass (173.21 GeV [17]) compared to the other quarks (2 MeV - 5 GeV) making its life time too small ($0.5 \cdot 10^{-24}$ s [17]) to combine with other quarks before decaying. This allows the study of several properties of this quark such as its mass, its spin, its helicity, etc.
- It is decaying exclusively (or almost exclusively, $V_{tb} = 1.021 \pm 0.032$ [17]) to a *W* boson and a *b* quark. This fact leads to a distinct signature. The possible presence of a lepton in the final state, resulting of the decay of the W boson, allows the studies of some of its properties as mentioned above.

The dominant top quark production modes at leading order (LO) are shown in Figure 1.1. The main production mode at hadron colliders is the pair production, shown in Figure 1.1a. The single top production, shown in Figure 1.1b, is possible through electroweak interaction with the intervention of a *W* boson. The pair production is slightly more than twice as frequent at the LHC for a centre of mass energy of 8 TeV. The high multiplicity final state following the creation of top quarks (leptons, (*b*) jets, neutrinos) and its high production rate makes these processes a relevant background for a wide range of other SM processes and for BSM searches.



(a)

(b)

Figure 1.1: LO top production diagrams. (a) top pair production. (b) single top production: t-channel (left), s-channel (centre) and *W* associated production (right).

Other important processes imply the creation of a **Z boson**. Two sub-processes are studied in addition to the inclusive *Z* production: the **Z+jets** production and the **Z+heavy flavour** (HF) jets production (jets originating from the production of a *c* or a *b* quark). More generically, the *Z* production is studied as a *Z*/$\gamma*$ production. When the *Z*/$\gamma*$ is decaying to a pair of leptons, this process is commonly called Drell-Yan production. Examples for such processes are visible in Figure 1.2. The main reason to separate the inclusive *Z*+jets and the *Z*+HF jets processes arises from the way to treat the production of HF quarks in hadron collisions. Indeed, the *c* and *b* quarks are heavier than the proton. This implies that it is not trivial to include them in the PDFs of the proton and two approaches can be followed at fixed order [18]. One possibility is to assume them to be massless. This case is the so-called '5F' scheme. The other possibility is to consider the *b* quarks to be only indirectly present in the proton and to be produced only as pairs through gluon splitting. This case is the so-

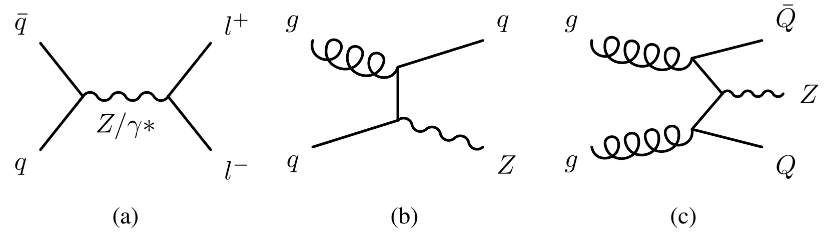


Figure 1.2: Examples of Z production diagrams at hadron collider: (a) inclusive Z production decaying to a pair of leptons (Drell-Yan) (b) Z + 1 jet production and (c) Z + HF jet production with $Q = c, b$.

called '4F' scheme. The same issue arises for all processes with an HF quark in the initial state. For example, the s-channel and the W associated productions of the single top are also in this configuration (see Figure 1.1b, the two rightmost diagrams). Discussions and studies on this topic can be found in [19] including some kinematic comparisons using the LHC data.

The production of the ***W* boson** is similar to the case described above for the Z boson, except for the W production with HF quarks. In this case, a diagram as shown in Figure 1.2c is not possible. The HF quarks can be produced in association with a W boson only through a gluon radiated by an initial state quark.

Di-boson productions are other ways to produce W and Z bosons at hadron colliders. The two main LO diagrams to produce ZZ, WZ and WW events are shown in Figure 1.3.

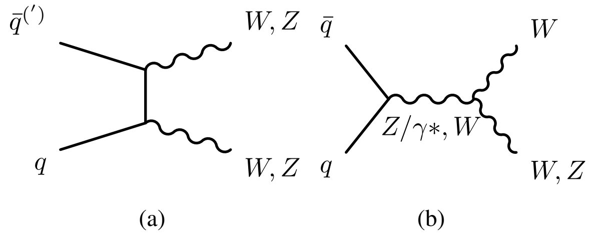


Figure 1.3: Main VV production diagrams at LO with $V = W, Z$.

This list of processes is not exhaustive but includes already a large part of the most common processes in hadron collisions. They also constitute the most relevant background for searches for new particles and new phenomena as the search of the Higgs boson.

1.2 Higgs physics at the LHC

The BEH mechanism was proposed [3–5] in order to explain the way the W and Z bosons acquire mass in the SM. It is assumed to be responsible for the spontaneous electroweak symmetry breaking in the SM. The masses of the W and Z bosons are the results of their interactions with a new field to which is associated a new boson, commonly called the Higgs boson. This field is an SU(2) doublet with four degrees of freedom, in which three of them will mix to the $W^{\pm}$ and Z bosons after symmetry breaking which makes these three bosons massive. The vacuum expectation value (vev) of this field is 246 GeV. The new boson is carrying this new interaction. In the SM, fermions acquire masses by their interactions with this boson through the Higgs Yukawa couplings.

Several production modes are possible in order to produce Higgs bosons at the LHC. They are visible in the left plot of Figure 1.4 for proton-proton collisions [20]. The dominant production mode is the gluon-induced process which is possible through a loop of top quarks. However, depending on its decay, this can lead to some difficulties in identifying such production. In the right plot of Figure 1.4 the possible decays of the Higgs boson and their corresponding branching ratio (BR) are shown [20]. For a 125 GeV Higgs particle, its decay to a pair of b quarks is dominant. At hadron colliders this decay is really challenging to observe. To probe it, sub-dominant production modes are needed. One of this is the associated production with a W or a Z boson. In this case, it is possible to take advantage of the leptons which might arise form the W or Z boson decays. Such search will be the topic of the analysis presented in Chapter 3.

The Higgs boson was already searched for at LEP and at the Tevatron but was only recently discovered at the LHC as mentioned previously. For now, all the tests performed on this newly discovered boson tend to confirm it is identical to the predicted SM Higgs boson. The only parameter not predicted by the theory was its mass. It is now measured to be 125 ± 0.21(stat.) ± 0.11(syst.) GeV [8].

Considering the different experimental results, there is not much room left to question the compatibility of the newly discovered boson with the SM Higgs boson. The decays to bosons ($\gamma\gamma$, ZZ, WW) are really clear [7, 8, 21]. The decays to fermions, even if the significance is lower, also show results within the expectations from the theory [6–8]. In this latter case, however, the uncertainties are still quite large, especially considering the decay to a pair of b quarks. This allows some leeway on the

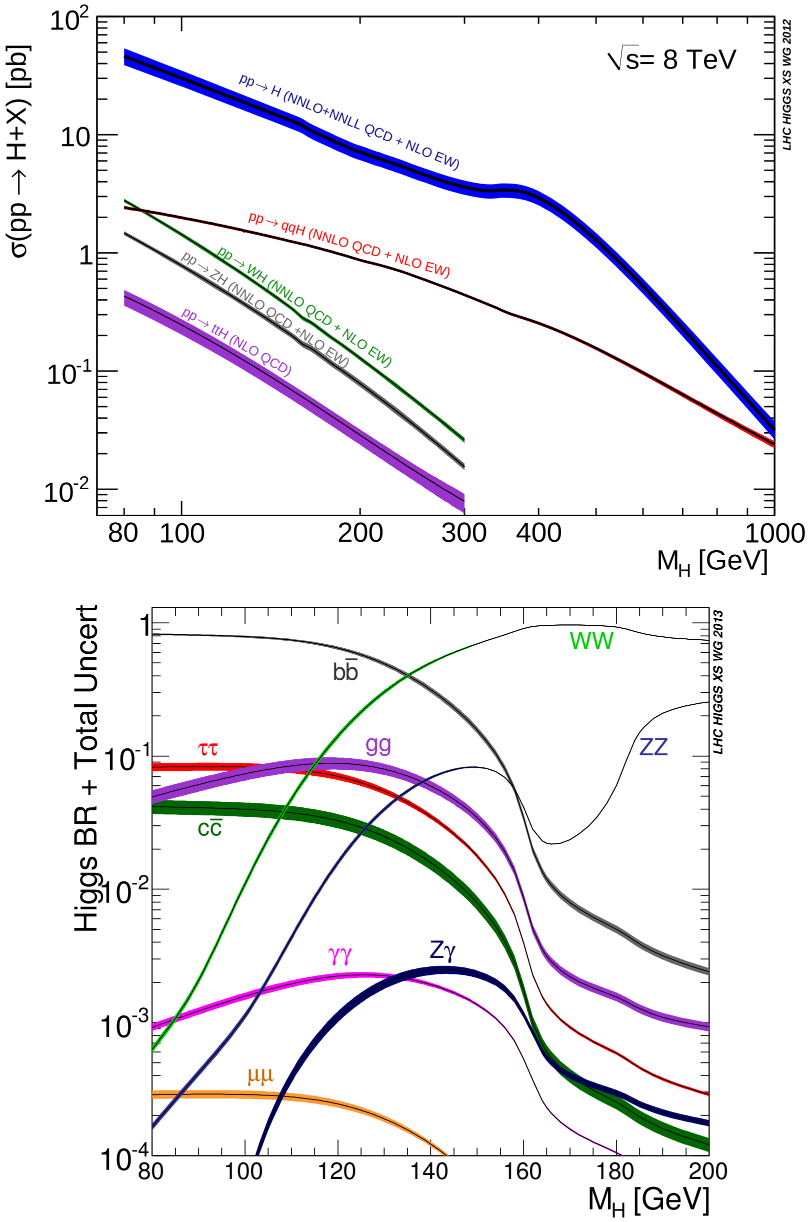


Figure 1.4: Top: Production cross sections at the LHC at $\sqrt{s} = 8$ TeV for the different Higgs boson production modes as a function of the Higgs boson mass. Bottom: Branching ratios of the different Higgs boson decays as a function of the Higgs boson mass.

possible unexpected behaviour of this boson. The study of possible deviations from the theoretical expectations is now one of the most lively topics in the Higgs sector. Indeed, such deviations might be an indication of physics beyond the SM.

If the discovery of the Higgs boson confirms the existing theory, this does not help to solve the remaining issues. Up till now no clear indication was given by the LHC results on which directions to explore. This is probably the most challenging part in particle physics for the coming years.

1.3 Extension of the scalar sector

In order to solve some of the known shortfalls of the SM, several models are proposing extensions of the SM scalar sector. A class of models which has been quite extensively studied is the addition of a second doublet [10]. These models are known as '2HDMs' which stands for two-Higgs-doublets models. Supersymmetric models, for example, especially in their minimal definition (MSSM), require such an extension [9]. More generally 2HDMs are simple extensions of the SM scalar sector which can help to explain the asymmetry between matter and antimatter in the universe [10, 22]. Axion models, which propose a solution to the strong CP problem in QCD and also propose a dark matter candidate, are another example of models in which a second doublet is needed [23]. Finally, it has recently been noted [24] that certain realisations of 2HDMs can accommodate the muon g-2 anomaly [25] without violating present theoretical and experimental constraints.

Different scalar structures can also be inferred as singlet [26] or 3HDM [27] but they will not be discussed in this thesis.

In 2HDMs, if CP invariance is imposed, five well-defined physical states are predicted: two neutral CP-even scalars (*h*, *H* ; *h* being the lightest one by convention), one neutral CP-odd pseudoscalar (*A*) and two charged scalars ($H^{\pm}$). The four masses, m_h, m_H, m_A and $m_{H^\pm}$ are free parameters of these models. The ratio $v_2/v1$ of the vev's for the two doublets is defined as *tan*(β) where β is the angle which allows to go in the Higgs basis where only one doublet has a non 0 vev v such as $v^2 = v_1^2 + v_2^2$. A second angle α parametrises the mixing between the neutral CP-even states *h* and *H*. In the alignment limit $\beta - \alpha = \pi/2$, the doublet with non zero vev matches the SM Higgs doublet while the other doublet is composed of the four new states. 2HDMs are generally parametrised as a function of *tan*(β) and *sin*$(\beta - \alpha)$ (or *cos*$(\beta - \alpha)$). The masses and these two parameters are the most relevant parameters considering the study which will be presented in Chapter 4.

Nevertheless in their most generic form 2HDMs have 14 free parameters:

$$\begin{aligned} V = {} & m_{11}^2 \Phi_1^\dagger \Phi_1 + m_{22}^2 \Phi_2^\dagger \Phi_2 - \left(m_{12}^2 \Phi_1^\dagger \Phi_2 + \text{h.c.} \right) \\ & + \tfrac{1}{2}\lambda_1 \left(\Phi_1^\dagger \Phi_1 \right)^2 + \tfrac{1}{2}\lambda_2 \left(\Phi_2^\dagger \Phi_2 \right)^2 + \lambda_3 \left(\Phi_1^\dagger \Phi_1 \right) \left(\Phi_2^\dagger \Phi_2 \right) + \lambda_4 \left(\Phi_1^\dagger \Phi_2 \right) \left(\Phi_2^\dagger \Phi_1 \right) \\ & + \left[\tfrac{1}{2}\lambda_5 \left(\Phi_1^\dagger \Phi_2 \right)^2 + \lambda_6 \left(\Phi_1^\dagger \Phi_1 \right) \left(\Phi_1^\dagger \Phi_2 \right) + \lambda_7 \left(\Phi_2^\dagger \Phi_2 \right) \left(\Phi_1^\dagger \Phi_2 \right) + \text{h.c.} \right], \end{aligned} \tag{1.1}$$

where ‘h.c.’ stands for the Hermitian conjugate. The parameters m_{11}^2, m_{22}^2, and $\lambda_{1,2,3,4}$ are real. In general, m_{12}^2 and $\lambda_{5,6,7}$ are complex. Imposing CP conservation and a $\mathbb{Z}2$ symmetry, to suppress flavour changing neutral currents, reduces significantly the number of free parameters. By considering $\lambda_6 = \lambda_7 = 0$ to avoid hard violation of the $\mathbb{Z}2$ symmetry, six physical parameters are left in the potential. A soft breaking of the $\mathbb{Z}2$ symmetry is generally allowed and represented by $m_{12} \neq 0$. This parameter is less relevant for the study which will be presented in Chapter 4 because it mainly contributes to the Higgs bosons self-coupling. In most cases, the 125 GeV Higgs boson is associated to the lightest neutral scalar of the model (*h*). The *H* boson mass depends on the mixing angle α defined with respect to the SM Higgs boson h_{125}. The mass of the other bosons are expressed as

$$m_{H^\pm}^2 = m_{12}^2 / cos\beta sin\beta - (\lambda_4 + \lambda_5) \cdot v^2/2$$

$$m_A^2 = m_{12}^2 / cos\beta sin\beta - \lambda_5 \cdot v^2 = m_{H^\pm}^2 + (\lambda_4 - \lambda_5) \cdot v^2/2.$$

Some custodial symmetry needs to be imposed in order to avoid large deviation of the ρ parameter, defined as

$$\rho = \frac{m_W^2}{m_Z^2 \cdot cos^2\theta_W}$$

with θ_W the weak-mixing angle, and estimated to be 1.00040 ± 0.00024 [17] from electroweak precision data. The custodial symmetry leads to the following cases [28, 29]:

- $\lambda_4 = \lambda_5$ in the ‘usual’ case with as consequences

$$m_{H^\pm} = m_A \text{ and } m_{12}^2 = (m_A^2 + \lambda_5 \cdot v^2) \cdot cos\beta sin\beta.$$

- $\lambda_4 = -\lambda_5$ in the ‘twisted’ case with as consequences

$$m_{H^\pm} = m_H \text{ and } m_{12}^2 = m_H^2 \cdot cos\beta sin\beta.$$

Models	Up quarks	Down quarks	Leptons
Type I	Φ_2	Φ_2	Φ_2
Type II	Φ_2	Φ_1	Φ_1
Lepton-specific (Type III)	Φ_2	Φ_2	Φ_1
Flipped (Type IV)	Φ_2	Φ_1	Φ_2

Table 1.1: Association of the two Higgs doublets, Φ_1 and Φ_2, to the quarks and leptons for the four 2HDMs types.

ξ_h^u	$\xi_h^{d,l}$	$\xi_h^{W,Z}$	ξ_H^u	$\xi_H^{d,l}$	$\xi_H^{W,Z}$	ξ_A^u	$\xi_A^{d,l}$
cos(α)/*sin*(β)	-*sin*(α)/*cos*(β)	*sin*(β-α)	*sin*(α)/sin(β)	*cos*(α)/*cos*(β)	*cos*($\beta - \alpha$)	*cot*(β)	*tan*(β)

Table 1.2: Modification factors to the Higgs Yukawa couplings with respect to the SM ones.

The custodial symmetries leads then to a degeneracy in mass between the charged Higgses and either the scalar *H* or the pseudoscalar *A*. The m_{12} mass is also linked to the mass of either the scalar *H* or the pseudoscalar *A*. In the MSSM limit instead $\lambda_4 = -g^2/2$ and $\lambda_5 = 0$ where g is the gauge coupling constant of SU(2). Because $m_W = g \cdot v/2$ (v = 246 GeV), it has as consequences

$$m_{H^{\pm}}^2 = m_A^2 + m_W^2 \text{ and } m_{12}^2 \ = \ m_A^2 \cdot cos\beta sin\beta.$$

In this case no explicit custodial symmetry is imposed but it arises from the assumed decoupling limit where the new states (*A*, *H*, $H^{\pm}$) are heavy compared to the *W* mass and can then be considered as degenerated.

Depending on the way the two doublets couple to the up and down quarks and the leptons, four types of 2HDMs which lead to natural flavour conservation are defined. These are listed in Table 1.1. The first two types (I and II) are the most common ones in the literature. In what follows, only the case of the type II model will be detailed. The reason for this is that this model is used to interpret the results presented in Chapter 4. This also corresponds to the MSSM configuration.

The couplings are generally expressed as a modification factor compared to the SM Higgs Yukawa couplings. Let ξ these factors, $\xi_{h,H,A}^{u,d,l,W,Z}$ are then the modification factors with respect to the SM Higgs Yukawa couplings for the three neutral scalars and pseudoscalar particles to the up (*u*) quarks, down (*d*) quarks, leptons (*l*) and the *W* and *Z* bosons. These factors are shown in Table 1.2. The case of $H^{\pm}$ will not be discussed here because it is not directly relevant for the results presented in Chapter 4. However, the couplings to fermions follow the same dependency as $\xi_A^{u,d,l}$ with respect to the SM.

Concerning the couplings modifier ξ_h^u, if *h* is assumed to be the 125 GeV boson, h_{125}, then this factor is constrained to be close to 1 by experimental measurements because no significant deviations have been observed in the production of this boson compared to the SM predictions. The same comment can be made on $\xi_h^{d,l}$ but the experimental constraints on $h_{125} \rightarrow bb$ and $h_{125} \rightarrow \tau\tau$ are weaker. The couplings modifier factors $\xi_h^{W,Z}$ are also strongly constrained by the experimental results as no significant deviations have been observed from the SM expectations for $h_{125} \rightarrow WW, ZZ$. This implies $sin(\beta - \alpha) \sim 1$ or similarly $cos(\beta - \alpha) \sim 0$. No $\xi_A^{W,Z}$ is shown because it is strongly suppressed and possible only at loop level, assuming CP is conserved.

The existence of new particles with respect to the SM leads to new possible production and decay modes. These depend both on the 2HDM type and the mass hierarchy. Some examples of new processes are shown in Figure 1.5. The mass hierarchy between the different Higgs bosons has consequences on the enabled processes and their importance. Typically, for $m_{A,H} \geq 2 \cdot m_t$, the decays *A*/*H* $\rightarrow t\bar{t}$ will tend to be dominant. It is important to keep in mind that this could change if $tan(\beta)$ differs significantly from 1. For a light *A* ($m_A < m_{h_{125}} + m_Z$), the decays to *bb* and $\tau\tau$ will start to be dominant. The *H* decays to *hh*/*AA*/*WW* and *ZZ* can play different roles depending on m_H, but also on the choice of the other parameters for the model.

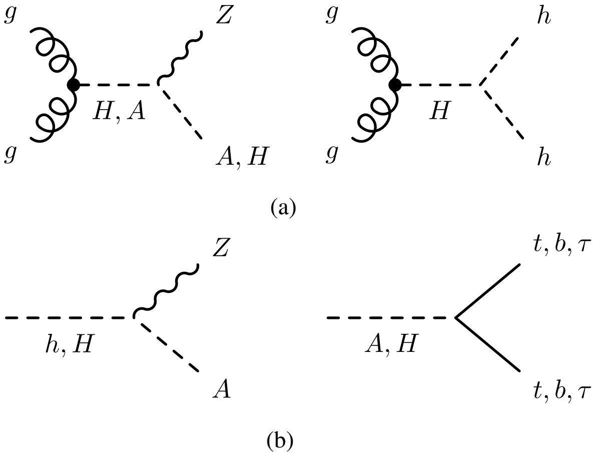


Figure 1.5: Examples of new processes in 2HDMs (a) and decays of the neutral Higgs bosons (b).

The processes *H*/*A* $\rightarrow$ *ZA*/*H* are interesting in the light of the recent results from the LHC. Indeed, it is one of the dominant process in the region of the phase space yet poorly constrained by the existing SM Higgs measurements [30] as visible in Figure 1.6. The case *A* $\rightarrow$ *ZH* is the least sensitive to the choice of $cos(\beta - \alpha)$ as *A* $\rightarrow$ *WW/ZZ/hh* are suppressed. The main competitive process is *A* $\rightarrow t\bar{t}$ in case $m_A \geq 2 \cdot m_t$. For the case *H* $\rightarrow$ *ZA*, the choice of $cos(\beta - \alpha)$ is more relevant. Due to the fact that the LHC data favour a small value for $cos(\beta - \alpha)$, the decays to *WW* and *ZZ* are expected to be negligible. The main competitive processes are then *H* $\rightarrow$ *hh/tt* for $m_H \geq 2 \cdot m_{h,t}$. The decay of the lightest *H/A* boson in *bb* starts to be interesting for masses lower than twice the top mass. Still, it depends to some extend on the choice of the model parameters. The decay to $\tau\tau$ is also interesting but it will be sub-dominant for type II 2HDMs for masses higher than twice the *b* mass. For types III and IV, there is some freedom to make the decay to $\tau\tau$ more competitive with respect to the decay to *bb*.

Here, the emphasis will be put on the processes *H/A* $\rightarrow$ *ZA/H*(*bb*). Choosing a final state where the *Z* decays to two leptons (muons or electrons) leads to a clean signal with a moderate background. This background comes mainly from *Z*+jets and $t\bar{t}$ processes described in Section 1.1. The masses of the different resonances are good variables to reduce the background contributions even more.

These processes can be decomposed in three parts: the production of the heavy Higgs boson, its decay to a *Z* boson and the lighter Higgs boson and the decay of this lighter Higgs boson to *bb*. The expected NNLO (next-to-next-to-leading order) cross sections for the first part are calculated using the SUSHI program [31] and varies roughly between few fb to few pb depending on the choice of the parameters. The expected BRs for the two other parts are obtained thanks to the 2HDMC calculator tool [32]. The product of the three terms and the BR(*Z* $\rightarrow ll$), with $l = e, \mu$, gives the final cross sections for these processes which can vary over several order of magnitude depending on the choice of the parameters. Theses cross sections can be as large as few hundreds of fb for m_H and m_A below 300 GeV.

In what follows, $m_{H^{\pm}}$ is required to be degenerated with the heaviest Higgs boson mass, m_H or m_A, as a consequence of the custodial symmetry discussed previously. The choice was to take precisely $m_{H^{\pm}} = m_{H,A}$ to avoid the decays of the heaviest Higgs boson which could include charged Higgses. The value of the m_{12} parameter is defined as $m_{12}^2 = m_{H^{\pm}}^2 \cdot cos\beta sin\beta$. This agrees with the custodial symmetries and the MSSM in the limit of heavy $H^{\pm}$.

All results presented below assume proton-proton collisions with a centre of mass energy of 8 TeV.

In Figure 1.7, the *H* and *A* production cross sections are shown as a function of $tan(\beta)$ and $cos(\beta - \alpha)$ in the ranges [0.1,10] and [-1,1], respectively, for two arbitrary choices of m_A and m_H (700 GeV for the heaviest boson, 300 GeV for the lightest boson).

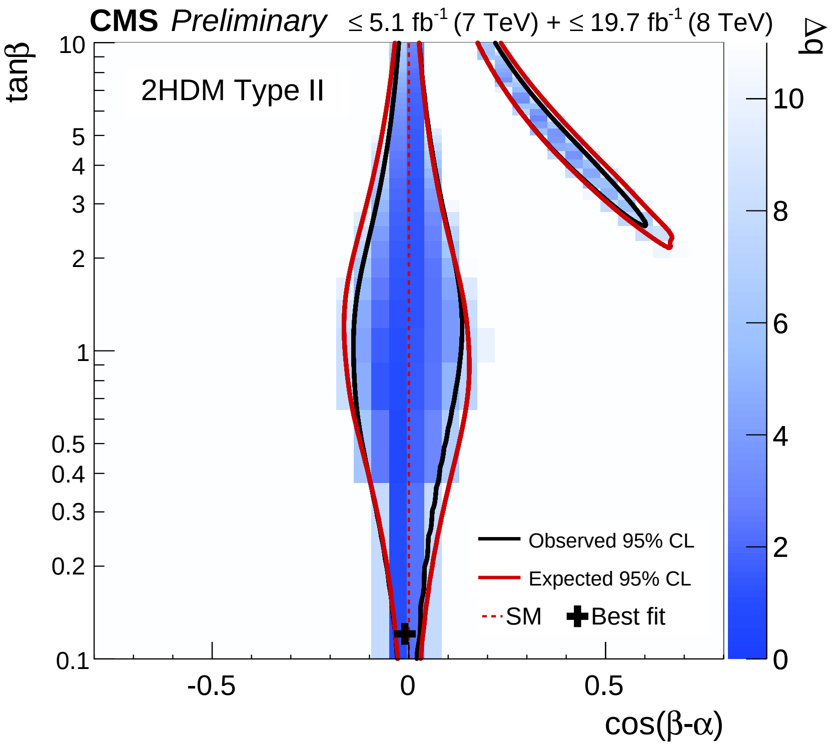


Figure 1.6: General constraints on the 2HDM parameter space obtained from the compatibility with the observed couplings of the Higgs boson when interpreted as the h. The lines show the contours which restrict the allowed parameter space at the 95% CL for a type II 2HDM. These contours have been obtained from an increase of the test statistic, q($cos(\beta-\alpha)$, $tan(\beta)$) as defined in [30] by Δq = 5.99 relative to the minimum in the $cos(\beta-\alpha)$-$tan(\beta)$ plane, corresponding to a 95% confidence region for a χ^2 function with two degrees of freedom. The observed constraints are shown in black. The expected constraints assuming just the SM Higgs sector are indicated by the red continuous line.

This choice of masses has been made for illustration purpose only. The amplitude of the cross sections is following the expected dependency as a function of β and α. Especially, the production of the *A* boson is independent of α then of *cos*$(\beta-\alpha)$. In the region *tan*$(\beta) \sim 1$ and *cos*$(\beta-\alpha) \sim 0$, the production cross sections for both processes are about the same order ($\sim$ [0.1,1] pb). It should be mentioned also that SUSHI assumed perturbativity in order to compute the production cross sections. This has as consequence that the cross sections computed for *tan*$(\beta) \lesssim 0.5$ should be taken with caution because this assumption becomes weaker when *tan*(β) reduces in value.

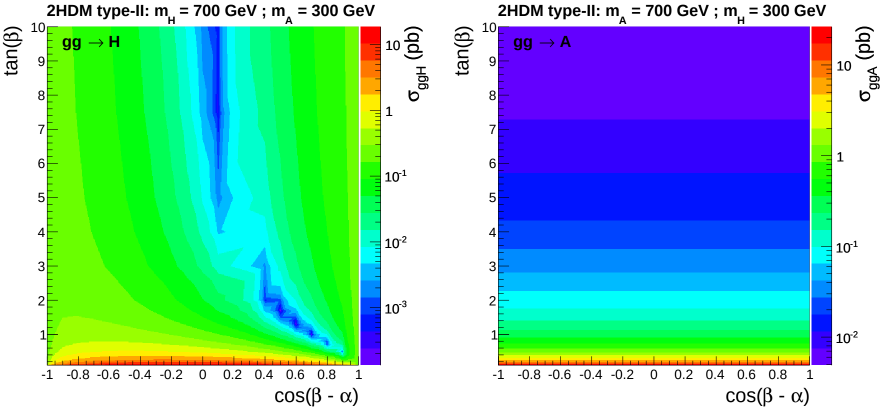


Figure 1.7: Theoretical NNLO cross sections for $gg \longrightarrow H/A$ as a function of *tan*(β) and *cos*$(\beta-\alpha)$ in type II 2HDM. In both cases the heavy Higgs boson has a mass of 700 GeV.

In Figure 1.8, the BRs for the decay of the heaviest Higgs boson, $H/A \longrightarrow A/H$, are also shown as a function of *tan*(β) and *cos*$(\beta-\alpha)$. The choice of masses is the same as in the previous case: $m_{H/A}$ = 700 GeV and $m_{A/H}$ = 300 GeV. The two processes have a different dependency on *tan*(β) and *cos*$(\beta-\alpha)$. This is mainly due to the different decays allowed for both bosons. In the case of the H boson, the decays to *hh*/*WW*/*ZZ* can play a significant role, especially for values of *cos*$(\beta-\alpha)$ far from 0. However for our region of interest, *tan*$(\beta) \sim 1$ and *cos*$(\beta-\alpha) \sim 0$, both $H/A \longrightarrow A/H$ decays are important with a BR $\gtrsim$ 40%.

In Figure 1.9, the BRs for the decay of the lightest Higgs boson, $A/H \longrightarrow$ *bb*, are also shown as a function of *tan*(β) and *cos*$(\beta-\alpha)$ for the same choice of masses. The dependency of the BR($A \rightarrow$ *bb*) in *cos*$(\beta-\alpha)$ comes from the other dominant decays for this choice of masses, $A \longrightarrow Zh$. Otherwise it is increasing as a function of *tan*(β) as expected from Table 1.2. The case $H \longrightarrow$ *bb* is much more complex

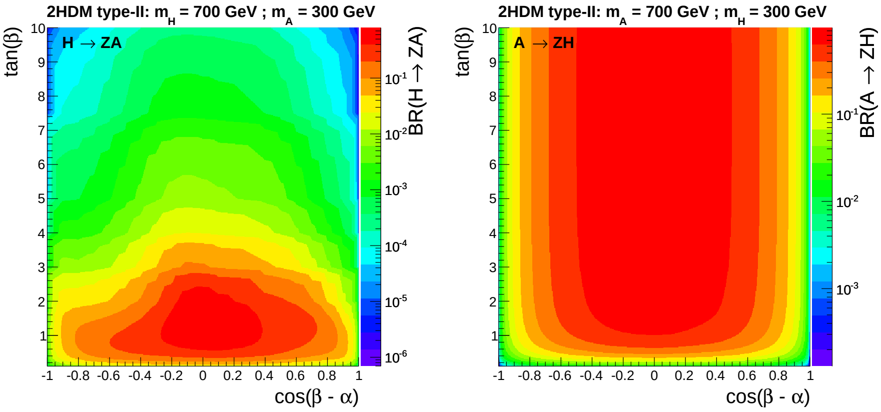


Figure 1.8: Branching ratio for $H/A \longrightarrow ZA/H$ as a function of $tan(\beta)$ and $cos(\beta - \alpha)$ in type II 2HDM. In both cases the heavy Higgs boson has a mass of 700 GeV and the light Higgs boson has a mass of 300 GeV.

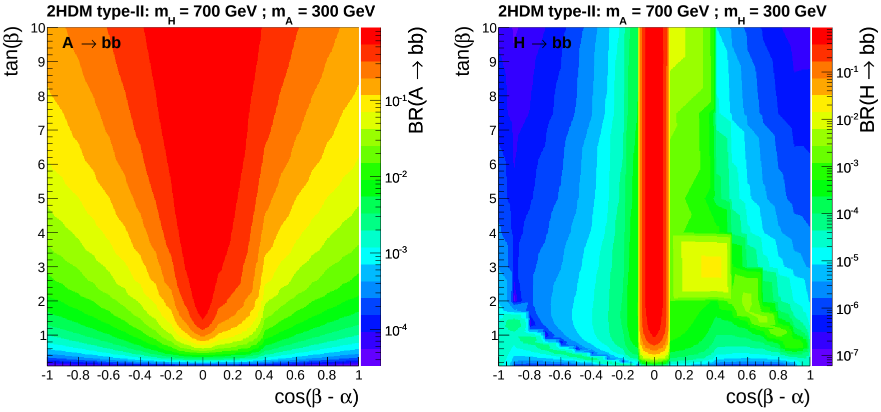


Figure 1.9: Branching ratio for $A/H \longrightarrow bb$ as a function of $tan(\beta)$ and $cos(\beta - \alpha)$ in type II 2HDM. In both cases the heavy Higgs boson has a mass of 700 GeV and the light Higgs boson has a mass of 300 GeV.

due to the variety of possible decays for a H boson of 300 GeV. Still, it can be noted that for $cos(\beta - \alpha) \sim 0$, the BR($H \longrightarrow bb$) $\sim$ 100%. This reflects in fact that decays to *WW* and *ZZ* go to 0. In both cases, for $tan(\beta) \sim 1$ and $cos(\beta - \alpha) \sim 0$, the BR($A/H \longrightarrow bb$) are high ($\gtrsim$ 10%).

The total cross sections for the processes $H/A \longrightarrow Z(ll)A/H(bb)$ are shown in the top plots of Figure 1.10 as a function of $tan(\beta)$ and $cos(\beta - \alpha)$. Considering the region of interest, $tan(\beta) \sim 1$ and $cos(\beta - \alpha) \sim 0$, both processes have their highest cross section in this region. They are similar in both cases, of the order of 1 fb. This region of higher cross sections is of course depending on the masses of the two Higgs bosons and can be slightly shifted and be more or less wide. In principle, it always includes the case $tan(\beta) \sim 1$ and $cos(\beta - \alpha) \sim 0$. These results and the fact that such region is poorly constrained by direct SM Higgs measurements show that these processes are interesting for further studies in the context of the search of physics beyond the SM.

A proper choice of $tan(\beta)$ and $cos(\beta - \alpha)$ is made to study the cross section dependency on m_A and m_H. Choosing $tan(\beta)$ = 1.5 and $cos(\beta - \alpha)$ = 0.01 allows to have a small $cos(\beta - \alpha)$ keeping $sin(\beta - \alpha)$ and $cos(\alpha)/sin(\beta) \sim 1$. On the basis of this choice of parameters the total cross sections as a function of m_A and m_H are visible in the bottom left plot of Figure 1.10. Below $m_{A,H} < 2 \cdot m_t$, the cross sections for such processes range from hundreds of fb for light $m_{H,A}$ down to 1 fb for $m_{H,A}$ close to 1 TeV. For higher $m_{A,H}$, the cross sections drop quickly due to the opening of the top pair decay mode. The bottom right plot in Figure 1.10 shows the theoretical width of the *H* boson for the case $H \longrightarrow Z(ll)A(bb)$ as a function of m_H for different m_A. The width is increasing with the difference of mass between the two bosons but does not take values larger than 25% of m_H in the most extreme case. For most of the cases, it stays below 15% of m_H or even much lower. The *A* boson, being a pseudoscalar and thus having less possibilities for decaying, is expected to have a smaller width than *H*.

The existence of the processes $H/A \longrightarrow Z(ll)A/H(bb)$ will be probed in Chapter 4 using the LHC data collected by the Compact Muon Solenoid detector (CMS) in 2012.

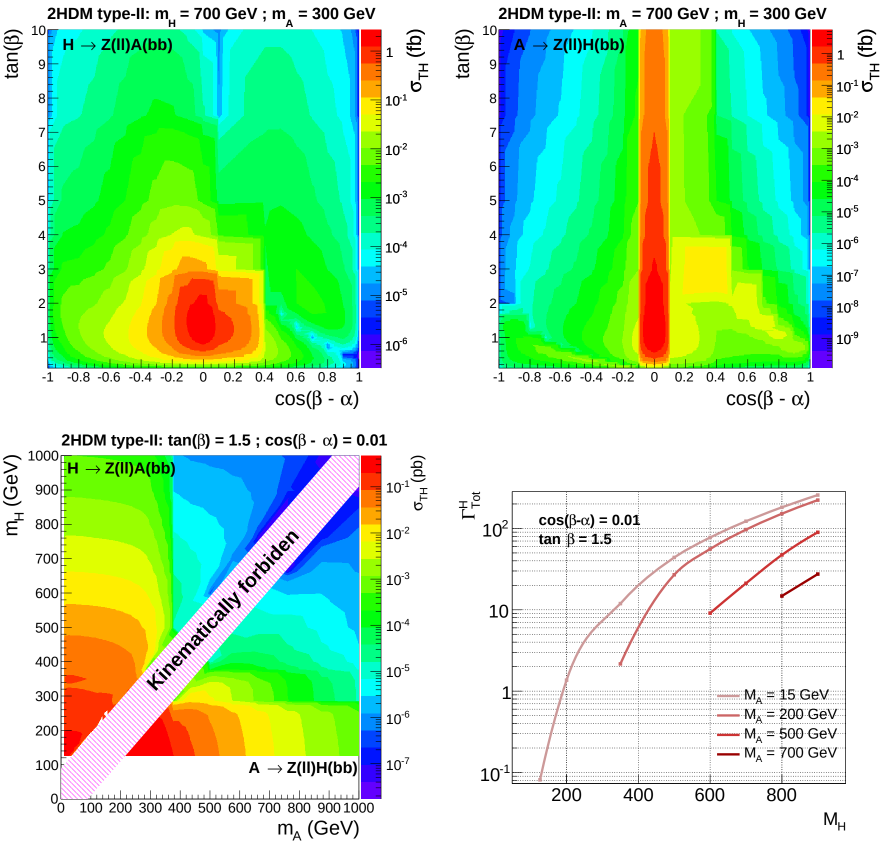


Figure 1.10: The top plots represent the theoretical NNLO cross sections for the processes $H/A \longrightarrow Z(ll)A/H(bb)$ as a function of $tan(\beta)$ and $cos(\beta - \alpha)$ for $m_{H,A} = 700$ GeV and $m_{A,H} = 300$ GeV. The left bottom plot represents the theoretical NNLO cross sections for $H \longrightarrow Z(ll)A(bb)$ and $A \longrightarrow Z(ll)H(bb)$ as a function of m_H and m_A in type II 2HDM with $cos(\beta - \alpha)$ = 0.01 and $tan(\beta)$ = 1.5. The right bottom plot represents the natural width (in GeV) of the H boson in function of m_H (in GeV) for different m_A and for the same choice of parameters.

Chapter 2

Experimental setup

2.1 The LHC

The proton-proton (*p*-*p*) collisions, used for the analyses which will be presented in the next chapters, were produced by the Large Hadron Collider at CERN near Geneva. It is a circular particle accelerator of 27 km of circumference situated on the Swiss-French border. It was built to study *p*-*p*, proton-lead and lead-lead collisions. The latter cases will not be detailed here as these collisions were not used in the work which will be discussed in the following chapters. Technical details about the LHC design can be found elsewhere [33].

The accelerator complex which is used to create the proton beams and to accelerate them to their nominal energy is visible in Figure 2.1 [34]. Protons are produced from hydrogen atoms and successively injected to the LINAC2, BOOSTER, PS, SPS and finally the LHC after having been accelerated to an energy of 50 MeV, 1.4 GeV, 25 GeV, 450 GeV, respectively. In the LHC, they are then accelerated to the nominal energy which was increased over time from 3.5 TeV (2010-2011) to 4 TeV (2012) and 6.5 TeV (2015-2016).

Two beams are circulating inside the LHC, each moving in an opposite direction into separated beam pipes. These beams are divided into bunches of protons. In 2012 each beam was made of 1380 bunches of 1.7 10^{11} protons each. The time between two bunch crossings was 50 ns. The instantaneous luminosity of the beams, which reflects the density of protons at the interaction point in a given time interval, reached a peak of 7.7 10^{33} cm^{-2}s^{-1} in 2012 and resulted in an average of 21 *p*-*p* collisions

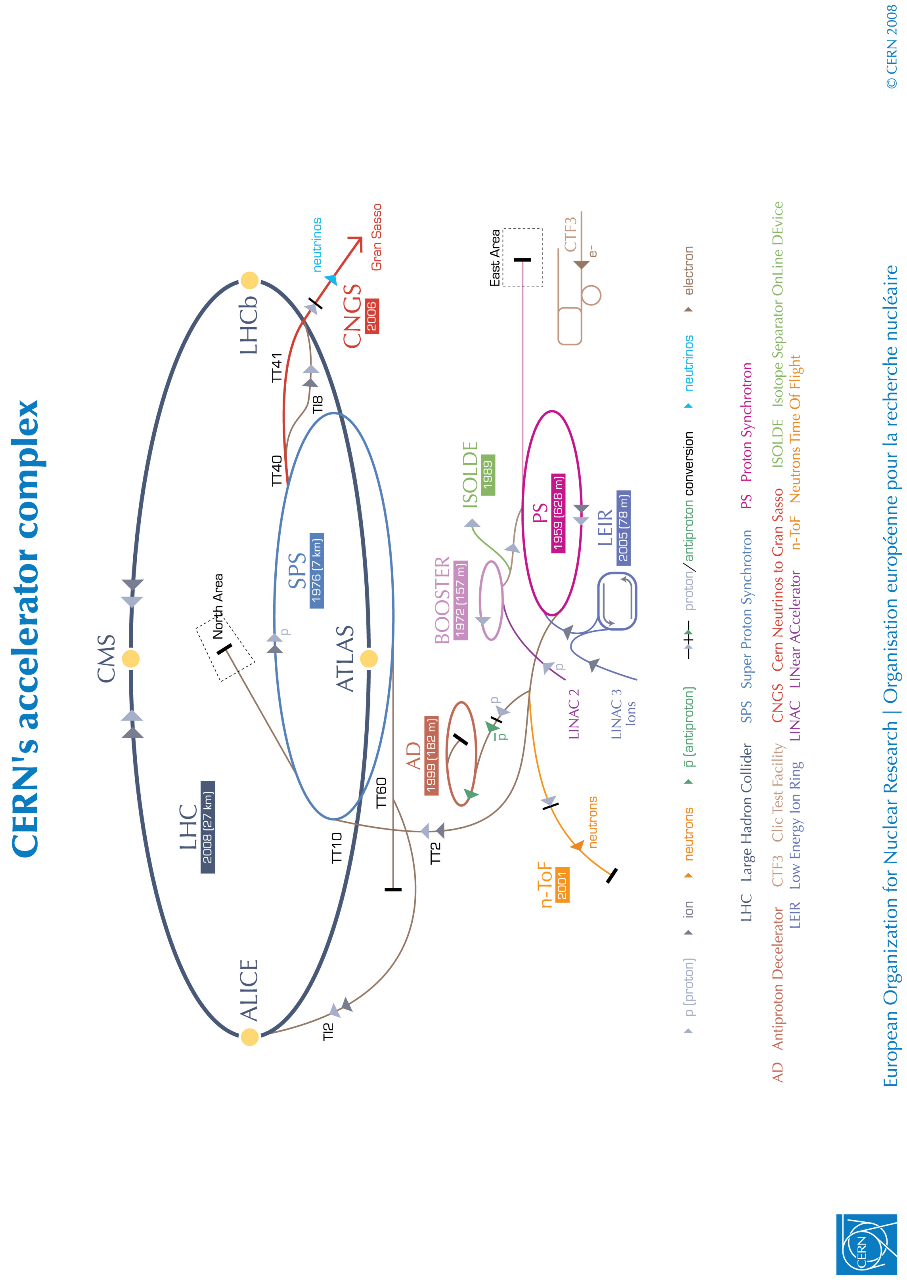


Figure 2.1: The CERN accelerator complex.

for each bunch crossing. These multiple and simultaneous interactions phenomena is referred as ‘pile-up’ (PU) and it is an undesired side effect of the high instantaneous luminosity. Please refer to [35] for the definition of the ‘instantaneous luminosity’ and the relevant parameter values entering its computation for the period 2010-2012 LHC operations.

Once the proton beams are stabilised, collisions are produced in four points. In these four points detectors are placed in order to record what is happening during these collisions. The LHCb detector was optimised to study CP violation and rare decays of B hadrons [36]. The ALICE detector was developed to study heavy-ions collisions [37]. In addition two generic purpose detectors were built: ATLAS [38] and CMS. The works presented in this thesis are based on the data collected by the CMS experiment.

2.2 CMS

The Compact Muon Solenoid (CMS) detector is a generic purpose detector situated along the LHC accelerator. A sketch is shown in Figure 2.2 [39]. The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the solenoid volume are located in concentric layers a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of one barrel and two endcap sections. The apparatus is completed by gas-ionisation muon chambers which surround the solenoid and are embedded in the steel flux-return yoke. Three technologies are used for these chambers: drift tubes, cathode strip chambers, and resistive plate chambers.

A coverage of the range in pseudo-rapidity $|\eta| < 2.5$ (2.4) is provided by the layers inside the solenoid volume (muon chambers). An extended coverage is provided by the forward calorimeter in the endcaps up to $|\eta| < 5.2$. Combining the energy measurement in the ECAL with the measurement in the tracker, the momentum resolution for electrons with $p_T \approx 45$ GeV from $Z \to ee$ decays ranges from 1.7% for non-showering electrons in the barrel region to 4.5% for showering electrons in the endcaps [40]. Matching muons to tracks measured in the silicon tracker results in a relative transverse momentum resolution for muons with $20 < p_T < 100$ GeV of 1.3–2.0% in the barrel and better than 6% in the endcaps [41].

The first level of the CMS trigger system uses information from the calorimeters and muon detectors to select the most interesting events. A high-level trigger processor farm decreases the event rate from approximately 100 kHz to 600 Hz before data storage. A more detailed description of the CMS detector, together with a definition of the coordinate system and kinematic variables, can be found in Ref. [42].

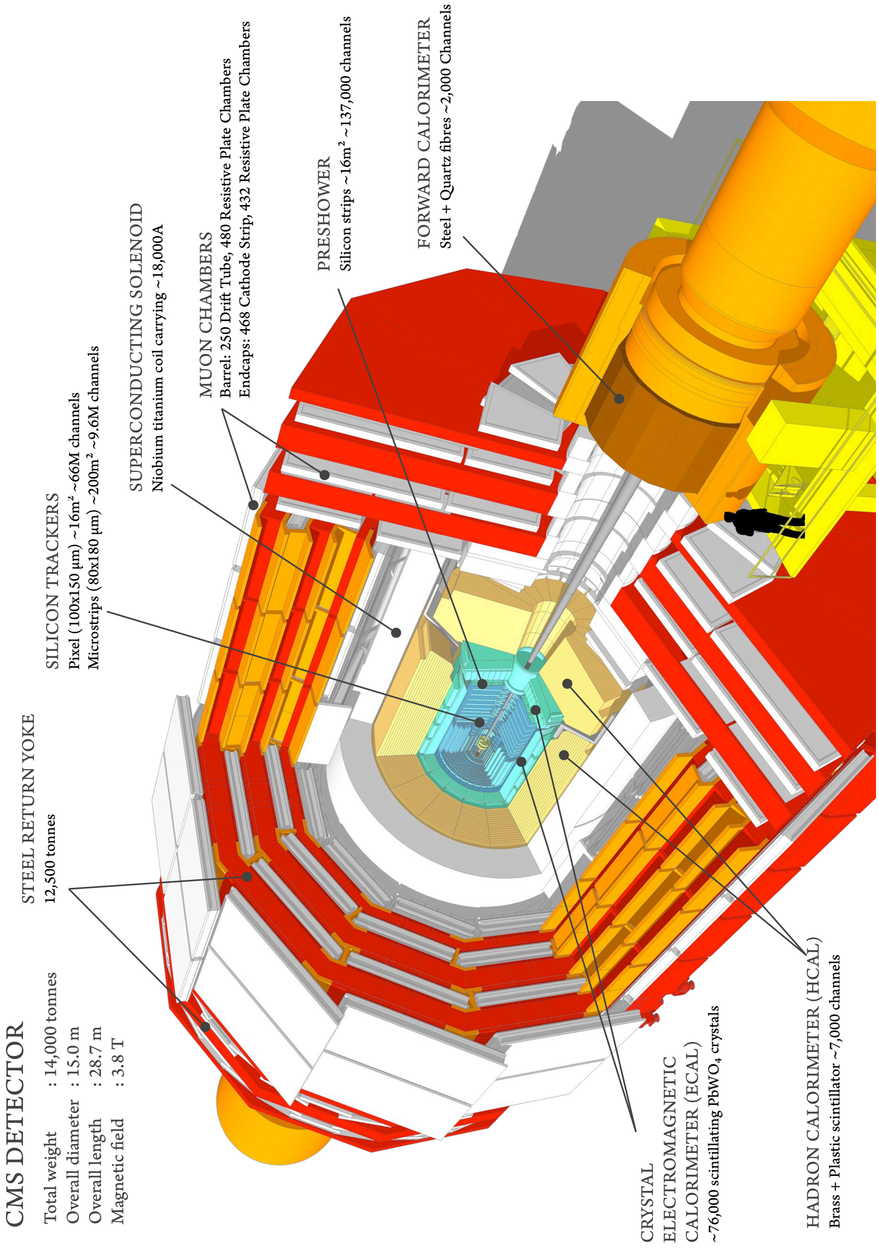

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m^2 ~66M channels
Microstrips (80x180 μm) ~200m^2 ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m^2 ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO$_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.2: Sketch of the CMS detector with view of the internal structure.

As discussed in Chapter 1, final states with *b* quarks are common in the SM (top decay, *Z* + HF production...) and are especially relevant for the studies of the Higgs boson and possible new Higgs bosons. Indeed, this quark is the heaviest fermion next to the top quark. Consequently this is the preferred decay of the SM Higgs boson for a mass of 125 GeV. In models beyond the SM it is also a relevant decay to potentially discover new Higgs bosons which might decay dominantly to a pair of *b* quarks depending on the model parameters. All this highlights the need to develop an experimental setup to differentiate the creation of *b* quarks from the creation of lighter quarks or gluons. The validation and the development of experimental tools applied by CMS to achieve this was part of the work delivered in parallel to the analysis presented in Chapters 3 and 4. The relevant elements to perform the identification of jets originating from *b* quarks (*b* jets) will be detailed in what follows.

The differentiation of *b* jets from other jets (b-tagging) can be done to some extent by taking advantage of the lifetime of the *B* hadrons which originate from the hadronisation of *b* quarks. In fact, *B* hadrons with a Lorentz factor $\gamma > 2$ can travel a few millimeters or more before decaying. With a good track and vertex reconstruction capabilities this property allows experimentally to separate them from prompt production and decays happening at the interaction vertex. At the same time, the B hadron decay products will travel through all the relevant detector components including the first layer situated a few centimeters away from the interaction point. This last point is important because the first hit is essential to get a good resolution on the position of the vertices and the origin of the tracks.

2.2.1 Tracker system

A longitudinal view of the tracker system of CMS is visible in Figure 2.3 [42].

The inner layers of the CMS detector are composed of silicon pixel modules. In the barrel three layers are placed between 4.4 and 10.2 cm from the beam line. In the endcaps two disks complete the inner tracking system, placed at z = ±34.5 cm and z = ±46.5 cm. The pixel cell size is 100×150 μm with a spatial resolution of about 15-20 μm. This small spatial resolution is crucial in order to get a good resolution on the tracks impact parameter - defined as the distance between the interaction vertex and the extrapolated origin of the track - and to be able to reconstruct vertices with low track multiplicity as the ones originating from the decay of *B* hadrons.

The tracker system is completed by silicon strip layers. In the barrel, a total of 10 layers are present up to a radius of 1.1 m from the beam line (TIB and TOB in Figure 2.3). In the endcap a total of 12 layers complete the coverage of the tracker system up to $|\eta| < 2.5$ (TID and TEC in Figure 2.3). The pitches of the micro-strips vary

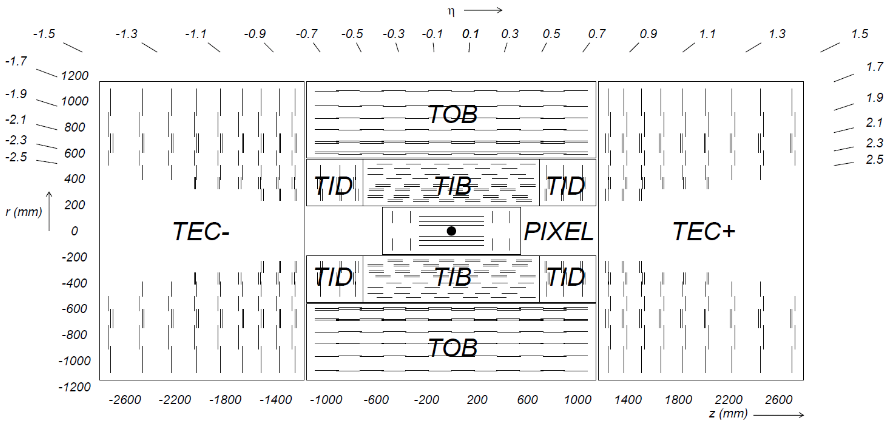


Figure 2.3: Longitudinal view of the tracker system of CMS.

depending on the layers from 80 μm to 184 μm leading to a single point resolution between 23 and 53 μm. Some of the layers carry a second micro-strip detector module which is mounted back-to-back with a stereo angle of 100 mrad in order to provide a measurement of the second coordinate (z in the barrel and r on the disks with r the distance in the plane x-y from the beam line). This allows a resolution of about few hundreds of μm on this coordinate.

Plans for future upgrades of the tracker system can be found in the references [43, 44].

2.2.2 Tracking, vertexing and alignment

Charged particles produced by the p-p collisions inside the CMS detector will interact with the CMS tracker system when traveling through it, creating electron-hole pairs by ionisation. The charge will drift toward the sensors creating a current which will be detected. Hits are reconstructed each time an interaction with a module is recorded. In order to reconstruct the full path of these particles, the hits need to be linked together. Based on the Kalman Filter [45], the CMS tracking algorithm [46] handles this reconstruction despite the high number of possible combinations. This is done by an iterative process in six steps starting from high p_T tracks from the interaction point to lower p_T and less central tracks, reducing the combinatorial process each time by removing hits associated to the reconstructed tracks. The typical resolution on the transverse (longitudinal) impact parameter of a track from a central muon or pion of 100 GeV is about 10 (30-40) μm. The resolution on the particle p_T is in most cases between 1 and 10 % depending on the p_T and η of the particle.

To be able to properly associate hits, the geometry of the tracker system needs to be precisely known and controlled during data taking. Details on the way the alignment of the detector modules is achieved can be found in [47]. A good knowledge of the module positions over the time is mandatory in order to take advantage of the good resolution of the CMS tracker system. The statistical accuracy on the module alignment is found to be better than 10 μm, i.e. smaller than the space resolution of the tracker modules.

Vertices can be reconstructed from the tracks. A vertex from a *p*-*p* interaction is called primary vertex (PV) as opposed to secondary or even tertiary vertices originating from the decay of particles produced by a *p*-*p* collision. The vertex corresponding to the *p*-*p* collision of interest is identified as 'the PV' when other primary vertices are associated to PU events. Details on the reconstruction of the PVs can be found in [46]. In order to perform this reconstruction some quality requirements are imposed on the tracks. These must have a minimum number of hits ($\geq$ 2 pixel layers, pixel+strip $\geq$ 5) and a reasonable normalised χ^2 from a fit to the trajectory ($<$20). To ensure that these tracks are compatible with a *p*-*p* collision, they are required to have a transverse impact parameter to the centre of the beam spot, the luminous region where the collisions take place, smaller than 5 times the transverse impact parameter uncertainty. This allows the removal of tracks which have a low probability of originating from the primary interaction. The cluster of the selected tracks is done using their z-coordinates at their point of closest approach to the centre of the beam spot. To deal with the high luminosity of the LHC and in order to properly separate close-by vertices, a deterministic annealing algorithm [48] is used. Candidate vertices with at least two tracks based on these clusters are then fitted using an adaptive vertex fitter [49]. The resolution on the vertex position depends on the number of associated tracks. It goes down to 10 (12) μm in x (z) for vertices with 50 tracks for events enriched in jets.

Secondary vertices (SVs) can be reconstructed in a similar way. From a *b*-tagging point of view, tracks associated to jets with a $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ cone of 0.3 around the jet axis, are used to perform this reconstruction. However an alternative method, the inclusive vertex finder [50], was developed with the nice feature of being independent from the jet reconstruction. This method was not used in the algorithm which will be presented later because it only started to be used as a standard reconstruction tool at the end of the 2012 data taking period. Still, it was used in an analysis at 7 TeV [50] and, since the 2015 data taking campaign, it has been used as default SV reconstruction for *b*-tagging purposes mainly because of its higher SV reconstruction efficiency (10-15 %) [51].

2.2.3 B-tagging algorithms

In order to select the most relevant tracks, only high quality tracks, with a $\Delta R < 0.5$ to the jet axis, are considered for the b-tagging algorithms. This selection is the following:

- $p_T > 1$ GeV
- $N_{pixel\ hits} \geq 2$
- $N_{hits} \geq 8$
- χ^2/n.d.o.f< 5 where n.d.o.f stands for the number of degrees of freedom in the fit of the track
- $|d_{xy}| < 0.2$ cm where d_{xy} is the transverse distance to the PV at the point of closest approach in the transverse plane
- $|d_z| < 17$ cm where d_z is the longitudinal distance to the PV at the point of closest approach in the transverse plane
- $D_{jet,track} < 0.07$ cm where $D_{jet,track}$ is the distance of the track to the jet axis defined as the distance of closest approach of the track to the jet axis
- $L_{decay} < 5$ cm where L_{decay} is called the decay length, defined as the distance between the point of closest approach of the track to the jet axis and the PV.

For the SV reconstruction, the cut on L_{decay} is completely relaxed and the cut on $D_{jet,track}$ is loosen to 0.2 cm. However, an additional requirement for these tracks is to pass the 'high-purity' criterion as defined in [46] to optimise the purity for each of the iterations in track reconstruction. This criterion is based on the normalised χ^2 of the track fit, the track length and the impact parameter information.

The main feature of tracks from *B*-hadron decays is that they appear more displaced with respect to the PV than other 'ordinary' tracks. The impact parameter of the selected tracks is one of the most obvious discriminating variables in which this displacement is visible. To reduce the pollution from displaced tracks reconstructed with a low precision, the impact parameter significance (IPS = IP/σ_{IP}) is preferred. This also allows a smaller dependency on the change of alignment conditions which will modify the impact parameter value and similarly, its uncertainty.

Simple algorithms were developed based on this variable called 'track counting high efficiency and high purity' (TCHE, TCHP) [52]. In these algorithms the tracks are ordered by decreasing order of IPS. The IPS of the second (third) track is used as discriminant for the high efficiency (purity) case. An algorithm using the information of the IPS of all the tracks in the jet was also developed and called 'jet probability' (JP) [52]. This tagger associates to each track originating from a *B* hadron a probability which is defined according to its IPS. The sum of the probabilities constitutes the discriminant. An alternative version of the JP algorithm is the JBP algorithm which considers at most only the first four tracks with the highest IPS. However none of these algorithms takes advantage of the possible presence of a SV.

The SVs, reconstructed as described above, are selected if:

- They share less than 65 % of their tracks with the PV and the significance of the radial distance between the PV and the SV exceeds 3 σ.
- They do not have a radial distance of more than 2.5 cm with respect to the PV, a mass compatible with the mass of K^0 or exceeding 6.5 GeV.
- They are in a cone of $\Delta R < 0.5$ around the jet direction.

The displacement of the SV is represented by its distance to the PV. As in the case of the IP, the uncertainty on the SV position (dominant with respect to the PV) can be quite large and needs to be taken into account in order to not be polluted by highly displaced SV with small precision. Defining the flight distance significance as the distance between the SV and the PV divided by its uncertainty gives a new interesting discriminating variable.

Simple algorithms based on this variable were also developed and are called 'simple secondary vertex high efficiency and high purity' (SSVHE, SSVHP) [52]. The first one uses SVs with at least two tracks while the second one uses SVs with at least three tracks. The limitation of these algorithms is that the SV reconstruction efficiency is about 65% for the first case and about 50% for the second case.

It is possible to make use of both types of information (track displacement and presence of an SV) by combining these variables using a multivariate technique. The combined secondary vertex (CSV) algorithm was developed for this purpose [52]. It is based on a likelihood method which combines a list of variables related to the tracks and the SVs. Three categories are defined depending if there is a SV, a 'pseudo-vertex' or no SV. A pseudo-vertex can be defined when no SV is found inside the jet and at least two tracks which are associated to this jet have an IPS > 2. These tracks can be combined in a pseudo-vertex allowing for the computation of a subset of secondary-vertex-based quantities even without an actual vertex fit. The complete list of variables used with their definition can be found in [52].

The performance of the algorithms is defined according to their abilities to select *b* jets for a given mistagging rate defined as the efficiency of selecting light jets (*u*, *d*, *s* and gluon jets). Standard working points (WPs) are defined for 0.1, 1 and 10% of mistagging efficiencies for the tight, medium and loose WPs, respectively. The performance curves at 7 TeV are shown in Figure 2.4. At 8 TeV the performance are sensibly the same. Jets originating from *c* quark are a specific case as *D* hadrons have a life time close to the *B* hadrons. This explains why these are more difficult to discriminate than light jets. Still, they will travel a shorter distance in average before decaying which results in slightly less displaced tracks and vertices. The JBP algorithm shows the best performance for the loose WP. For the medium WP the CSV and the JBP are competitive whereas the CSV is the best for the tight WP. The *b*-jet efficiency for the medium WP is about 65% for the CSV algorithm which will be used in the analyses presented in Chapters 3 and 4.

An improved version of the CSV, called 'CSVv2', was made available for the 2015 data taking [51].

2.2.4 Impact of the alignment on b-tagging capabilities

To illustrate the importance of a good knowledge of the detector geometry we will discuss an example which shows the impact of a misalignment of one sub-detector on the b-tagging capabilities. This example is taken from one of the studies performed during the 2012 data taking period to validate the impact on the b-tagging observables of the new alignment conditions derived from data.

During the 2012 data taking period, in November, the two half sections of the pixel detector moved relative to each other. This movement led to a misalignment of about 120 μm along the z axis. It affected several runs (over the period of a few days) before being identified, corrected and injected in the detector conditions which were used to reconstruct the data. The impact on the relevant b-tagging variables have been investigated. As discussed before, the IP of the reconstructed tracks is crucial for the b-tagging performance. In Figure 2.5, the transverse (2D) and 3D IP of the tracks are shown for data passing a trigger with high momentum jets ('JetHT' dataset). The blue histogram corresponds to a reference run before the movement (Run 204577). The red histogram corresponds to a run affected by the misalignment of the pixel detector (Run 208307). The biggest difference is visible for the 3D IP which includes the z coordinate. Less tracks with IP ~ 0 are observed, consistent with a degradation of the IP resolution. The uncertainty on the IP is only marginally affected as the uncertainty on the pixel modules position was unchanged explaining the large effect on the 3D IPS (bottom left plot of Figure 2.5). From this observation, as most of the jets in this sample are light jets, it is clear that the misalignment led to light jets with more displaced tracks which made them more *b*-jets like. This is reflected in the jet CSV values which are shifted toward higher values as can be seen from the bottom right

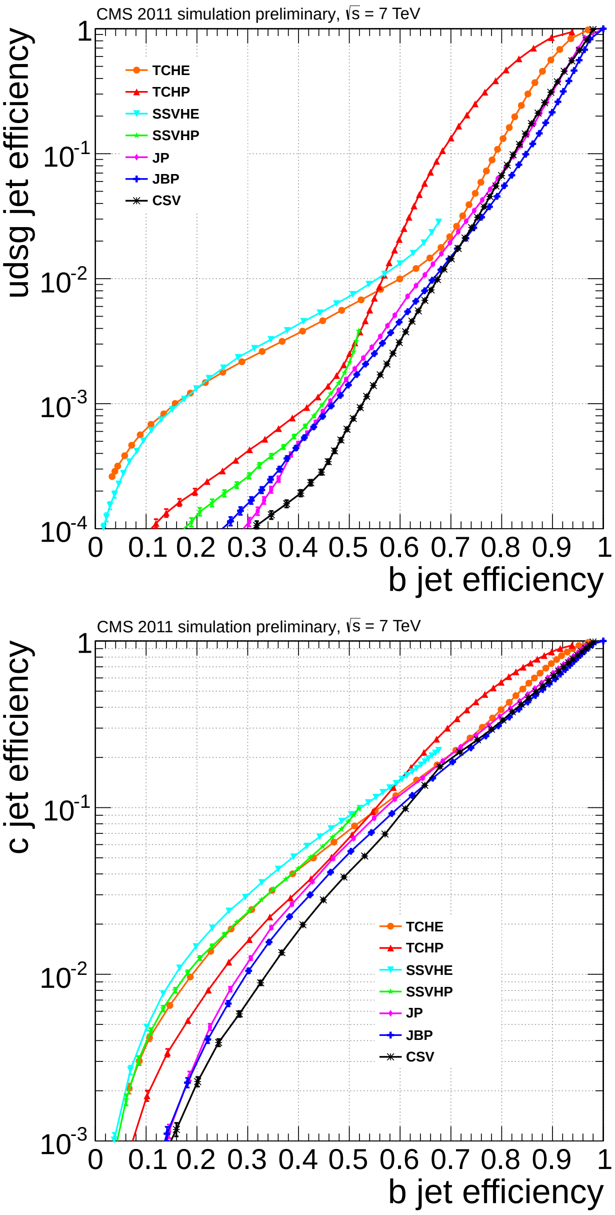


Figure 2.4: Performance curves obtained from simulation (multi-jet events) for the algorithms described in the text. Light-jets (left) and c-jets (right) misidentification probabilities as a function of the b-jet efficiency for jets with $p_T > 60$ GeV.

plot of Figure 2.5. The peak at 0 which corresponds to the region enriched in light jets is clearly reduced, meaning that the light jets tend to be effectively more b-like. Even if, from data, it is difficult to judge and quantify the loss of performance, it is obvious that the performance was degraded by the misalignment. After re-reconstruction of the same data with the adequate detector alignment, the change between the runs almost disappears as can be seen in Figure 2.6. To facilitate the comparison the ratios have the same y-axis range as in Figure 2.5. After the correction of the alignment, both runs look compatible with a ratio close to 1. Residual differences might come from slightly different run conditions (slightly more jets with lower p_T in the Run 208307). In any case a perfect agreement is not expected because the alignment constants are derived as an average over a given data taking period.

This example clearly illustrates the importance to monitor the evolution of the detector geometry with time to avoid the risk that the performance of the object reconstruction deteriorates and impacts the CMS physics program.

2.3 Data and simulation

A summary of the amount of data delivered by the LHC to CMS is shown in Figure 2.7. For the end of 2016 about 40 fb^{-1} are expected to be delivered at 13 TeV. The data used for this thesis were taken during the 2012 data taking period. The data taking efficiency, which corresponds to the percentage of data recorded by CMS, was about 93% during this period. The data usable for the analyses, which will be presented later, corresponds to an integrated luminosity of about 19.5-19.8 fb^{-1} at 8 TeV. The data collected in 2011 at 7 TeV are combined in the first analysis with the 2012 data to improve the sensitivity to the signal. These data correspond to an integrated luminosity of 5 fb^{-1}.

In order to compare the data with what the theory predicts, events can be generated using Monte-Carlo (MC) tools and passed through detector simulation programs to simulate the detector response. Event generators use PDFs to estimate the probability of interaction of the different elements of the protons and the fraction of the energy they are bringing. Available sets of PDFs and recommendations can be found in [53–55]. Matrix element interaction can be generated by a dedicated program at fixed order (LO or NLO). The showering steps can be performed by the same tool or by a dedicated program interfaced with the matrix element generator. At the end of the last step, only stable particles remain (particles which will not decay before crossing the first detector layer). The main generators and showering programs used by CMS with some of their characteristics are listed in Table 2.1. This list is not exhaustive but contains the main generation tools which will be used later. In 2012, the main matrix

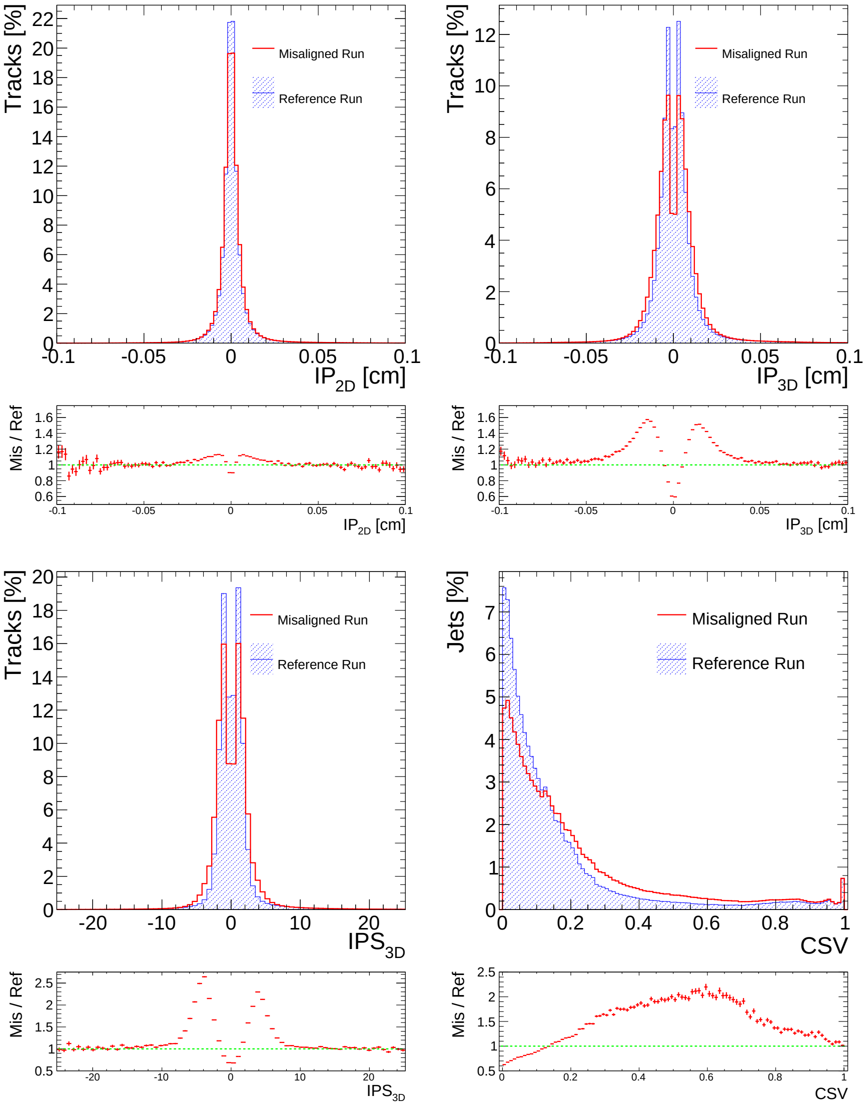


Figure 2.5: From the top left to the bottom right: 2D and 3D track IPs, 3D IPS and the jet CSV discriminant. In blue (red) the distributions are shown for a reference (misaligned) run before (after) the movement of the pixel detector described in the text.

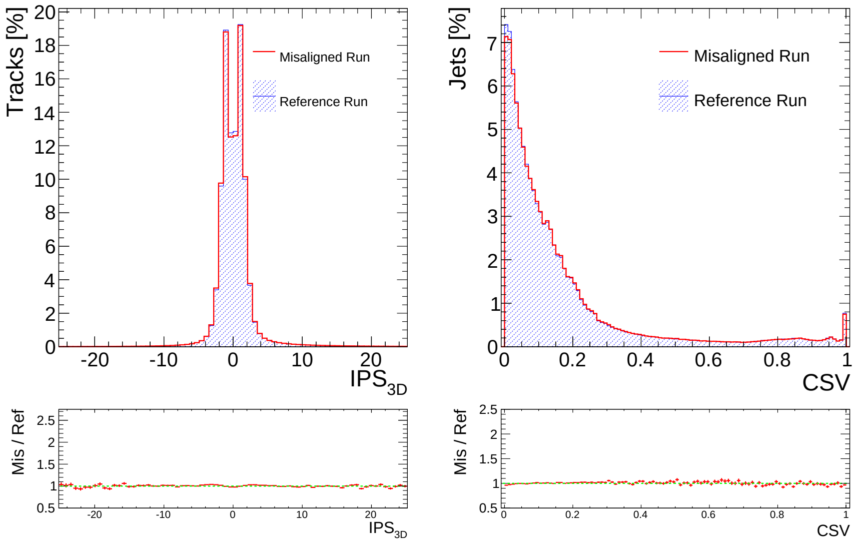


Figure 2.6: Track 3D IPS (left) and jet CSV discriminant (right) for a reference run (blue) before misalignment of the pixel detector and for a run (red) affected by the misalignment, both after being reconstructed with the proper alignment.

element generator used was `MadGraph` (MG) [56] interfaced with `Pythia 6` [57]. It has the advantages of generating LO matrix element processes up to several jets (up to four for Z+jets events) and of being well known and validated by the collaboration. More recently the usage of `aMC@NLO` [58] interfaced with `Pythia 8` [59] tends to replace the use of MG. `aMC@NLO` has as advantage to generate matrix element at NLO. One drawback of this generator, however, is the longer computing time for generating events especially with extra jets. One specific characteristic of this tool is to produce events with negative weight. Consequently more events need to be generated in order to recover comparable statistics to the approach using MG. `Powheg` [60] allows for a set of available processes to also generate matrix element at NLO but without this feature of negative weights. It is used for several processes, especially for Higgs samples. The hadronistation is made with the `Herwig++` [61] showering program. In a few cases `Pythia` is also used as a standalone generator especially when the simulation of extra jets is less relevant (for example di-boson processes). Indeed: `Pythia` is known to produce a softer spectrum for extra jets than generally observed. Finally `TAUOLA` [62] is, when relevant, interfaced to the showering program in order to improve the simulation of the τ decays.

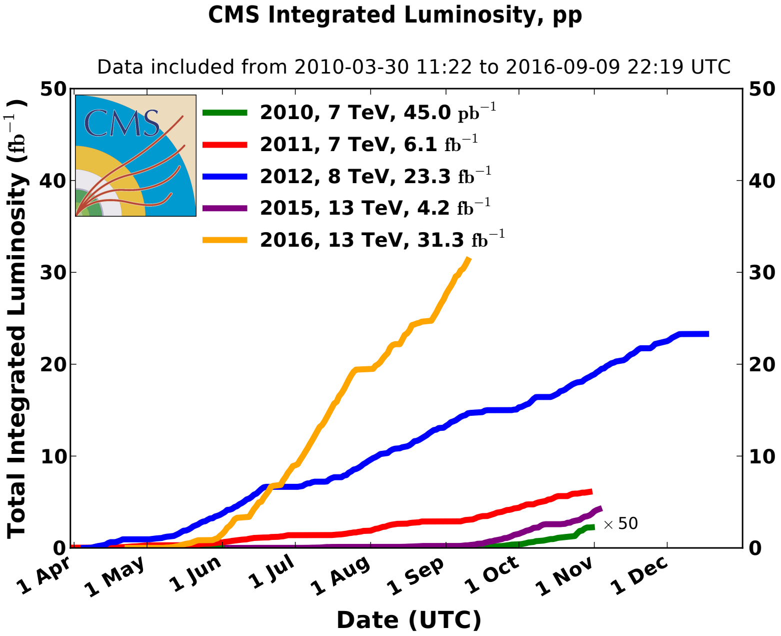


Figure 2.7: Cumulative luminosity versus day delivered to CMS during stable beams and for *p-p* collisions. This is shown for 2010 (green), 2011 (red), 2012 (blue), 2015 (purple) and 2016 (yellow) data taking periods.

Once events are generated, in order to compare them to the data, the response of the detector to the passage of the produced particles needs to be simulated. To do this, CMS is using the `Geant 4` toolkit [63] in order to reproduce the detector geometry and the interactions with the detector materials. The output of this detailed simulation, which can take up to a minute for each event, mimics the output received from the data (same format with the same low-level information) which allows the use of the same reconstruction procedure. It also contains the information on the generated particles for each event. Information about the detector conditions (not working modules, alignment, beam spot, etc) are exploited to get a realistic simulation of the detector performance. However the reconstruction of the simulated samples are generally done once or twice during a data taking period therefore the average conditions for the data can differ slightly from the injected ones in the reconstruction of the simulated events. Effects from PU are simulated by adding the information from events chosen randomly from a simulated minimum bias sample (sample of soft inelastic collisions) after the simulation step. The number of events to be added is also chosen randomly following a predefined probability function which is supposed to be close to the final

Generators	Hadronisation	Order of computation	Processes
MadGraph	N	LO	Automatic for different set of models
Pythia	Y	LO	Set of built-in processes
Powheg	N	NLO	Set of built-in processes
Herwig++	Y	LO/NLO	Set of built-in processes
aMC@NLO	N	NLO	Automatic for different set of models

Table 2.1: List of the main event generators used by the CMS collaboration to simulate Monte-Carlo events at 8 TeV with some of their main characteristics.

PU distribution integrated over the whole data taking period. This is done event-by-event using a Poissonian distribution with as mean the number of expected interactions randomly chosen from this probability function. Residual differences between simulation and data are corrected using correction factors which allow to reweight the simulated events so as to make the simulated samples closer to the actual data.

The data recorded and analysed by CMS already led to more than 530 publications including about 150 publications on the SM (including top physics) and almost 70 about Higgs physics, including BSM Higgs searches. The level of agreement of the data with the SM is summarised in Figure 2.8. No significant deviations have so far been observed.

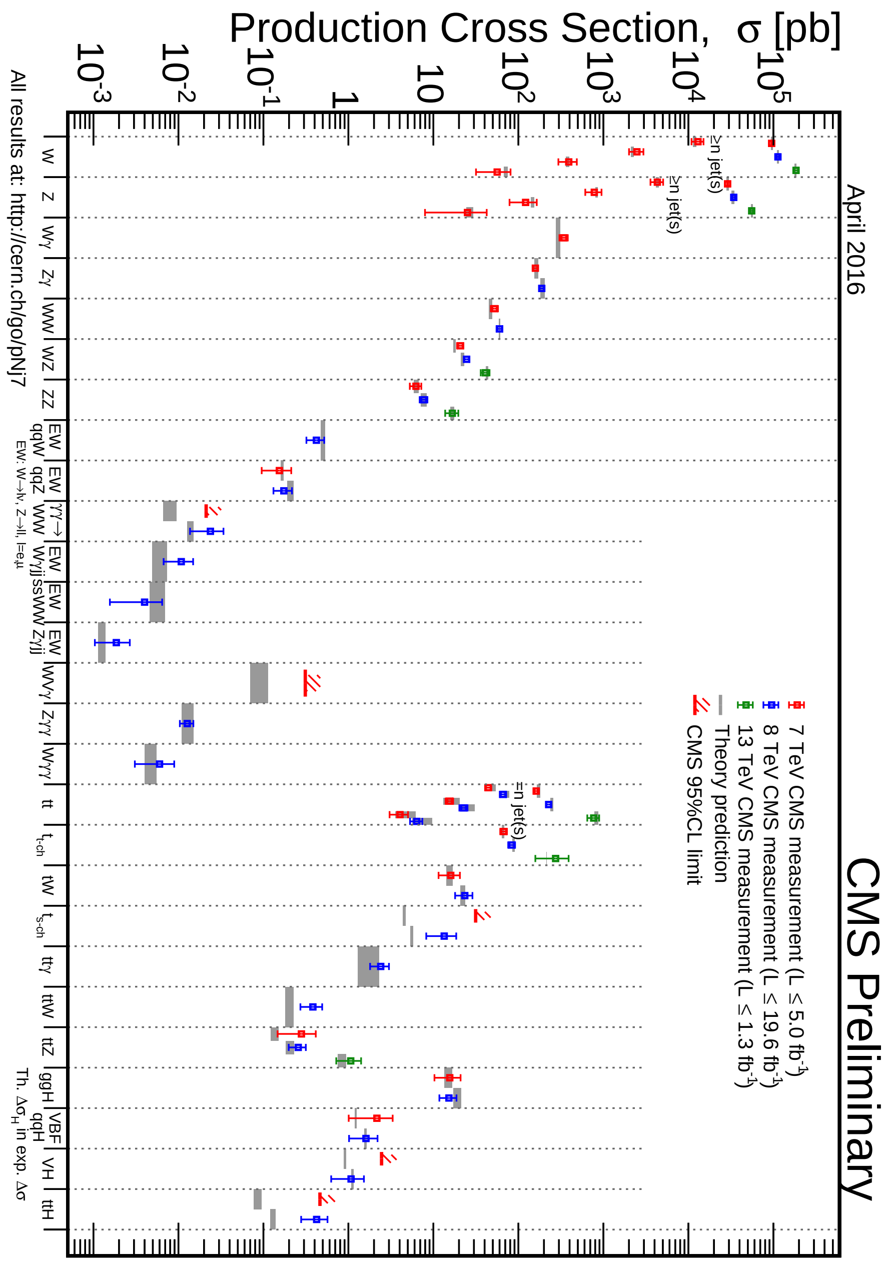


Figure 2.8: Summaries of CMS cross section measurements compared to the theory expectations.

Chapter 3

SM Higgs search

As discussed in Chapter 1, the discovery of the Higgs boson [1, 2] was possible using the di-boson channels (ZZ, WW, $\gamma\gamma$). Even so, at the time of the discovery, the decays into fermions was not confirmed. Despite it has the highest BR at 125 GeV the $h \to bb$ decay is difficult to observe at the LHC. Indeed, the direct production of the Higgs boson decaying to a pair of b quarks has the disadvantage of having as background the huge production of di-jet events ($gg \to bb$ cross section is $\sim 10^8$ larger). The possible production modes for observing the $h \to bb$ decay at the LHC are then Vh (with $V = W, Z$), $t\bar{t}h$ and vector boson fusion. Another challenge of this decay is the poorer mass resolution with respect to $\gamma\gamma$ or ZZ decays due to the experimental limit on the jet energy resolution. In what follows, to avoid confusion with BSM Higgs bosons, the SM Higgs boson will be denominated as h_{125}.

3.1 Analysis strategy

Before the start of the Run 2 LHC, hints of $h_{125} \to bb$ decay had been observed at the Tevatron and at the LHC [64–66]. The choice made for this analysis was to look for the *Z*(*ll*)h_{125}(*bb*) production mode driven by the expertise acquired with the *Z*+*bb* cross section measurement [67] also described in details in [19]. Only the $Z \to ll$ decay is considered with $l = e, \mu$. In order to improve the sensitivity to this process, advanced techniques can be used such as multivariate analysis (MVA), mass regression, etc. However such techniques cannot guarantee to make use of all the available information by themselves. The originality of this analysis is the usage of a matrix element (ME) technique. Such technique has the advantage to use all the kinematic information from the events. Combination of several advanced techniques can then be used to get the best sensitivity. This has successfully been used, for example, in the $t\bar{t}h_{125}$ with $h_{125} \to bb$ search [68]. Here this analysis is an extension of the analysis performed at 7 TeV and detailed in [69]. I will focus, amongst others, on the issues faced due to the increasing PU during the 2012 data taking and the optimisation performed to reduce their impact on the results. This analysis intended to complement the published CMS analysis [66]. A combination with the 7 TeV data collected in 2011 will also be shown.

The final state is composed of two leptons (electron or muon) and two *b* quarks. The constraints are:

- Two same flavour opposite sign leptons with an invariant mass close to the *Z* mass.
- Two *b* jets with an invariant mass close to the Higgs mass hypothesis.
- No neutrino, so no physical missing transverse energy (E_T^{miss}).

The main backgrounds are *Z*+jets, $t\bar{t}$ and *ZZ*. The presence of the two leptons ensures a clean event signature thanks to the good ability of the CMS detector to reconstruct leptons. This guarantees a negligible contamination from pure multi-jets events. The presence of *b* jets can be tagged using the algorithms described in Section 2.2.3. This is used to reduce the *Z*+jets contribution as well as the small *WZ* contribution. The absence of neutrinos makes it possible to reject $t\bar{t}$ events as well as events from smaller contributions such as *tW* and *WW* processes by selecting events with low E_T^{miss}. Constraining the di-lepton mass further reduces these contributions.

In order to enhance the discrimination power of this search between the signal and the backgrounds, the main specificity of this analysis is the use of a ME technique based on the `MadWeight` program [70]. More details on this technique and its possible application are given in [69]. In short, several ME weights are computed corresponding to different background and signal hypotheses. Each of these ME weights give an indication for an event to be compatible with the tested hypothesis. For each event, this represents an estimation of the probability that this event was produced by a given process, based on the observed kinematics of the final state particles of the event and the predictions from the `MadGraph` event generator. This probability is written

$$P(p^{vis}|\vec{\alpha}) = \frac{1}{\sigma_{\vec{\alpha}}} \int dx_1 dx_2 f(x_1) f(x_2) \int d\Phi |M_{\vec{\alpha}}(p)|^2 TF(p, p^{vis}). \tag{3.1}$$

where $f(x_1)$ and $f(x_2)$ are the parton distribution functions which describe the energy spectrum of the incoming partons. $|M_{\vec{\alpha}}(p)|$ is the probability amplitude of the hard scattering between the partons for the considered process. TF(p, p^{vis}) is the transfer function allowing to go from the reconstructed objects to the initial partons. $\sigma_{\vec{\alpha}}$ is the cross section of the process in the integrated phase space. The unnormalised probability, defined as $W = \sigma_{\vec{\alpha}} P(p^{vis}|\vec{\alpha})$, corresponds to the so-called ME 'weight'. As an ME weight is a small number (typically between 10^{-10} to 10^{-30}), the variable $W^{-log} = -log_{10}(W)$ is defined. The more an event is compatible with a given hypothesis the smaller is W^{-log}.

These ME weights have the advantage that they encompass the full kinematic information available in the event by using the quadri-vector of the two leptons and two jets which are reconstructed and selected. The integration is performed using the available phase space given by the TFs. These TFs reflect the resolution on the measured energies and momenta of the reconstructed objects. The integration starts with the better known particles, in this case the leptons, to explore the available phase space. The energy-momentum $(E - p)$ conservation constrains this integration by fixing the energy of the jets. In some cases this can lead to an over-constrained system and some constraints had to be relaxed in order to let the program float within the available phase space for the jets. This is the case when the bb resonance has a small width. In such case, the m_{bb} will often not match the resonance and fall outside the peak. Relaxing the $E-p$ conservation allows to explore the phase space for the jets and to find the best combination which matches the resonance peak. Another possibility to mitigate this effect is to relax the constraint on the physical width by enlarging the width to a value which gives a smooth distribution of W^{-log} when testing a hypothesis on events generated with this hypothesis. In this configuration the events outside the peak get back inside the peak. This might be seen as a way to absorb in the width the experimental resolution on the mass of the resonance. As both choices do not give completely correlated results, both computations are used. An alternative to these solutions is to allow a variation of η_{jet} by adding a TF to take into account the difference in the direction between the generated parton and the jet directions. This configuration was not used for this analysis. All the ME weights used in this search are listed in Table 3.1.

In addition, events are categorised according to the multiplicity of jets. This choice is driven by the different sensitivity observed in the two categories: exactly 2 jets $(2j)$ and at least 3 jets $(3j)$. As shown in [69] this difference is attributed to the poorer resolution on m_{bb} in the $3j$ category. Final state radiations are responsible for this observation. This categorisation also makes sense in view of using a ME technique. Indeed: final state radiations are not treated by the ME technique used for this analysis. This implies a loss of power of this technique in such configuration. A dedicated ME weight can in principle be used for such topology but it is time consuming and it has been shown that there is no clear gain compared to the choice made for this analysis [71]. Here, in order to get part of the lost sensitivity back, additional kinematic

ME weights	hypotheses
W_{gg}	$gg \longrightarrow Z + bb$
W_{qq}	$qq \longrightarrow Z + bb$
$W_{t\bar{t}}$	$t\bar{t} \longrightarrow llbb\nu\nu$
W_{ZZ0}	$ZZ \longrightarrow llbb$
W_{ZZ3}	$ZZ \longrightarrow llbb$ with no $E - p$ conservation constraint
W_{Zh0}	$Zh_{125} \longrightarrow llbb$ with relaxed h_{125} width constraint (∼6 MeV to 2.5 GeV)
W_{Zh3}	$Zh_{125} \longrightarrow llbb$ with no $E - p$ conservation constraint

Table 3.1: List of ME weight hypotheses used in the analysis.

information is used in the $3j$ category. Considering the closest jet (j) in ΔR to one of the two selected b-tagged jets, the following variables are defined:

- $\Delta R(b, j)$: the smallest ΔR between one of the selected b-tagged jets and j.
- m_{bbj}: the invariant mass of the system composed of the two selected b-tagged jets and j.

The ME weights and these new variables are then used as input to a cascade of Neural Networks (NNs). At first three intermediate NNs are trained. Each of these discriminates one background from the signal hypothesis making use of the ME weights of the signal and the considered background. These three NNs are then used as input to a final NN trained to discriminate all the backgrounds from the signal. The shape of the last NN is used as discriminator to look for the presence of the signal. These NNs are built using the TMultiLayerPerceptron class defined in ROOT [72]. This cascade of NNs is illustrated in Figure 3.1 for the category $2j$. In the category $3j$, the only difference is the use of $\Delta R(b, j)$ and m_{bbj} in the three intermediate NNs.

In order to be more sensitive to the signal and to be able to control the background normalisation, two orthogonal regions are defined. The first one, called the signal region (SR), is enriched in signal by selecting a window around the expected Higgs boson mass. The second one, called the control region (CR), is depleted in signal by selecting the complementary region in m_{bb}. The background normalisation is extracted from the CR. The SR is then used to search for the presence of the signal. Shapes are taken from simulation. Reweighting procedures are used in order to improve the data description of the simulation. These reflect the approximation of the simulation and the evolution of the detector during the data taking period. The average performance is measured both in data and in simulation in order to estimate the efficiencies. Correction factors are derived to cover the residual differences. Dependency on the p_T and η of the reconstructed objects are taken into account when relevant. Events

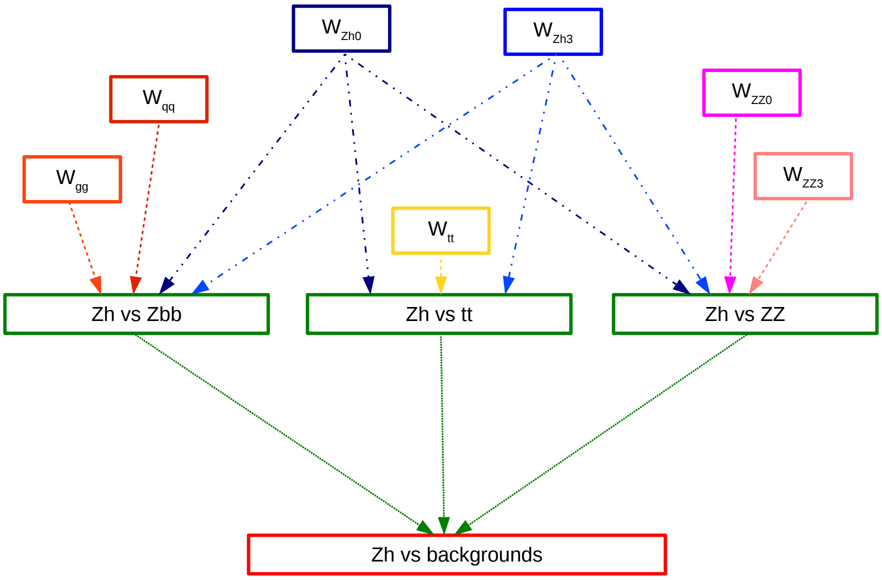


Figure 3.1: Illustration of the cascade of NNs.

in simulation are then reweighted one by one according to these correction factors in order to reproduce the data in the best possible way. This has been done for the trigger efficiency [40, 73], the lepton identification and isolation [40, 74, 75] and for the *b*-jet identification [76]. A similar procedure is followed to reproduce the PU multiplicity observed in data.

The data in the SR has been masked to perform the optimisation of the search in order to avoid biases from the possible presence of the signal. The method has been first applied to the search of the *ZZ* process. This test is used to validate the method. The data in the SR can be unmasked after these steps are completed. If no significant excess is observed, upper limits can be derived on $\mu = \sigma/\sigma_{SM}$, where σ_{SM} is the expected cross section from the SM.

3.2 Setup

The search is based on the standard CMS analysis framework [77].

3.2.1 Samples

The 8 TeV data collected by CMS in 2012 are used corresponding to an integrated luminosity of 19.5 fb^{-1}. The datasets used for this search contain the events selected by the di-muon and di-electron triggers. These triggers require the presence of at least two electrons or two muons with p_T greater than 8 GeV and at least one with a p_T greater than 17 GeV. Loose identification and isolation criteria are required with respect to the one used later in the analysis.

In order to describe the backgrounds and the signal kinematics, events have been simulated using MC event generators. The full list of samples used are shown in Table 3.2.

Data

Di-electron
Di-muon

MC

Samples	Cross section in pb (Calculation order)	Number of events	% of events used
Z+jets: inclusive	3503.7 (NNLO), 2950 (LO)	30459503	100%
Z+jets: $50 < p_T^{ll} < 70$ GeV†	93.8 (LO)	4930773	100%
Z+jets: $70 < p_T^{ll} < 100$ GeV†	52.3 (LO)	1413395	100%
Z+jets: $p_T^{ll} > 100$ GeV†	34.1 (LO)	2662137	100%
Z+jets: $p_T^{ll} > 180$ GeV†	4.6 (LO)	1555476	100%
$t\bar{t} \to ll\nu\nu bb$	27.3 (NNLO)	12119013	30%
$t\bar{t} \to l\nu jjbb$	109.2 (NNLO)	25414818	100%
ZZ	8.2 (CMS)	9799908	70%
Zh_{125}	0.0249 (NNLO QCD + NLO EW)	999462	20%
Zbb‡	76.8 (LO)	14129304	20%
$t\bar{t}^{\ddagger}$	245.8 (NNLO)	6923750	100%

Table 3.2: List of samples used in the analysis with their associated cross sections in pb, the number of events generated and the % of events used in this analysis. The order of precision of the theoretical cross sections is shown in parenthesis. When CMS is specified, the theoretical cross section is rescaled to the best CMS measurement at the time of the analysis. The † symbol refers to the MC samples not used for the NNs training. The ‡ symbol refers to the MC samples used only for the NNs training.

For some samples for which a high event selection efficiency is expected, only a fraction of the events are used in order to be less penalized by the time needed to compute the ME weights. These fractions are chosen in order to keep a reasonable statistic for the search. The contributions from *WW*, *WZ* and single top are neglected as in [67]. The *ZZ* sample is generated with `PYTHIA 6` with the *Z2** tune [78] and interfaced with `TAULA`. The *Z*+jets and $t\bar{t}$ samples are generated with `MadGraph 5` interfaced with `PYTHIA 6` with the *Z2** tune. Four Z+jets samples have been produced in order to improve the MC statistics in the region where the sensitivity to the signal is expected to be higher, e.g. for $p_T \gtrsim 50$ GeV. These samples generated with a cut on the generated p_T^{ll} are listed in Table 3.2. The five *Z*+jets samples are merged using weights based on the relative LO cross sections, extracted from `MadGraph`, in the different bins in p_T^{ll} reported in Table 3.3. The signal sample is generated with `POWHEG` interfaced with `HERWIG++`. NLO electroweak corrections [79] to the Zh_{125} production as a function of the p_T of the *Z* boson have been applied to this sample. The generated events are simulated within the CMS detector using the `GEANT 4` toolkit as mentioned in Section 2.3.

Bins	Relative LO cross section
$p_T^{ll} < 50$ GeV	93.96%
50 GeV $< p_T^{ll} <$ 70 GeV	3.18%
70 GeV $< p_T^{ll} <$ 100 GeV	1.71%
100 GeV $< p_T^{ll} <$ 180 GeV	1.00%
$p_T^{ll} > 180$ GeV	0.15%

Table 3.3: Relative LO cross section in the different bins in p_T^{ll} used to merge the different *Z*+jets samples.

In what follows, and except if another definition is explicitly mentioned, the *Z*+jets sample is subdivided according to the flavour of the two selected *b*-tagged jets: *Z+bb*, *Z+bx*, *Z+xx* where *x* corresponds to non *b* jets (*u*, *d*, *s*, gluon and *c* jets). The flavour of the jets are obtained by using the available generator information. A matching in ΔR is performed between the reconstructed jets and the generated partons before the hadronisation step. A *b* jet is then defined by a jet matching at least one *b* parton.

3.2.2 Selection and object reconstruction

Leptons are reconstructed using particle-flow (PF) techniques [80–82]. Electrons are identified using the cut-based medium working point defined in [40]. Muons are identified using the tight definition defined in [74]. The isolation of the leptons is performed using a cone of $\Delta R < 0.3$ (0.4) for electrons (muons). The pile-up contamination is subtracted from the cone by means of techniques exploiting charged deposits inside the cone itself. The isolation criterion is $I_{\text{rel}} = I_{\text{abs}}/p_T < 0.15$ (0.2) for the electrons (muons) where:

$$I_{\text{abs}} = \sum_{\text{CH}} p_T + \max\left(\sum_{\text{NH}} p_T + \sum_{\gamma} p_T - \sum_{\text{PU}} p_T, 0\right) \tag{3.2}$$

The $\sum_{\mathrm{i}} p_T$ is the scalar sum of the transverse momenta of the charged hadrons originating from the primary vertex (i=CH), the neutral hadrons (i=NH) and the photons (i=γ). The term $\sum_{\mathrm{PU}} p_T$ is the contribution from PU to the neutral component. It is computed in a different way for electrons and muons. For electrons it is estimated by $\sum_{\mathrm{PU}} p_T = \rho A_{eff}$ where ρ is the median energy density in the detector [83, 84] and A_{eff} the effective area of the isolation cone in the (η, ϕ) plane rescaled to take into account the effective neutral contribution to ρ [40]. For muons it is estimated by $\sum_{\mathrm{PU}} p_T = 0.5 \cdot \sum_{\mathrm{CH,PU}} p_T$ where $\sum_{\mathrm{CH,PU}} p_T$ is the sum of the transverse momenta of the charged hadrons not originating from the PV. The factor 0.5 corresponds to an estimated average of neutral to charged particles produced in the hadronisation process and is measured using simulated events. Electron energy corrections [85] and muon momentum scale corrections [86] are also applied for a better data-MC matching of the lepton kinematics.

Jets are clustered from the four-momenta of the particles reconstructed by the PF algorithm, using the FASTJET software package [87]. The anti-k_T jet clustering algorithm [88] is used with a distance parameter of radius R = 0.5. Corrections to the jet energy scale (JES) are applied both on data and MC to account for detector effects and PU contamination [89]. The PU contribution to the jet energy is obtained using the ρ parameter defined above. This contribution is then subtracted from the jet energy.

Identification of b jets is done via the CSV algorithm described in Section 2.2.3. At the medium working point (WP) the b-tagging efficiency is about 65-75% for b jets with p_T in the range 80-150 GeV (for a mistagging efficiency of 1% for light jets). This WP was chosen for this search because it allows to have a sufficiently pure sample, without losing too much in efficiency, in particular when requiring two b-tagged jets.

The E_T^{miss}, defined as the magnitude of the vector sum of the transverse momenta of all observed bodies in an event, is based on PF reconstructed objects. Corrections for JES, PU contamination and modulation in ϕ are applied [90]. The 'E_T^{miss} significance' [91] is used in the selection because this variable is more discriminating than the E_T^{miss} itself and, at the same time, leads to a smoother E_T^{miss} distribution as will be shown later in Figure 3.10. As defined, the significance offers an event-by-event assessment of the likelihood that the observed E_T^{miss} is consistent with zero, given the reconstructed content of the event and known measurement resolutions of the CMS detector.

Particles are required to originate from the PV, reconstructed as described in Section 2.2.3. The PV is characterised by the largest quadratic sum of the p_T of its constituent tracks.

The reconstruction and selection of the *llbb* events require the presence of two reconstructed lepton candidates of the same flavour and with opposite-sign forming an invariant mass pair in the range [76,106] GeV. This significantly reduces the contamination by $t\bar{t}$ and non-resonant *Z*+jets events. In case of multiple Z candidates, the lepton pair with the closest invariant mass to the Z mass is chosen. At least two *b*-tagged jets are required in order to suppress backgrounds with no *b* jets, especially the *Z*+light-jets events. The two *b*-tagged jets with the highest CSV discriminator value are chosen to form the h_{125} boson candidate. A cut to select events with low E_T^{miss} significance helps reducing the contamination from processes, here mainly $t\bar{t}$, with true E_T^{miss} coming from the production of neutrinos. The cuts defining the selection are listed in Table 3.4.

$p_T^{\mu,e} > 20$ GeV
$\|\eta_\mu\| < 2.4$, $\|\eta_e\| < 2.5$
76 GeV $< m_{ll} <$ 106 GeV
$p_T^{ll} > 20$ GeV
$p_T^{jet} > 20$ GeV
$\|\eta_{jet}\| < 2.4$
$\Delta R(l, j) > 0.5$
$CSV_b > 0.679$
$p_T^{b1} > 40$ GeV, $p_T^{b2} > 25$ GeV
E_T^{miss} significance < 10

Signal Region

$n_{jets} = 2$	$n_{jets} \geq 3$
80 GeV $< m_{bb} <$ 150 GeV	50 GeV $< m_{bb} <$ 150 GeV

Control Region

$n_{jets} = 2$	$n_{jets} \geq 3$
$m_{bb} <$ 80 GeV or $m_{bb} >$ 150 GeV	$m_{bb} <$ 50 GeV or $m_{bb} >$ 150 GeV

Extended Control Region

Control Region less the m_{ll} requirement
60 GeV $< m_{ll} <$ 120 GeV

Training Signal Region

Signal Region less the CSV_b requirement
$max(CSV_{b1}, CSV_{b2}) > 0.679$
$min(CSV_{b1}, CSV_{b2}) > 0.244$

Table 3.4: Event selection for the objects in the (training) signal and (extended) control regions where b corresponds to the selected *b*-tagged jets, b_1 and b_2 the leading and sub-leading selected *b*-tagged jets and $l = e, \mu$.

One important specificity to note in the selection, and which differs from the 7 TeV analysis [69], is the cut on the p_T of the two b-tagged jets forming the h_{125} candidate and on the p_T^{ll}. The cut is raised from 20 GeV to 40 and 25 GeV for the leading and sub-leading selected b-tagged jets. A new cut at 20 GeV is introduced on the p_T^{ll}. The motivation for these modifications comes from the different PU conditions. The mean of PU interactions by bunch-crossing increases from $\sim$10 in 2011 to an average of 21 in 2012. The mean number of interactions per bunch crossing during the 2012 data taking period can be seen in Figure 3.2. Jets coming from PU are therefore expected to be more numerous. This increases the background contributions. This was the main issue identified in this analysis as a result of the increase of the number of PU interactions.

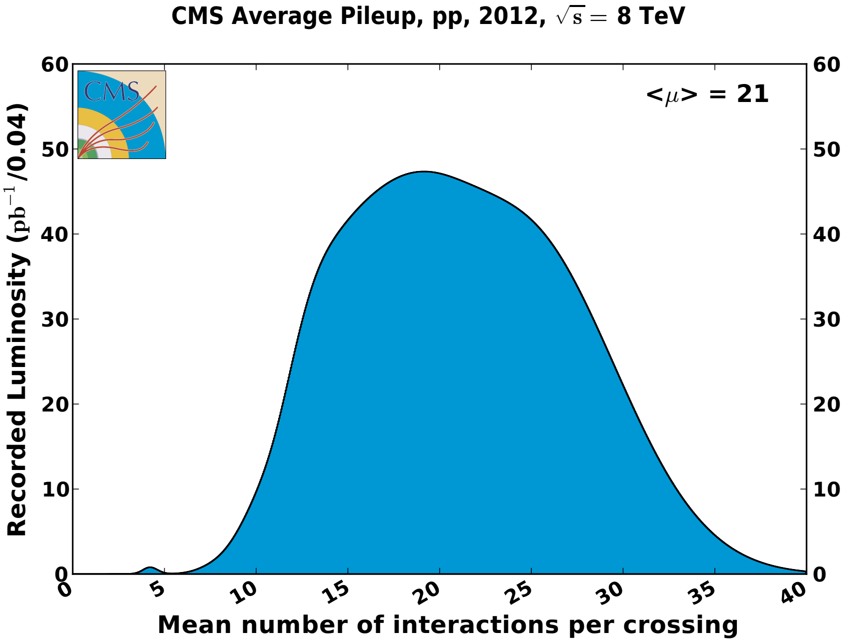


Figure 3.2: Mean number of interactions per bunch crossing during the 2012 data taking period.

In order to rely as little as possible on the simulation of PU, it is worth to reduce its presence as much as possible. To investigate the level of contamination from PU in our events, a study was performed using the available information from the generators in the simulated events. No generator content is available from the PU interactions. This implies that what cannot be matched to a generated object has a high probability to come from a PU interaction. In this way, the absence of a generator jet associated to a reconstructed jet indicates with high probability a jet which originates from a PU interaction. To match the generator jets to the reconstructed ones, a simple ΔR matching within a cone of 0.4 is used. It has been observed that the *Z*+jets events are especially sensitive to the presence of PU jets. This contamination can be seen in Figure 3.3 where it is represented by the brown contribution. The cut on the *b*-tagged jets and di-lepton p_T are chosen in order to make this contribution negligible keeping a similar sensitivity to the signal. Together, these cuts remove $\sim 45\%$ of background and $\sim 15\%$ of signal events. The technique developed here to identify PU jets has been adopted by the group in charge of the *b*-jet identification studies where a similar problem appeared especially for high PU conditions in the perspective of the upgrade of the CMS detector.

3.3 Analysis

3.3.1 Background normalisation

To extract the normalisation of the main background processes, a fit is performed in the CR extended by increasing the m_{ll} window, as defined in Table 3.4. This increases the sensitivity of the fit to the $t\bar{t}$ contribution. Four contributions are estimated:

- Zxx: events with no *b* jets reconstructed in the acceptance.
- Zb(b)j: events with at least one *b* jet reconstructed in the acceptance and with at least one extra jet reconstructed in the acceptance.
- Zbb: events with exactly two *b* jets and without another jet reconstructed in the acceptance.
- $t\bar{t}$: all $t\bar{t}$ events.

The unusual parametrisation of the *Z*+*b*-jets processes is chosen to consider the modelling of extra jet especially to consider the impact of the NLO contributions. This is the consequence of the way *b* quarks are produced at the LHC. Indeed: they are produced in pair through gluon splitting (see as example the diagram of Figure 1.2c). The production of *Z*+*b* jets+extra jets is illustrated in Figure 3.4. The final state is always composed of two *b* quarks and at least one extra parton. This explains why the contributions *Z*+*bx* in both categories ($2j$ and $3j$) and *Z*+*bb* in the category $3j$ are assumed to originate from the same process and are then estimated as a single process. The ZZ contribution is too small to be extracted from the data; therefore it is fixed and rescaled to the best available CMS measurement at the moment of the analysis [92].

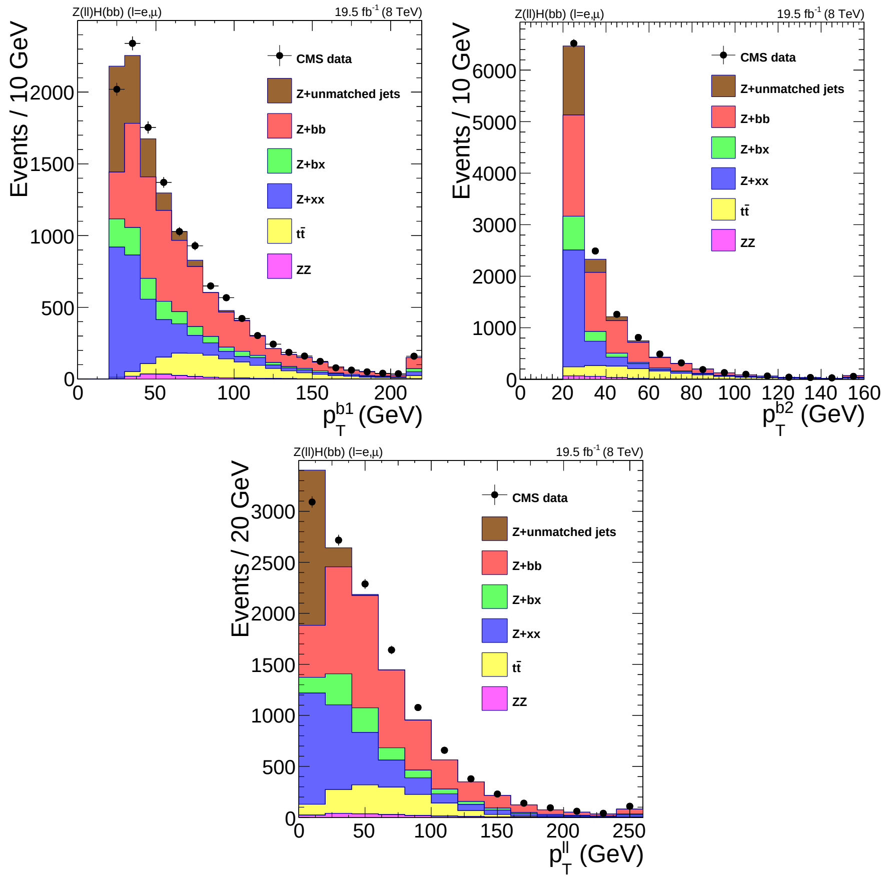


Figure 3.3: Comparisons in the CR+SR of data and simulation for the p_T^{b1} (top left), the p_T^{b2} (top right) and the p_T^{ll} (bottom) before cutting on $p_T^{b1,b2,ll}$. Simulation samples are normalised to the theory expectation. The Z+unmatched jets contribution is defined as the Z+jets events where none of the two selected b-tagged jets match a generator jet.

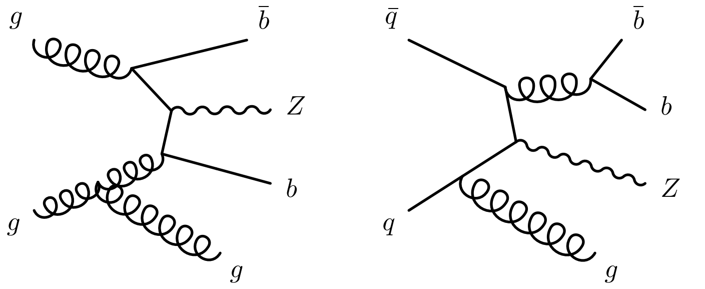


Figure 3.4: Examples of diagrams for the production of *Z*+*b* jets+one extra jet.

In order to be sensitive to these different contributions, 2D shapes are built using:

- The CSV product of the two selected *b*-tagged jets: this variable is sensitive to the flavour of the selected jets.
- A NN discriminating $t\bar{t}$ and *Z*+jets processes: this variable is sensitive to the fraction of $t\bar{t}$ and *Z*+jets events.

The NN discriminating the *Z*+jets and $t\bar{t}$ events takes as input the ME weights corresponding to the two processes, corresponding to a total of three inputs. Two layers of two and four neurons have been used to perform the training. The evaluation of the NN output is shown in Figure 3.5 for *Z*+jets and $t\bar{t}$ events. This NN gives a nice discrimination of the two processes.

The contributions are estimated with a simultaneous fit of four categories:

- Electron channel - $2j$.
- Electron channel - $3j$.
- Muon channel - $2j$.
- Muon channel - $3j$.

The resulting scale factors (SFs) are shown in Table 3.5. They are compatible with the ones obtained at 7 TeV.

The plots showing the input variables after the fit are shown in Figure 3.6 as 1D projection of the 2D templates. The binning is driven by the MC statistics in order to avoid empty bins in the 2D templates.

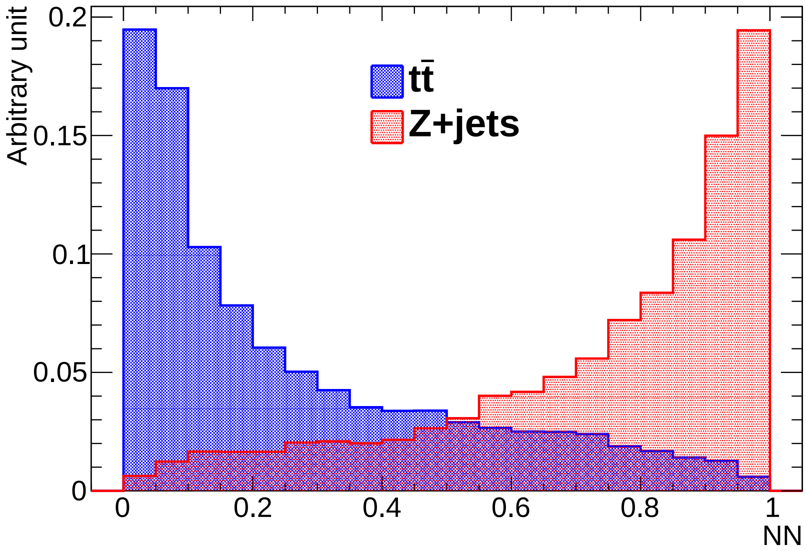


Figure 3.5: Output of the NN discriminating the Z+jets and $t\bar{t}$ processes. Both samples are taken from simulation and normalised to the unity.

The signal is normalised to the NNLO cross section [20].

The final yields in the CR and SR for the different categories are shown in Table 3.6. In the CR, the data and background yields agree well by construction and the signal yields are negligible (less than 2 expected signal events over more than 2800 predicted background events). In the SR, the yields agree in the $2j$ category but not in the $3j$ category where significantly more events are predicted than observed. The impact of the PU contamination is suspected to bias the jet categorisation and then the application of the SFs measured in the CR to the SR. This topic will be discussed in more

SF_Zbb	1.12 ± 0.05
SF_Zb(b)x	1.27 ± 0.05
SF_Zxx	1.08 ± 0.11
SF_tt	0.94 ± 0.03

Table 3.5: The background scale factors as estimated from the 2D fit.

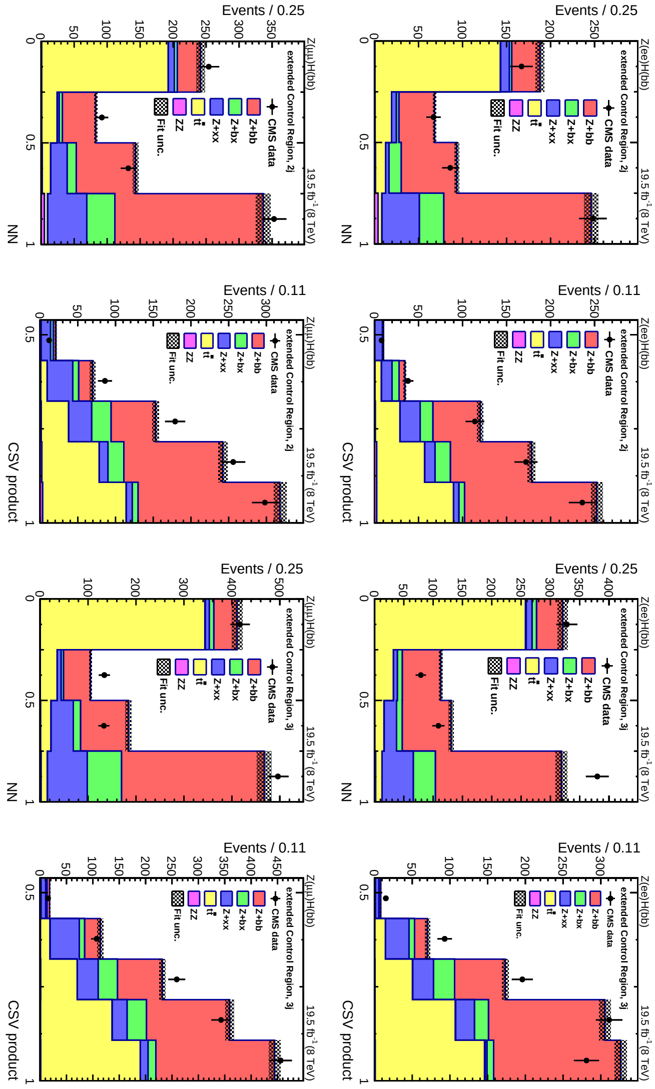


Figure 3.6: Variables used in the fit procedure for the estimation of the background normalisation scale factors with the binning used for the fit. The first two columns show the results for the $2j$ category, while the last two columns illustrate the results in the $3j$ category. The first row refers to the electron channel and the second row to the muon channel. Events are selected in the Extended Control Region. The backgrounds are normalised to the results of the fit. The uncertainty from the fit is shown as a hatched band.

Category	Data	total Bkg	ZZ	$t\bar{t}$	Z+xx	Z+bx	Z+bb	Zh_{125}
Control Region								
$2j$	1173	1178 ± 35	14 ± 1	217 ± 6	155 ± 16	111 ± 13	681 ± 27	0.60 ± 0.04
$3j$	1663	1653 ± 40	6.0 ± 0.5	381 ± 7	231 ± 19	157 ± 13	879 ± 32	0.69 ± 0.04
Signal Region								
$2j$	875	882 ± 32	45 ± 1	122 ± 4	102 ± 13	109 ± 15	504 ± 25	9.7 ± 0.2
$3j$	2056	2258 ± 51	64 ± 1	347 ± 7	382 ± 26	198 ± 17	1267 ± 40	10.8 ± 0.2
Full Region								
All	5767	5971 ± 80	129 ± 2	1067 ± 13	870 ± 38	575 ± 29	3330 ± 63	21.8 ± 0.2

Table 3.6: Data yields in the CR and SR for the $2j$ and $3j$ categories and in the full region. The yields are compared with the expectation from different processes based on simulation after the full normalisation. The 'total Bkg' numbers represent the sum of the background processes.

details in Section 3.4. In total, around 22 events are expected to be observed from the signal over almost 3000 background events.

In the following, the kinematic modelling is checked in the CR. The plots in the SR are available in Appendix A.

Figure 3.7 shows the m_{ll} and p_T^{ll} observables related to the lepton kinematics. The agreement for m_{ll} is within the expectation for prompt data, i.e. without the best alignment condition. A good agreement is observed for p_T^{ll} within the statistical uncertainties.

Figure 3.8 shows the p_T of the two selected jets. A good agreement is also observed here.

Figure 3.9 shows the b-tagging discriminant of the two selected jets. The product of these two observables is used to extract the background normalisation. The agreement is within the statistical uncertainties.

Figure 3.10 shows the E_T^{miss} and the E_T^{miss} significance. Both observables show a good agreement. The discrepancy observed in the $3j$ category (right plots) is assumed to come from a statistical fluctuation because it is not present in the $2j$ category. Nothing is visible also in the SR (see A.4).

Figure 3.11 shows the jet multiplicity in the CR and SR. These plots confirmed the observation from the yields in Table 3.6, i.e. a good agreement except in the SR for $n_j \geq 3$. The largest discrepancy is visible for $n_j = 4$ with a data ratio over the estimated background of $\sim 88\%$. In the SR, the signal is also shown overlaid for comparison with the backgrounds. In this case it is normalised to 50 times its theoretical cross section. This will be also the case for all the following plots where the signal is expected to have a non negligible contribution.

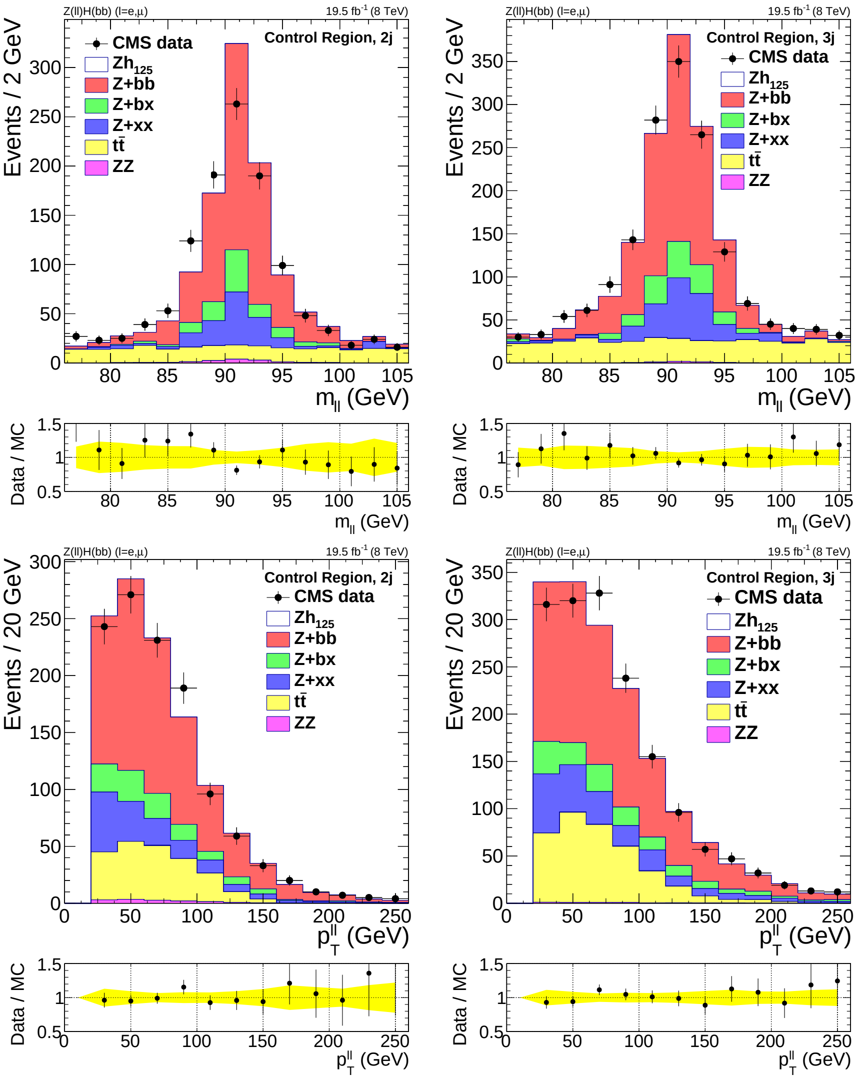


Figure 3.7: Comparisons of data and simulation in the CR of the m_{ll} (top) and the p_T^{ll} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the ratio, the yellow band represents the statistical uncertainty from simulation.

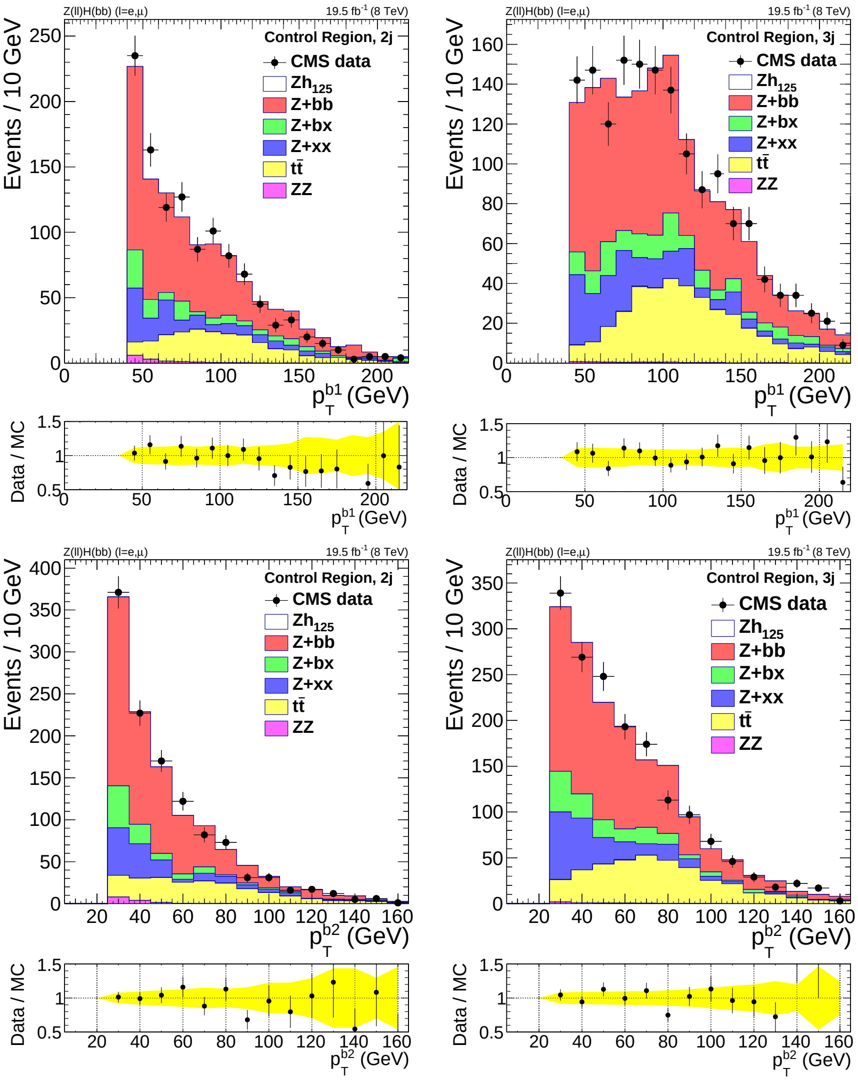


Figure 3.8: Comparisons of data and simulation in the CR of the p_T^{b1} (top) and for the p_T^{b2} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

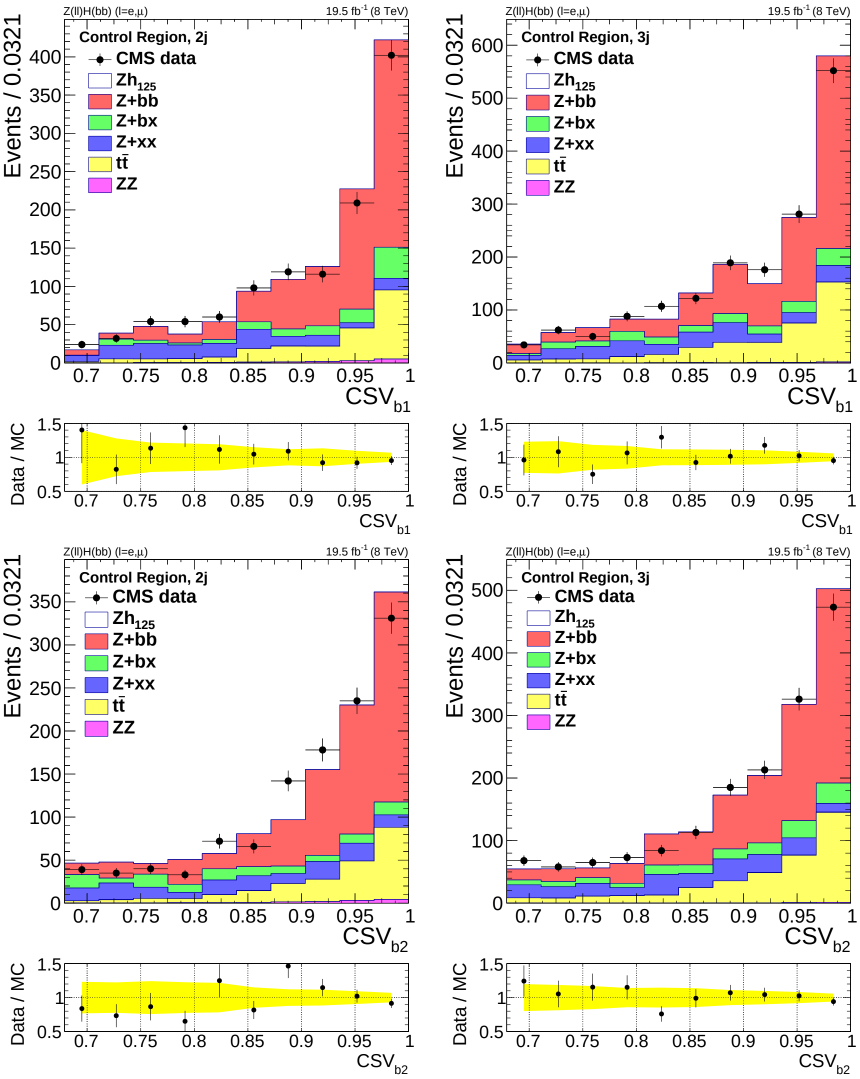


Figure 3.9: Comparisons of data and simulation in the CR of the CSV_{b1} (top) and for the CSV_{b2} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the ratio, the yellow band represents the statistical uncertainty from simulation.

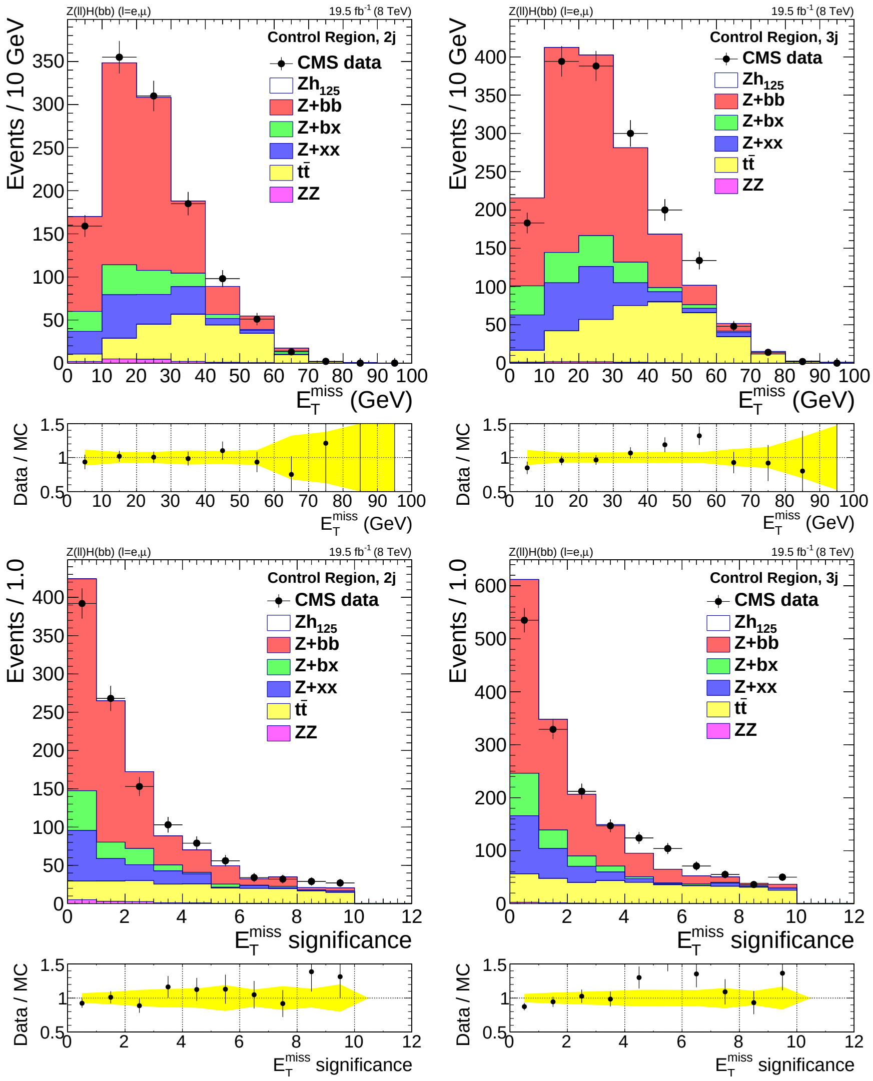


Figure 3.10: Comparisons of data and simulation in the CR of the E_T^{miss} (top) and for the E_T^{miss} significance (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the ratio, the yellow band represents the statistical uncertainty from simulation.

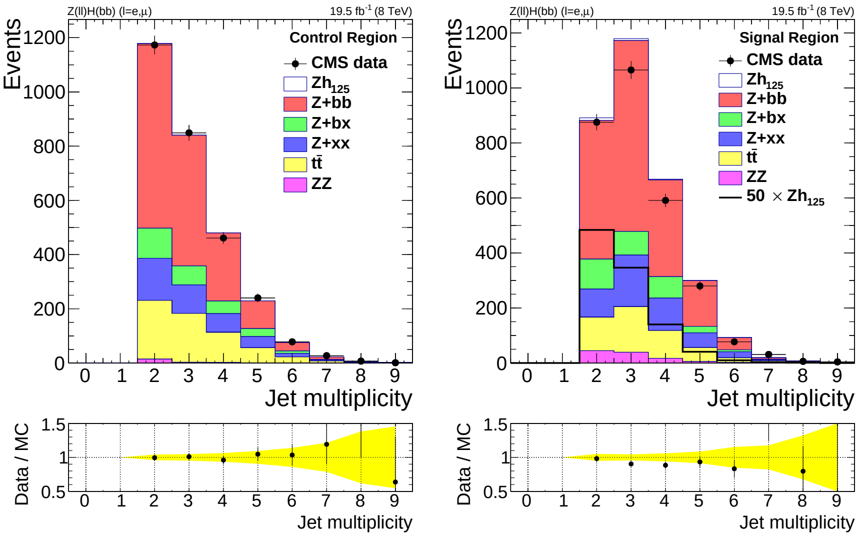


Figure 3.11: Comparisons in the CR (left) and in the SR (right) of data and simulation for the jet multiplicity observable. The bin $n_j = 2$ and the bins $n_j \geq 3$ have been selected with different cut on m_{bb} according to the SR and CR definition for the $2j$ and $3j$ categories. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

Figure 3.12 shows the m_{bb} in the $2j$ and $3j$ categories. In the $2j$ category (left plot), the data agree well with both background and signal-plus-background predictions. In the $3j$ category (right plot), a good agreement is observed in the region corresponding to the CR. However inside the SR some tensions are observed. The bin with m_{bb} between 50 and 75 GeV shows the largest discrepancy.

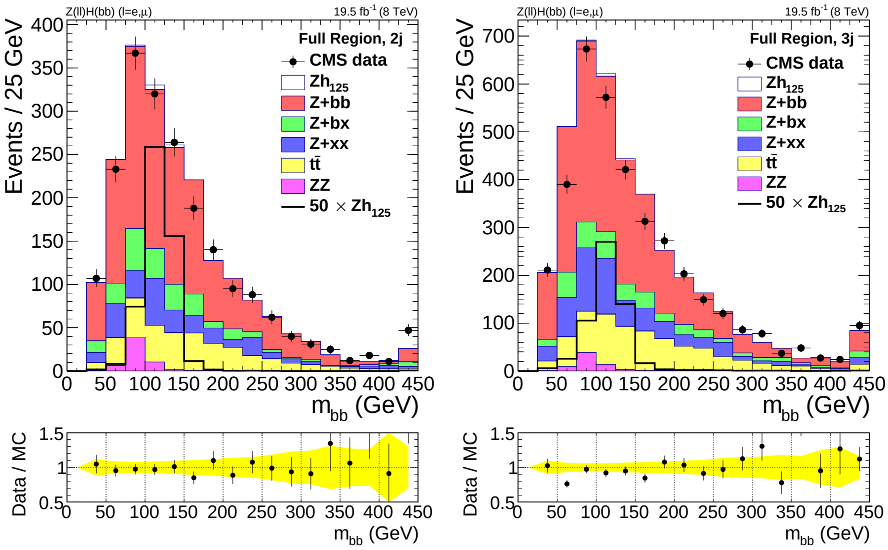


Figure 3.12: Comparisons in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the m_{bb} observable. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

Other kinematics are shown in Appendix A. They also show a good agreement in the CR and some tensions in the $3j$ category in the SR.

3.3.2 Final discriminants

To further improve the discrimination between the signal and the backgrounds, two final discriminants are built. The optimisation is performed separately for the categories $2j$ and $3j$. The construction of the final discriminants is done as described in Section 3.1 and illustrated in Figure 3.1 using different NN configurations from 1 to 4 layers made of up to 12 neurons.

Unfortunately, it was realised late in the study of the results that the implementation of the computation of m_{bbj} and $\Delta R(b, j)$, used as inputs in the $3j$ category, was wrong. This means these variables does not bring to the NNs additional information to help discriminating the different hypotheses. Due to time constrains, it was not possible to reprocess all the data and MC samples and repeat the trainings of the NNs which were using this information. This led to a reduction of discrimination power of the NNs in the $3j$ category but no bias is expected because it affects data and simulation in the same way.

Due to the limited amount of simulated events entering in the SR, it was decided to relax the b-tagging criteria for the events used in the training as defined in Table 3.4. Instead of using two b-tagged jets with the Medium WP, only one has to pass this requirement. For the second b-tagged jet, only the Loose WP is required. This resulted in a slight improvement in the discrimination power of the final discriminants. The improvement is attributed mainly to the gain in statistics ($+60\%$ for *Z+bb* and $\times 6$ for *Z+xx*) making the trainings more robust with respect to eventual over-trainings. In principle, as no b-tagging information is directly used in the training, the b-tagging requirement could have been ignored in order to gain statistics. To keep the proper fraction of Z+jets events after b-tagging, weights reflecting the b-tagging efficiencies could have been applied. This was not done as the ME weights were not computed for events passing looser b-tagging selection because it would have resulted in a large number of additional events and thus of computing time.

Another issue of the limited MC statistics is the use of the same events in the training and test of the NNs and in the analysis. It is partially resolved by the choice of a looser b-tagging requirement as the events added in the training are not used in the analysis. For Z+jets and $t\bar{t}$ events, as shown in Table 3.2, the usage of some samples only in the training and test of the NNs and some only for the analysis reduces even more possible biases. This issue could have also been partially solved by processing all the available statistics for the samples where only a fraction of the events were used.

The ME weights are shown from Figure 3.13 to Figure 3.16 in the SR. The plots for the CR are available in Appendix A (from Figure A.9 to Figure A.12).

It can be seen that, as expected, the ME weights are able to distinguish to some extent the different processes. For example in Figure 3.13, a different behaviour is clearly visible for the Z+jets events, peaking at low values, and the $t\bar{t}$ events peaking at higher values. In Figure 3.14, the tails at high values are due to events poorly compatible with the $t\bar{t}$ hypothesis. As expected the contribution from $t\bar{t}$ events is negligible in these tails. In Figure 3.15 the majority of the non-*ZZ* events get higher values compared to the *ZZ* events. Especially in the $2j$ category, the *ZZ* and Zh_{125} processes appear to be well separated. A similar conclusion can be drawn from Figure 3.16 for the Zh_{125} hypotheses. It is interesting to note that in the CR, especially for W^{-log}_{Zh3} (Figure A.12), the events clearly get higher values. This means that m_{bb} plays a determinant role in the computation of the signal ME weights. However the discrimination visible in the SR shows that the ME weights are using more information than only this observable.

The intermediate NNs are shown in Figures 3.17 and 3.18. Each time the signal and the background which is discriminated peak at high and low values, respectively. This clearly shows the discriminatory power of these NNs. The other backgrounds however behave in a less trivial way depending on the training. Overall, the agreement with the data is good except from the surplus of MC in the $3j$ category. The rightmost bin for the $NN^{3j}_{Zhvst\bar{t}}$ (top right plot in Figure 3.18) as well as the second bin of the $NN^{2j}_{ZhvsZ+jets}$ (top left plot in Figure 3.17) have been checked. In the first case this is related to the surplus of MC visible for $W^{-log}_{t\bar{t}}$ (right plot of Figure 3.14) around 23-24 and in the overflow bin. This can be due to the case where the sub-leading b-tagged jet is originating from PU because it corresponds to events with $p^{b2}_{T} < 35$ GeV. In the second case, the missing events in the simulation behave in the same way as the brown contribution in Figure 3.3. One possible explanation is that such events, biased by the PU, migrated more often in simulation than in data in the $3j$ category. This will be discussed later in Section 3.4.

The final discriminants are shown in Figure 3.19. The excess at low values is mainly due to the excess visible in the second bin of the $NN^{2j}_{ZhvsZ+jets}$.

3.3.3 Systematics

In order to consider the uncertainties on the background and signal predictions, different systematics on the yields and on the shapes of the final observables have been considered:

- **Luminosity**: the uncertainty on the luminosity at the time of the analysis was 4.4% [93]. This affects the normalisation of the signal and the *ZZ* background.
- **Signal cross section**: the uncertainty on the total signal cross section is 4% [20]. This number accounts for both the scale and the PDFs uncertainties. An extra 5% (2%) uncertainty is assigned to the NLO EWK (QCD) correction as a function of the p_T of the *Z* boson [79].
- **Lepton reconstruction and trigger efficiencies**: a flat uncertainty of 2% is assigned to the trigger and lepton reconstruction efficiency for both electrons and muons. Uncertainties between electrons and muons are assumed to be uncorrelated. This uncertainty is only applied to the signal process. For the *Z*+jets and $t\bar{t}$ backgrounds, the scale variation is assumed to be absorbed by the normalisation derived in section 3.3.1. For the *ZZ* background, this uncertainty is

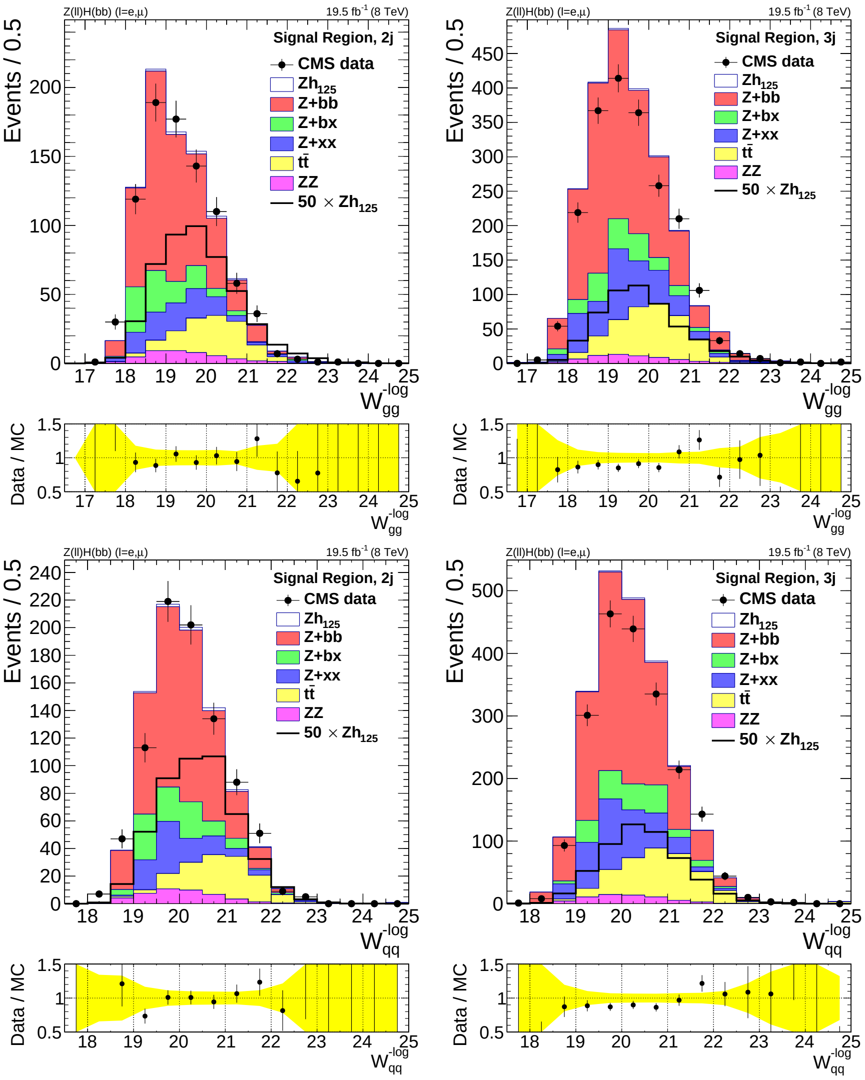


Figure 3.13: Comparisons in the SR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the Z+jets process. The top (bottom) plots represent W_{gg}^{-log} (W_{qq}^{-log}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

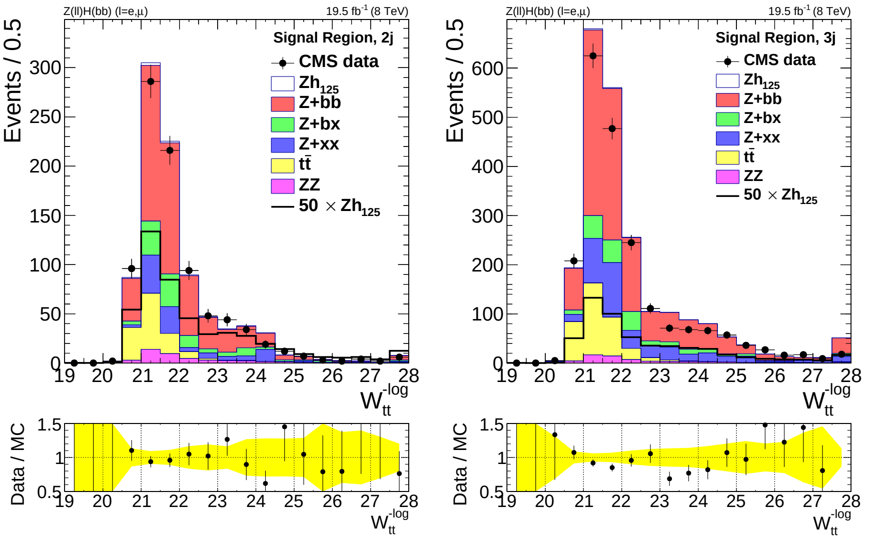


Figure 3.14: Comparisons in the SR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for $W_{t\bar{t}}^{-log}$. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

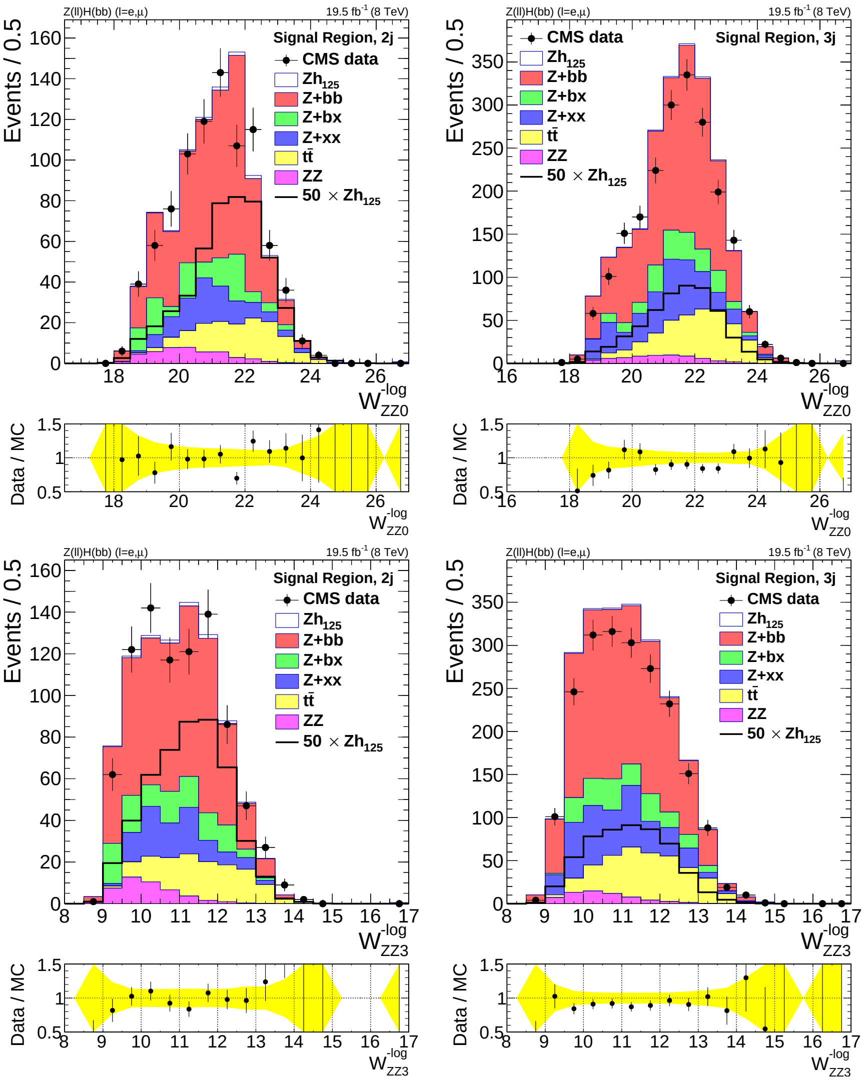


Figure 3.15: Comparisons in the SR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the ZZ process. The top (bottom) plots represent W_{ZZ0}^{-log} (W_{ZZ3}^{-log}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

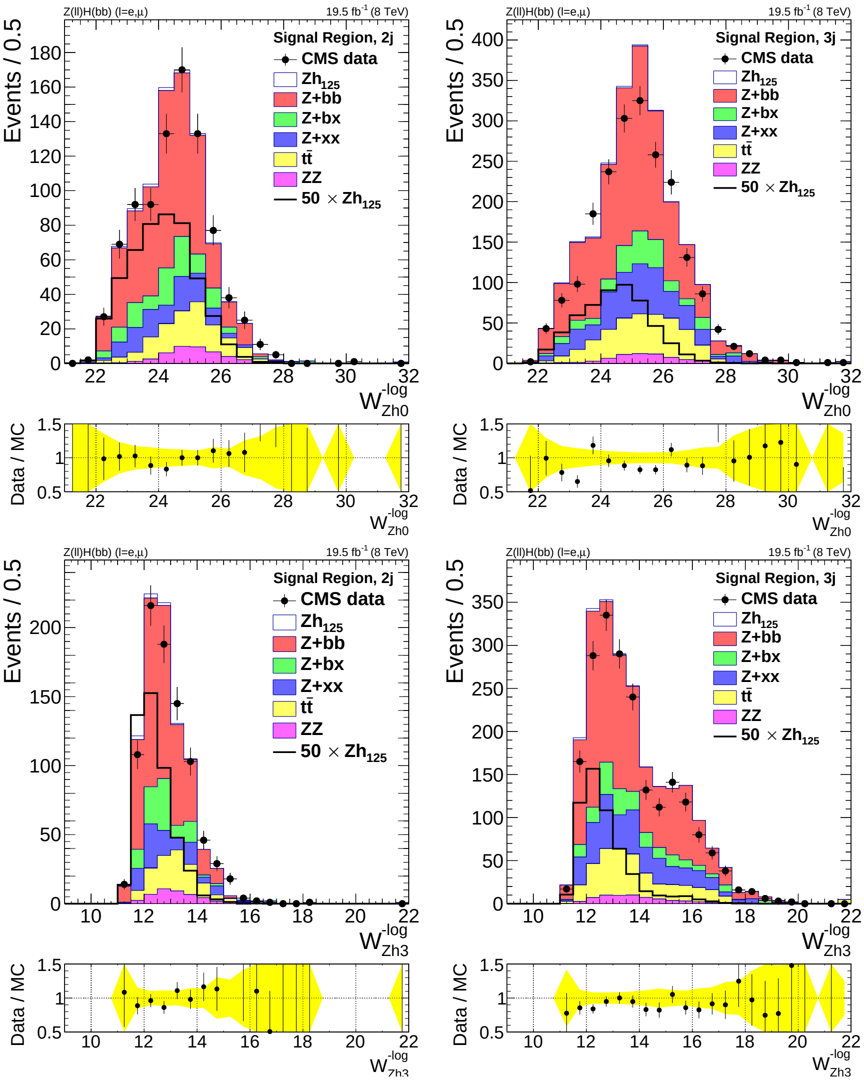


Figure 3.16: Comparisons in the SR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the signal process. The top (bottom) plots represent W_{Zh0}^{-log} (W_{Zh3}^{-log}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

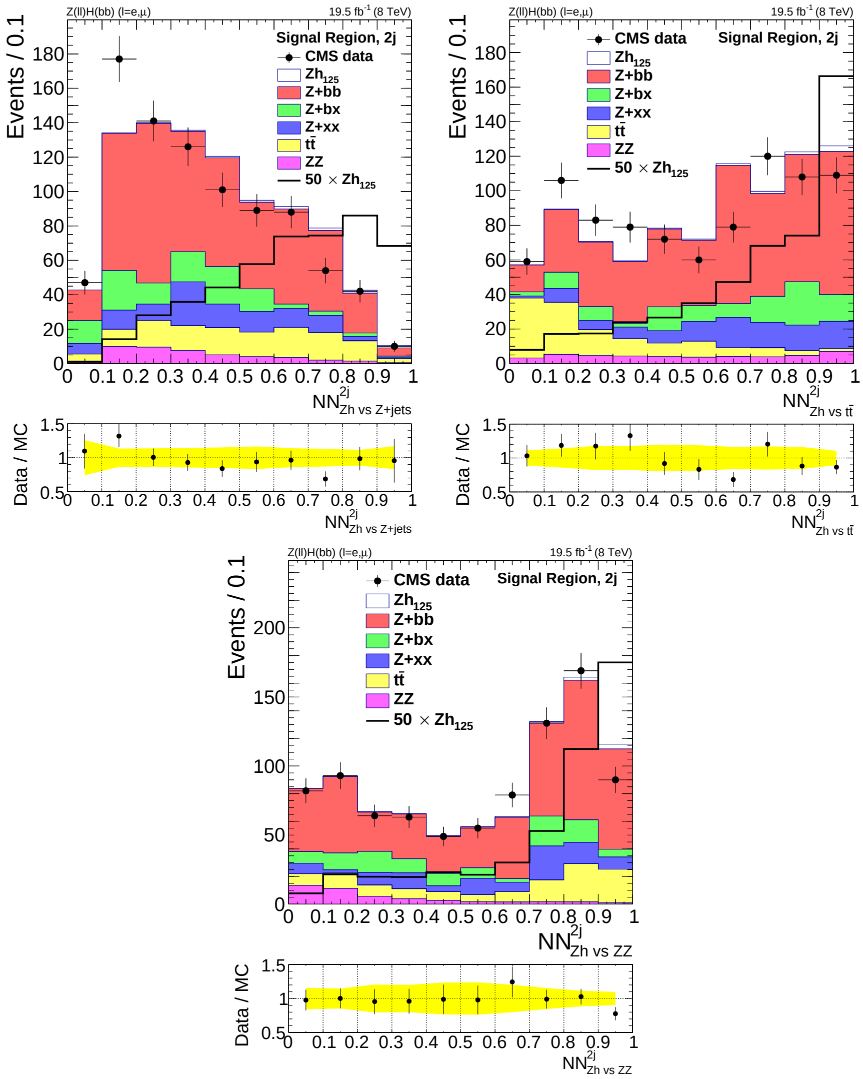


Figure 3.17: Comparisons in the $2j$ category of data and simulation for the intermediate NNs in the SR. The top right, top left and bottom plots correspond to the NN discriminating Zh_{125} versus Z+jets, $t\bar{t}$ and ZZ processes, respectively. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

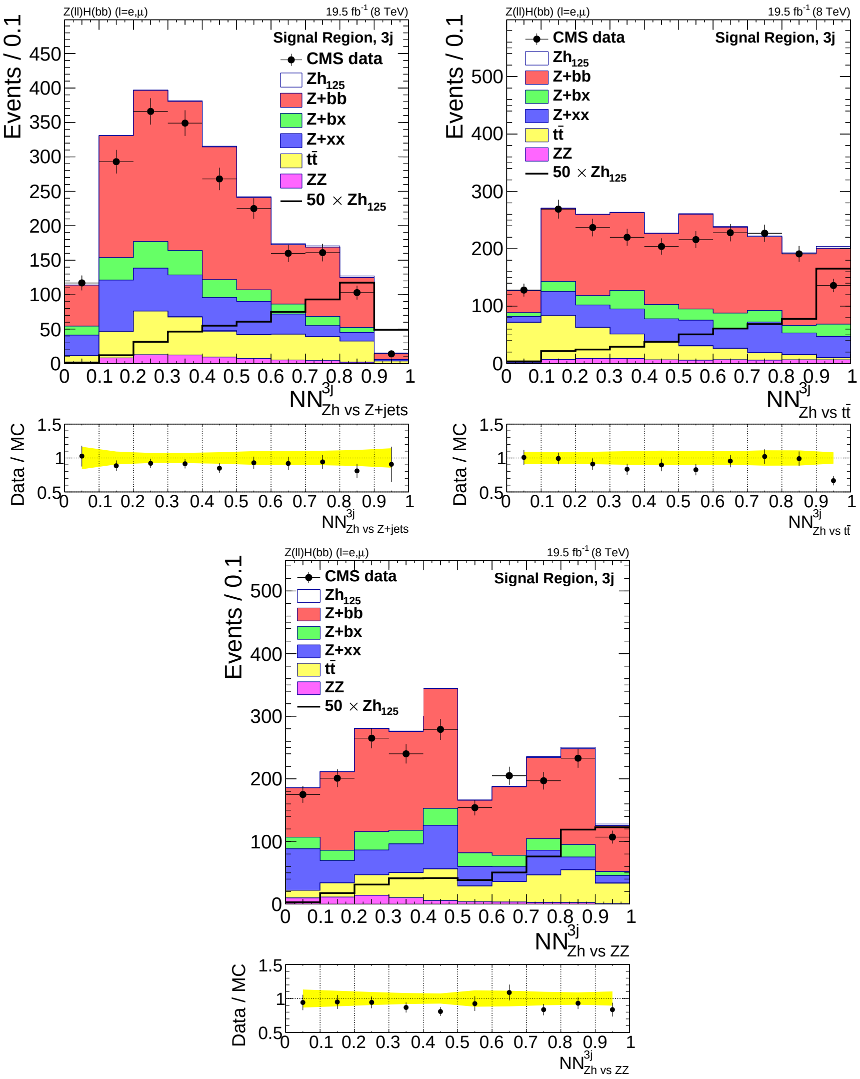


Figure 3.18: Comparisons in the $3j$ category of data and simulation for the intermediate NNs in the SR. The top right, top left and bottom plots correspond to the NN discriminating Zh_{125} versus Z+jets, $t\bar{t}$ and ZZ processes, respectively. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

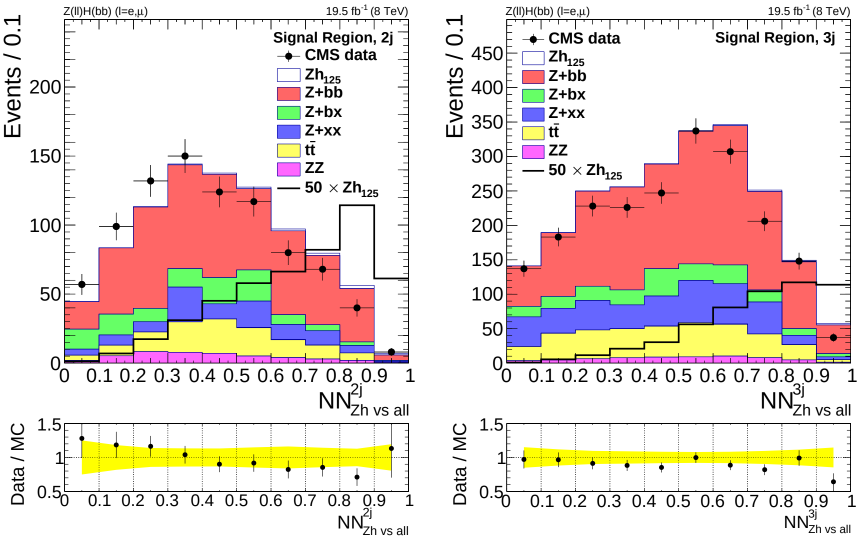


Figure 3.19: Comparisons in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the final discriminant observables in the SR. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its theoretical cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

included in the normalisation uncertainty coming from the CMS measurement used for its normalisation.

- ***b*-tagging and mistagging efficiencies**: the correction factors associated to the b-tagging selection are varied up and down according to their uncertainties as described in [94, 95]. The variations are performed separately for heavy flavour jets (b and c) and for light jets. The former affects the normalisation of the signal and ZZ process by about 5%. The reason why b and c-jets are considered together is due to the fact that measuring c-jet mistagging efficiency in data is quite challenging and therefore the modelling of the c-tagging efficiency is assumed to be close to the b-tagging efficiency because c jets and b jets have similar properties. The background fit has been repeated using the up and down variations in order to assess the effect of the b-tagging and mistagging uncertainties on the normalisation of the $t\bar{t}$ and Z+jets processes. The shape uncertainties are taken into account for all the processes.

- **Jet Energy Scale**: the JES uncertainty is evaluated by applying jet-energy corrections that describe one standard deviation variation with respect to the default corrections. The event selection and the evaluation of the ME weights are performed again after the application of the up and down variations. Rate and shape effects are considered both for signal and background processes. Similarly to the b-tagging and mistagging efficiencies uncertainties, the background fit has been repeated in order to better constrain the normalisation of the $t\bar{t}$ and Z+jets processes. An uncertainty of 5% is found for the signal normalisation.

- **Jet Energy Resolution**: this analysis has been performed without applying the default jet-energy smearing to the simulated events in order to match the measured jet-resolution in data [96]. To evaluate the jet-energy-resolution uncertainty, the event selection for the signal samples is repeated after doubling the default smearing. This uncertainty leads to a yields uncertainty of 4-6% for the different processes.

- **Background Fit**: the statistical uncertainties associated to the four scale factors extracted from the background fit have been considered. The correlation and the covariance matrices of the fit are used to obtain a set of uncorrelated systematic uncertainties according to a procedure described in [69].

- ***ZZ* normalisation**: an uncertainty of 15% is assigned to the ZZ normalisation uncertainty. This corresponds to the uncertainty from the CMS cross section measurement [92] available at that time.

- **Monte Carlo Statistic**: the limited size of the generated Monte Carlo samples represents an important source of uncertainty. To account for this effect, alternative shapes that vary exclusively the contents of one of the bins of the discriminants are introduced for each process. The considered bin is multiplied by factors representing $+/-$ one standard deviation of a Poisson distribution centred around the number of predicted events. This means that the up and down fluctuations are not symmetrical, especially when it is populated by only a small number of events. The statistical uncertainties which correspond to the different bins are included only for the most sensitive bins corresponding to a discriminant value higher than 0.5. They are assumed to be uncorrelated between themselves and between the processes.

The impact of the systematics on the final limit is shown in Table 3.7. This impact is measured by removing one systematic uncertainty at a time and recomputing the limit. The difference is interpreted as the degradation due to one systematic uncertainty. The main source of uncertainty affecting the limit is the MC statistic uncertainties, and especially the one for the *Z+bb* events. This uncertainty can be reduced by using a larger background sample. The second most important source of uncertainty is the background normalisation. This uncertainty can only be reduced by having a larger data sample which will be the case in the second run of the LHC. This will allow either to more precisely estimate the backgrounds, or to tighten the selection to reduce the background contributions and then the impact of their uncertainties on the final results.

Systematics	degradation (%)
-MC statistical unc.	15
-*Z+bb*	7
-*Z+xx*	1.8
-$t\bar{t}$	1.8
-*Z+bx*	0.9
-*ZZ*	$\ll 0.1$
-Zh_{125}	$\ll 0.1$
-Background norm.	1.8
-b-tag *b*, *c*-jets SFs	0.9
-JER	0.9
-signal cross section	0.9
-JES	$\ll 0.1$
-b-tag light-jets SFs	$\ll 0.1$
-Luminosity	$\ll 0.1$
-Lepton SFs	$\ll 0.1$

Table 3.7: Breakdown of the systematics on the final limits.

3.4 Results

The results of the search are expressed as a 95% confidence-level (CL) upper limit on $\mu = \sigma/\sigma_{SM}$. In order to compute this limit, the CLs criterion is used [97, 98]. The standard tool within the CMS collaboration, called `combine` [99] and based on `RooStats` [100], is exploited to perform this computation. The asymptotic method [101] is used here because it is faster and gives a fairly good approximation. More complex and accurate methods have been tried but have been shown to give really close results (at the % level). The results are extracted from the shape of the final discriminants, made of 20 equal bins, using as nuisance parameters the systematic uncertainties described in the previous section.

Before showing the results on the Zh_{125} process, we will present the sensitivity of the analysis to the *ZZ* process.

3.4.1 *ZZ* observation

As a test of the method the *ZZ* process can be searched for using the same approach and topology. From the yields in Table 3.6 approximatively 5 times more events are expected from the *ZZ* process than from the Zh_{125} process. Still, it remains a small process compared to the Z+jets and $t\bar{t}$ processes. It also behaves more like the Z+jets process than the Zh_{125} process does. It is therefore challenging to observe in this final state. Because this process can lead to final states with jets not necessary originating from *b* quarks, the strategy developed for the Zh_{125} process does not guarantee to be the most optimal. For example, categories of events with no *b*-tagging or with a lower *b*-tagging requirement could be considered to increase the sensitivity of the analysis. However, this is not what this test is intended for, therefore such categories are not considered. In what follows, the same procedure which is used for computing the limits on the Zh_{125} process is used and discussed for the search of the *ZZ* process.

In principle, the Zh_{125} and *ZZ* processes differ only from the mass of the *Z* and h_{125} bosons, neglecting effects from spin and parity. In order to adapt the analysis to the *ZZ* process, the SR definition is modified to match the *Z* mass as shown in Table 3.8. The yields for this region are shown in Table 3.9. The impact of the presence or absence of the Zh_{125} process as background was studied. This has no impact on the results as expected from the yields which are ten times smaller for this process than for the *ZZ* signal. This is mainly due to the cut on m_{bb} for defining the SR of this search which removes half of the events from the Zh_{125} process. All the results in this section include the Zh_{125} contribution as a background process.

Signal Region	
$n_{jets} = 2$	$n_{jets} \geq 3$
$45 < m_{bb} < 115$	$15 < m_{bb} < 115$

Table 3.8: Definition of the Signal Regions (SR) for the ZZ search analysis.

Category	Data	total Bkg	Zh_{125}	$t\bar{t}$	Z+xx	Z+bx	Z+bb	ZZ
2j	838	801 ± 30	4.4 ± 0.1	105 ± 4	111 ± 14	99 ± 13	481 ± 23	55 ± 1
3j	1642	1740 ± 45	5.7 ± 0.1	230 ± 6	320 ± 24	156 ± 15	1028 ± 35	60 ± 1

Table 3.9: Data yields in the SR for the $2j$ and $3j$ categories for the *ZZ* search analysis. The yields are compared with the expectation from different processes based on MC after the full normalisation. The 'total Bkg' represents the sum of the background processes. The Zh_{125} yields are included in the total Bkg yields. The *ZZ* process is the signal therefore its yields are not included in the total Bkg yields.

New NNs were trained to discriminate *ZZ* from Z+jets and $t\bar{t}$ processes. As only two backgrounds have to be considered, only two intermediate NNs are needed. These NNs can be seen in Figures 3.20 and 3.21 for the intermediate and final discriminants, respectively. The $3j$ category seems to be poorly sensitive to the presence of the *ZZ* signal. On the contrary: in the $2j$ category, the presence of the *ZZ* signal appears to be important in order to get a better description of the enriched signal regions ($NN \gtrsim$ 0.5).

The results are shown in Table 3.10 both for the 2012 data and for the combination of the 2011 and 2012 data. The combination of the two datasets is performed by a simultaneous estimation of the parameter of interest (e.g. μ). All systematics are taken uncorrelated between the 7 TeV and 8 TeV analyses. The final results is dominated by the 2012 dataset which is approximatively four times larger. As can be seen from the combined results, this analysis would be able to exclude the *ZZ* process from the SM at 95% CL (expected upper limit on $\mu < 1$). However as expected, this is not the case and a slight excess of 1.18 standard deviation (s.d.) is observed compared to the background-only hypothesis. This excess is lower than the expectation based on simulation but gives a value of μ compatible with 1 within the systematic uncertainties. To compute the significance of the excess and to fit the observed μ, a Profile Likelihood method and a Maximum Likelihood method have been used using the CMS `combine` tool. After this analysis was done, the *ZZ* cross section measurement from CMS was updated using the full 2012 data and a lower cross section by $\sim$ 10% was measured [102]. In conclusion this analysis shows sensitivity to the *ZZ* process and

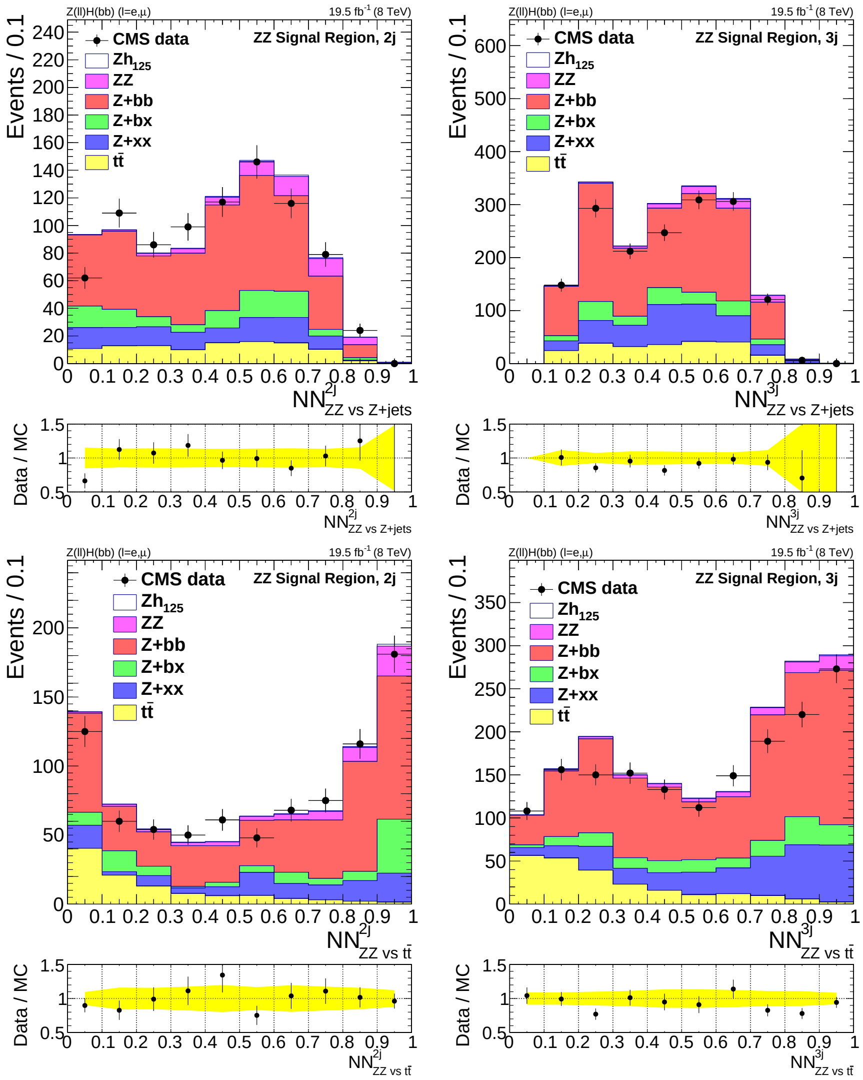


Figure 3.20: Comparisons in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the intermediate NNs in the ZZ SR. The top and bottom plots correspond to the NN discriminating Zh_{125} versus Z+jets and $t\bar{t}$, respectively. Simulation samples are normalised using the SFs shown in Table 3.5. The ZZ signal is shown in pink on top of the $t\bar{t}$ and Z+jets events. It is also included in the data/MC ratio. In the ratio, the yellow band represents the statistical uncertainty from simulation.

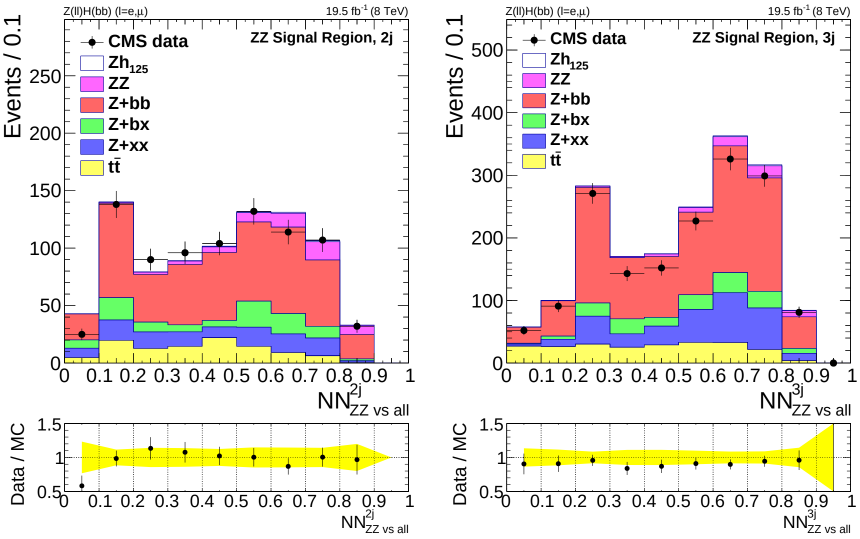


Figure 3.21: Comparisons in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the final NNs in the ZZ SR. Simulation samples are normalised using the SFs shown in Table 3.5. The ZZ signal is shown in pink on top of the $t\bar{t}$ and Z+jets events. It is also included in the data/MC ratio. In the ratio, the yellow band represents the statistical uncertainty from simulation.

gives results compatible with other CMS measurements. This established the analysis strategy and justifies its application to the Zh_{125} search.

Datasets	2012	2011+2012
Expected upper limit ($\times\mu^{+1s.d.+2s.d.}_{-1s.d.-2s.d.}$)	$1.00^{+0.46+1.05}_{-0.30-0.48}$	$0.88^{+0.39+0.90}_{-0.26-0.42}$
Expected upper limit with signal injection ($\times\mu$)	1.92	1.82
Observed upper limit ($\times\mu$)	1.66	1.43
Observed significance (s.d.)	1.23	1.18
Fit of μ	$0.73^{+0.49}_{-0.46}$	$0.57^{+0.48}_{-0.47}$

Table 3.10: Results of the *ZZ* search: upper limits on $\mu = \sigma_{Meas}/\sigma_{CMS}$ with σ_{Meas} the observed cross section for the *ZZ* signal and σ_{CMS} the expected cross section from the CMS measurement [92], observed significance of the excess and best fit of μ.

3.4.2 *Zh*$_{125}$ limits

In order to not be biased by the data, and in addition to the *ZZ* analysis, three steps had been followed.

The definition of the strategy and the optimisation were done totally blindly meaning without looking at the data in the SR. In this step the expected sensitivity to the signal was extracted in order to judge of the capability of the analysis. Only when no major improvement was expected, a partial unmasking of the data was performed meaning that the left-part of the discriminants in Figure 3.19 has been unmasked. This corresponds to a region poor in signal by definition and so this enables to cross-check the modelling of the final discriminants and the background normalisation. The tool to compute the limits has been used to refit the backgrounds in this region and check the impact of the systematics on the fit. No major effect has been observed.

Finally after the robustness of the analysis has been confirmed, the data in the SR were totally unmasked. As no significant excess was observed, this led to the final expected and observed limits in Table 3.11. With the combination of the 2011 and 2012 data, an upper limit on μ of 2.8 is expected and of 1.6 is observed. The observation is compatible within 1 s.d. with the background-only hypothesis and within 2 s.d. with the presence of the Zh_{125} signal. In conclusion this analysis is unable to conclude on the presence or absence of the Zh_{125} process.

The expected sensitivity of this analysis which focused on a first application of the full ME method is about 45% less important than the result only based on MVA techniques

Datasets	2012	2011+2012
Expected upper limit ($\times\mu^{+1s.d.+2s.d.}_{-1s.d.-2s.d.}$)	$3.4^{+1.5+3.4}_{-1.0-1.6}$	$2.8^{+1.3+2.8}_{-0.8-1.4}$
Expected upper limit with signal injection ($\times\mu$)	4.3	3.7
Observed upper limit ($\times\mu$)	1.9	1.6

Table 3.11: Results of the Zh_{125} search: upper limits on $\mu = \sigma_{Meas}/\sigma_{SM}$ with σ_{Meas} the observed cross section for the Zh_{125} signal and σ_{SM} the expected cross section from theory prediction.

presented in the published CMS analysis [66] and about 10% better with respect to the m_{bb} analysis cross-check presented in the same paper. This comparison is based on the expected upper limit on μ. To be fair, it is relevant to mention several differences between these analyses:

- A b-jet energy regression is performed in [66] in order to get a better m_{bb} resolution. This led to 20% improvement of the m_{bb} resolution. For the analysis presented here, this would imply a smaller SR window while keeping the same signal yields. This would also imply a better discriminatory power of the ME weights for the *ZZ* and Zh_{125} hypotheses. Only considering the former, a rough estimation gives at least 10% improvement in the sensitivity.
- In [66], additional *Z*+jets samples were used, reducing the statistical MC uncertainties for events with important hadronic activity. It is the main systematic uncertainty of the analysis presented here. For this reason few % on the sensitivity would be gained by the use of these additional events. Approximatively four times more events would be available for events with at least two hard partons generated. So a reduction by a factor two of the statistical uncertainty could be expected.
- Better PU treatment on jet kinematics is used in [66]. As already discussed, the presence of PU can increase the background contributions but also impact their modelling. Therefore it is clear that the analysis presented here would benefit from the same treatment. However it is not trivial to estimate the impact it would have on the sensitivity.

For what concerns the observed limit, the compatibility between the two analyses can be questioned. Nevertheless, it is not trivial although the same datasets are used. Indeed, the selections are quite different implying only a partial overlap between the two analyses. Unfortunately, a complete check of the degree of this overlap has not been done. The following conclusion can still be inferred:

- The analysis described in [66] shows a nice agreement with the expectation from the presence of a Higgs boson at 125 GeV.
- This is not the case here where the observation is only compatible within two standard deviations with the presence of a Higgs boson at 125 GeV.
- A lower observed limit than expected from simulation is consistent with the yields in Table 3.6 and the deficit of data observed in the $3j$ category.

From these results, it might be interesting to better understand the origin of the disagreement on the $3j$ category. The most plausible hypothesis relates to the fact that the contamination from PU interactions was insufficiently mitigated. This is supported by several observations. Considering the ratio of the number of events between the two categories $R_{2j/3j}$ for both 7 TeV and 8 TeV analyses, the following is found:

$$\left(\frac{R^{8TeV}_{2j/3j}}{R^{7TeV}_{2j/3j}}\right)_{MC} \sim 0.65 \qquad \left(\frac{R^{8TeV}_{2j/3j}}{R^{7TeV}_{2j/3j}}\right)_{data} \sim 0.80$$

From an experimental point of view, the only relevant difference between the 7 and 8 TeV analyses is the increase of the PU conditions. Knowing this, two remarks can be made from these numbers:

- There are significantly more events in the $3j$ category than in the $2j$ category for the 8 TeV analysis with respect to the 7 TeV analysis. This cannot be explained by the difference in the centre of mass energy. This means that events migrate from the $2j$ category to the $3j$ category. This would be the case for example if the categorisation is biased by the presence of PU jets in the selected events.
- MC and data do not evolve in the same way. The effect on the MC is more significant. Referring to the PU identification study done in CMS [103], it is clear that the PU-jets rate is not well modeled. Looking especially to the variables discriminating the PU jets from the other jets, it appears that more PU jets are predicted than observed. This is going in the direction of what is observed here.

This is not invalidating the results presented above but it is legitimate to think that an extra-systematic uncertainty could have been considered to take into account possible disagreement in the PU modelling. Such uncertainty would allow to properly consider the possible migration between the two categories. This would, in principle, lead to a better agreement between the observed and expected limits. This brings the focus on the importance and the difficulty to properly deal with PU at the LHC. This is a topic which will continue to be important for the coming years with the High Luminosity

LHC in target where on average at least 140 PU interactions are expected for each bunch crossing. To avoid possible biases from the PU modelling as well as additional systematics, the best mitigation techniques are necessary. This is an important lesson learnt from this exercise.

3.5 Outlook on the use of a ME technique

In summary, this ME technique performs well but it would benefit from several improvements before it can be more widely used as a completely generic tool by high energy physics experimentalists. In the following, the experience on the use of a ME technique in this analysis is briefly discussed.

Beside the advantages discussed in the analysis strategy several items need to be presented in order to tackle the weak points of the method used here. This analysis used the `MadWeight` program in order to compute the ME weights. This has the advantage to be generic and allows to compute ME weights for all the basic processes available in `MadGraph`. This fits the needs of this analysis perfectly. During the analysis some improvements have been made to the program by the authors allowing a faster processing of the events and a better job submission to computing clusters. Despite these improvements further developments are required. These have to focus on computational processing time, being able to deal with larger amounts of data and also on improving the interface with standard experimental tools:

- **Faster computation**: with the `MadWeight` version used (revision 258), approximatively 12 hours were needed to process one sample of approximatively 110 000 events for one hypothesis on the Louvain Tier2 computing infrastructure (with an average of 300 jobs running in parallel). It means $\mathcal{O}(2)$ min by event to compute a ME weight (here the example is based on the $t\bar{t}$ hypothesis). To complete the analysis, all the samples in Table 3.2 have to be processed and most of them have a similar amount of events. To compute for all the events all the ME hypotheses, it took approximatively one month for this analysis. Furthermore this had to be done again twice to compute the systematics related to the JES. This means almost three months of continuous processing. This clearly weighed on the analysis.

  During an analysis process, the events have to be reprocessed several times: new objects definition (e.g. new PU mitigation algorithm), new jet energy corrections, new selection, improved detector conditions (e.g. alignment), new analysis techniques (e.g. mass regression), etc. Some of these affect the selection of events or the objects in the events implying that the ME weights have to be recomputed. In this context the time to obtain new ME weights freezes the analysis in fixed state. Indeed: because any change will add time over and above the reprocessing time it is harder to make the decision to start a new reprocessing. Also, computing the ME weights will often conflict with the new reprocessing for computing resources. This is an issue because the analysis would not benefit from the latest improvements from the collaboration without an important cost in time and person power. The same problem arises for testing new techniques which can complement the ME technique. This is even more dramatic considering more inclusive analyses with more events and more processes, or considering the LHC program where significantly more data is expected in the coming years. Finally, in case there is a need to cross-check an extra control region, the ME weights are not necessary available. This could cost an extra month to process them. As examples for the analysis presented here, extra control regions might be defined by reverting the E_T^{miss} significance cut or the b-tagging requirements to control better the $t\bar{t}$ or Z+jets backgrounds.

An improvement by a factor of $\mathcal{O}(3)$ in time would be the minimum to meet the requirements for this analysis, meaning the amount of time to process the ME weights will be of the same order of the other important steps of the analysis. This might be challenging but without changes on this point, the solution would either be to reduce the number of tested hypotheses or to choose a use case with a smaller phase space.

- **Better user interface**: the process leading from an event to the processing and the use of the ME weights for this same event was quite complex. Without going into technical details, the event had to be processed and the relevant information had to be stored in specific format (LHCO). The LHCO file has to be read by the `MadWeight` program which writes a text file with the ME weights. The ME weights are then added back to the event content. This can lead to several issues, e.g. the ME weights could be associated to the wrong events.

  To increase the usability, the tool would benefit from a more simple and customisable interface. A nice improvement would be the possibility to compute the ME weight as a plug-in, directly usable in the user code.

Chapter 4

BSM Higgs bosons search

After the observation of the Higgs boson and the confirmation of the BEH mechanism, simple extensions of the SM in the Higgs sector such as 2HDMs provide interesting signatures to probe the existence of BSM physics. In this context and as discussed in the Chapter 1, the processes $H \longrightarrow Z(ll)A(bb)$ and $A \longrightarrow Z(ll)H(bb)$ are interesting. In addition a search for this process can lead to the discovery of two new particles: *H* and *A*. Such search is the subject of the analysis presented in this chapter. It is relevant to mention that the *H* boson is in general not the SM Higgs of 125 GeV but in case $m_H = 125$ GeV then *H* can be interpreted as the SM h_{125} boson. A similar analysis has been made with the $A/H \rightarrow \tau\tau$ [104]. Both results and their combination have been summarized in [105, 106]. It establishes the first experimental results for a search for these processes at the LHC. This analysis have been also pursued at 13 TeV with the 2015 data corresponding to an integrated luminosity of 2.3 fb^{-1} [107]. A special case has been studied by CMS by another analysis in the same final state: $A \longrightarrow Z(ll)h_{125}(bb)$ in the context of MSSM [108]. The analysis presented here has the objective to be more inclusive and more generic with the idea to facilitate a possible recasting of the analysis for a wide range of models.

In the following, the notation refers to *A* as the lighter particle and *H* as the heavier but they can be exchanged without any changes except in the few cases which will be highlighted.

4.1 Analysis strategy

In this analysis the masses of the two new resonances are unknown. In order to keep the analysis generic and the results usable in many models, the analysis strategy does not rely on possible model-dependent theoretical constraints or indirect experimental constraints. As will be discussed later, however, these are relevant for the choice of the benchmark model which is used for the interpretation of the results. The following physical and experimental constraints are present:

- m_b, the mass of a *b* quark is $\sim$ 5 GeV implying $m_A \gtrsim$ 10 GeV.
- m_Z, the mass of the Z boson imposes $m_H \gtrsim m_A + m_Z$.
- The fixed jet cone size, here 0.5, implies for high m_H and low m_A (e.g. for $m_H \gtrsim 5 \times m_A$) that the two *b* jets start to overlap. This implies a loss in the efficiency of reconstruction. This efficiency is almost null for $m_H \gtrsim 10 \times m_A$. To illustrate this, Figure 4.1 shows the average ΔR between the two leading jets as a function of m_{jj} and m_{lljj} for Z+jets events using the CMS simulation. The correlation is clear between the two masses and the $\Delta R(j,j)$. When the difference between the two masses increases, the $\Delta R(j,j)$ decreases. As expected for an average $\Delta R(j,j) < 1$, almost no events are predicted. For the signal a larger boost is expected implying a smaller $\Delta R(j,j)$ in average compared to the Z+jets background. Specific techniques making use of jet substructure grooming techniques can recover events in this specific topology but they are only considered for the future perspectives of this analysis.

In order to cover the entire available phase space, a scan in the 2D plane m_{bb} - m_{llbb} is performed. The scan is done with a granularity reflecting the experimental resolution on the two masses. The theoretical width of the two new particles is assumed to be smaller. A simple cut and count analysis is performed to rely as little as possible on the model description. It makes the reinterpretation of the results possible in other models. In case no significant excess is visible, limits are set on $\sigma \times BR$.

A type II 2HDM benchmark model is defined in order to reinterpret the results with *cos*$(\beta - \alpha)$ = 0.01 and *tan*(β) = 1.5 (see Section 1.3). For this benchmark model, limits are set on μ as a function of m_A and m_H. For a given pair of masses, the results are also reinterpreted in terms of *cos*$(\beta - \alpha)$ and *tan*(β).

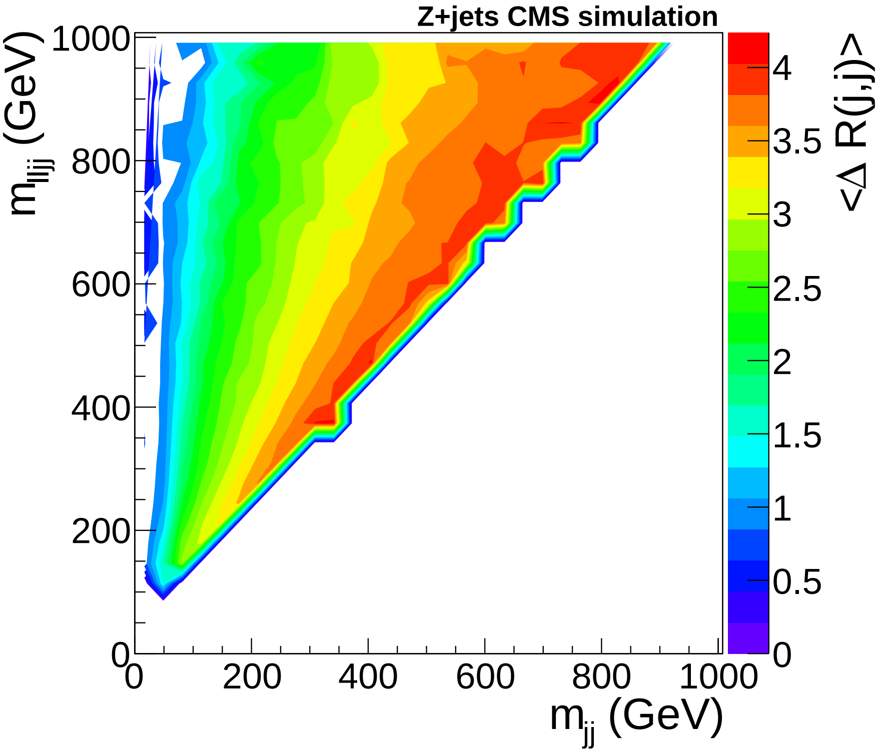


Figure 4.1: Average ΔR between the two leading jets as a function of m_{jj} and m_{lljj} for *Z*+jets events using the CMS simulation.

4.2 Setup

The analysis setup is based on the SM Higgs analysis described in Chapter 3 and inspired by the CMS *Z*+*bb* measurement [67].

4.2.1 Samples

As for the previous analysis, this study uses the dataset collected in 2012 at 8 TeV, but this time with an improved reconstruction including the best knowledge of the detector conditions during the full 2012 data taking period. This allows for the recovery of few missing data leading to an integrated luminosity of 19.8 fb^{-1}.

The simulated MC samples, used to describe the data, remain unchanged. The samples listed on Table 3.2, except the ones with a ‡, are used for this analysis too. The full statistic of each sample is used. For what concerns the *Z*+jets samples two additional exclusive samples are used. These samples target specifically events with large hadronic activity, defined as $HT = \Sigma_i \, p_T^i$ where i runs over the partons produced by the hard scattering interaction. In order to merge these two samples with the other *Z*+jets samples, the procedure described in Section 3.2.1 was extended to consider this extra dimension. As the differential cross sections were only known as a function of p_T^{ll} or as a function of HT, it was chosen to simply apply this procedure twice. This means first to merge the inclusive *Z*+jets sample and the p_T^{ll} exclusive samples and then merge this new weighted *Z*+jets sample with the exclusive HT samples. This allows to derive event weights in each 2D bin defined by the p_T^{ll} and HT bins. Other ways can be used to perform this merging but this method has the advantage of being simple and reliable. The outcome of this merging procedure can be seen in Figure 4.2. As expected, by construction for the HT variable, the merged sample perfectly follows the expectation from `MadGraph` (green curve). This is however not the case for the p_T^{ll} variable where small differences are visible. Nevertheless, these are not larger than the differences between the inclusive sample and the `MadGraph` predictions. Finally the blue curve gives a feeling of how much the statistic is improved by adding the exclusive samples. For example, in the highest HT and p_T^{ll} bins, they add around 65 times more events than what is present in the inclusive sample. More technical details about the computation of the event weights for the merging procedure and the weights are presented in Appendix B. Some other small contributions are also added in this analysis (*tW*, *WZ*). These additional samples are described in Table 4.1. In this analysis, the Zh_{125} process is now considered as background.

Samples	Cross-section in pb	Number of events
Z+jets: 200 GeV $< HT <$ 400 GeV	19.7 (LO)	3789889
Z+jets: $HT >$ 400 GeV	2.8 (LO)	1703863
tW	23.3 (CMS)	991118
WZ	36.6 (CMS)	10000283

Table 4.1: Additional MC samples with their cross sections and the number of generated events.

4.2.2 Signal production

To simulate the signal samples, only the case *H* $\longrightarrow$ *Z*(*ll*)*A*(*bb*) was considered. The model description was obtained using the `2HDMC` calculator. `MadGraph 5` was used

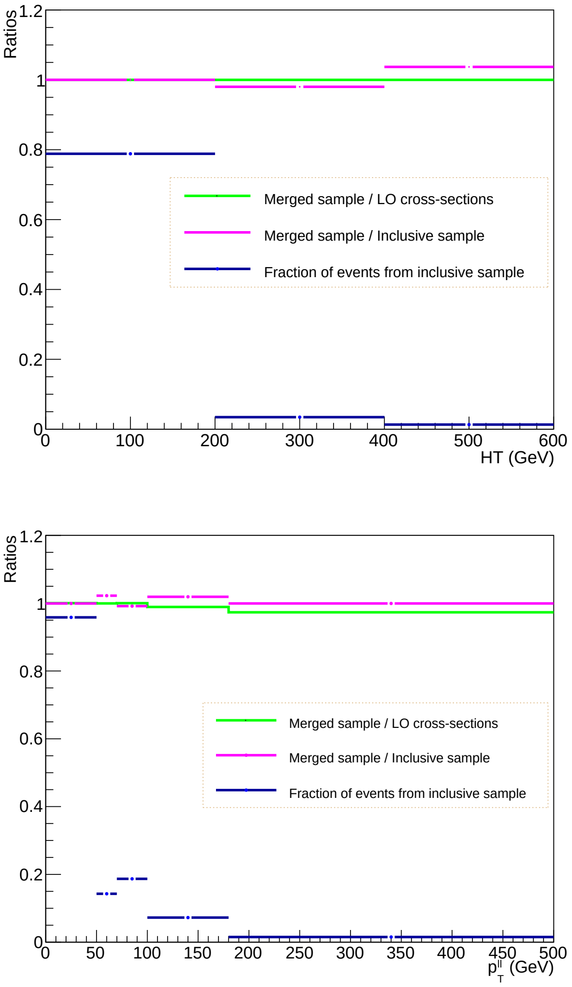


Figure 4.2: The top (bottom) plot shows the HT (p_T^{ll}) variable. Both are defined using generator particles and correspond to the variables used to define the exclusive Z+jets samples and to perform the merging of these samples. The green histogram represents the ratio of the differential cross sections of the merged Z+jets sample and the expectation from MadGraph. The pink histogram represents the ratio of the differential cross sections of the merged Z+jets sample and the inclusive Z+jets sample. The blue histogram represents the fraction of events coming from the inclusive sample and present in the merged Z+jets sample.

for the event generation. The signal cross sections have been computed at NNLO based on SUSHI 1.4.1. The branching ratios have been obtained using 2HDMC. The model parameters are listed in Table 4.2. Only m_A and m_H parameters have been varied among the various generated signal samples. Almost 400 samples of 100 000 events each were generated over the full phase space. DELPHES 3 [109] was used to simulate the detector response resulting in a faster handling of these samples. In order to check the consistency between the DELPHES and CMS simulations, 13 samples have also been simulated through the official CMS simulation based on GEANT 4. For these samples only 25 000 events were used. The parametrisation of DELPHES was done using the known lepton and b-tagging efficiencies in CMS. This has been validated on several masses to check the good agreement between the DELPHES and CMS simulations as shown in Figure 4.3. The observed agreement is within the expected ability of DELPHES to reproduce full detector simulation. Thanks to this, the DELPHES samples were used to check the dependency of the resolution on m_{bb} and m_{llbb} as a function of m_A and m_H. This is shown in Figure 4.4 as a function of m_A. The resulting resolution is pretty stable as a function of the two masses. That motivates the choice of a constant resolution value (R) in order to define the binning of the SRs. This value is 15% which corresponds to the red lines in Figure 4.4. The factor 3 in the y-axis of the plots is driven by the choice of the width of the mass windows used to define the SRs. This will be described in Section 4.2.3. It should be kept in mind that when $m_{H,A} \rightarrow 1$ TeV the width of these particles tend to be on the same level of R (see Figure 1.10). This means that this choice of a constant value of R for all mass hypotheses have some limitations in this limit.

$m_A \in [10, 1000]$ GeV
$m_H \in [100, 1000]$ GeV
$m_H > m_A + m_Z$
$m_{H^\pm} = m_H$
$m_h = 125$ GeV
$tan(\beta)$=1.5
$cos(\beta - \alpha)$=0.01
$m_{12}^2 = m_{H^\pm}^2 \cdot cos\beta sin\beta$
type II

Table 4.2: Parameters of the model used to generate the signal samples.

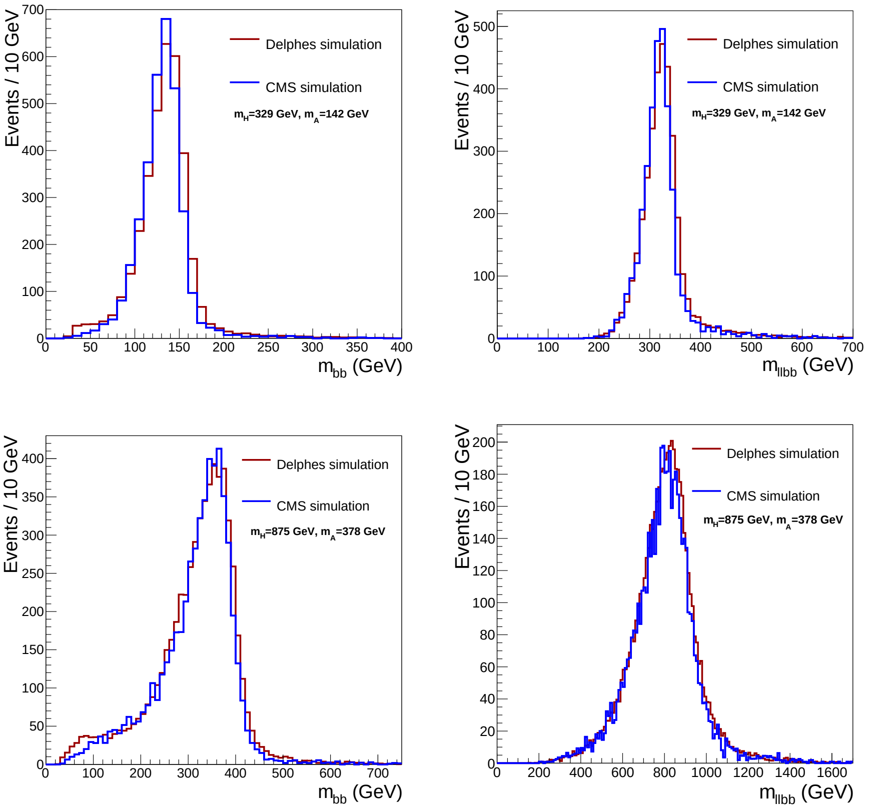


Figure 4.3: The left (right) plots shows the m_{bb} (m_{llbb}) observable. On the top plots, the signal sample is generated with $m_H = 329$ GeV and $m_A = 142$ GeV. On the bottom plots, the signal sample is generated with $m_H = 875$ GeV and $m_A = 378$ GeV. The red histograms are produced using the DELPHES simulation and the blue ones using the CMS simulation.

The numbers of expected signal events have been derived in two steps:

- The signal acceptance and efficiency map was obtained as a function of m_A and m_H using the samples simulated through DELPHES.
- This map was corrected by comparing the signal efficiencies in several points of the phase space with the CMS simulation.

These number are derived based on the selection which will be described in Section 4.2.3 and summarized in Table 4.4.

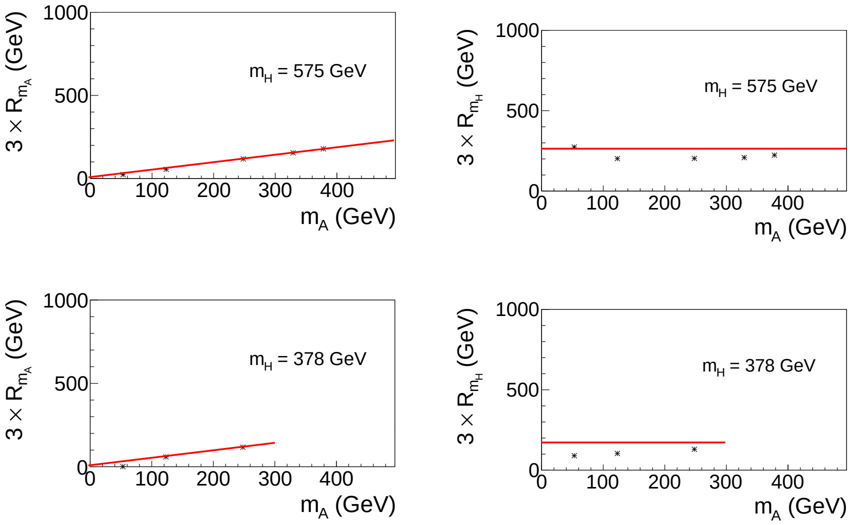


Figure 4.4: The left plots show the resolution on m_{bb} as a function of m_A for $m_H = 378$ GeV (bottom) and for $m_H = 575$ GeV (top). The right plots show the resolution on m_{llbb} as a function of m_A for $m_H = 378$ GeV (bottom) and for $m_H = 575$ GeV (top). The red lines correspond to a $3 \times 15\% \times m_{A(H)}$ in the left (right) plots.

The signal acceptance and efficiency map derived with `DELPHES` can be seen in Figure 4.5. The average value is around 10% over the full phase space. This value degrades when the difference of mass between the A and the H bosons increases due to the boost of the A boson which results in the overlap of the b jets. For low A and H masses (bottom left angle of the plot), the thresholds on the p_T of the reconstructed objects also play an important role. Indeed, the cut at 30 GeV on the jets p_T reduces the ability to see signal events in this region. The b-tagging efficiency also enters in this result as it is p_T-dependent with a maximum efficiency for intermediate p_T (50 to 200 GeV). The last factor playing a role here is the reconstructed width of the resonances which is slightly dependent on the masses. However this last effect is smaller than the other effects. For example, the b-tagging can change the signal efficiency by 40% where the effect of the width is not larger than 20%, both considering extreme cases. A few samples for the $A \longrightarrow ZH$ process were produced to check the possible differences in the acceptance and reconstruction efficiency with respect to the $H \longrightarrow ZA$ process. The results were compatible within the statistical uncertainties. This map is therefore considered valid for both processes. This result is expected be-

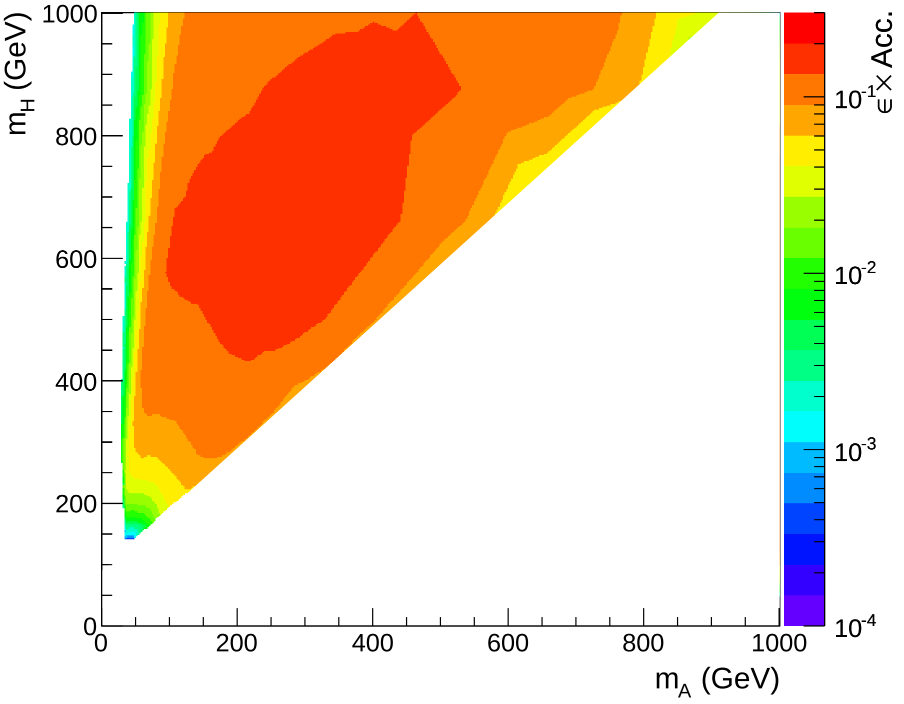


Figure 4.5: The signal acceptance and reconstruction efficiencies obtained with the `DELPHES` simulation as a function of m_A and m_H.

cause the analysis is based on the invariant masses of the two resonances and therefore the effect of the spin can be expected to be negligible. It might not have been the case if a ME method or a MVA tool was used.

Table 4.3 shows the ratios of the efficiencies between the CMS and `DELPHES` simulations with the statistical uncertainty. The ratios are close to 0.9 in most cases. These ratios are then used to derive the map of Figure 4.6 covering the full phase space. This map is used to obtain the final expected efficiencies for the signal shown in Figure 4.7 by rescaling the `DELPHES` efficiency map of Figure 4.5. This analysis is mostly efficient for signals with $m_H \in [500, 900]$ GeV and $m_A \in [100, 400]$ GeV.

m_A (GeV)	m_H=142 GeV	m_H=200 GeV	m_H=329 GeV	m_H=575 GeV	m_H=875 GeV
30-35	0.41 ± 0.24	–	0.83 ± 0.10	–	–
50	–	0.84 ± 0.05	–	–	–
70	–	–	–	0.96 ± 0.02	0.91 ± 0.04
90	–	0.80 ± 0.03	–	–	–
142	–	–	0.93 ± 0.02	–	0.93 ± 0.02
378	–	–	–	0.89 ± 0.02	0.91 ± 0.01
575	–	–	–	–	0.89 ± 0.02
761	–	–	–	–	0.85 ± 0.02

Table 4.3: The ratios $\epsilon_{CMS}/\epsilon_{DELPHES}$ of the efficiencies between the CMS and DELPHES simulations for few representative m_A and m_H mass points.

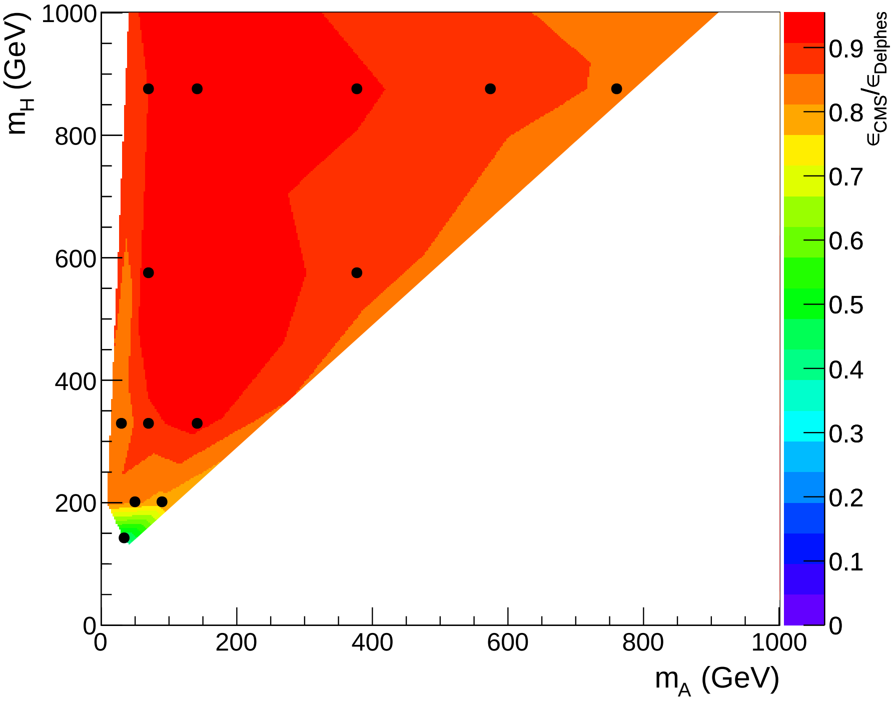


Figure 4.6: Ratio of the signal efficiencies for the CMS and DELPHES simulations as a function of m_A and m_H. The dots represent the samples used to derived this ratio map.

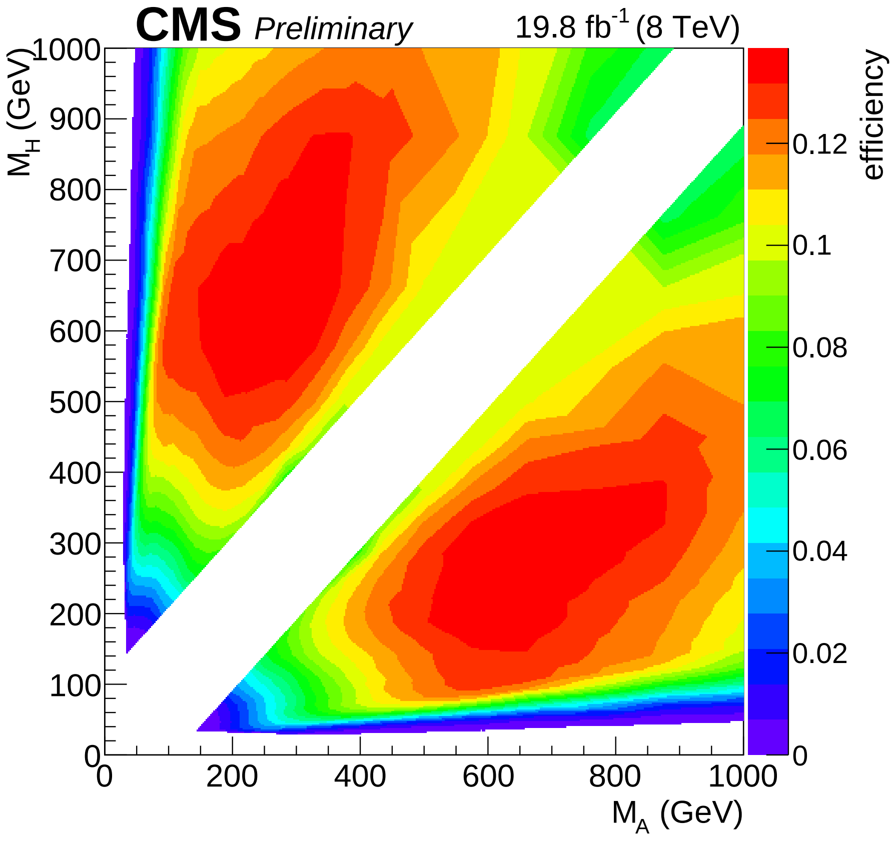


Figure 4.7: Signal efficiencies as a function of m_A and m_H including the detector acceptance, obtained with the `DELPHES` simulation and corrected to match the CMS simulation.

4.2.3 Selection

This analysis has been designed to be as generic as possible to not depend on the choice of the signal. This justifies that the cuts are chosen to be inclusive and that no specific optimisation has been performed. The list of cuts are listed in Table 4.4. However, for some cuts, it has been checked whether they are close to be optimal for different signal mass hypotheses as shown in Figure 4.8. Indeed, for the cut at 10 on the E_T^{miss} significance, the figure of merit used here ($\sqrt{B+S} - \sqrt{B}$) is close to its maximum for the three tested signals. For this plot, the signal samples are normalised to correspond to 1% of the predicted background yields. Changing the signal normalisation does not change the conclusion of this study. A more precise tuning of the cut value would lead to a more model-dependent selection and results. The signal regions are defined by cuts on m_{bb} and m_{llbb} in order to select a window containing about 75% of the signal events. In other words, with R the resolution on m_{bb} and m_{llbb}, the cuts are defined as $\pm 1.5 \times R \times m_{H,A}$.

$p_T^{\mu,e} >$ 20 GeV
$\lvert\eta_\mu\rvert < 2.4$, $\lvert\eta_e\rvert < 2.5$
76 GeV $< m_{ll} <$ 106 GeV
$p_T^{jet} >$ 30 GeV
$\lvert\eta_{jet}\rvert < 2.4$
$\Delta R(l, j) > 0.5$
$CSV_b > 0.679$
$n_b \geq 2$
E_T^{miss} significance < 10
Signal Regions
$0.775 \times m_A < m_{bb} < 1.225 \times m_A$
$0.775 \times m_H < m_{llbb} < 1.225 \times m_H$
Z+jj region
no CSV cut
$e\mu$+*bb* region
no E_T^{miss} significance cut

Table 4.4: Event selection for the objects in the inclusive and signal regions where b refers to the selected b-tagged jets, n_b is the multiplicity of selected b-tagged jets, E_T^{miss} is the missing transverse energy and $l = e, \mu$. For the definition of the *Z+jj* region and $e\mu$+*bb* region, only the differences with respect to the inclusive region are reported.

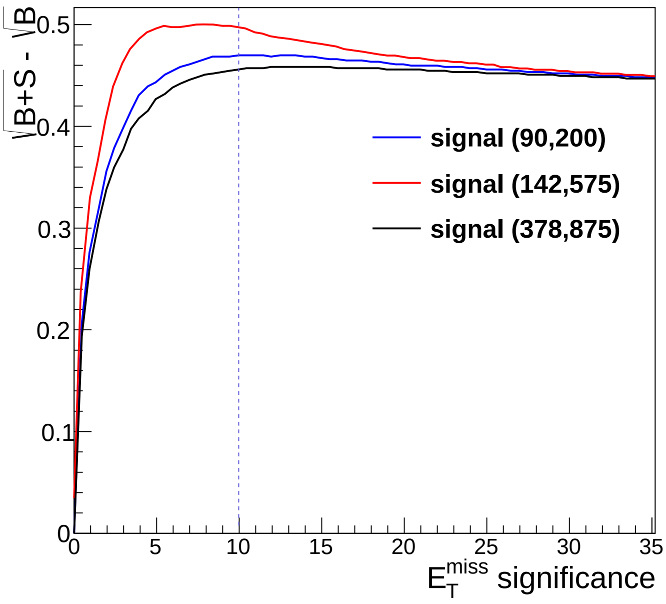


Figure 4.8: Figure of merit for three signal hypotheses depending on (m_A, m_H) as a function of the E_T^{miss} significance cut. In all three cases the signal is normalised in order to have $N_{sig} = 0.01 \times N_{bkg}$ where N_{sig} (N_{bkg}) is the number of selected events for the signal (background). The vertical line represents the cut at 10 used in this search.

One major improvement with respect to the Zh_{125} analysis described in Chapter 3 is the use of the charged hadron subtraction (CHS) technique for the jet clustering. This technique removes charged PF particles which are not associated to the PV. If they do not contribute to a vertex, they are kept. This leads to a significantly lower contamination from PU in the jets [110]. This justifies a lower cut on the leading b-tagged jet p_T and no cut on p_T^{ll} for this analysis. The choice of 30 GeV on the p_T^{jet} is still driven by the remaining PU contribution. Below 30 GeV, the PU contamination is still non negligible and further techniques would have been needed to be sure to control this contribution [103]. It has been decided to not use such techniques because no explicit gain in sensitivity was expected whereas additional corrections and systematics would have been needed. The region that would have benefited by a lower p_T^{jet} is the region with low m_A and $m_H \sim m_Z + m_A$.

Another change with respect to the Zh_{125} analysis described in Chapter 3 is the application to the simulated jets of the default jet-energy smearing in order to match the measured jet-resolution in data [96].

No advanced techniques, such as MVA or ME, have been used in this analysis. These techniques would help to improve the sensitivity of the analysis to a specific signal but, at the same time, they would lead to more model dependent results. In addition, because the number of possible signal hypotheses to be tested is large (few hundreds according to the mass of the two new particles), the use of such techniques would be complex. They are however interesting in case evidence of a signal is observed. This is particularly relevant when this analysis will be repeated with the data collected in 2016.

4.3 Analysis

4.3.1 Background estimation

In order to estimate the normalisation of the main backgrounds, no dedicated control region is defined. The main reason for this is that the signal can be anywhere. Furthermore, the choice of an inclusive selection, similar to the *Z+bb* cross section measurement, implies that any signal which can be present is expected to be negligible. Otherwise it would have led to a significant deviation from the SM expectation in the SM measurement. This argument is supported by the fact that the signal is supposed to pop up in a small window in the m_{bb} - m_{llbb} plane.

In order to estimate the *Z*+jets and $t\bar{t}$ background normalisation, the same strategy described in section 3.3.1 is followed. However, here, no NN based on ME weights is used. Instead, the m_{ll} observable is used as discriminating variable between the $t\bar{t}$ and *Z*+jets processes. The m_{ll} cut is also relaxed (60 GeV $< m_{ll} <$ 120 GeV). The projections of the fit outcome are shown in Figure 4.9. In the legend, the *Zh* entry refers to the Zh_{125} process. This applies to all the plots from this analysis. The resulting SFs are presented in Table 4.5. This table also shows two cross-checks which were made in order to validate the possible impact of the presence of a signal on the fit. These cross-checks assumed a signal with m_A = 70 GeV and m_H = 350 GeV. In the first one, the fit is performed in the CR for this signal defined as the complementary region of the SR as defined in Table 4.4. The results are compatible within the uncertainty from the fit. In the second one, the fit is performed by injecting the signal with an arbitrary cross section of 20 fb corresponding to $\sim$ 40 signal events. In this case, the SFs slightly decrease but remain compatible with the one derived without signal injection.

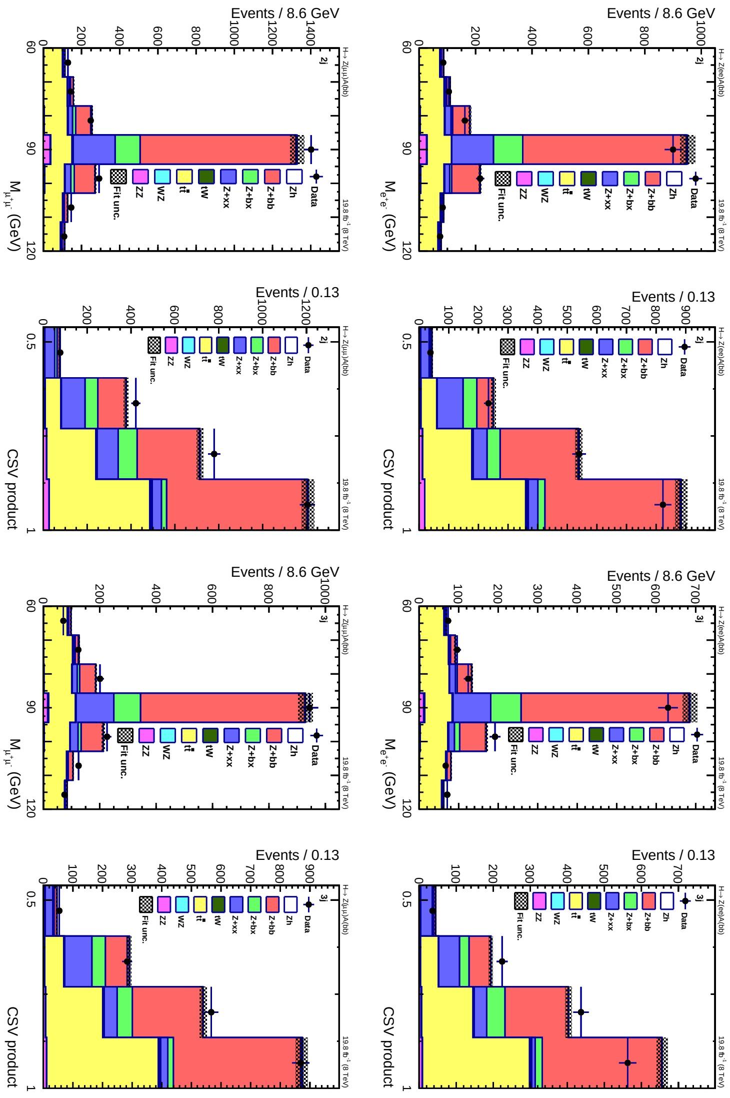


Figure 4.9: Input variables of the fit performed to estimate the background normalisation scale factors. The binning is the same as the one used for the fit. Events are selected after the inclusive selection with an enlarged m_{ll} window. The first row refers to the electron channel and the second row to the muon channel. The first two columns show the results for the $n_j = 2$ exclusive region, while the last two columns illustrate the results in the inclusive $n_j > 2$ region. The backgrounds are normalised using the results of the fit. The uncertainty from the fit is shown as a hatched band.

These checks show that the fit of the background normalisation has a small sensitivity to the presence of a signal and that defining dedicated CRs is not necessary.

SF	Inclusive region	Control region	Inclusive region with signal injected
SF_Zbb	1.16 ± 0.04	1.14 ± 0.04	1.15 ± 0.04
SF_Zb(b)x	1.27 ± 0.05	1.31 ± 0.05	1.25 ± 0.05
SF_Zxx	1.27 ± 0.10	1.31 ± 0.10	1.28 ± 0.10
SF_tt	1.04 ± 0.03	1.03 ± 0.03	1.03 ± 0.03

Table 4.5: The background scale factors as estimated from the 2D fit in the inclusive region, in a specific control region defined by the region outside the signal window where the signal corresponds to $m_A = 70$ GeV and $m_H = 350$ GeV, and in the inclusive region after injecting a signal with $m_A = 70$ GeV and $m_H = 350$ GeV with a cross section of 20 fb.

Comparing the SFs obtained in this search to the ones in Table 3.5 the results are compatible for SF_Zbb and SF_Zb(b)x. For SF_tt and SF_Zxx some differences are observed which may come from the slightly different selection and the better handling of the PU, especially for the *Zxx* contribution which was observed to be more sensitive to the presence of PU. Small contributions as *ZZ*, *WZ*, and *tW* are normalised to the best CMS measurements existing at the time of the analysis [102, 111, 112]. The Zh_{125} process is normalised to the theoretical expectation [113].

In all the plots which are shown in the following, unless information on the normalisation is specified, the background samples will be normalised as described above (including the Appendices C and D).

4.3.2 Modelling of m_{llbb}

During the analysis process, a disagreement was observed between the data and the simulation in the modelling of the m_{llbb} observable (Figure 4.10). Up to 700 GeV a clear trend is visible between the data and the simulation both from the shapes and the ratio. Below this value, the data show a higher mass than expected giving a large excess between 450 GeV and 700 GeV. Because this is one of the two main observables of this search, this triggered multiple checks.

The first check consisted of looking at regions enriched in *Z*+light jets and in $t\bar{t}$. For this purpose the *Z+jj* and the $e\mu$+*bb* regions have been defined as described in Table 4.4. For the $e\mu$+*bb* region, the data have been selected by a trigger requiring at least one muon and one electron. Events were kept if the two leptons with the highest

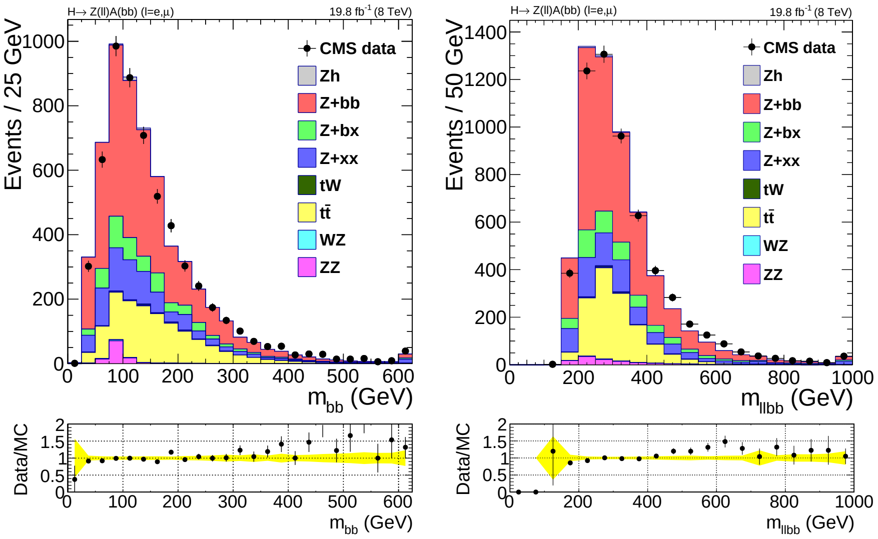


Figure 4.10: Comparisons of the data to the expectation from the simulation of the backgrounds for the two main observables of the search. The left (right) plot shows the m_{bb} (m_{llbb}) observable. Simulation samples are normalised including the SFs presented in Table 4.5. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bin in both plots includes the overflow.

p_T have a different flavour. These two leptons form the di-lepton pair have been used to compute the m_{lljj} observable. Figure 4.11 shows the m_{lljj} and m_{llbb} observables for the *Z+jj* and the $e\mu$+*bb* regions. In order to check the effect of small changes in the selection, in the $e\mu$+*bb* region, the *b*-tagging requirements and the E_T^{miss} significance cut have been varied. Also an additional $t\bar{t}$ region was defined by changing the selection in Table 4.4 by inverting the E_T^{miss} significance requirement. The conclusion of these studies was the presence of a clear discrepancy in the *Z+jj* region while no significant deviation was present in the $t\bar{t}$ enriched regions.

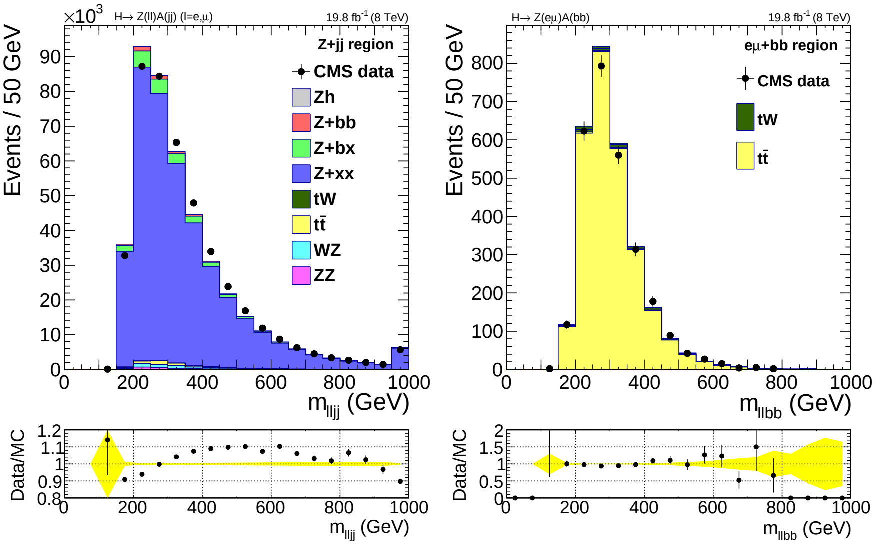


Figure 4.11: Comparisons of the data to the expectation from the simulation of the backgrounds. The left plot shows the m_{lljj} observable in the *Z+jj* region. The right plot shows the m_{llbb} observable in the $e\mu$+*bb* region where only the *tW* and $t\bar{t}$ processes are shown. All backgrounds with a real *Z* (*Z*+jets, *ZZ*, *WZ*, Zh_{125}) are negligible in this specific region. Simulation samples are normalised to their expected cross sections from theory. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bin in both plots includes the overflow.

Taking advantage of the observation of the discrepancy in the *Z+jj* region, the events were split in function of m_{lljj} without risk of being biased by the presence of a signal. This additional check allowed to verify if any obvious issue was visible either in data or in simulation. Indeed, differences are also clear in other observables but no sign of selection bias or missing MC contribution were observed. An example of such discrepancy is presented in Figure 4.12 where the $\Delta R(l, l)$ observable shows important differences between data and simulation depending on m_{lljj}.

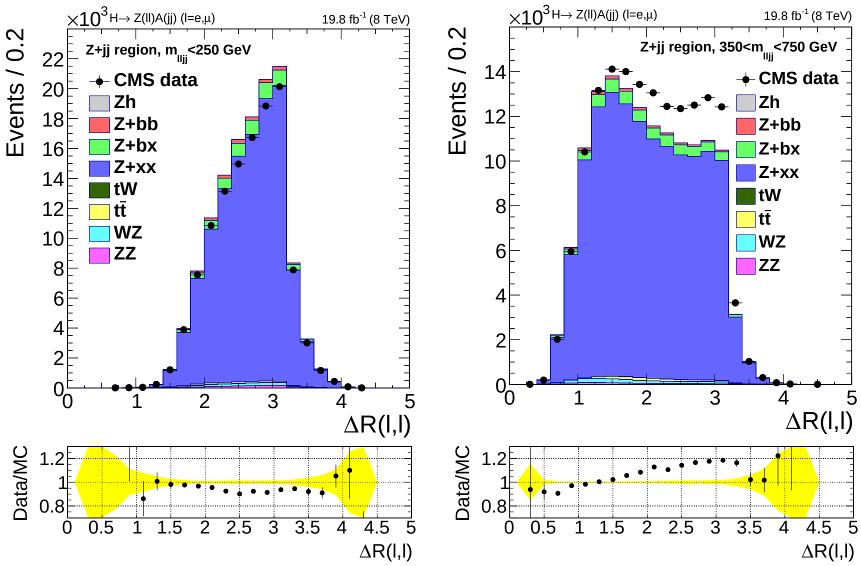


Figure 4.12: Comparisons in the *Z+jj* region of the data to the expectation from the simulation of the backgrounds showing the $\Delta R(l,l)$ for $m_{lljj} < 250$ GeV (left) and for 350 GeV $< m_{lljj} < 750$ GeV (right). Simulation samples are normalised to their expected cross sections from theory. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation.

The last check was to compare several MC generators available at the time of the analysis for the Z+jets process. Two generators were compared to `MadGraph+Pythia 6`: `MadGraph+Pythia 8` and `aMC@NLO+Pythia 8`. The first one considered possible mis-modelling in the underlying events and/or multiple parton interactions while the second one addressed possible missing higher order contributions. In the following, only the comparison with `aMC@NLO` will be discussed because no significant difference has been observed between `MadGraph+Pythia 8` and `MadGraph+Pythia 6`. For these studies, `DELPHES` was used to simulate the detector response because no official CMS simulation was available yet. The left plot of Figure 4.13 shows the comparison for m_{lljj} defined similarly to the left plot of Figure 4.11. These comparison show that the effect of NLO contributions are not negligible. They go in the direction of what is observed in data. Several checks on other observables showed that the NLO prediction behaves similarly to the data. As an example, and similarly to the right plot of Figure 4.12, the right plot of Figure 4.13 shows the $\Delta R(l,l)$ in the mass range [350,750] GeV. The discrepancy between the two generators is really close to the difference observed between the data and the LO simulation.

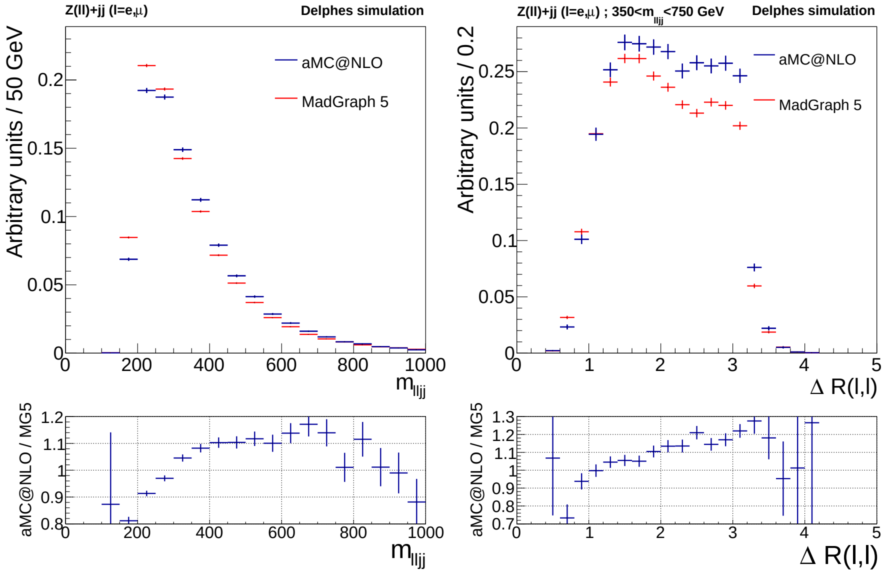


Figure 4.13: Comparison between the MadGraph and aMC@NLO simulations of the m_{lljj} observable (left) and of the $\Delta R(l,l)$ observable (right) for the Z+jets process. Both samples have been reconstructed using the DELPHES simulation. The events are required to be in the Z+*jj* region defined in Table 4.4. Samples are normalised to unity in this region to compare only shape differences. For the $\Delta R(l,l)$ observable, the events are also required to be in the mass window 350 GeV $< m_{lljj} <$ 750 GeV in order to compare with Figure 4.12.

Reweighting functions have been derived in order to test the impact on the modelling of the NLO contribution. This has been done by fitting the NLO/LO ratio for the m_{lljj} observable shown in Figure 4.14. A separated fit was performed for the Z+*xx* events and for the Z+*bx* plus Z+*bb* events. The Z+*bb* events were added to the Z+*bx* events as the statistic was too limited to make a proper fit of the Z+*bb* contribution alone. In addition the fit of the Z+*bx* contribution alone gives a function which reasonably agrees with the Z+*bb* contribution. This justifies the sum of these two contributions. The fits were realised using a polynomial of third degree which appeared to be the lowest polynomial function giving a reasonable result. The fitted functions are shown on Table 4.6. The Figure 4.15 shows the effect of the reweighting in the modelling of m_{llbb}. The improvement is clear with a ratio data over MC flatter than in Figure 4.10 and, with this adjustment, a good agreement within the statistical uncertainty is observed. A slight improvement is also visible, especially for $m_{bb} > 400$ GeV. The uncertainty resulting from this reweighting procedure is discussed later in Section 4.3.4. Two signal samples are shown for illustration purpose. The first one - in red - corresponds to ($m_A = 70$, $m_H = 329$) GeV. The second one - in purple - corresponds to ($m_A = 575$, $m_H = 875$) GeV. The same signal samples are also used later in Figures 4.17 and 4.18.

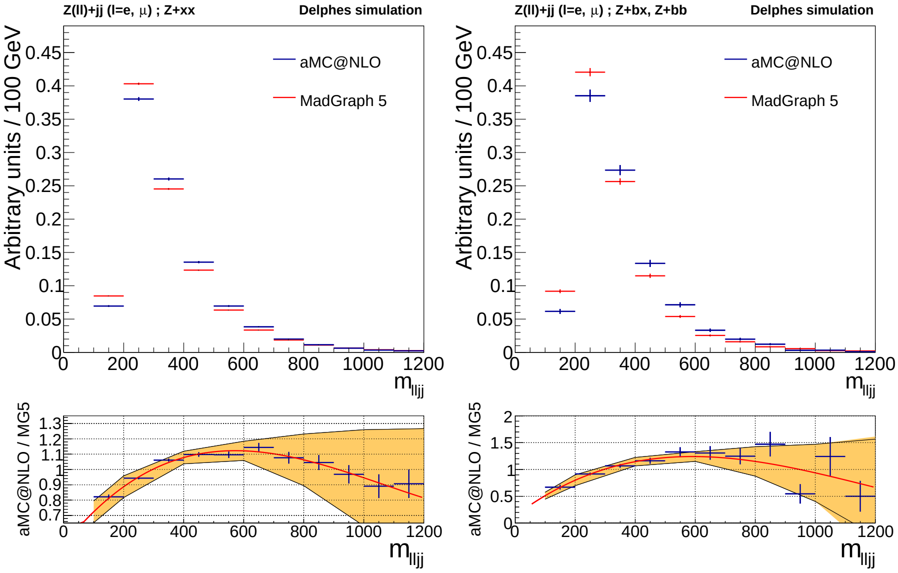


Figure 4.14: Comparison between the MadGraph and aMC@NLO simulations of the m_{lljj} observable for the Z+jets process. Both samples have been reconstructed using the DELPHES simulation. The left (right) plot corresponds to the *Z+xx* (*Z+bx* plus *Z+bb*) component of the *Z*+jets process. The events are required to be in the *Z+jj* region defined in Table 4.4 Samples are normalised to unity in this region to compare only shape differences. The ratios are fitted using a third order polynomial function (red line). The orange band in the ratio corresponds to the fits uncertainty.

This reweighting procedure has some limitations in order to properly take into account the full NLO effects. Indeed there is no guarantee that the correlations with the other observables are properly propagated. The possibility to use a 2D reweighting method in order to get a better description of the full event has been studied. However one issue in that case was the lack of available statistics to perform a smooth bin to bin reweighting. A second issue was the difficulty to find a simple function which can fit such a 2D ratio. Finally, due to these difficulties and because no other observables were found to provide a better modelling of the m_{llbb} and m_{bb} observables

Component	χ^2/d.o.f	p0	p1	p2	p3
Z+xx	7.56/10	0.51 ± 0.04	0.0025 ± 0.0002	(-29.4 ± 4.4)$\cdot 10^{-07}$	(9.2 ± 2.3)$\cdot 10^{-10}$
Z+bx & *Z+bb*	13.29/10	0.14 ± 0.13	0.0042 ± 0.0009	(-4.7 ± 1.6)$\cdot 10^{-06}$	(13.4 ± 9.0)$\cdot 10^{-10}$

Table 4.6: Third degree polynomial parameters for the fit of the NLO/LO ratio for *Z+xx* events on one side and for *Z+bx* plus *Z+bb* events on the other side.

than the simple 1D third-order polynomial reweighing functions, this possibility was discarded.

In this context the fact that the analysis relies only on two variables and does not use advanced techniques such as MVA techniques is an advantage. Indeed, in this way any remaining possible discrepancies in the correlation between other observables and m_{llbb} have less impact on the results of this search. The fact that the shapes of the m_{bb} and m_{llbb} are not used in the SRs has the advantage that the modelling of these observables is less critical but has the inconvenience that the shapes cannot be used to constrain the systematic uncertainty coming from this mis-modelling.

As only the shape differences are used to derive the reweighting functions and thanks to the small correlation between the m_{llbb} and the two variables used for the fit of the background normalisation, a negligible impact from the reweighting of the *Z*+jets sample has been observed on the backgrounds SFs. The results presented in Section 4.3.1 already contain the correction of m_{llbb}.

Once the analysis was finalized, the `aMC@NLO` *Z*+jets sample simulated in `DELPHES` was also available with the official CMS simulation and reconstruction. In order to check the consistency of the strategy described above, the left plot of Figure 4.16 shows, in the *Z+jj* region, the m_{lljj} observable replacing the `MadGraph 5` *Z*+jets sample by the `aMC@NLO` sample. The `aMC@NLO` *Z*+jets sample is normalised to match the number of events expected by the `MadGraph 5` *Z*+jets sample in the *Z+jj* region in order to see only the effect from the shape difference. The same is done for the right plot showing the $\Delta R(j,j)$ after selecting events in the mass window 350 GeV $< m_{lljj} < 750$ GeV. Both plots show a better agreement for these two observables with respect to Figures 4.11 left and 4.12 right. It is a clear indication of the importance of NLO contributions in the m_{lljj} observable and support the strategy used in this analysis. An alternative reweighting function was also derived using the CMS simulated samples but variations on the final yields in each signal regions were found to be within the systematic uncertainty from the reweighting procedure.

In all the plots which will be shown in the following sections, the *Z*+jets sample will be reweighted by the functions of Table 4.6 (this includes the Appendices C and D). A point to be raised is that this reweighting has a little effect on lower level kinematic

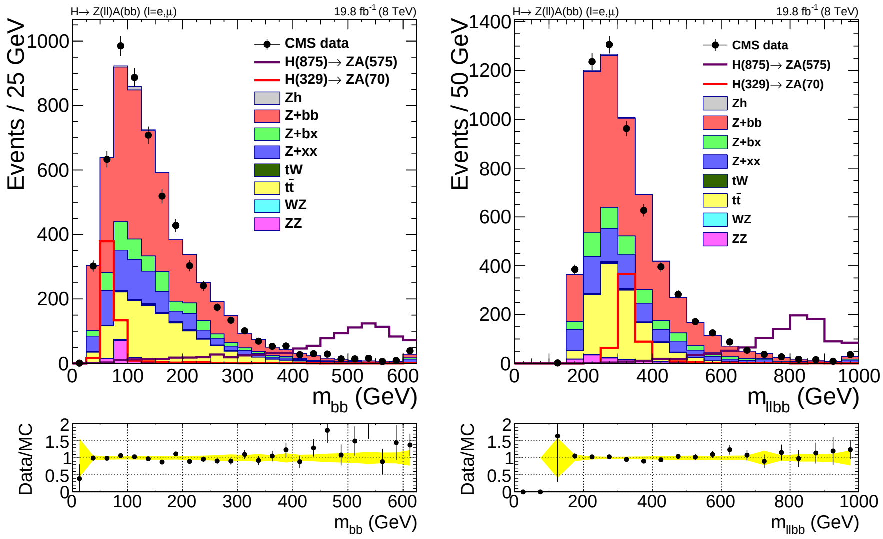


Figure 4.15: Comparisons of the data to the expectation from the simulation of the backgrounds for the two main observables of the search. The left (right) plot shows the m_{bb} (m_{llbb}) observable. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bin in both plots includes the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

observables as the p_T^{ll} for example. The effect on m_{bb} is rather small too and is mainly visible for $m_{bb} > 400$ GeV which explains, for example, why this issue was not observed in the analysis described in Chapter 3. For the same reason, no significant impact from this observation is expected in the same analysis.

4.3.3 Kinematic comparisons

The yields for the backgrounds are shown in Table 4.7 after the full normalisation and reweightings described in the previous sections. As expected from the fit of the background normalisation a good agreement in the yields between data and MC is observed for the combination of the two channels. It is possible to compare the last row of this Table with the last row of Table 3.6. In both tables the yields are close to each other for the data even though the selection is slightly different. The differences in the MC yields give an approximate idea of the increase of the background contributions due to PU contamination in the previous analysis.

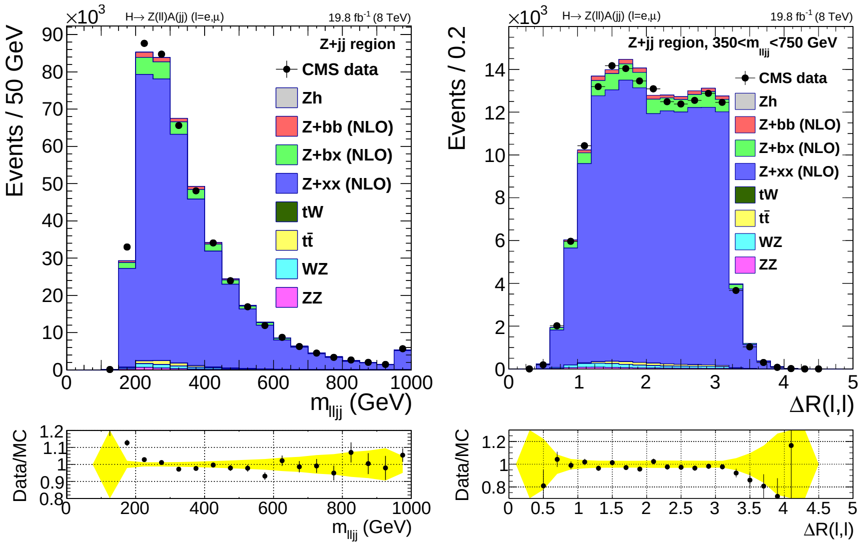


Figure 4.16: Comparisons of the data to the expectation from the simulation of the backgrounds in the *Z*+*jj* region. The left plot shows the m_{lljj} observable. The right plot shows the $\Delta R(l,l)$ in the mass window 350 GeV $< m_{lljj} <$ 750 GeV. The *Z*+jets sample in both plots was generated with `aMC@NLO`. It is normalised to match the *Z*+jets sample generated with `MadGraph 5` used in Figure 4.11 left. Other samples are normalised to their expected cross sections from theory. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bin in both plots includes the overflow.

In what follows, several relevant kinematics comparisons between data and simulation will be shown. Two signal samples are also added to the plots for illustration purpose, including the plots in Figure 4.15. They represent different types of signals but do not encompass all possible cases. The first sample - in red - corresponds to (m_A = 70, m_H = 329) GeV. It is a sample with a relatively important boost for the *A* boson due to the difference of mass with respect to the *H* boson. The second sample - in purple - corresponds to (m_A = 575, m_H = 875) GeV. For this sample the *A* is not boosted but the decay products are more energetic due to the higher masses for the two new bosons. Both samples are normalised to a cross section of 300 fb. This choice is arbitrary but allows to clearly see the shape of the two signals. The reconstructed mass of the *A* and *H* candidates can be seen in Figure 4.15.

Figures 4.17 and 4.18 show several relevant kinematic observables for this search. On the top plots of Figure 4.17, the p_T^{ll} and p_T^{bb} observables are well described. The signals show higher p_T than the backgrounds in this inclusive region. The red signal

Channel	*Z+bb*	*Z+bx*	*Z+xx*	*tW*	$t\bar{t}$	*WZ*	*ZZ*	Zh_{125}	tot. MC	Data
Electron	1284±34	212±16	317±23	9±1	547±5	4±1	48±1	9.2±0.1	2431±44	2321
Muon	1755±39	282±18	463±28	19±3	723±6	7±1	65±1	12.3±0.1	3327±51	3455
Combined	3039±51	494±24	780±37	28±4	1270±8	12±1	113±1	21.5±0.1	5758±68	5776

Table 4.7: Data and MC yields in the inclusive region. Efficiency corrections and background normalisation are applied to the simulated backgrounds. Statistical uncertainties are quoted.

peaks around 130 GeV when the purple signal peaks at higher value but with a wider distribution.

On the bottom plot of Figure 4.17, the jet multiplicity is properly modeled for up to four jets. For higher jet multiplicity, the agreement is still good within the statistical uncertainties even if a trend is observed. It is expected to get a better modelling up to 4 jets due to the way events are generated with `MadGraph`. Indeed, events are generated for *Z*+jets events with up to 4 hard partons. Extra partons are added by the parton shower (here `PYTHIA 6`). These additional partons are generally softer which is compatible with the observation of more extra jets in data. For the purple signal, the jet multiplicity peaks at 3. This might be explained by the fact that in order to create a 875 GeV resonance, the existence of an energetic initial state radiation is required.

On the top plots of Figure 4.18, the $\Delta R(l,l)$ and $\Delta R(b,b)$ observables are also well described. For the signals, they depend on the mass of the intermediate resonances and their p_T. The smaller is the ratio m/p_T the smaller is the ΔR. It should be noted that there are almost no events for $\Delta R(l,l) < 0.5$. This means some improvements are mandatory either in the lepton reconstruction when they are too close to each other or in the definition of the lepton isolation with respect to the presence of another lepton in the isolation cone. This can start to be a real problem for $m_H > 1$ TeV. This is not the case in this analysis but it is something to keep in mind for future improvements when going to more boosted topologies. The sharp drop for $\Delta R(b,b) = 0.5$ is due to the finite size of the jet cone of 0.5. For the red signal, it is clear that the signal efficiency would benefit of dedicated techniques to recover events in this region.

On the bottom plot of Figure 4.18, the $\Delta\Phi(ll,bb)$ is well modeled too and the signals peak at high values as expected. Additional kinematic observables are shown in Appendix C.

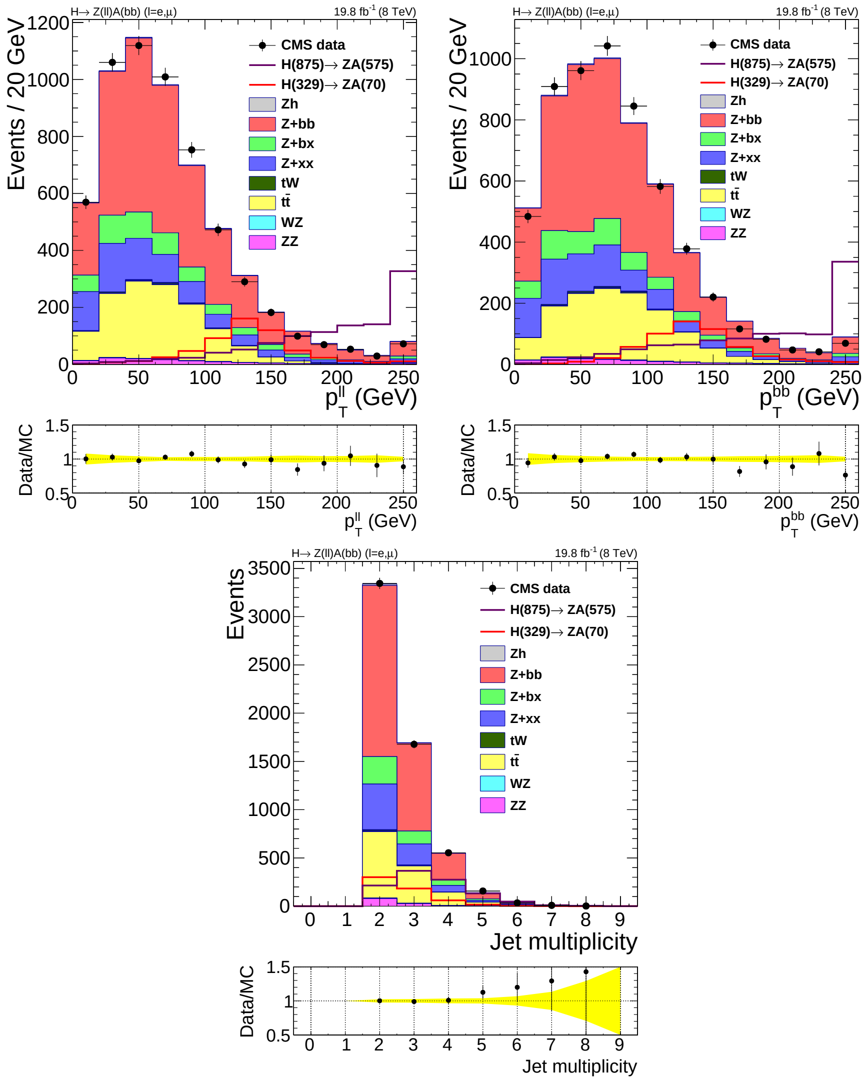


Figure 4.17: Comparisons of the data to the expectation from the simulation of the backgrounds for, from the top left to the bottom, the p_T^{ll}, p_T^{bb}, and n_j observables. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bins include the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

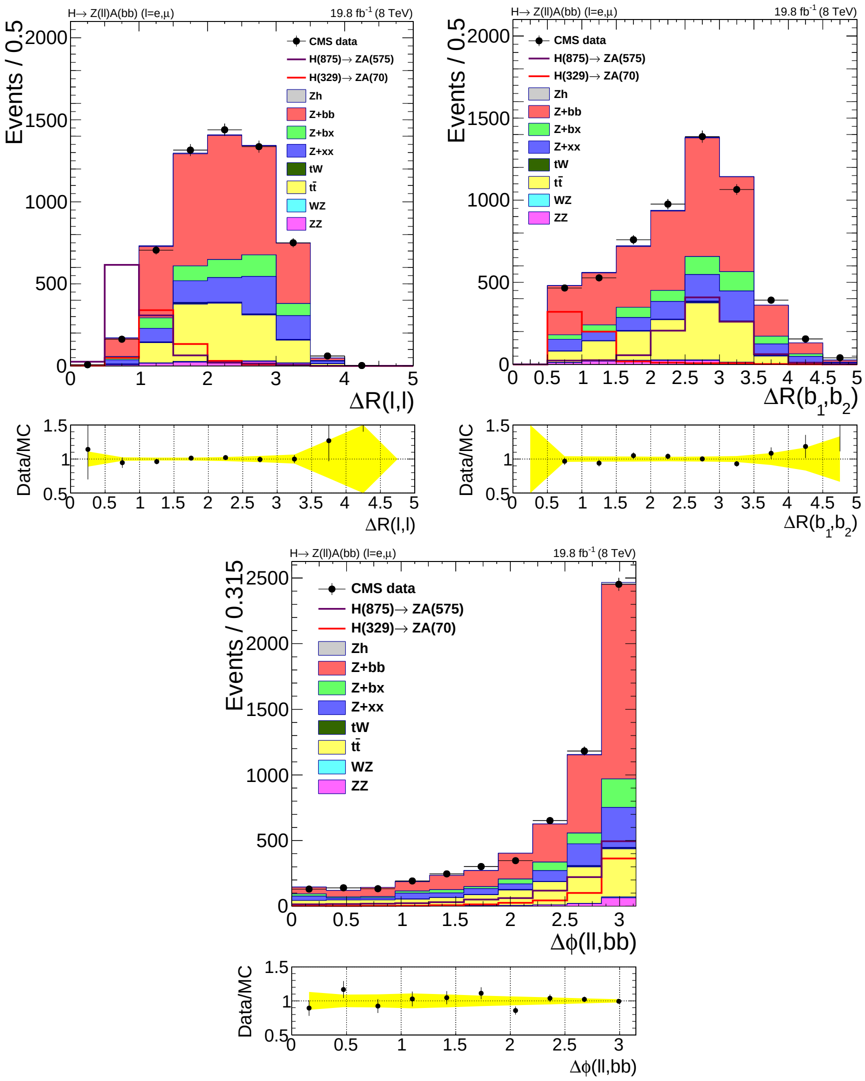


Figure 4.18: Comparisons of the data to the expectation from the simulation of the backgrounds for, from the top left to the bottom, the $\Delta R(l,l)$, $\Delta R(b,b)$ and $\Delta\Phi(ll,bb)$ observables. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bins include the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

4.3.4 Systematics

Systematic uncertainties affecting the estimated rates of signal and background processes can bias the outcome of this search. In this analysis, the impact of the systematic uncertainties on the background rates is reduced, as a fit to the data is performed in order to correct the normalisation of the main background processes. This is used to constrain the uncertainties on the b-tagging efficiency, mistagging rate and JES/JER for the $t\bar{t}$, *Z+bb*, *Z+bx*, and *Z+xx* processes. The statistical uncertainty from the fit is also taken into account. Uncertainties on the reweightings of the *Z*+jets events to correct for NLO effects on m_{llbb} are also considered. The normalisation of the di-boson, Zh_{125} and *tW* processes and the corresponding uncertainties are also included. Additional sources of systematical uncertainties are considered for the signal, such as the theoretical uncertainties on the cross section and the signal efficiencies uncertainty. Finally, uncertainties on the integrated luminosity and on the lepton reconstruction and trigger efficiencies are taken into account. A detailed list of all the systematic uncertainties considered in this search is given below:

- **Luminosity**: The uncertainty on the luminosity affects the normalisation of the signal and the di-boson and *tW* backgrounds. It is estimated to be 2.6% [114].

- **Lepton reconstruction and trigger efficiency**: A flat uncertainty of 3% is assigned to the lepton trigger plus reconstruction and isolation efficiency both for electrons and muons. This uncertainty has been derived by variating this efficiency by its known uncertainty. It is quite stable over the full phase space and 3% corresponds to the maximum of variation observed. Uncertainties between electrons and muons are assumed to be uncorrelated.

- **b-tagging and mistagging efficiency**: The correction factors associated to b-tagging and mistagging are varied up and down according to their uncertainties. The variations are performed separately for heavy flavour jets (b and c) and for light jets. The effect of the b-tagging uncertainty on the signal normalisation is estimated to be 4-6%. In order to assess the effect of the b-tagging and mistagging efficiency uncertainties on the normalisation of the different backgrounds, the background fit has been repeated using the up and down variations. For all the background processes the rate uncertainties are found to be smaller compared to the corresponding statistical uncertainty from the fit. The uncertainty is estimated to be close to 5% for the Zh_{125}, di-boson and *tW* backgrounds.

- **Jet Energy Scale**: The jet-energy-scale uncertainty is evaluated by applying jet-energy corrections that describe one standard deviation variations with respect to the default correction factors. Here also the background fit has been repeated using the up and down variations. As result, an uncertainty close to 3% has been derived.

- **Jet Energy Resolution**: To evaluate the jet-energy-resolution uncertainty, the event selection for the signal samples is repeated after removing and doubling the default smearing. Here also the background fit has been repeated using the up and down variations. For the signals, systematics are found to be of the order of 1-2%. It is interesting to note that this uncertainty is small for the signal due to the shape of the signal and the choice of the signal window boundaries. Indeed, migrations in/out the signal windows concern only the tail of the signal peak.

- **Signal cross section**: The uncertainty on the total signal cross section has been evaluated by changing the renormalisation and factorisation scale and using a different set of PDF. Running `MadGraph+SysCalc` and using the `CT10NLO` PDF set, an uncertainty of 5% was found over the entire signal mass spectrum. The factorisation scale μ_F and the renormalisation scale μ_R has been varied using the values 0.25/ 0.5/ 1.0 $\times m_H$. An uncertainty of 6% was found over the entire mass spectrum.

- **Z+jets and $t\bar{t}$ background normalisations**: The uncertainties on the background scale factors resulting from the fit have been considered for *Z*+jets and $t\bar{t}$ processes. A set of uncorrelated uncertainties are derived based on [69].

- **Di-boson and *tW* background normalisations**: An uncertainty of 11% is assigned to the *ZZ* normalisation. This value comes from the propagation of the uncertainties from the CMS cross section measurement of the *ZZ* process [102]. The uncertainty for the *WZ* sample is found to be smaller (6%) and are taken from the cross section measured by CMS as well [111]. The uncertainty for the *tW* process is taken from the CMS-measured cross section [112] and found to be of the order of 23%. These backgrounds are not expected to play a relevant role in the final results, given the small fraction of events passing the selection.

- ***Zh*$_{125}$ normalisation**: An uncertainty for the *Zh*$_{125}$ process is taken from the theoretical predictions [113]. It is found to be of the order of 7%. Also this uncertainty is expected to have a small impact on the final results, given the small fraction of events from this process in the total number of expected events.

- **Systematics on the Z+jets modelling**: A systematic uncertainty has been associated to the reweighting method described in section 4.3.2. For this the values of the four parameters of each of the two fits together with their uncertainties and the covariance matrix were used. The systematic computation consisted in varying the fitted function within the parameters uncertainties. But, given that the errors on the fit parameters are correlated, a decorrelation is performed beforehand via the diagonalisation of the covariance matrix. This operation provides the matrix for the change of the basis of the fitted parameters. A new vector of parameters (*p'*) is consequently defined on this basis. New values of *p'* are generated according to a Gaussian distribution centered around *p'* with a σ corresponding to the uncertainty on *p'*. A set of curves for each of these parameter sets is obtained, with the normalisation kept fixed, so that only the shape is affected. The curves corresponding the $\pm 1\,\sigma$ variation of the parameters are shown in Figure 4.19 for the *Z+xx* events on one side and for the *Z+bx* plus *Z+bb* events on another side. The final effect on the yields is checked in different bins of m_{lljj}. This varies up to 10 - 15% for $m_{lljj} < 600$ GeV, while it goes up to 30 - 55% for large mass values (around 1 TeV), as can be seen in the plots of Figure 4.19.

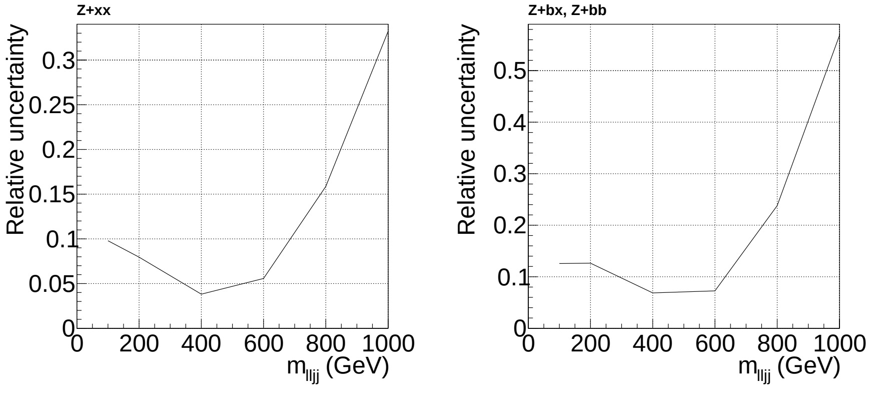


Figure 4.19: Relative uncertainty from the fit, shown in Figure 4.14, of the ratio of `aMC@NLO` and `Madgraph 5` as a function of m_{lljj}. On the left (right), the plot corresponds to the *Z+xx* (*Z+bb* and *Z+bx*) process.

- **Systematics on the signal efficiency**: The systematic uncertainty on the signal yields from the signal efficiency and acceptance in the m_H-m_A plane can be seen in Figure 4.20. The uncertainties for the 13 reference points match the statistical uncertainties from the samples simulated using the CMS simulation. For the others points where the efficiencies have been extrapolated or interpolated, the uncertainty has been smoothed. Outside the region where no CMS simulation reference points are available, e.g for high masses, a flat 10% has been assigned, which reflects the expected limits of the `DELPHES` parametric simulation in resembling CMS performances. In the lower corner of the triangle the uncertainty derived varies up to 50%, which is a reasonable value if compared with the expected CMS vs DELPHES ratio in this specific m_A-m_H mass region.

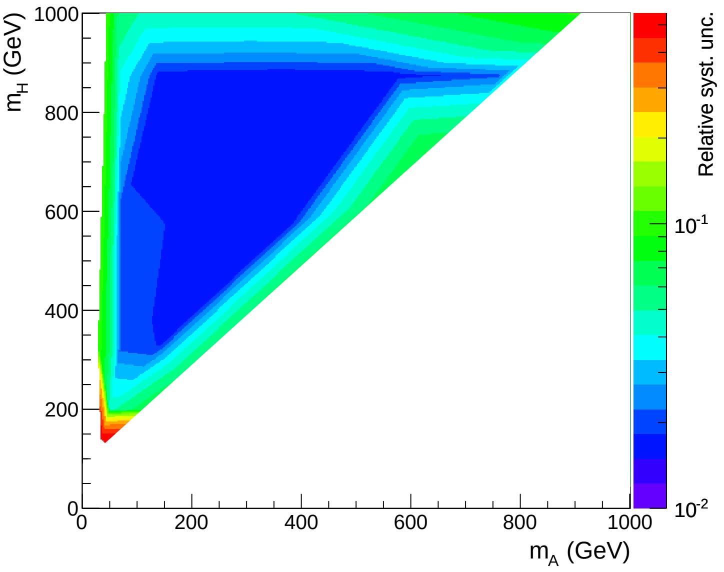


Figure 4.20: Visualisation of the systematic uncertainty on the signal efficiency plus acceptance in the m_A-m_H plane.

4.4 Results

The same tools which are described in Section 3.4 were used in order to quantify the compatibility of the data with the expectation from the backgrounds and express it as a limit on a possible signal. A scan is performed in the m_{bb}-m_{llbb} plane by sliding the centre of the signal windows by $R \times m_{A,H}$ in both directions in order to ensure a fine granularity which allows to not miss the possible presence of a signal. The observed p-value is shown in Figure 4.21 as a function of the two masses. The most significant excesses are for $(m_A = 575, m_H = 662)$ GeV and $(m_A = 93, m_H = 256)$ GeV with a local (global) significance of 2.9 (1.9) and 2.6 (1.5), respectively. These excesses have been studied and the conclusion was that both can be compatible with a

Source	Uncertainty [%]
Luminosity	2.6
Lepton ID/Isolation/Trigger	3
Jet ES/resolution	1-3
B-tagging and mistagging efficiency	4-6
Bkg. normalization (*ZZ*)	11
Bkg. norm. (*Z*+jets and $t\bar{t}$)	< 8
Bkg. norm. (*tW*, *WZ* and Zh_{125})	6-23
Z+jets bkg. modelling	4-55
Signal efficiency extrapolation	3-50
Signal modelling (PDF, scale)	5-6

Table 4.8: Summary of systematic uncertainties on the yields of the signal and background processes.

fluctuation of the background. Details on the first excess can be found in Appendix D. The second excess raised more interest because this excess is also compatible with the presence of a signal, considering the benchmark model used in the following for the reinterpretation of the results in the context of type II 2HDM. The study of this excess is discussed in the following section.

4.4.1 Interesting excess

In the scan of the m_{bb}-m_{llbb} plane showed in Figure 4.21, an excess with a local significance of 2.6 is observed for the bin centred at m_{bb} = 93 GeV and m_{llbb} = 286 GeV. This bin is defined by 72 GeV $< m_{bb} <$ 114 GeV and 222 GeV $< m_{llbb} <$ 350 GeV. Some studies have been performed to check the data in this bin and to verify the consistency of the excess both with background-only and signal-plus-background hypotheses. To do this a signal sample was generated with m_A and m_H chosen to fit the excess. This corresponds to m_A $=$ 104 GeV and m_H $=$ 270 GeV. Figure 4.22 shows the m_{bb} (left) and the m_{llbb} (right) observables after applying the cut 222 GeV$< m_{llbb} <$ 350 GeV in the first case and the cut 72 GeV $< m_{bb} <$ 114 GeV in the second case. The top plots show the data comparisons to the simulated backgrounds. The bottom plots show the data with the background yields subtracted bin by bin. In the latter case, the distributions are fitted by a simple Gaussian function (red lines). For m_{bb}, the best fit gives a reconstructed mass of 99 GeV. For the m_{llbb}, the best fit gives a reconstructed mass of 263 GeV. Taking into account the small shift between the reconstructed and generated mass observed (2 - 5%), this leads to the

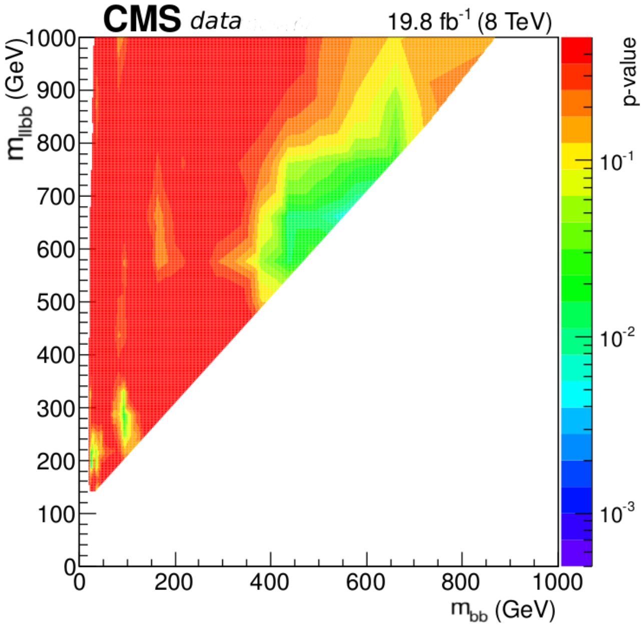


Figure 4.21: Observed p-value as a function of m_{bb} and m_{llbb}.

choice of the signal described above. On the top plots the signal is added on top of the background. It is normalised to the NNLO SUSHI cross section for the model parameters listed in Table 4.2. On the bottom plots the signal is compared to the subtracted data. The four plots show that for these two observables the observed excess is compatible with the tested signal both in shape and in amplitude. This unexpected result raises the interest for this search.

Other kinematic observables are presented in Figures 4.23 and 4.24. The excess is mainly visible at low p_T, low E_T^{miss} and $\Delta\phi(ll, bb) \sim \pi$. Most of the bins showing an excess are compatible with the background-only expectation within two standard deviations. While it might be only a statistical fluctuation in this region of the phase space, the signal-plus-background hypothesis matches well the data for most of the distributions. This is e.g. the case for $|cos\theta_{b1}|$ with θ_{b1} the helicity angle defined as the angle between the *bb* system (in the *llbb* rest frame) and the leading *b* jet (in the *bb* rest frame). Nevertheless, a less clear conclusion can be drawn from the p_T^{ll} and p_T^{bb} observables.

From this study the conclusion is that the excess is not significant enough to separate the background-only and signal-plus-background hypotheses. However it is crucial to

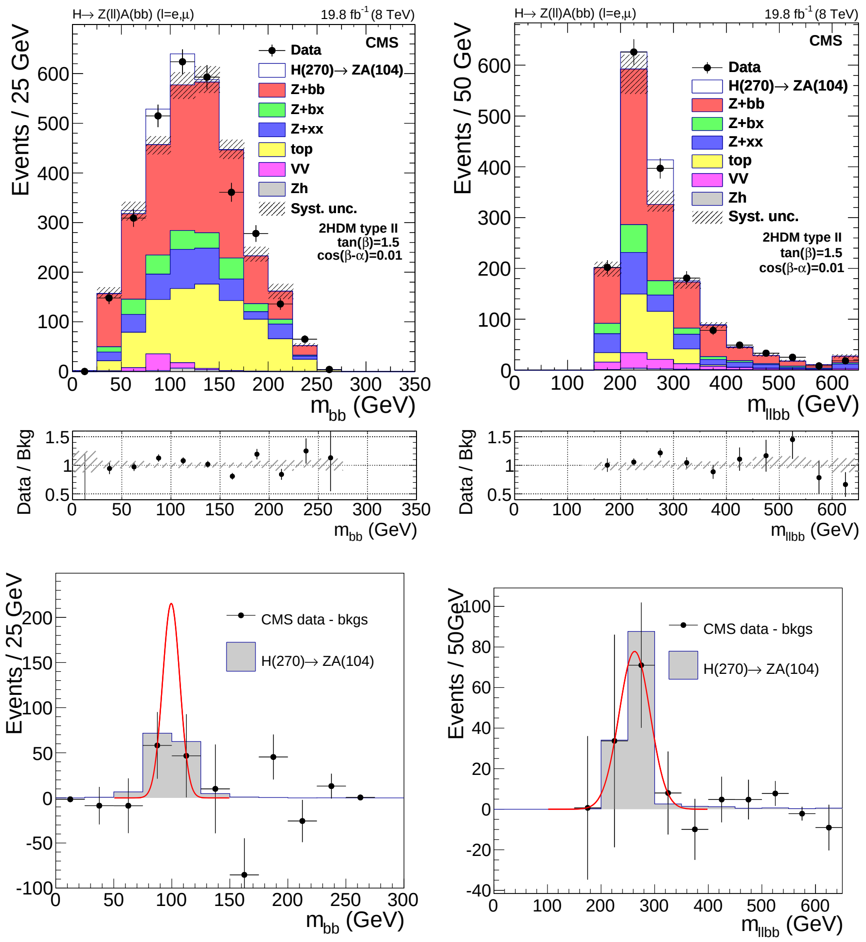


Figure 4.22: The top plots show the comparisons of the data to the expectation from the simulation of the backgrounds. The left (right) plots shows the m_{bb} (m_{llbb}) observable after applying the cut 222 GeV $< m_{llbb} <$ 350 GeV (72 GeV $< m_{bb} <$ 114 GeV). The hashed band on the sum of the backgrounds and on the ratio represent the systematic uncertainty on the sum of the backgrounds. On top of the backgrounds, a signal is added corresponding to $m_A = 104$ GeV and $m_H = 270$ GeV. It is normalised to the NNLO cross section for the model parameters listed in Table 4.2. The last bins include the overflow. The bottom plots show the comparisons of the data after subtraction of the expected backgrounds to the simulation of the signal.

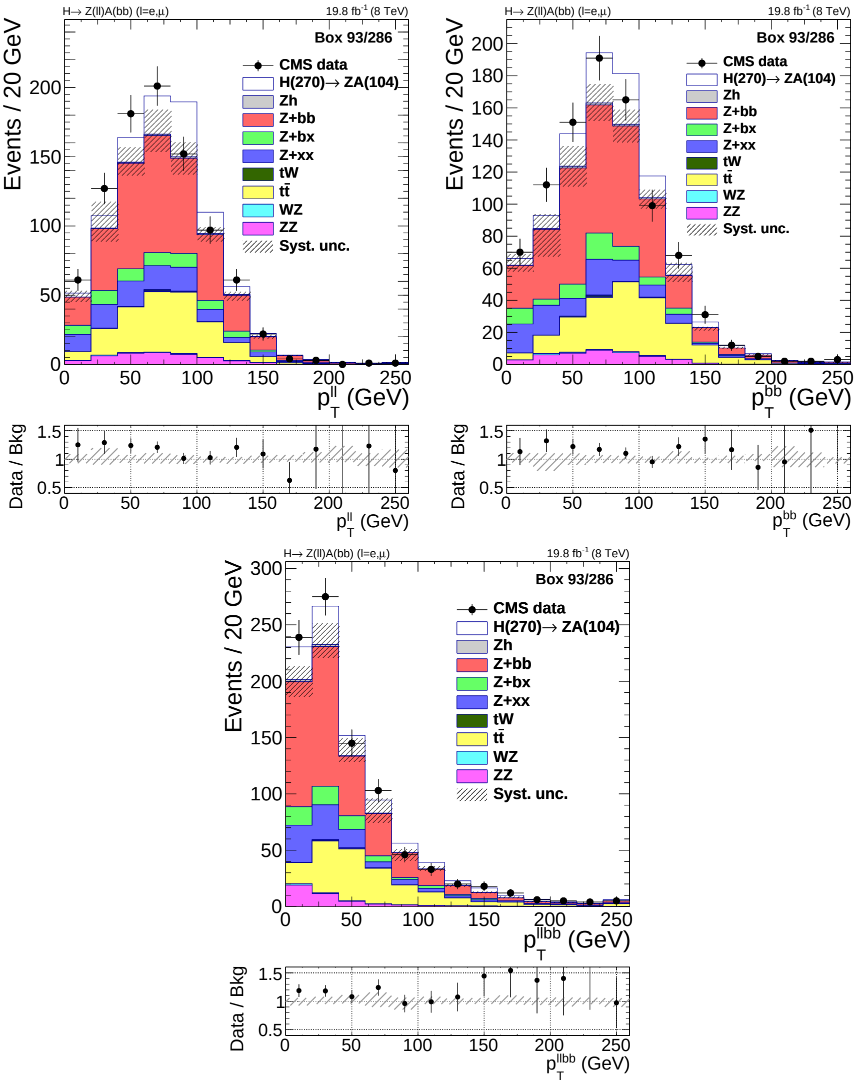


Figure 4.23: Comparisons of the data to the expectation from the simulation of the backgrounds in the signal region centred on m_{bb} = 93 GeV and m_{llbb} = 286 GeV. From the top left to the bottom, the plots represent the p_T^{ll}, the p_T^{bb} and the p_T^{llbb}. The hashed band on the sum of the backgrounds and on the ratio represent the systematic uncertainty on the sum of the backgrounds. On top of the backgrounds, the signal is added normalised to the NNLO cross section for the model parameters listed in Table 4.2. The last bins include the overflow.

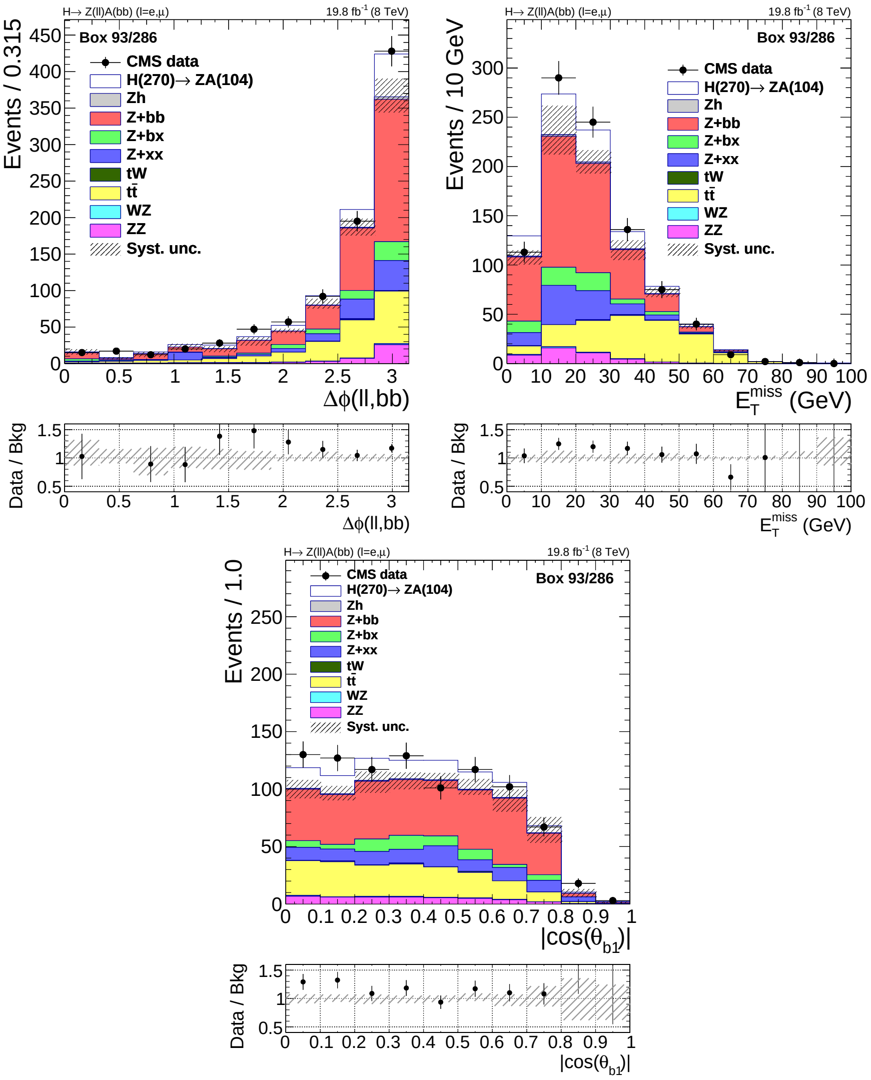


Figure 4.24: Comparisons of the data to the expectation from the simulation of the backgrounds in the signal region centred on m_{bb} = 93 GeV and m_{llbb} = 286 GeV. From the top left to the bottom, the plots represent the $\Delta\phi(l,l)$, the E_T^{miss} and $|cos\theta_{b1}|$. The hashed band on the sum of the backgrounds and on the ratio represent the systematic uncertainty on the sum of the backgrounds. On top of the backgrounds, the signal is added normalised to the NNLO cross section for the model parameters listed in Table 4.2. The last bins include the overflow.

pursue this analysis with the data collected in 2016 in order to confirm or discard the signal-plus-background hypothesis.

4.4.2 Generic limits

Awaiting more data to draw a clearer conclusion, upper limits are set on the number of expected signal events and on $\sigma \times BR$ for a hypothetic signal. These results are the key points of this analysis because they can allow to recast and reinterpret the limits in other models which respect the few assumptions made in this analysis. The list of models which can be probed can be extended by taking into account properly the possible differences in the signal efficiency. In this last case however the sensitivity might be reduced significantly.

The expected and observed upper limits can be seen in Figure 4.25. The top plots representing the limits on the number of events can be used to derive limits on any models also with different widths for the new particles. For particles with much wider widths than the ones probed in this analysis the only drawback is a lesser sensitivity to such particles as the signal windows will, relatively, contain a lower fraction of signal events. The bottom plots are derived from the top plots by taking into account detector acceptance and reconstruction efficiency for the signal events. Using the signal efficiency map in Figure 4.7, limits on the cross sections for any signal are obtained assuming $\Gamma_i < R \times m_i$ with Γ_i the width of the two new resonances, m_i their masses and R the experimental resolution. As can be seen from the bottom right plot representing the observed exclusion limits on the signal cross sections, this analysis can exclude signals with $\sigma \times BR$ down to a few fb for $m_H \gtrsim$ 600 GeV and $m_A \in [100, 400]$ GeV. It is interesting to note the limitation of this analysis to exclude signal with low m_A (below 40 to 70 GeV depending on m_H) due to the overlap of the *b* jets.

4.4.3 Model dependent interpretation

Now considering the model parameters listed on Table 4.2, the limits presented in Figure 4.25 are interpreted as limits on μ for this model. These limits are shown in Figure 4.26. The phase space with $\mu < 1$ is excluded for this model. This region is delimited by the solid line in the bottom plot. This corresponds approximatively to $m_A \in [50, 250]$ GeV and $m_H \in [200, 650]$ GeV for the process $H \longrightarrow ZA$ and $m_H <$ 250 GeV and $m_A <$ 700 GeV for the process $A \longrightarrow ZH$. The excess discussed in Section 4.4.1 is in a region close to the exclusion limit where the observed excluded region is smaller than expected.

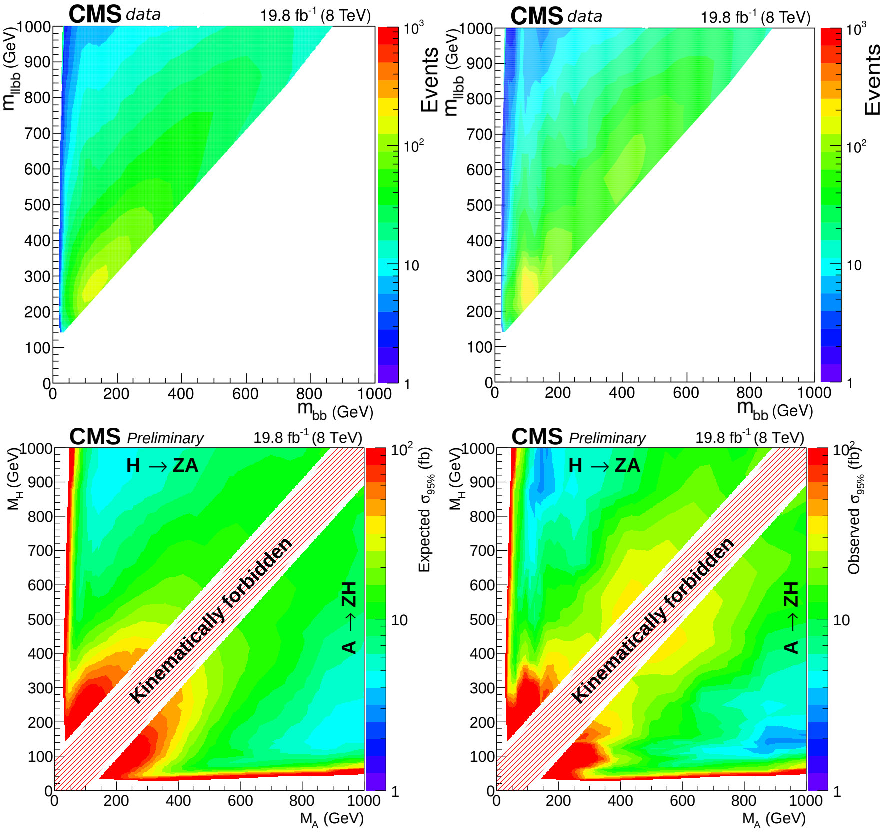


Figure 4.25: The top plots show the exclusion limits on the number of events within the acceptance as a function of m_{bb} and m_{llbb}. The bottom plots show the exclusion limits on the signal cross section as a function of m_A and m_H. The left (right) plots correspond to the expected (observed) limits from simulation.

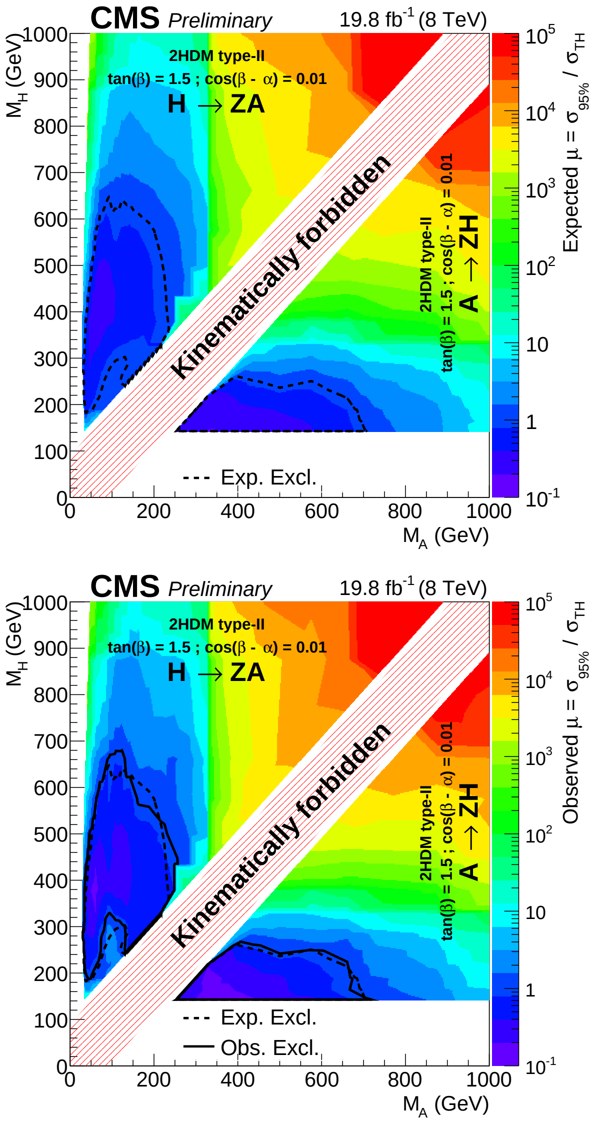


Figure 4.26: Limits on μ as a function of m_A and m_H for the model described in Table 4.2. The top (bottom) plot represents the expected (observed) limits from simulation. In both plots the dashed contour delimits the expected region to be excluded. In the bottom plot the solid line delimits the region currently excluded by the data.

For fixed m_A and m_H, limits are also put on $cos(\beta - \alpha)$ and $tan(\beta)$. Figure 4.27 shows the limits on these parameters for type II 2HDM with $m_A = 150$ GeV and $m_H = 350$ GeV. As expected from the plots in Figure 1.10, this search is sensitive to the region with $cos(\beta - \alpha) \sim 0$ and $tan\beta \sim 1$. For $tan\beta < 1$, the limits should be taken with caution as the signal cross section was computed assuming perturbativity.

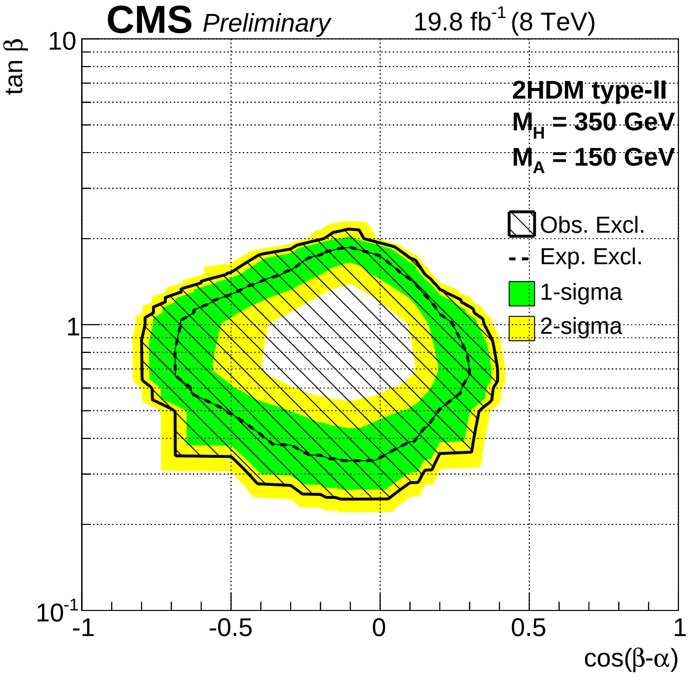


Figure 4.27: Limits on $cos(\beta - \alpha)$ and $tan(\beta)$ for type II 2HDM with $m_A = 150$ GeV and $m_H = 350$ GeV. The dashed line delimits the region which is expected to be excluded. The green and yellow bands represent the ± 1 and 2 standard deviations respectively from this limit. The solid line and the hashed region correspond to the excluded phase space.

4.5 Perspectives on low mass resonances

As already discussed, the region of low m_{bb} suffers from the experimental limitation from standard jet reconstruction. Indeed, low m_{bb} means the presence of boosted overlapping b jets. The typical signature of collimated boosted jets is the formation of a so-called *fat jet*. In previous years several studies were done in order to study such jets [115–117]. Two ingredients are relevant for the search presented here. The first one is the mass of the reconstructed fat jet. For this purpose grooming techniques have recently been developed such as trimming [118], filtering [116], pruning [119, 120], etc. They allow to remove soft contributions, including PU, to the fat jets and to get the two original prong sub jets. The groomed jet mass have been studied by the CMS collaboration in [121]. The mass is well reconstructed which allows a better discrimination from purely QCD jets. The second relevant ingredient is the b-tagging of such fat jets which are in this case originating form two b quarks. Several choices can be made depending on the need of efficiency and purity. The simplest case is to tag the fat jet [76]. This results in a good efficiency but a rather low purity because the fat jets originating from a single b quark are not suppressed. To improve the purity it is possible to apply the b-tagging on the sub jets [76]. This generally gives better performances but it is important to be careful on the possibility to share tracks among the sub jets. The last possibility, still under development, is to make a dedicated tagger for double b-tagging [122]. This is expected to give the overall best performance.

In this analysis, tools available and tested by the CMS experiment have been used. Several CMS analyses have used these tools for W-tagging [123–125], for Z-tagging [123, 125] or also for top-tagging[126–129]. For defining the fat jets, the Cambridge-Aachen [130, 131] algorithm was used with a cone radius of 0.8. The substructure of the fat jets is retrieved by pruning. The pruning consists in reclustering the fat jet constituents but rejecting soft objects. For this, two cuts are defined:

$$z_{ij} = \frac{min(p_{Ti}, p_{Tj})}{p_{Ti} + p_{Tj}} < z_{cut} = 0.1$$

$$\Delta R(i,j) > D_{cut} = \frac{m_J}{p_T}$$

where i and j are the two constituents to be merged and m_J and p_T are the mass and the transverse momentum of the original fat jet. If these two requirements are fulfilled then the softer particle is removed. Two exclusive sub jets are required from the clustering. The sub jets invariant mass gives the mass of the lightest new boson candidate.

The CSV b-tagging algorithm is applied to the two resulting sub jets. In order to avoid tracks to be shared between the two sub jets only tracks belonging to the PF elements of each sub jet are used for tagging each sub jets independently. This was not yet the standard recommendation at the time of the analysis but by now it is the recommended technique. This allows to go lower in ΔR between the two sub jets and avoids an increase of the mistagging efficiency when only one of the sub jets is a b jet.

In the following only a few plots will be shown to demonstrate the feasibility of reconstructing the signal in this region but also the difficulties of understanding the backgrounds.

To make this study, the same selection is used as for the standard analysis replacing the jets by the sub jets. In Figure 4.28 the $\Delta R(j, j)$ is shown in the *Z+jj* region. The left plot shows the difficulty to perform the same search due to a poor modelling of the data. A clear disagreement can be seen which cannot only be explained by a problem of normalisation. The issue is not understood but no deep investigation was done due to a lack of time. The same issue is also present after applying b-tagging. This is the reason why it was not possible to directly extend the standard analysis in this region.

The right plot shows the same observable but this time after selecting events with $p_T^{ll} > 180$ GeV. This is a standard cut applied by analysis looking to boosted jets. In this case, the agreement is much better. This means the issue mainly concerns soft events. However such cut can lead to two issues for this analysis. The first one is that some signal will not be accessible. Only the cases with $m_H \gtrsim 500$ GeV can lead to enough boost to be selected with a reasonable efficiency. The second issue is the statistic: after b-tagging only few tens of events remain after this selection representing only 10% of the available data. Therefore, with this data, this is not the best region to constrain the signal.

Waiting for a better understanding and description of this region, it is important to show the possibility of the grooming and b-tagging techniques to reconstruct and select signal events. The Figure 4.29 shows on the top the m_{bb} (left) and m_{llbb} (right) observables for several signal samples which are expected to be present in this topology. The samples are not normalised but for each of them, the same amount of events were simulated (25 000). This means that differences in the total yields are mainly due to reconstruction efficiencies. For example, the sample with $m_A = 15$ GeV and $m_H = 350$ GeV suffers from the really small angle between the two *b* jets ($\Delta R \sim 0.15$) showing the limitation of the substructure techniques. The sample $m_A = 70$ GeV and $m_H = 350$ GeV is another special case. Indeed, in this case, the $\Delta R(b, b)$ is between 0.5 and 1.0 implying that part of the events will be well reconstructed with the standard analysis and the other part needs this special treatment in order to be selected. This introduces the problem of reconciling both analyses to avoid overlap and at the same time to obtain the best sensitivity. One possibility would be to simply perform the standard analysis and to apply this boosted analysis only to the events which do not pass the standard analysis selection. This is simple but not necessary the most efficient for the signal. Another possibility would be to perform both analyses and define a quality criteria in case of conflict to help decide which case is the most signal-like. This is something which needs more study. What is also clear from the same plots is that the mass of the new bosons are well reconstructed with a

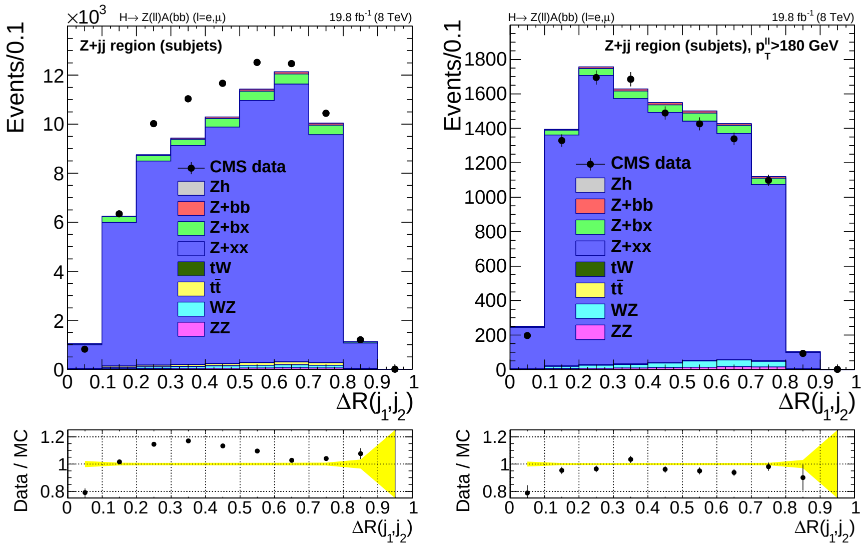


Figure 4.28: Comparisons of the data to the expectation from the simulation of the backgrounds in the *Z+jj* region for the $\Delta R(j, j)$ observable where j represent the sub jets. The right plot shows only events with $p_T^{ll} > 180$ GeV. Simulation samples are normalised to the theoretical expectation. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation.

resolution close to the standard jet reconstruction (e.g. about 10%) and with a good reconstruction efficiency (e.g. about 11% of events selected for $m_A = 30$ GeV and $m_H = 350$ GeV).

On the bottom plots of Figure 4.29, the shapes from the backgrounds are also shown and compared to one signal ($m_A = 30$ GeV and $m_H = 350$ GeV). On the left plot which represents m_{bb}, the resonances from the *ZZ* and *Z*h_{125} backgrounds are also visible. What was less expected was the peak at low m_{bb} for the *Z+bx* background. This peak is made up of two peaks, the first one close to 6 GeV and the second one close to 9 - 10 GeV. Because these masses are quite close to the masses of the mesons composed of at least one b-quark, this means that the jet substructure technique is able to see the B mesons decay products. From the two plots, it looks like combining both reconstructed masses information will lead in signal region with almost no backgrounds.

In conclusion, this region looks promising for extending the standard analysis but additional work is needed in order to properly understand the background modelling.

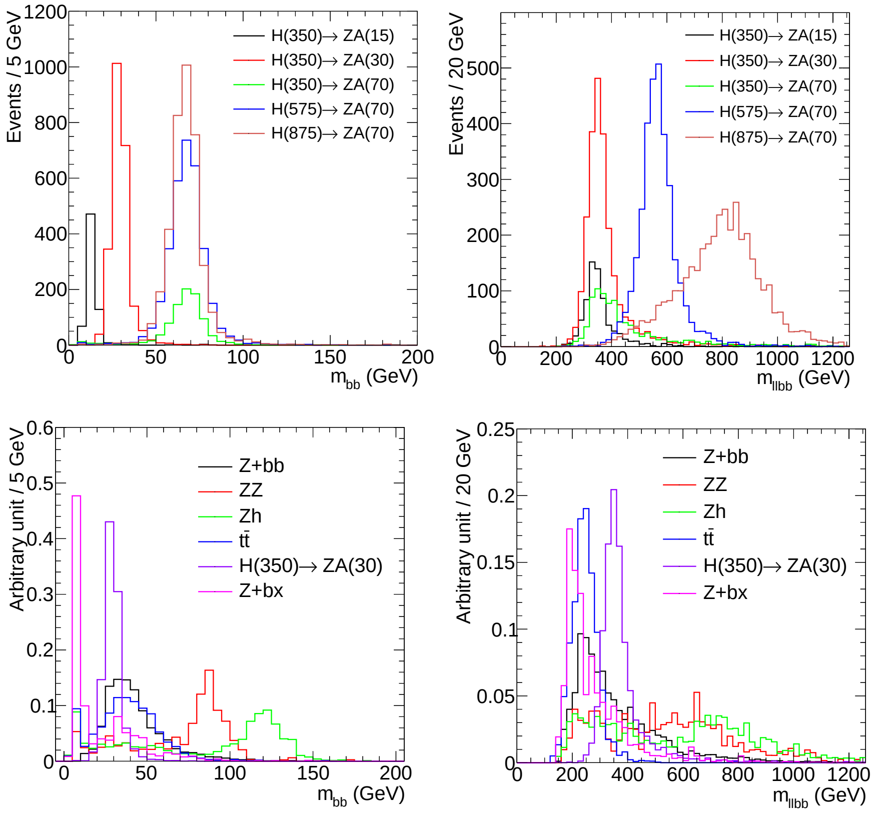


Figure 4.29: The left (right) plots show the m_{bb} (m_{llbb}) observable. The top plots show the number of signal events reconstructed and selected according to the selection defined in Table 4.4 for five signal samples. The bottom plots compare the shapes of the two reconstructed masses for the main backgrounds after the same selection with a signal sample.

Conclusion

Two analyses have been conducted using the data delivered by the LHC and collected by the CMS experiment in 2012 at $\sqrt{s} = 8$ TeV corresponding to an integrated luminosity of 19.5-19.8 fb^{-1}.

The aim of the first analysis was to search for the SM Higgs boson decaying into a pair of b quarks, in the final state $Z(ll)h_{125}(bb)$. A Matrix Element (ME) technique based on the `MadWeight` program was tested in order to improve the sensitivity to the signal. Several hypotheses have been considered for the signal and background processes. The ME weights have been combined in several neural networks in order to benefit from all these hypotheses. The shapes of the final discriminants have been used to look for the potential presence of the signal. This strategy showed to bring additional discriminatory power with respect to the use of a single kinematic variable such like m_{bb}. An outlook on the usage of the ME technique have also been discussed, especially concerning the improvements which could increase the interest for this technique. The results from this analysis have been combined with the results obtained at $\sqrt{s} = 7$ TeV with the data collected in 2011 and corresponding to an integrated luminosity of 5 fb^{-1}. The strategy was tested to highlight the presence of the well-known ZZ process. A slight excess of 1.2 standard deviations (s.d.) was, in fact, observed. The amplitude of the ZZ process obtained from this excess was compatible with the expectation from direct CMS measurements and measured to be $\mu = \sigma/\sigma_{CMS} = 0.57^{+0.48}_{-0.47}$. On the other hand no evidence of the $Z(ll)h_{125}(bb)$ process was observed. The obtained upper limit on μ is 1.6 meaning the signal hypothesis cannot be excluded. This upper limit on this process is compatible with the background-only hypothesis and only compatible with the presence of the signal within 2 s.d. This analysis also emphasizes the challenge that the mitigation of PU interactions represents. Several observations led to the conclusion that PU interactions reduced the capability of this analysis to exploit the event information in an optimal way. This resulted in some disagreement between the simulation and the data in the signal regions. The *Z*(*ll*)h_{125}(*bb*) search showed to be particularly sensitive to these effects. This explained, at least partially, the differences observed with respect to the CMS published analysis [66]. Concerning the perspective for discovering this decay, this final state can be combined to other sensitive final state, mainly *Z*($\nu\nu$)h_{125}(*bb*) and *W*(*l*ν)h_{125}(*bb*). Despite the Run 1 LHC combination of ATLAS and CMS for this decay [14], the significance of the observed excess did not reach the 3 s.d. necessary to state the evidence of this decay. The data collected in 2016 will allow to draw a clear conclusion for this search.

The second analysis aimed to search for yet undiscovered particles in the context of tow-Higgs-doublets models (2HDMs). A search for the $H/A \longrightarrow Z(ll)A/H(bb)$ processes was performed using a scan in the $m_{bb} - m_{llbb}$ plane. It establishes the first experimental results for a search for these processes at the LHC. A simple cut and count analysis was setup by defining SRs with windows around the two masses based on the experimental resolution. For a first time for a CMS analysis based on existing data, `DELPHES` was used to simulate the response of the detector for the signal events. In the analysis process, a strategy was proposed to cure the missing NLO contributions in the simulation of the *Z*+jets events. Two excesses have been reported with a local (global) significance of 2.9 (1.9) and 2.6 (1.5), respectively. The signal-plus-background hypothesis was shown to be poorly compatible with the first excess. The second excess is however well compatible with the signal-plus-background hypothesis both in shape and in amplitude (based on the tested model and the NNLO cross section from `SUSHI`). The mass of the new resonances would be $m_A = 104$ GeV and $m_H = 270$ GeV. This hypothesis can already be tested with the data collected in 2016. An optimised analysis can be done to increase the sensitivity to the possible signal. This can allow to confirm or discard this hypothesis. Still, upper limits on the presence of the tested signals were set in several ways allowing for reinterpretation in different models. The limits on the number of events can be used for recasting the results for signals with really different efficiency map than the one of the tested signals. The limits on the signal cross sections can be used to probe specific models for which the efficiency map is the same as the one of the tested signals. Upper limit on $\sigma \times BR$ down to a few fb for $m_H \gtrsim 600$ GeV and $m_A \in [100, 400]$ GeV are set. These limits were interpreted in a specific type II 2HDM. The region $m_A \in [50, 250]$ GeV and $m_H \in [200, 650]$ GeV for the process $H \longrightarrow ZA$ and $m_H < 250$ GeV and $m_A < 700$ GeV for the process $A \longrightarrow ZH$ are excluded for this model. For fixed m_A and m_H, limits were also put on *cos*($\beta - \alpha$)and *tan*(β). This search is sensitive to the region $cos(\beta - \alpha) \sim 0$ and $tan\beta \sim 1$. With the 2016 data this analysis might allow to push even more the exclusion limits and put even more stringent constraints on possible beyond standard models (BSM) physics. Finally, a perspective for this analysis has been studied in the boosted regime when the difference of mass between the *H* and *A* bosons starts to be significant. This study showed a good capability in reconstructing signal events and the masses of the *H* and *A* bosons. However it was observed that more work is needed in order to better control the background modelling especially in the soft regime ($p_T^{ll} < 180$ GeV). Such study can be conducted on the 2016 data in order to explore a still uncovered phase space.

In conclusion, these two analyses face several topics which are relevant for the future of the LHC physics program: development of new techniques to improve the sensitivity to rare and new processes, PU mitigation in increasingly challenging PU conditions and the search for BSM physics.

Appendix A

Additional kinematic comparisons for the SM Higgs search

Additional kinematic comparisons are shown for the SM Higgs search presented in Chapter 3. From Figure A.1 to Figure A.4, the same observables presented in Section 3.3.1 are shown in the SR. Good agreement is observed in the $2j$ category, within statistical uncertainties. The overestimation of the backgrounds is visible in the $3j$ category. Looking at the E_T^{miss} significance plot in the $3j$ category (Figure A.4, bottom right), the disagreement between data and simulation concerns mainly the Z+jets enriched region (low value). Another observation is the excess (deficit) visible in the $2j$ ($3j$) categories at low p_T for the p_T^{b1} and p_T^{b2} observables (Figure A.2). This tends to confirm the PU hypothesis proposed in Section 3.4.2 to explain the disagreement on the final results.

The Figures A.5, A.6, A.7 and A.8 show the p_T^{bb}, the $\Delta R(b_1, b_2)$, the $\Delta R(l_1, l_2)$ and the $\Delta\phi(ll, bb)$ observables both in the CR and SR. This completes the set of relevant observables showed in chapter 3.

From Figure A.9 to Figure A.12, as in Section 3.3.2 the ME weights are shown but this time in the CR. The agreement is within the statistical uncertainties.

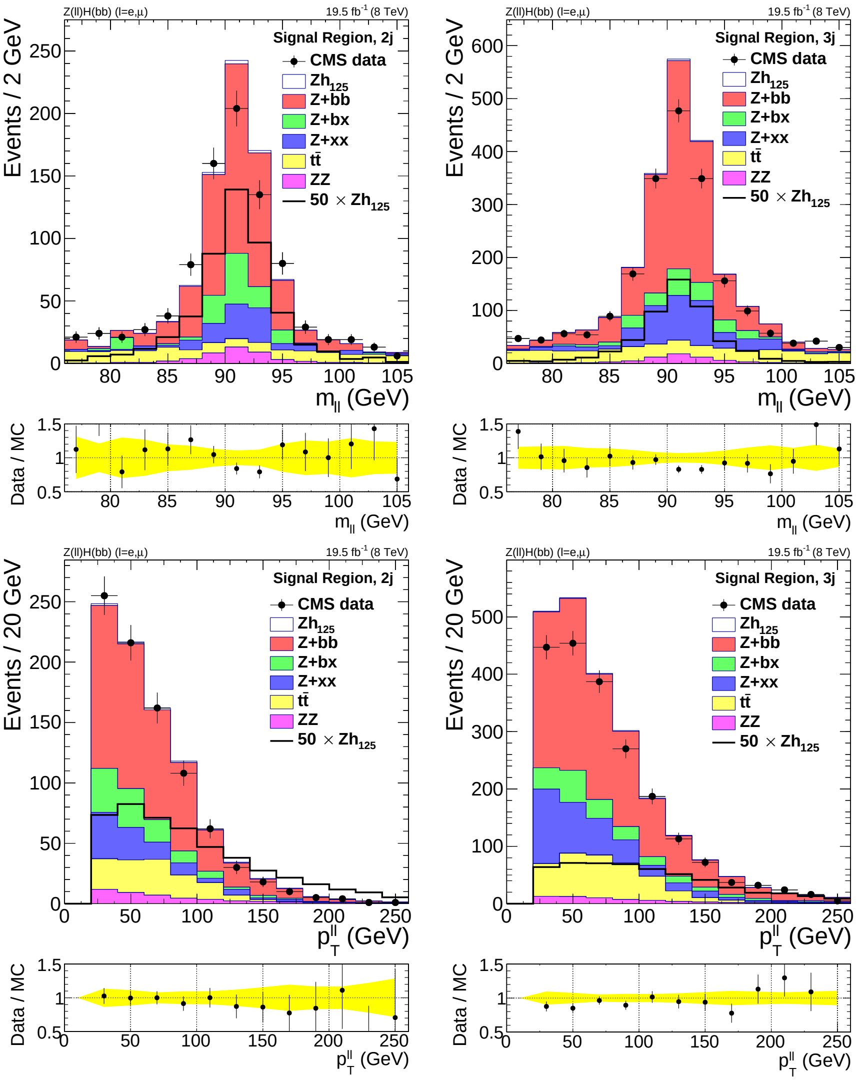


Figure A.1: Comparisons in the SR of data and simulation for the m_{ll} (top) and p_T^{ll} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

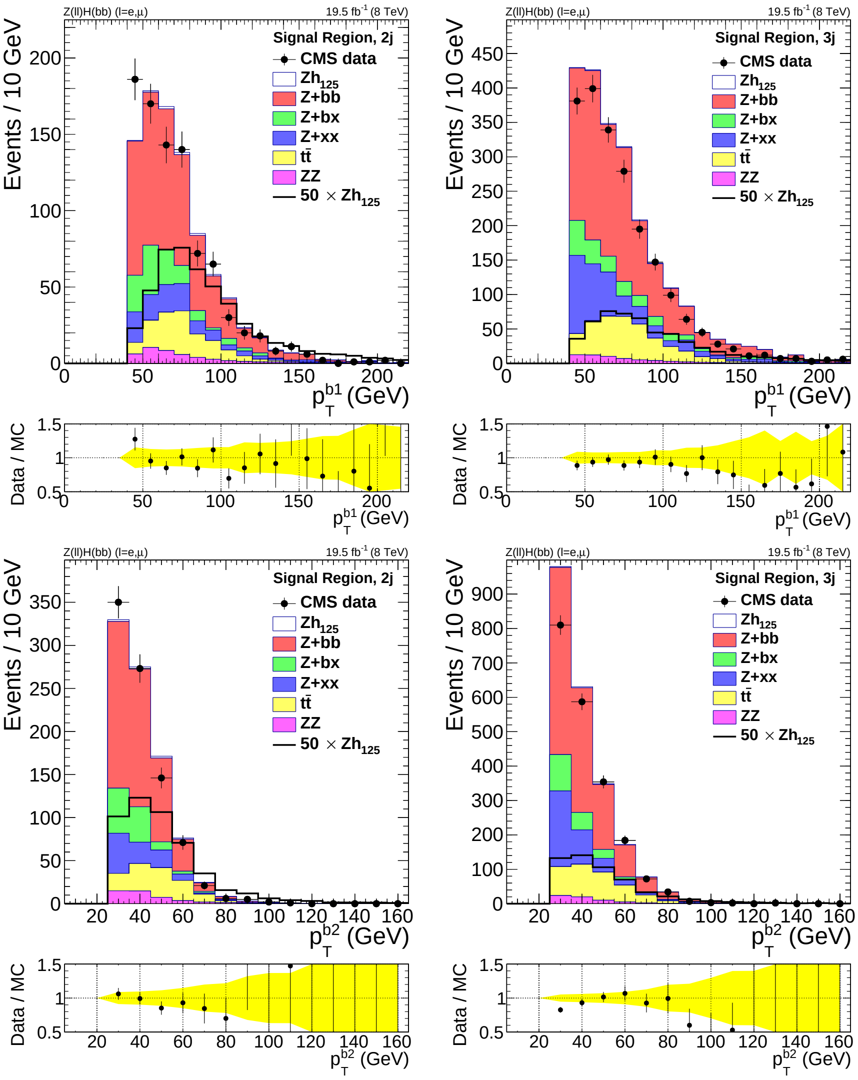


Figure A.2: Comparisons in the SR of data and simulation for the p_T^{b1} (top) and the p_T^{b2} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

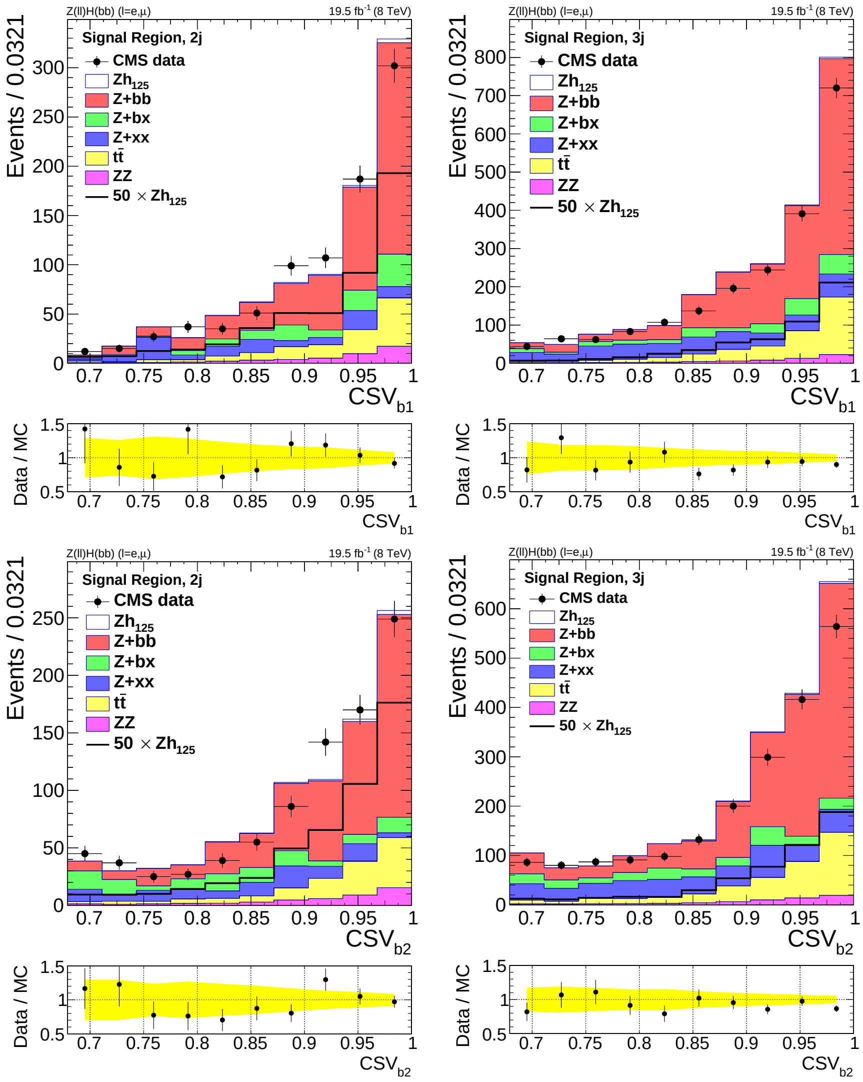


Figure A.3: Comparisons in the SR of data and simulation for the CSV_{b1} (top) and CSV_{b2} (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

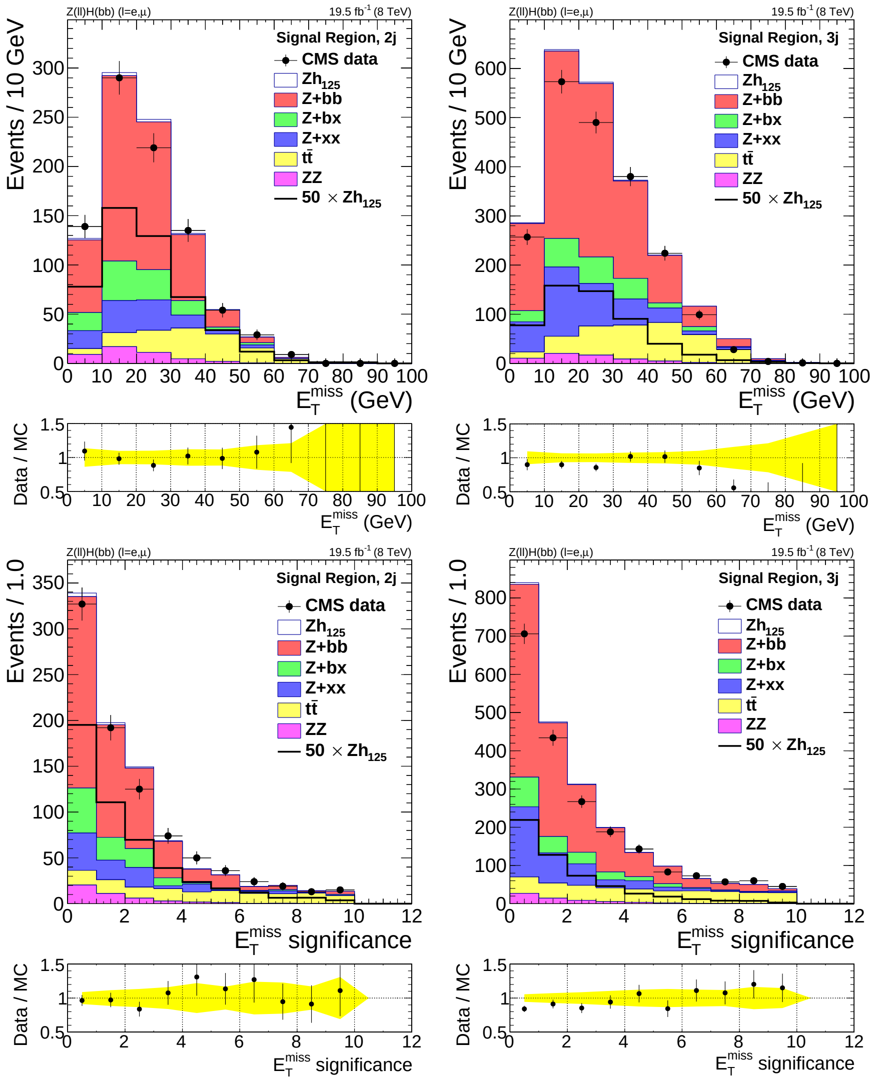


Figure A.4: Comparisons in the SR of data and simulation for the E_T^{miss} (top) and the $E_T^{miss} significance$ (bottom) observables. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

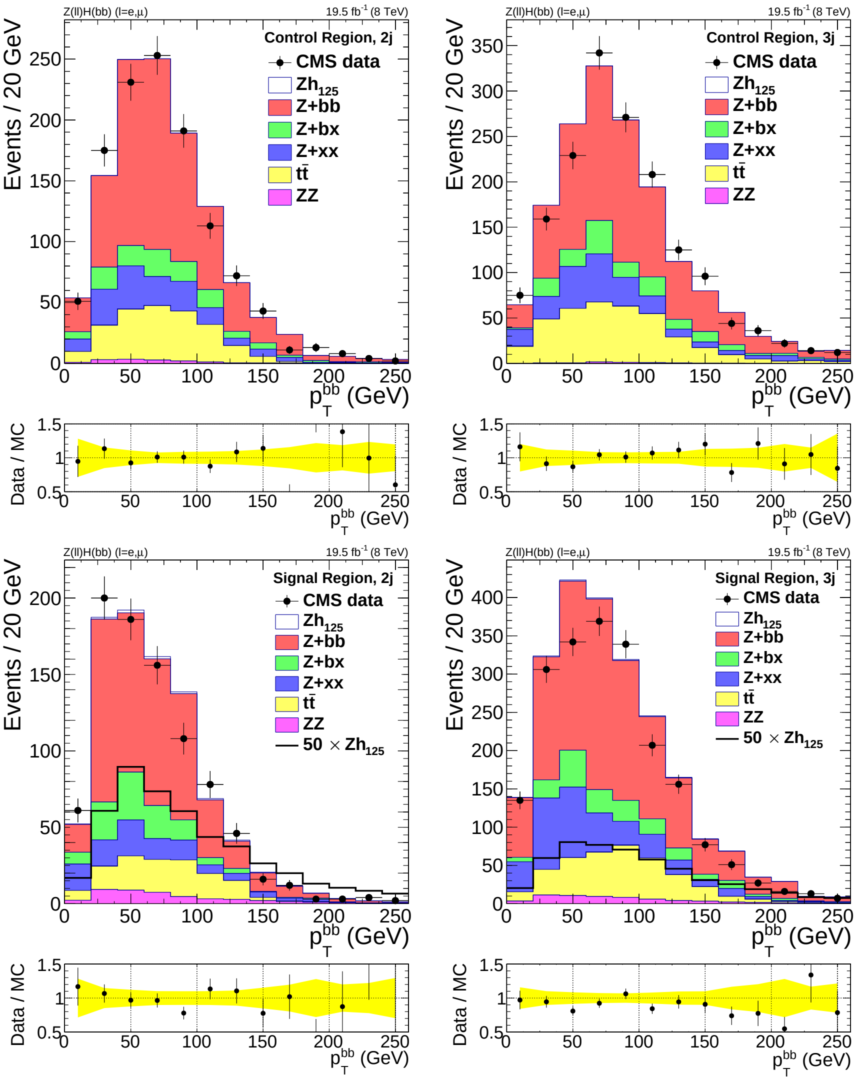


Figure A.5: Comparisons in the CR (top) and in the SR (bottom) of data and simulation for the p_T^{bb} observable. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

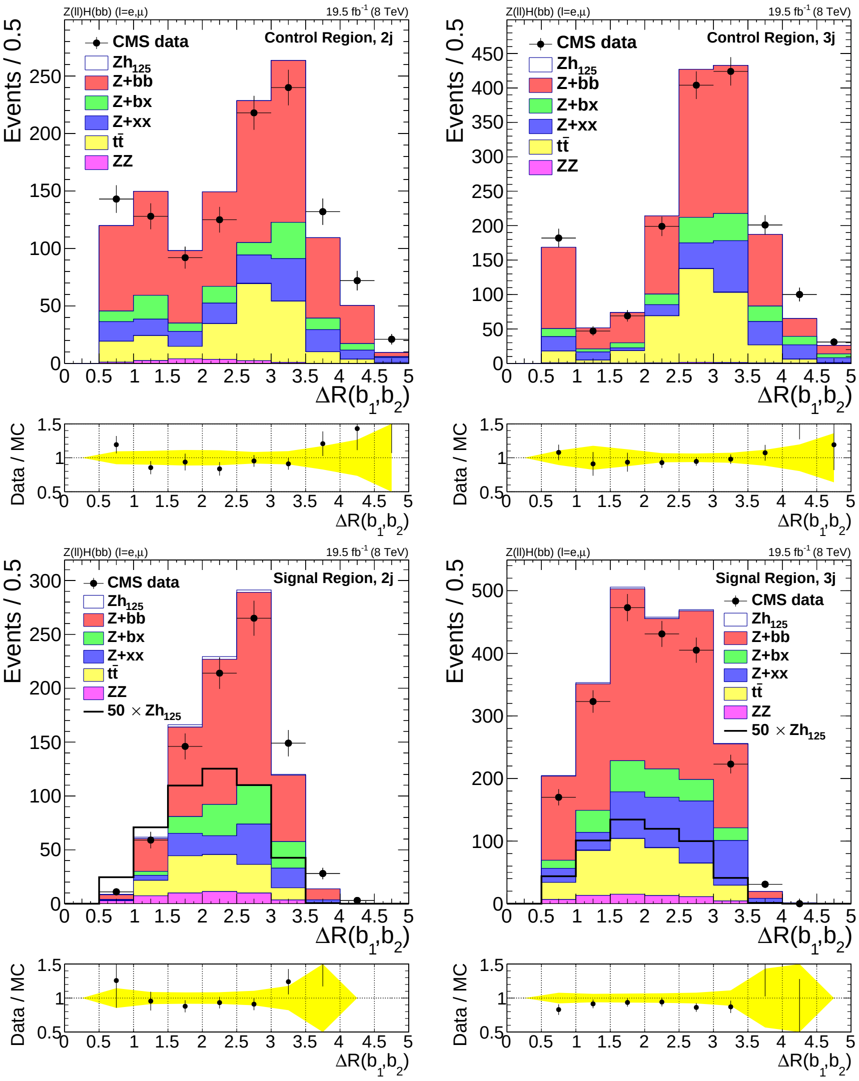


Figure A.6: Comparisons in the CR (top) and in the SR (bottom) of data and simulation for the $\Delta R(b_1, b_2)$ observable. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

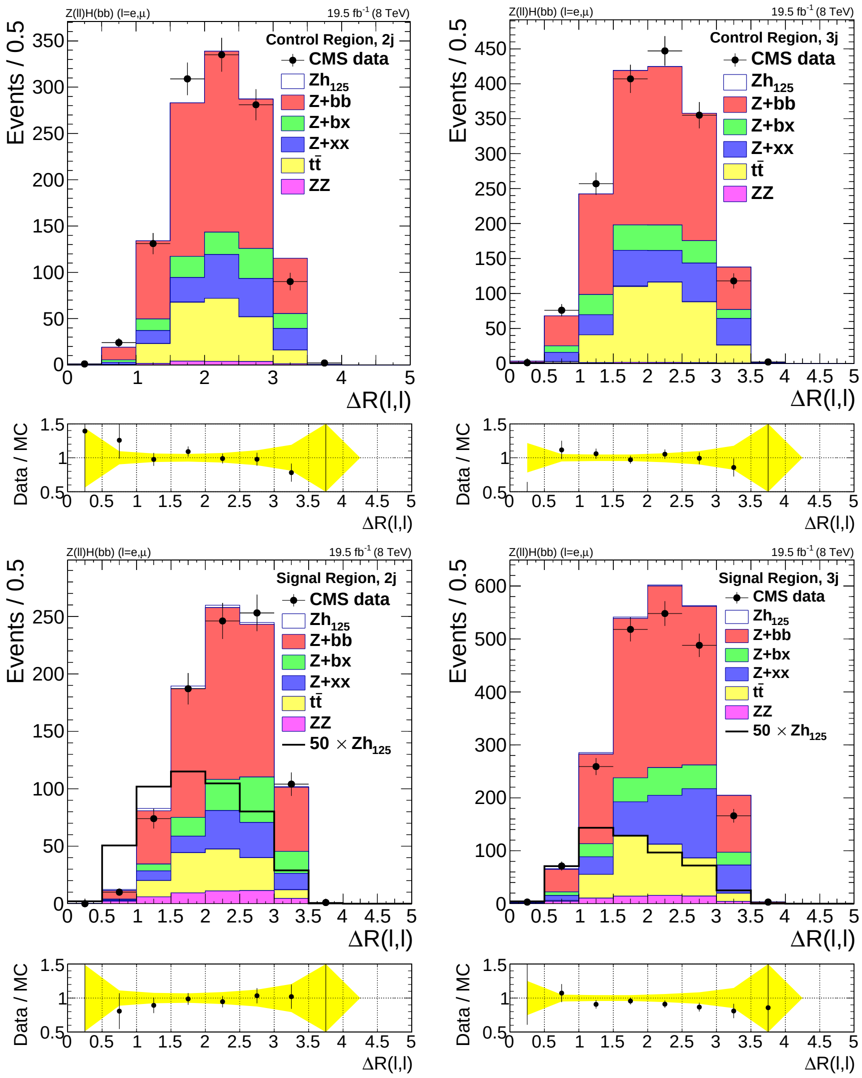


Figure A.7: Comparisons in the CR (top) and in the SR (bottom) of data and simulation for the $\Delta R(l,l)$ observable. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

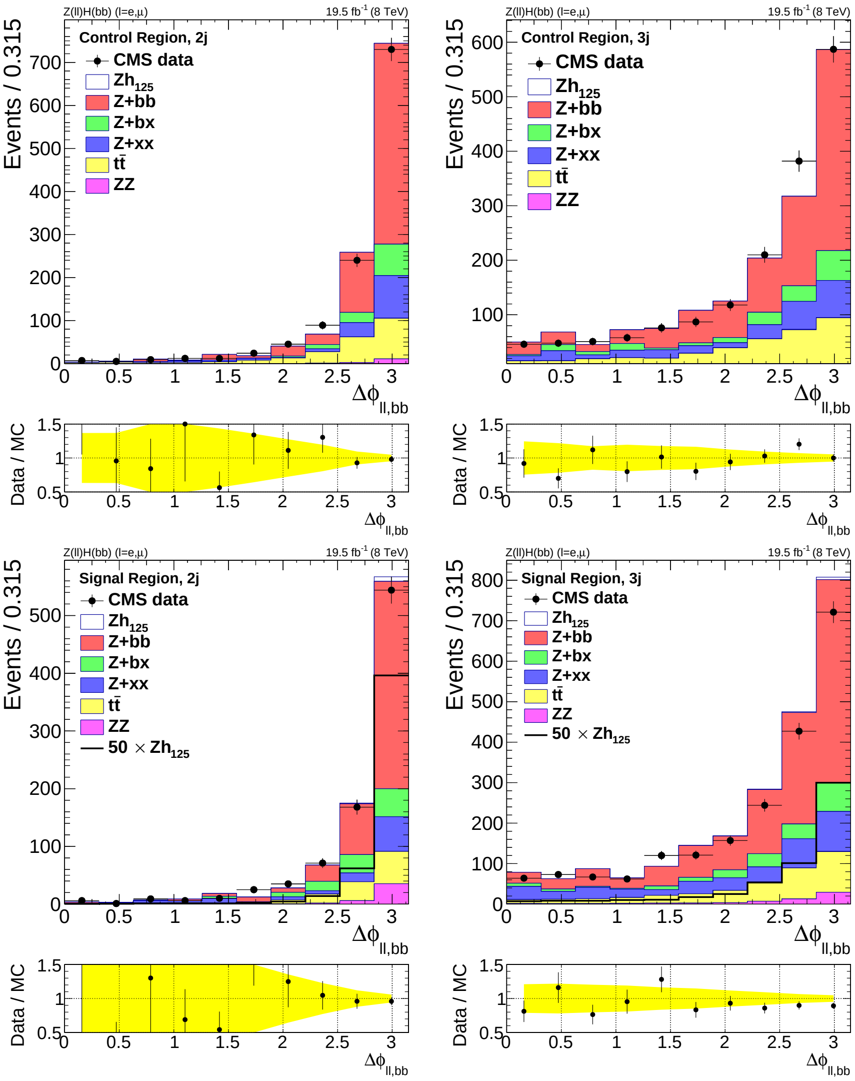


Figure A.8: Comparisons in the CR (top) and in the SR (bottom) of data and simulation for the $\Delta\phi(ll, bb)$ observable. The left plots correspond to the $2j$ category and the right plots to the $3j$ category. Simulation samples are normalised using the SFs shown in Table 3.5. In the SR, the signal is also showed separately normalised to 50 times its cross section. In the ratio, the yellow band represents the statistical uncertainty from simulation.

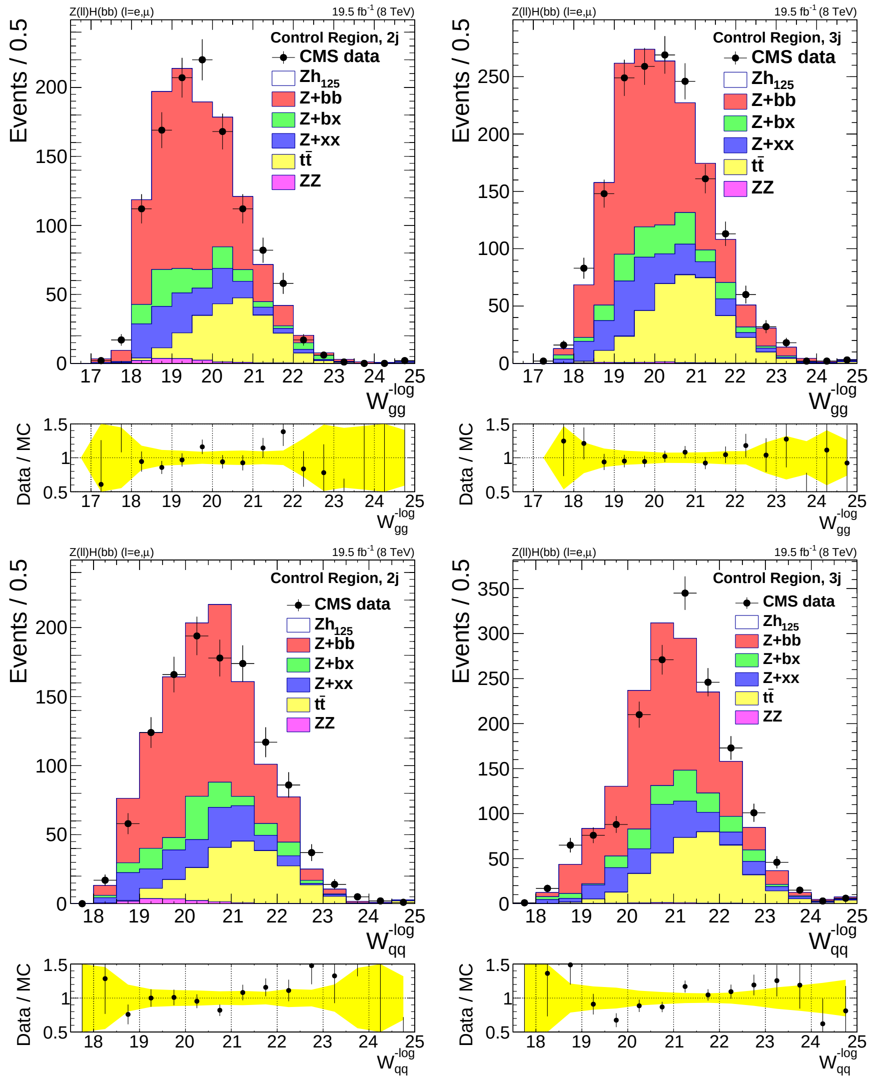


Figure A.9: Comparisons in the CR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the Z+jets process. The top (bottom) plots represent W_{gg}^{-log} (W_{qq}^{-log}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

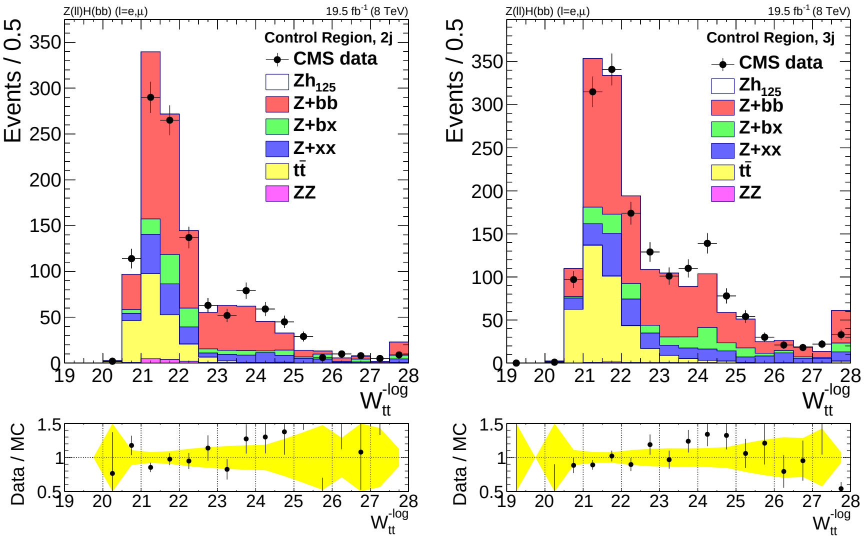


Figure A.10: Comparisons in the CR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for $W_{t\bar{t}}^{-log}$. Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

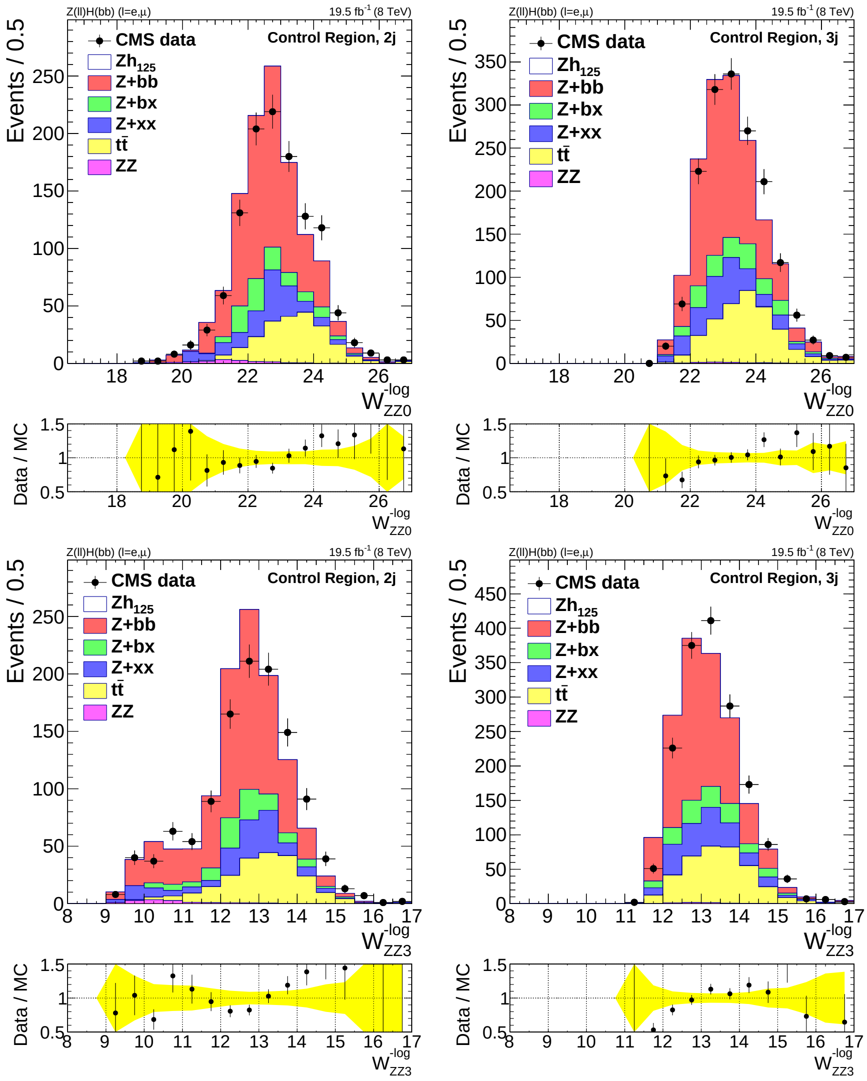


Figure A.11: Comparisons in the CR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the ZZ process. The top (bottom) plots represent W^{-log}_{ZZ0} (W^{-log}_{ZZ3}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

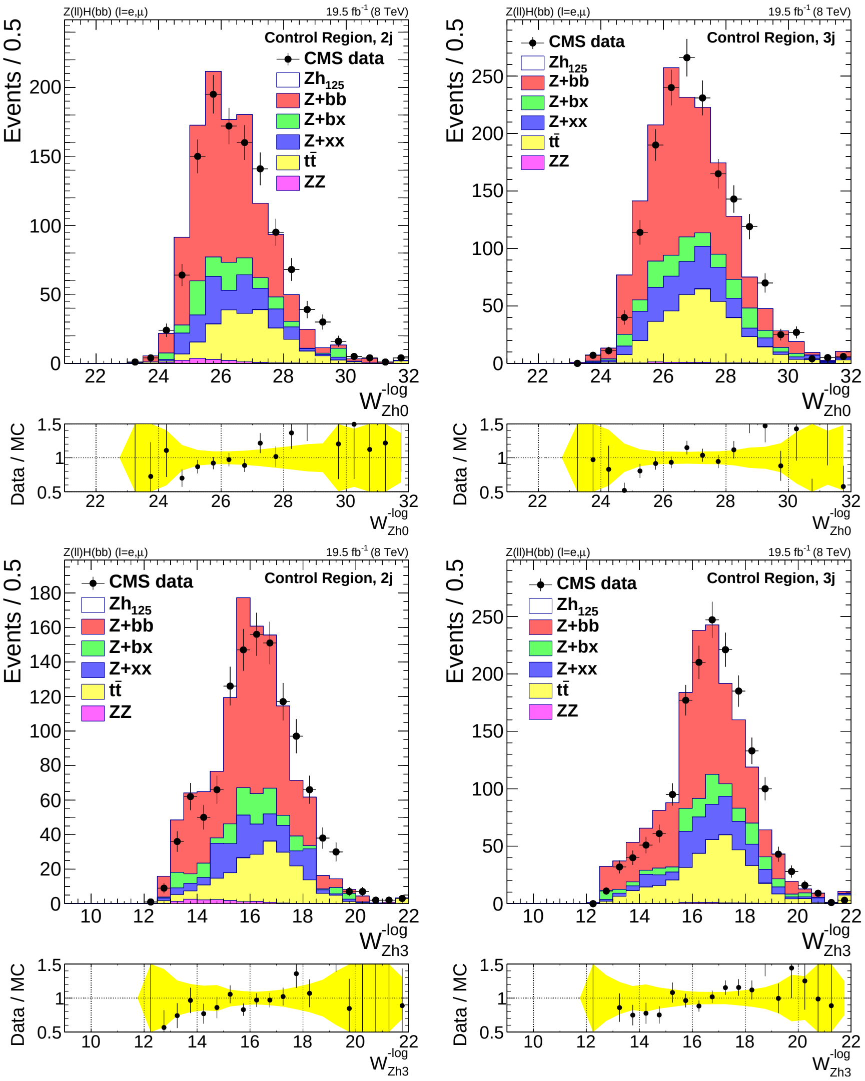


Figure A.12: Comparisons in the CR in the $2j$ category (left) and in the $3j$ category (right) of data and simulation for the ME weights related to the signal process. The top (bottom) plots represent W^{-log}_{Zh0} (W^{-log}_{Zh3}). Simulation samples are normalised using the SFs shown in Table 3.5. The signal is also shown separately normalised to 50 times its cross section. The last bin includes the overflow. In the ratio, the yellow band represents the statistical uncertainty from simulation.

Appendix B

Z+jets event weight computation

A set of exclusive Z+jets samples, binned according to different ranges of p_T^{ll} and HT, has been used for modelling the background coming from the Z+jets process in regions of the phase space poorly populated by the inclusive sample (see Section 4.2.1). These samples have been combined according to a reweighting procedure which accounts for the differential cross section of this process (see Table B.1) and the effective number of events processed (see Tables 3.2 and 4.1).

Sample	LO cross section (pb)
Z+jets: inclusive	2950
Z+jets: $50 < p_T^{ll} < 70$ GeV	93.8
Z+jets: $70 < p_T^{ll} < 100$ GeV	50.31
Z+jets: $p_T^{ll} > 100$ GeV	34.1
Z+jets: $p_T^{ll} > 180$ GeV	4.56
Z+jets: $200 < HT < 400$ GeV	19.73
Z+jets: $HT > 400$ GeV	2.826

Table B.1: `MadGraph` LO cross sections for the different simulated Z+jets samples.

First, the p_T^{ll} binned samples (50-70, 70-100, $>$100 and $>$180 GeV) are combined with the inclusive sample. The following bins are defined: 0-50, 50-70, 70-100, 100-180 and $>$180 GeV. The cross section, σ_i, in each bin is derived from Table B.1. Defining N_i^{pt} as the number of events generated for each exclusive sample, the weight for each

i-th p_T-binned sample is determined by:

$$w_i^{pt} = \frac{\sigma_i}{\sigma_{incl}} \times \frac{N_{tot}^{pt}}{N_i^{pt}}$$

where σ_{incl} is the cross section of the inclusive sample and N_{tot}^{pt} is the sum of the events from the five samples merged together.

Using a similar approach, the HT-binned Z+jets samples are also reweighted in order to be properly combined with the final sample. Three bins are defined: 0-200, 200-400 and $>$ 400 GeV. The weights account for the previous treatment of the events belonging to the p_T-binned samples, and thus the weights are defined as:

$$w_j^{HT} = \frac{\sigma_j}{\sigma_{incl}} \times \frac{N_{tot}^{HT}}{N_j^{HT-pt}}$$

where $N_j^{HT-pt} = N_j^{HT} + \sum_i N_{ij}^{pt} \cdot w_i^{pt}$, i.e. this accounts for the weighted numbers of events from the five first samples merged together which fall into the *j-th* HT bin, and N_{tot}^{HT} is the total number of events summing up all the samples.

The final weight for each event is then extracted as the ratio between the weighted numbers of events in the two-dimensional *i-th* p_T - *j-th* HT bin (with weights computed as before) and the total number of generated events in that bin, namely:

$$w_{ij} = \frac{(N_{ij}^{pt} \cdot w_i^{pt} + N_j^{HT}) \cdot w_j^{HT}}{N_{ij}^{pt} + N_{ij}^{HT}}$$

The w_{ij} are shown on Table B.2.

p_T (GeV) / $HT(GeV)$	0-50	50-70	70-100	100-180	$>$ 180
0-200	1.526	0.252	0.421	0.180	0.037
200-400	0.078	0.061	0.068	0.055	0.025
$>$ 400	0.026	0.024	0.025	0.023	0.015

Table B.2: Event weights for the different bins in p_T and HT of the merged Z+jets sample.

Appendix C

Additional kinematic comparisons for the BSM Higgs boson search

To complete the set of observables in Figures 4.17 and 4.18, Figures C.1, C.2 and C.3 are showing the p_T^{b1}, p_T^{b2}, CSV_{b1}, CSV_{b2}, E_T^{miss}, E_T^{miss} significance, m_{ll} and p_T^{llbb} observables. A good agreement is visible for all of them. In the legend, the *Zh* entry refers to the Zh_{125} process.

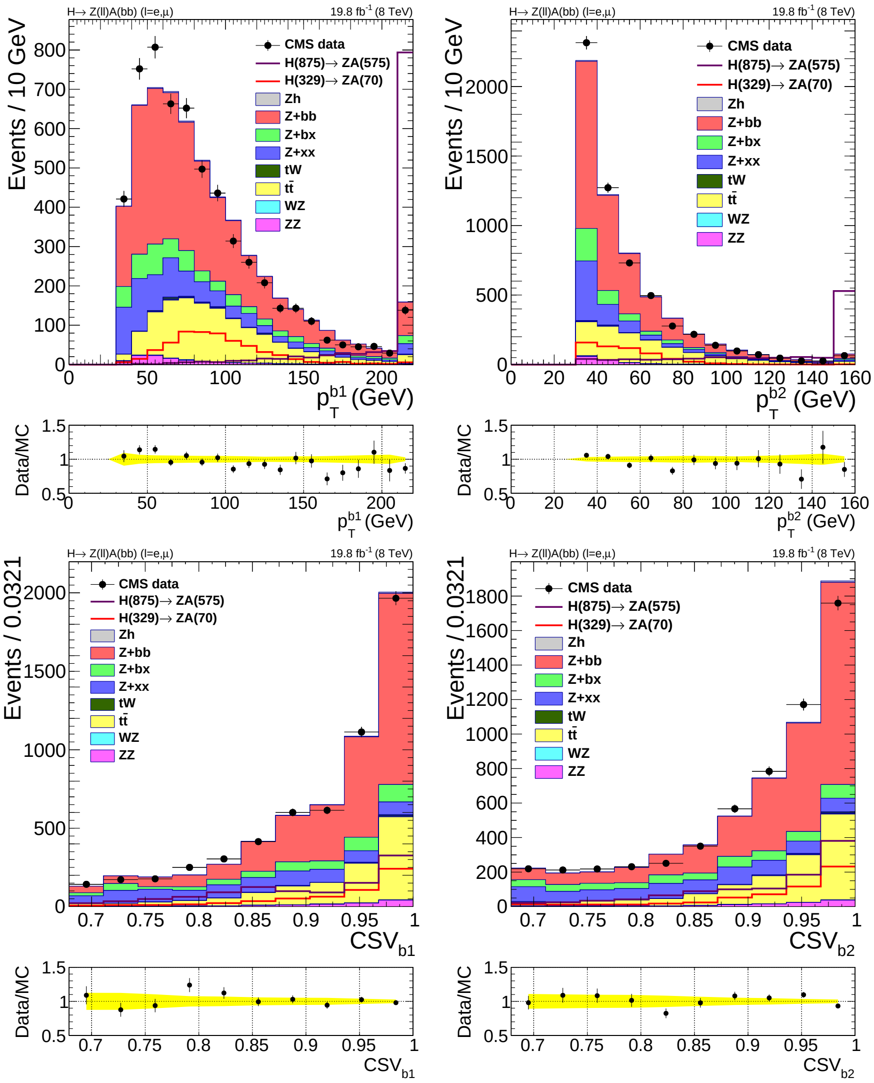


Figure C.1: Comparisons of the data to the expectation from the simulation of the backgrounds. The top (bottom) plots show the p_T^{b1} and p_T^{b2} (CSV_{b1} and CSV_{b2}) observables. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bins include the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

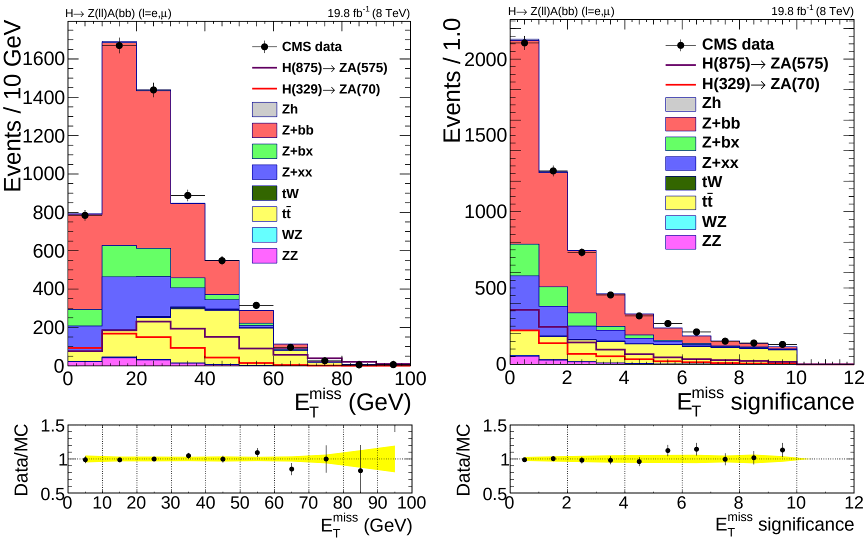


Figure C.2: Comparisons of the data to the expectation from the simulation of the backgrounds. The left (right) plot shows the E_T^{miss} (E_T^{miss} significance) observable. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bins include the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

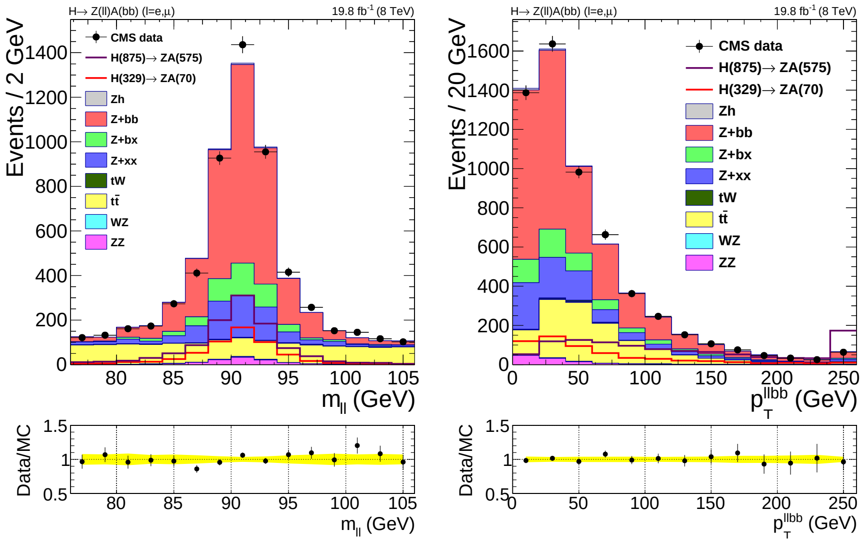


Figure C.3: Comparisons of the data to the expectation from the simulation of the backgrounds. The left (right) plot shows the m_{ll} (p_T^{llbb}) observable. On the ratio, the yellow band corresponds to the statistical uncertainty from simulation. The last bins include the overflow. Two signal samples, normalised to a cross section of 300 fb, are superimposed upon the background.

Appendix D

Study of the excess at high masses

In the scan of m_{bb}-m_{llbb}, shown in Figure 4.21, an excess with a significance of 2.9 s.d. is observed for the bin centred at m_{bb} = 575 GeV and m_{llbb} = 662 GeV. This bin is defined by a cut on m_{bb} between 446 and 704 GeV and a cut on m_{llbb} between 513 and 811 GeV. This corresponds to the largest excess observed. However several arguments disfavoured the signal hypothesis in this region. First, this bin is at the edge of the allowed kinematic regime ($m_{llbb} \approx m_{bb} + m_Z$). Also as shown in Figure 1.10, the probed signal has an extremely small cross section in this region due to the dominant (almost exclusive) decay to $t\bar{t}$. Still, some studies have been performed to check the data in this bin and check the consistency of the excess both with the background and signal-plus-background hypotheses. Comparisons between data and simulation are shown in Figures D.1 and D.2. No odd behaviour is observed in the data in this bin and the comparisons indicate that this excess can be compatible with a statistical fluctuation. A signal sample has been generated with masses corresponding to m_A = 500 GeV and m_H = 662 GeV and normalised to 3000 times the NNLO SUSHI cross section. This signal is displayed on top of the sum of the backgrounds in the same plots and also in Figure D.4.

The data are shown in Figure D.3 after subtracting the expectation from the simulated backgrounds. For m_{bb} the excess is fitted with a Gaussian function. The mean corresponds to m_{bb} = 445 GeV. Concerning m_{llbb}, no fit is performed as the excess is at the edge of the distribution. This exercise is not conclusive.

In order to further check the compatibility of the excess with the signal hypothesis, other kinematic comparisons are shown in Figure D.4. The conclusion is that the chosen signal hypothesis seems disfavoured for most of the variables. This is especially

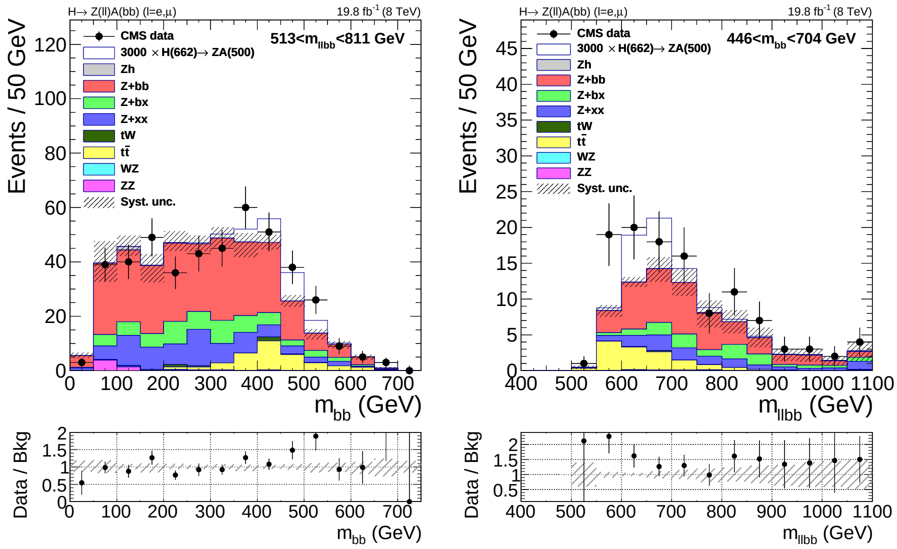


Figure D.1: Comparisons of the data to the expectation from the simulation of the backgrounds. The left (right) plots shows the m_{bb} (m_{llbb}) observable after applying the cut 513 GeV $< m_{llbb} <$ 811 GeV (446 GeV $< m_{bb} <$ 704 GeV). The hashed band on the sum of the backgrounds and in the ratio represents the systematic uncertainty. The signal added on top of the backgrounds corresponds to m_A = 500 GeV and m_H = 662 GeV. It is normalised to 3000 times the NNLO cross section for the model parameter listed in Table 4.2. The last bin includes the overflow.

true for the Centrality observables defined as $\Sigma p_T^i/\Sigma E_i$ in the *llbb* system rest frame with i running over the two leptons and the two b-tagged jets. All this strengthens the hypothesis that this excess is purely statistical.

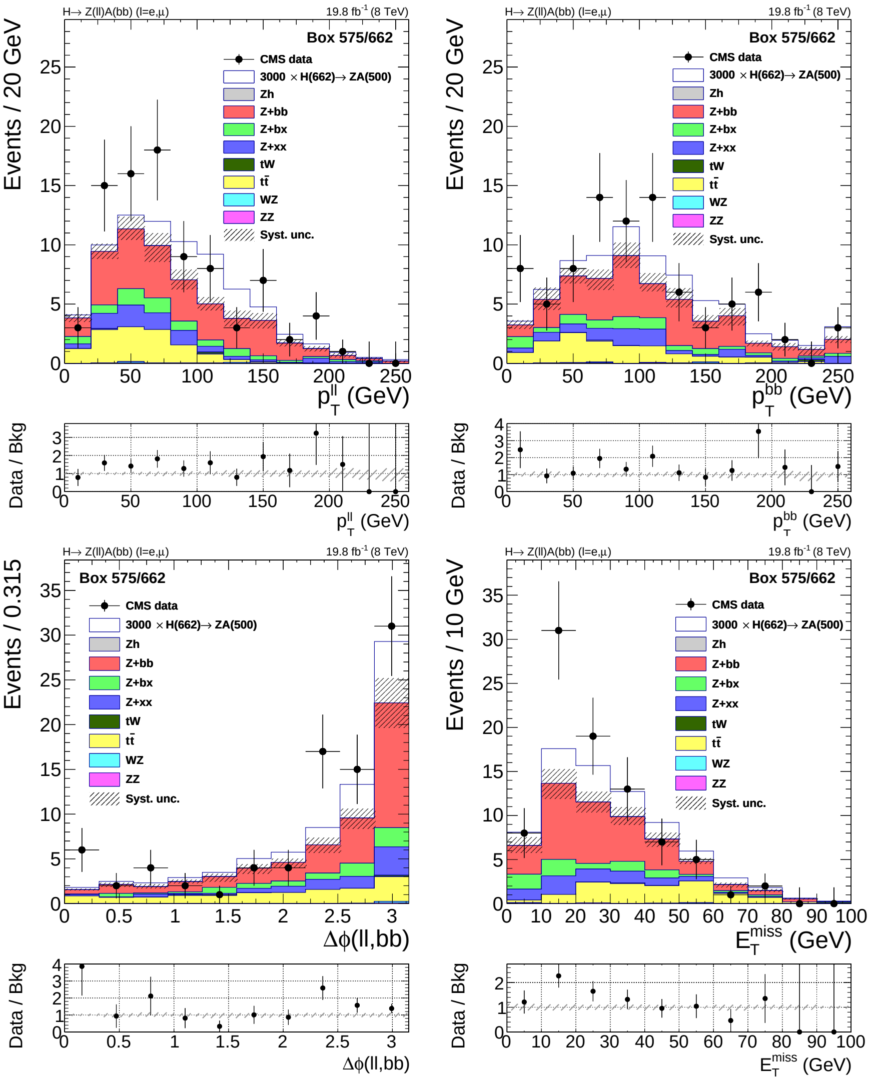


Figure D.2: Comparisons of the data to the expectation from the simulation of the backgrounds in the signal region centred on m_{bb} = 575 GeV and m_{llbb} = 662 GeV. The plots represent, from the top left to the bottom right, the p_T^{ll}, the p_T^{bb}, the $\Delta\phi(ll, bb)$ and the E_T^{miss} observables. The hashed band on the sum of the backgrounds and in the ratio represents the systematic uncertainty. The signal added on top of the backgrounds corresponds to $m_A = 500$ GeV and $m_H = 662$ GeV. It is normalised to 3000 times the NNLO cross section for the model parameter listed in Table 4.2. The last bins include the overflow.

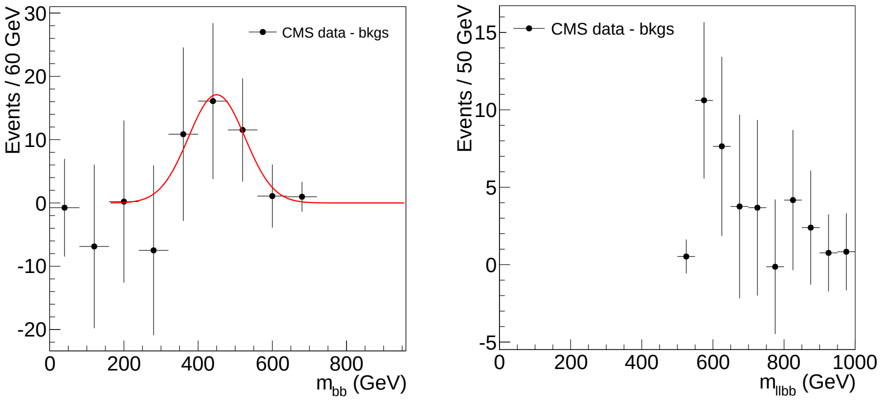


Figure D.3: Data with background estimated from simulation subtracted for m_{bb} (m_{llbb}) for the left (right) plot after cutting on 513 GeV $<$ m_{llbb} $<$ 811 GeV (446 GeV $<$ m_{bb} $<$ 704 GeV). The red curve represent the best fit by a Gaussian function for m_{bb}.

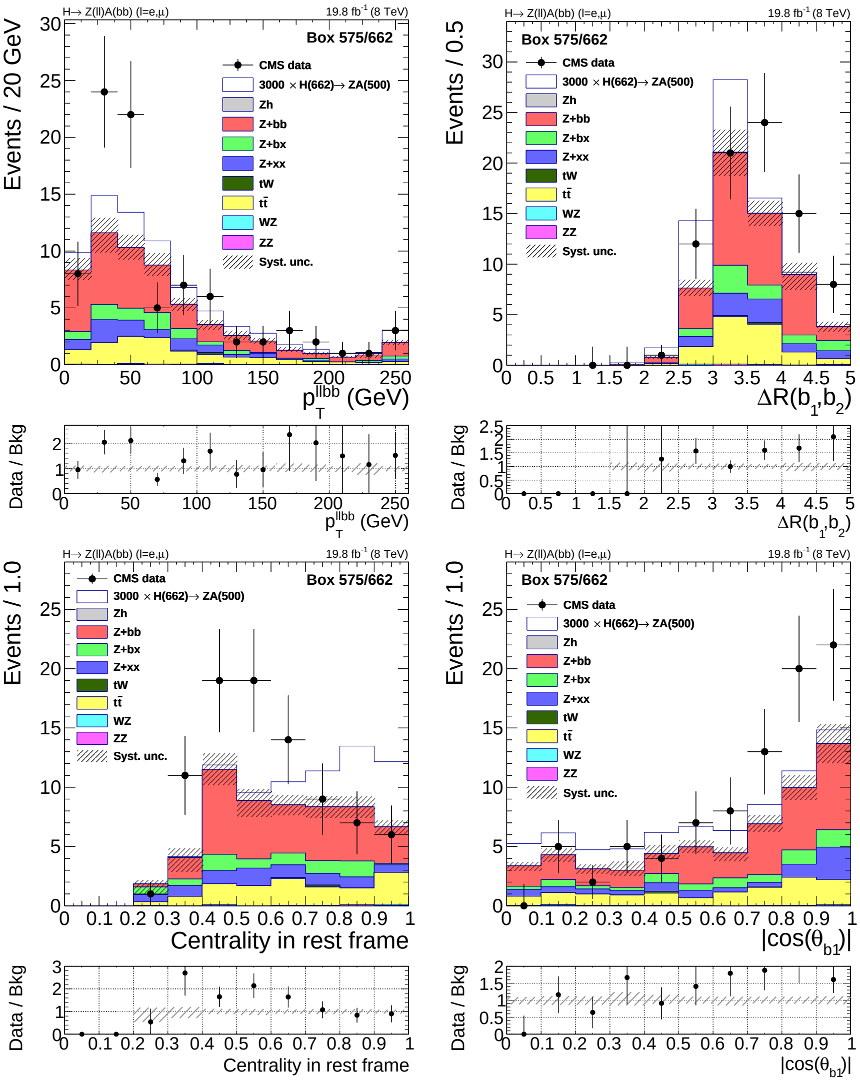


Figure D.4: Comparisons of the data to the expectation from the simulation of the backgrounds in the signal region centred on m_{bb} = 575 GeV and m_{llbb} = 662 GeV. The plots represent, from the top left to the bottom right, the p_T^{llbb}, the $\Delta R(b, b)$, the Centrality and the $|cos\theta_{b1}|$ (defined in Section 4.4.1) observables. The hashed band on the sum of the backgrounds and in the ratio represents the systematic uncertainty. The signal added on top of the backgrounds corresponds to $m_A = 500$ GeV and $m_H = 662$ GeV. It is normalised to 3000 times the NNLO cross section for the model parameter listed in Table 4.2. The last bins include the overflow.

Bibliography

[1] CMS Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, *Phys. Lett.* **B716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[2] ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, *Phys. Lett.* **B716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[3] F. Englert and R. Brout, “Broken Symmetry and the Mass of Gauge Vector Mesons”, *Phys. Rev. Lett.* **13** (1964) 321–323, doi:10.1103/PhysRevLett.13.321.

[4] P. W. Higgs, “Broken symmetries, massless particles and gauge fields”, *Phys. Lett.* **12** (1964) 132–133, doi:10.1016/0031-9163(64)91136-9.

[5] P. W. Higgs, “Broken Symmetries and the Masses of Gauge Bosons”, *Phys. Rev. Lett.* **13** (1964) 508–509, doi:10.1103/PhysRevLett.13.508.

[6] CMS Collaboration, “Evidence for the direct decay of the 125 GeV Higgs boson to fermions”, *Nature Phys.* **10** (2014) 557–560, doi:10.1038/nphys3005, arXiv:1401.6527.

[7] ATLAS Collaboration, “Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment”, *Eur. Phys. J.* **C76** (2016), no. 1, 6, doi:10.1140/epjc/s10052-015-3769-y, arXiv:1507.04548.

[8] ATLAS, CMS Collaboration, “Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS

Experiments", *Phys. Rev. Lett.* **114** (2015) 191803, doi:10.1103/PhysRevLett.114.191803, arXiv:1503.07589.

[9] S. P. Martin, "A Supersymmetry primer", doi:10.1142/9789812839657_0001, 10.1142/9789814307505_0001, arXiv:hep-ph/9709356. [Adv. Ser. Direct. High Energy Phys.18,1(1998)].

[10] G. C. Branco et al., "Theory and phenomenology of two-Higgs-doublet models", *Phys. Rept.* **516** (2012) 1–102, doi:10.1016/j.physrep.2012.02.002, arXiv:1106.0034.

[11] S. F. Novaes, "Standard model: An Introduction", in *Particles and fields. Proceedings, 10th Jorge Andre Swieca Summer School, Sao Paulo, Brazil, February 6-12, 1999.* 1999. arXiv:hep-ph/0001283.

[12] D. H. Perkins, "Introduction to high-energy physics; 4th ed.". Cambridge Univ. Press, Cambridge, 2000.

[13] D. J. Griffiths, "Introduction to elementary particles; 2nd rev. version". Physics textbook. Wiley, New York, NY, 2008.

[14] ATLAS, CMS Collaboration, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 045, doi:10.1007/JHEP08(2016)045, arXiv:1606.02266.

[15] CMS Collaboration, "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV", *Phys. Rev.* **D92** (2015), no. 1, 012004, doi:10.1103/PhysRevD.92.012004, arXiv:1411.3441.

[16] ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", *Phys. Lett.* **B726** (2013) 120–144, doi:10.1016/j.physletb.2013.08.026, arXiv:1307.1432.

[17] Particle Data Group Collaboration, "Review of Particle Physics", *Chin. Phys.* **C38** (2014) 090001, doi:10.1088/1674-1137/38/9/090001.

[18] F. Maltoni, G. Ridolfi, and M. Ubiali, "b-initiated processes at the LHC: a reappraisal", *JHEP* **07** (2012) 022, doi:10.1007/JHEP04(2013)095, 10.1007/JHEP07(2012)022, arXiv:1203.6393. [Erratum: JHEP04,095(2013)].

[19] L. Céard, "First measurement of the associated production of a Z boson with b jets at the LHC". PhD thesis, Louvain U., 2015.

[20] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", doi:10.5170/CERN-2013-004, arXiv:1307.1347.

[21] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV", *JHEP* **06** (2013) 081, doi:10.1007/JHEP06(2013)081, arXiv:1303.4571.

[22] G. C. Dorsch, S. J. Huber, K. Mimasu, and J. M. No, "Echoes of the Electroweak Phase Transition: Discovering a second Higgs doublet through $A_0 \to ZH_0$", *Phys. Rev. Lett.* **113** (2014), no. 21, 211802, doi:10.1103/PhysRevLett.113.211802, arXiv:1405.5537.

[23] J. E. Kim, "Light Pseudoscalars, Particle Physics and Cosmology", *Phys. Rept.* **150** (1987) 1–177, doi:10.1016/0370-1573(87)90017-2.

[24] A. Broggio et al., "Limiting two-Higgs-doublet models", *JHEP* **11** (2014) 058, doi:10.1007/JHEP11(2014)058, arXiv:1409.3199.

[25] F. Jegerlehner and A. Nyffeler, "The Muon g-2", *Phys. Rept.* **477** (2009) 1–110, doi:10.1016/j.physrep.2009.04.003, arXiv:0902.3360.

[26] T. Robens and T. Stefaniak, "Status of the Higgs Singlet Extension of the Standard Model after LHC Run 1", *Eur. Phys. J.* **C75** (2015) 104, doi:10.1140/epjc/s10052-015-3323-y, arXiv:1501.02234.

[27] I. P. Ivanov and C. C. Nishi, "Symmetry breaking patterns in 3HDM", *JHEP* **01** (2015) 021, doi:10.1007/JHEP01(2015)021, arXiv:1410.6139.

[28] J. M. Gerard and M. Herquet, "A Twisted custodial symmetry in the two-Higgs-doublet model", *Phys. Rev. Lett.* **98** (2007) 251802, doi:10.1103/PhysRevLett.98.251802, arXiv:hep-ph/0703051.

[29] S. de Visscher et al., "Unconventional phenomenology of a minimal two-Higgs-doublet model", *JHEP* **08** (2009) 042, doi:10.1088/1126-6708/2009/08/042, arXiv:0904.0705.

[30] CMS Collaboration, "Summary results of high mass BSM Higgs searches using CMS run-I data", Technical Report CMS-PAS-HIG-16-007, 2016.

[31] R. V. Harlander, S. Liebler, and H. Mantler, "SusHi: A program for the calculation of Higgs production in gluon fusion and bottom-quark annihilation

in the Standard Model and the MSSM", *Comput. Phys. Commun.* **184** (2013) 1605–1617, doi:10.1016/j.cpc.2013.02.006, arXiv:1212.3249.

[32] D. Eriksson, J. Rathsman, and O. Stal, "2HDMC: Two-Higgs-Doublet Model Calculator Physics and Manual", *Comput. Phys. Commun.* **181** (2010) 189–205, doi:10.1016/j.cpc.2009.09.011, arXiv:0902.0851.

[33] O. S. Bruning et al., "LHC Design Report Vol.1: The LHC Main Ring", (2004).

[34] C. Lefèvre, "The CERN accelerator complex. Complexe des accélérateurs du CERN", (Dec, 2008).

[35] M. Lamont, "Status of the LHC", *J. Phys. Conf. Ser.* **455** (2013) 012001, doi:10.1088/1742-6596/455/1/012001.

[36] LHCb Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008) S08005, doi:10.1088/1748-0221/3/08/S08005.

[37] ALICE Collaboration, "ALICE: Physics performance report, volume I", *J. Phys.* **G30** (2004) 1517–1763, doi:10.1088/0954-3899/30/11/001.

[38] ATLAS Collaboration, "ATLAS: Technical proposal for a general-purpose p p experiment at the Large Hadron Collider at CERN", (1994).

[39] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", *J. Phys. Conf. Ser.* **513** (2014) 022032, doi:10.1088/1742-6596/513/2/022032, arXiv:1311.4942.

[40] CMS Collaboration, "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV", *JINST* **10** (2015), no. 06, P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[41] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071.

[42] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[43] CMS Collaboration, "CMS Technical Design Report for the Pixel Detector Upgrade", Technical Report CERN-LHCC-2012-016. CMS-TDR-11, 2012.

[44] CMS Collaboration, "CMS Phase II Upgrade Scope Document", Technical Report CERN-LHCC-2015-019. LHCC-G-165, CERN, Geneva, Sep, 2015.

[45] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. Meth.* **A262** (1987) 444–450, doi:10.1016/0168-9002(87)90887-4.

[46] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.

[47] CMS Collaboration, "Alignment of the CMS tracker with LHC and cosmic ray data", *JINST* **9** (2014) P06009, doi:10.1088/1748-0221/9/06/P06009, arXiv:1403.2286.

[48] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression and related Optimisation Problems", *Proceedings of the IEEE* **86** (1998) doi:10.1109/5.726788.

[49] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, "Adaptive vertex fitting", *J. Phys.* **G34** (2007) N343, doi:10.1088/0954-3899/34/12/N01.

[50] CMS Collaboration, "Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV", *JHEP* **03** (2011) 136, doi:10.1007/JHEP03(2011)136, arXiv:1102.3194.

[51] CMS Collaboration, "Identification of b quark jets at the CMS Experiment in the LHC Run 2", Technical Report CMS-PAS-BTV-15-001, 2016.

[52] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[53] M. Botje et al., "The PDF4LHC Working Group Interim Recommendations", (2011). arXiv:1101.0538.

[54] S. Alekhin et al., "The PDF4LHC Working Group Interim Report", (2011). arXiv:1101.0536.

[55] NNPDF Collaboration, "Parton distributions with LHC data", *Nucl. Phys. B* **867** (2013) 244, doi:10.1016/j.nuclphysb.2012.10.003, arXiv:1207.1303.

[56] J. Alwall et al., "MadGraph 5 : Going Beyond", *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[57] T. Sjöstrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 physics and manual", *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[58] J. Alwall et al., “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[59] T. Sjostrand, S. Mrenna, and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1”, *Comput. Phys. Commun.* **178** (2008) 852–867, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[60] S. Alioli, P. Nason, C. Oleari, and E. Re, “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”, *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[61] M. Böhr et al., “Herwig++ physics and manual”, *The European Physical Journal C* **58** (2008), no. 4, 639–707, doi:10.1140/epjc/s10052-008-0798-9.

[62] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, “The tau decay library TAUOLA: Version 2.4”, *Comput. Phys. Commun.* **76** (1993) 361–380, doi:10.1016/0010-4655(93)90061-G.

[63] GEANT4 Collaboration, “GEANT4: A Simulation toolkit”, *Nucl. Instrum. Meth.* **A506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[64] CDF, D0 Collaboration, “Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron”, *Phys. Rev. Lett.* **109** (2012) 071804, doi:10.1103/PhysRevLett.109.071804, arXiv:1207.6436.

[65] ATLAS Collaboration, “Search for the bb decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector”, *Journal of High Energy Physics* **2015** (2015), no. 1, doi:10.1007/JHEP01(2015)069.

[66] CMS Collaboration, “Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks”, *Phys. Rev. D* **89** (Jan, 2014) 012003, doi:10.1103/PhysRevD.89.012003.

[67] CMS Collaboration, “Measurement of the production cross sections for a Z boson and one or more b jets in pp collisions at $\sqrt{s} = 7$ TeV”, *JHEP* **06** (2014) 120, doi:10.1007/JHEP06(2014)120, arXiv:1402.1521.

[68] CMS Collaboration, "Search for a standard model Higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method", *The European Physical Journal C* **75** (2015), no. 6, doi:10.1140/epjc/s10052-015-3454-1.

[69] A. Pin, "The Matrix Element Method at the LHC : a search for the associated production of Higgs and Z bosons". PhD thesis, Cathol. U. Louvain (main), 2013.

[70] P. Artoisenet, V. Lemaitre, F. Maltoni, and O. Mattelaer, "Automation of the matrix element reweighting method", *JHEP* **12** (2010) 068, doi:10.1007/JHEP12(2010)068, arXiv:1007.3300.

[71] B. Francois, "Méthode des éléments de matrice à l'ordre α_s pour la recherche du boson de Higgs au LHC", Master's thesis, 2013.

[72] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework", *Nucl. Instrum. Meth.* **A389** (1997) 81–86, doi:10.1016/S0168-9002(97)00048-X.

[73] CMS Collaboration, "Double Muon Trigger efficiency in 2012 data", Technical Report CMS-DP-2014-038, Oct, 2014.

[74] CMS Collaboration, "Muon ID and Isolation Efficiencies in 2012 RunAB", Technical Report CMS-DP-2012-025, Oct, 2012.

[75] CMS Collaboration, "Electron performance with 19.6 fb^{-1} of data collected at $\sqrt{s} = 8$ TeV with the CMS detector.", Technical Report CMS-DP-2013-003, Mar, 2013.

[76] CMS Collaboration, "Performance of b tagging at $\sqrt{s}$=8 TeV in multijet, ttbar and boosted topology events", Technical Report CMS-PAS-BTV-13-001, CERN, Geneva, 2013.

[77] W. Adam et al., "PAT: The CMS Physics Analysis Toolkit", *J. Phys. Conf. Ser.* **219** (2010) 32017, doi:doi:10.1088/1742-6596/219/3/032017.

[78] CMS Collaboration, "Study of the underlying event at forward rapidity in pp collisions at $\sqrt{s} = 0.9, 2.76,$ and 7 TeV", *Journal of High Energy Physics* **2013** (2013), no. 4, doi:10.1007/JHEP04(2013)072.

[79] A. Denner, S. Dittmaier, S. Kallweit, and A. Muck, "Electroweak corrections to Higgs-strahlung off W/Z bosons at the Tevatron and the LHC with HAWK", *JHEP* **03** (2012) 075, doi:10.1007/JHEP03(2012)075, arXiv:1112.5142.

[80] CMS Collaboration, “Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV”, Technical Report CMS-PAS-PFT-10-002, CERN, Geneva, 2010.

[81] CMS Collaboration, “Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET”, Technical Report CMS-PAS-PFT-09-001, CERN, 2009. Geneva, Apr, 2009.

[82] CMS Collaboration, “Particle-flow commissioning with muons and electrons from J/Ψ and W events at 7 TeV”, Technical Report CMS-PAS-PFT-10-003, CERN, 2010. Geneva, 2010.

[83] M. Cacciari and G. P. Salam, “Pileup subtraction using jet areas”, *Phys. Lett.* **B659** (2008) 119–126, `doi:10.1016/j.physletb.2007.09.077`, `arXiv:0707.1378`.

[84] M. Cacciari, G. P. Salam, and G. Soyez, “The Catchment Area of Jets”, *JHEP* **04** (2008) 005, `doi:10.1088/1126-6708/2008/04/005`, `arXiv:0802.1188`.

[85] CMS Collaboration, “2012 ECAL detector performance plots”, Technical Report CMS-DP-2013-007, Mar, 2013.

[86] CMS Collaboration, “Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV”, *JINST* **7** (2012) P10002, `doi:10.1088/1748-0221/7/10/P10002`, `arXiv:1206.4071`.

[87] M. Cacciari, G. Salam, and G. Soyez, “FastJet user manual”, *The European Physical Journal C* **72** (2012), no. 3, `doi:10.1140/epjc/s10052-012-1896-2`.

[88] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-k_t jet clustering algorithm”, *JHEP* **04** (2008) 063, `doi:10.1088/1126-6708/2008/04/063`, `arXiv:0802.1189`.

[89] CMS Collaboration, “Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS”, *JINST* **6** (2011) P11002, `doi:10.1088/1748-0221/6/11/P11002`, `arXiv:1107.4277`.

[90] CMS Collaboration, “Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s}$ = 8 TeV”, *JINST* **10** (2015), no. 02, P02006, `doi:10.1088/1748-0221/10/02/P02006`, `arXiv:1411.0511`.

[91] CMS Collaboration, “Missing transverse energy performance of the CMS detector”, *JINST* **6** (2011) P09001, `doi:10.1088/1748-0221/6/09/P09001`, `arXiv:1106.5048`.

[92] CMS Collaboration, “Measurement of W^+W^- and ZZ production cross sections in pp collisions at $\sqrt{s}$ = 8 TeV”, *Phys. Lett.* **B721** (2013) 190–211, doi:10.1016/j.physletb.2013.03.027, arXiv:1301.4698.

[93] CMS Collaboration, “CMS Luminosity Based on Pixel Cluster Counting - Summer 2012 Update”, Technical Report CMS-PAS-LUM-12-001, CERN, Geneva, 2012.

[94] CMS Collaboration, “Performance of the b-jet identification in CMS”, Technical Report CMS-PAS-BTV-11-001, CERN, Geneva, 2011.

[95] CMS Collaboration, “b-Jet Identification in the CMS Experiment”, Technical Report CMS-PAS-BTV-11-004, CERN, Geneva, 2012.

[96] CMS Collaboration, “Jet Energy Resolution in CMS at $\sqrt{s}$=7 TeV”, Technical Report CMS-PAS-JME-10-014, CERN, Geneva, 2011.

[97] A. L. Read, “Modified frequentist analysis of search results (the CLs method)”, CERN Report CERN-OPEN-2000-005, 2000.

[98] T. Junk, “Confidence level computation for combining searches with small statistics”, *Nucl. Instrum. Meth.* **A434** (1999) 435–443, doi:10.1016/S0168-9002(99)00498-2, arXiv:hep-ex/9902006.

[99] LHC Higgs Combination Group, ATLAS Collaboration, CMS Collaboration, “Procedure for the LHC Higgs boson search combination in Summer 2011”, Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug, 2011.

[100] L. Moneta et al., “The RooStats Project”, *PoS* **ACAT2010** (2010) 057, arXiv:1009.1003.

[101] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics”, *Eur. Phys. J.* **C71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z, arXiv:1007.1727.

[102] CMS Collaboration, “Measurement of the $pp \to ZZ$ production cross section and constraints on anomalous triple gauge couplings in four-lepton final states at $\sqrt{s}$ =8 TeV”, *Phys. Lett.* **B740** (2015) 250–272, doi:10.1016/j.physletb.2016.04.010,10.1016/j.physletb.2014.11.059, arXiv:1406.0113. [Erratum: Phys. Lett.B757,569(2016)].

[103] CMS Collaboration, “Pileup Jet Identification”, Technical Report CMS-PAS-JME-13-005, CERN, Geneva, 2013.

[104] L. Perrini, “Search for Higgs bosons decaying to tau leptons with the CMS experiment at the LHC”. PhD thesis, Louvain U., 2015.

[105] CMS Collaboration, “Search for H/A decaying into Z+A/H, with Z to ll and A/H to fermion pair”, Technical Report CMS-PAS-HIG-15-001, CERN, Geneva, 2015.

[106] CMS Collaboration, “Search for Neutral Resonances Decaying into a Z Boson and a Pair of b Jets or Tau Leptons”, *Phys. Lett.* **B759** (2016) 369–394, doi:10.1016/j.physletb.2016.05.087, arXiv:1603.02991.

[107] CMS Collaboration, “Search for H to Z(ll)+A(bb) with 2015 data”,.

[108] CMS Collaboration, “Search for a pseudoscalar boson decaying into a Z boson and the 125 GeV Higgs boson in *llbb* final states”, *Phys. Lett.* **B748** (2015) 221–243, doi:10.1016/j.physletb.2015.07.010, arXiv:1504.04710.

[109] J. de Favereau et al., “DELPHES 3: a modular framework for fast simulation of a generic collider experiment”, *Journal of High Energy Physics* **2014** (2014), no. 2, doi:10.1007/JHEP02(2014)057.

[110] CMS Collaboration, “Pileup Removal Algorithms”, Technical Report CMS-PAS-JME-14-001, CERN, Geneva, 2014.

[111] CMS Collaboration, “Measurement of WZ and ZZ production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$ in final states with b-tagged jets”, *Eur. Phys. J.* **C74** (2014), no. 8, 2973, doi:10.1140/epjc/s10052-014-2973-5, arXiv:1403.3047.

[112] CMS Collaboration, “Observation of the associated production of a single top quark and a W boson in *pp* collisions at $\sqrt{s}$ =8 TeV”, *Phys. Rev. Lett.* **112** (2014), no. 23, 231802, doi:10.1103/PhysRevLett.112.231802, arXiv:1401.2942.

[113] CMS Collaboration, “Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks”, *Phys. Rev. D* **89** (Jan, 2014) 012003, doi:10.1103/PhysRevD.89.012003.

[114] CMS Collaboration, “CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update”, Technical Report CMS-PAS-LUM-13-001, CERN, Geneva, 2013.

[115] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, “WW scattering at the CERN LHC”, *Phys. Rev. D* **65** (May, 2002) 096014, doi:10.1103/PhysRevD.65.096014.

[116] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider", *Phys. Rev. Lett.* **100** (Jun, 2008) 242001, doi:10.1103/PhysRevLett.100.242001.

[117] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, "Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks", *Phys. Rev. Lett.* **101** (2008) 142001, doi:10.1103/PhysRevLett.101.142001, arXiv:0806.0848.

[118] D. Krohn, J. Thaler, and L.-T. Wang, "Jet trimming", *Journal of High Energy Physics* **2010** (2010), no. 2, doi:10.1007/JHEP02(2010)084.

[119] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Techniques for improved heavy particle searches with jet substructure", *Phys. Rev. D* **80** (Sep, 2009) 051501, doi:10.1103/PhysRevD.80.051501.

[120] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches", *Phys. Rev. D* **81** (May, 2010) 094023, doi:10.1103/PhysRevD.81.094023.

[121] CMS Collaboration, "Studies of jet mass in dijet and W/Z + jet events", *JHEP* **05** (2013) 090, doi:10.1007/JHEP05(2013)090, arXiv:1303.4811.

[122] CMS Collaboration, "b-tagging in boosted topologies", Technical Report CMS-DP-2015-038, Aug, 2015.

[123] CMS Collaboration, "Search for heavy resonances in the W/Z-tagged dijet mass spectrum in pp collisions at 7 TeV", *Phys. Lett.* **B723** (2013) 280–301, doi:10.1016/j.physletb.2013.05.040, arXiv:1212.1910.

[124] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons", *Journal of High Energy Physics* **2014** (2014), no. 12, doi:10.1007/JHEP12(2014)017.

[125] CMS Collaboration, "Search for a Higgs Boson in the Mass Range from 145 to 1000 GeV Decaying to a Pair of W or Z Bosons", *JHEP* **10** (2015) 144, doi:10.1007/JHEP10(2015)144, arXiv:1504.00936.

[126] CMS Collaboration, "Boosted Top Jet Tagging at CMS", Technical Report CMS-PAS-JME-13-007, CERN, Geneva, 2014.

[127] CMS Collaboration, "Search for resonant $t\bar{t}$ production in lepton+jets events in pp collisions at $\sqrt{s} = 7$ TeV", *Journal of High Energy Physics* **2012** (2012), no. 12, doi:10.1007/JHEP12(2012)015.

[128] CMS Collaboration, “Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state”, *Journal of High Energy Physics* **2012** (2012), no. 9, doi:10.1007/JHEP09(2012)029.

[129] CMS Collaboration, “Searches for new physics using the $t\bar{t}$ invariant mass distribution in pp collisions at $\sqrt{s}$=8 TeV”, *Phys. Rev. Lett.* **111** (Nov, 2013) 211804, doi:10.1103/PhysRevLett.111.211804.

[130] M. Wobisch and T. Wengler, “Hadronization corrections to jet cross-sections in deep inelastic scattering”, in *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999.* 1998. arXiv:hep-ph/9907280.

[131] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, “Better jet clustering algorithms”, *JHEP* **08** (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.