Search for rare processes with a Z+bb signature at the LHC, with the matrix element method

Camille Beluffi

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE Physique et Chimie-Physique

Institut pluridisciplinaire Hubert Curien (IPHC)

THÈSE présentée par :

Camille BELUFFI

soutenue le : **14 octobre 2015**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Physique des particules élémentaires

Recherche de processus rares avec la signature Z+bb au LHC, à l'aide de la Méthode des Éléments de Matrice

THÈSE dirigée par :

M. GELÉ Denis	Chargé de recherche HDR, université de Strasbourg
M. LEMAÎTRE Vincent	Professeur, université de Catholique de Louvain-la-Neuve

RAPPORTEURS :

Mme VANDER DONCKT Muriel	Assistant professeur, université Claude Bernard Lyon I
M. LUCOTTE Arnaud	Chargé de recherche HDR, LPSC, Président

AUTRES MEMBRES DU JURY :

M. CORTINA Eduardo	Professeur, université de Catholique de Louvain-la-Neuve
M. DELAERE Christophe	Professeur, université de Catholique de Louvain-la-Neuve
M. SCODELLARO Luca	Chargé de recherche, université de Cantabria
M. TOSI Silvano	Chargé de recherche, université de Gênes

A Frédéric,
A Mamie

Tant que la science progressera, tant que les mathématiques continueront à découvrir des mondes incroyables où deux fois deux n'auraient jamais l'idée d'égaler quatre, de nouvelles idées surgiront chez ceux qui laissent leur esprit vagabonder, passeport à la main, aux frontières du Possible.

—Arthur C. Clarke, *Odyssées*

Où vais-je ? Je ne sais. Mais je me sens poussé
D'un souffle impétueux, d'un destin insensé.

—Victor Hugo, *Hernani*

Remerciements

Après quatre années de travail, ce manuscrit voit le jour. Mais il s'agit pour moi bien plus que de pages rassemblées: de nombreuses personnes laissent leur marque dans ce livre, et j'aimerais les en remercier.

Tout d'abord, je tiens à remercier mes directeurs de thèse, Vincent et Denis, pour avoir accepté de diriger mon projet de recherche, et m'avoir accompagné durant ces quatre années. J'ai la sensation d'avoir eu le privilège d'effectuer cette thèse guidée par deux très grands scientifiques, passionnés par leur domaine, et je suis sûre que de cet apprentissage, certaines qualités de travail me resterons toujours.

Je voudrais également exprimer ma gratitude aux membres de mon jury, pour avoir eu le courage de venir jusqu'à Louvain-la-Neuve à deux reprises. Merci d'avoir lu ma thèse avec autant de sérieux, les corrections que vous y avez apporté m'ont été d'une grande aide.

Je tiens à remercier également Jérémy qui m'a d'abord encadrée en stage de Master puis dans le cadre de ma cotutelle. Merci d'avoir cru en moi et de m'avoir appris à être plus autonome et critique envers mon travail. I would also like to warmly thank Jesus for answering all my silly questions any time I had one, and being such a nice co-office mate !

J'aimerais remercier le groupe CMS de Strasbourg pour avoir accepté la cotutelle et m'avoir chaleureusement accueilli à chacun de mes déplacements. Je voudrais également remercier du fond du coeur tous les scientifiques que j'ai côtoyé à CP3: si il y a bien une chose dont je peux attester, c'est qu'à CP3, une aide est toujours apportée à

ceux qui la demande ! En particulier un grand merci à mon groupe de travail: Alexandre, Adrien, Arnaud, Brieuc, Christophe, Miguel, Tristan, Roberto, les Sébastien et Olivier. Merci également à Loïc, Pierre et Andrea pour l'aide précieuse qu'ils m'ont apporté en fin de rédaction. Et bien sur, un immense merci à Jérôme, Pavel et Juan pour avoir toujours résolu mes soucis d'ordinateur avec autant de gentillesse et de bonne volonté. Enfin, une attention toute particulière pour Ginette et Luc, qui auront pris soin de moi pendant tout ce temps. Je regretterai ces séances de natation avec vous Ginette !

Une pensée très forte pour mon club de plongée avec qui j'ai découvert ce sport formidable et qui m'ont aidé à oublier les soucis de la semaine bien plus d'une fois le vendredi soir !

Ces dernières années ont été riches en rencontres: de formidables personnes ont embelli mon quotidien et j'aimerais les en remercier. JB, Laetitia et Alban Choubidou, un grand merci pour ces après midi jeux de société et tous ces bons moments ! Sameh, merci pour ton soutien et ton amitié, et merci de m'avoir appris à réussir les macarons ! Rhys, I hope you will accept this formal high-five, as a way to thank you for the kind welcomes and good times we had in Prague. Angélique et William, j'attends avec impatience de fêter tout cela avec vous autour d'un crock-pot (ou tzatziki...)! Merci aussi à Sylvain et Jules pour ces bons moments à faire de la corde à sauter (et si c'est mal interprété tant pis pour vous :p). Merci Jo, je te dois toujours une fière chandelle ;) Un petit coucou à Ben et Emma, avec la promesse que oui, on viendra voir votre "dolmen" dans les volcans ! Et bien sûr merci à Juliiiiie qui m'a sauvé alors que je risquais ma vie en escaladant la cathédrale.

Ceci m'amène à remercier mon groupe de rafounes Strasbourgeoises, sans qui je ne serai sans doute même pas arrivée jusqu'en fin d'école d'ingénieurs: un énorme merci et pleins de poutoux à GG, Hélène, Anaïs, Rozenn, Bug, Delphine, Itia, Benoît, Nico, Damian, Julien, Nathalie et Rémi. Un poutou encore plus baveux pour GG, Hélène et Anaïs, qui ont pris soin de moi au moment où j'en avais le plus besoin, et qui ont accepté de manger mes rainbow cupcakes. Ces années passées à Strasbourg font parties des plus belles que j'ai vécu, grâce à vous. Merci de m'avoir soutenue, bichonnée, et de m'avoir appris qu'on est beaucoup plus fort quand on n'est pas seul. MUT MUT !!!

Une autre personne que j'aimerais remercier ici est Anne-Sophie Cordan, qui m'a beaucoup aidé pendant mes deux anneés chargées à l'ENSPS, pour former mon projet ERASMUS avec Pierre Graebling, et grâce à qui j'ai pu décrocher un stage au CERN.

Thanks a lot to all my Swedish-Lund friends, with who I had such great time in Sweden, under the rain in Norway and buried in the snow in Poland: Carolin, Marius, Marion, Petra, Erik, Bastiaan, Boris and also Mei, Ayano and Rachel that I had the chance to see again in Japan !

Évidemment, un très grand merci à Élodie pour toutes ces années d'amitié, ces bons souvenirs et folles anecdotes à travers l'Europe. J'ai toujours pu compter sur toi en cas de besoin où lorsqu'il fallait trouver des surnoms immondes aux personnes que je n'appréciais pas, et je ne te remercierai jamais assez pour cela !

Merci à ma famille de m'avoir soutenu pendant cette thèse, et d'avoir été là dans les moments de doute. Je sais qu'il n'a pas toujours été évident pour vous de comprendre où je voulais en venir avec mes études, mais vous m'avez toujours fait confiance malgré tout et cru en moi.

Enfin, je voudrais remercie la personne que j'aime le plus au monde, et sans qui cette thèse contiendrait 4 fois plus de fautes d'orthographe: Frédéric, merci pour tout, ton soutien, ta bonne humeur quotidienne, ta présence, ton stock de photos de panda roux pour les moments de déprime. Tu es de loin la chose la plus belle chose qui me soit arrivée aux cours de ces dernières années.

Contents

Introduction

Since the very first collisions started to occur at the LHC, the main goal of the CERN has remained the same: to test the validity of the Standard Model. This theory, elaborated in the 60's, has been very successful so far, with related discoveries that have continuously confirmed the theoretical predictions. On July the 4th, 2012, the last missing piece of the Standard Model was discovered: the Higgs boson. But, despite this big success, several questions remain opened: why are the fermions divided in three generations? Why is there more matter than anti-matter in the Universe? What is the origin of the Dark Matter? Is there a deepest understanding of the Higgs interactions than the one given by the Brout, Englert and Higgs mechanism? Those questions suggest that the Standard Model is only a specific case, belonging to a more extended theory. In this context, it is necessary to accurately measure the free parameters of the Standard Model to look for hints pointing towards a even more general theory. In particular, the coupling between the Higgs boson and the fermions, which has not been experimental proven yet, is of main interest.

This thesis focuses on testing the Standard Model by studying various properties of events recorded by the CMS detector at a center-of-mass energy of 8 TeV, presenting a final state with two leptons, two b jets and no transverse missing energy. This topology (named $llbb$ in the following) can be induced by several processes and experimental signatures that should be carefully handled, such as the b jets. To identify these jets, sophisticated algorithms are needed. One of them, called "Jet Probability", had to be calibrated during the data taking periods and constituted an important part of the present work. In particular, a new framework needed to be developed and was a key tool for the study of Jet Probability at high p_T.

An important work has been dedicated to the understanding of the $llbb$ topology with an additional restriction that both leptons come from a Z boson. In this context, a detailed study of the production of a Higgs boson in association with a Z boson, which decays in two leptons, while the Higgs decays into a pair of b quarks, has been performed. A first analysis has been designed, using the Matrix Element Method and the brand new 2012 reprocessed data. The Jet Probability tagger was chosen for b-tagging, in order to compare its performance with the mostly used tagger, the Combined Secondary Vertex. This analysis also paved the way to the development of a model-independent search of signatures beyond the Standard Model using the same final state.

In the first chapter, a brief description of the theoretical model of fundamental interactions between elementary particles and aspects of particle physics are presented, leading to a discussion on the Brout-Englert-Higgs mechanism, and the recent observation of the Higgs particle. The Matrix Element Method, used in the analyses described in this thesis, is also introduced in this chapter. The experimental context is exposed in the second chapter, with a presentation of the LHC and the CMS detector. General particle reconstruction and identification is described together with the particle flow algorithm, leading to the elaboration of one of the key ingredient required by the Matrix Element Method, namely the transfer functions. These functions allow to reconstruct the partons kinematics including all theoretical and experimental effects. In the third chapter, a detailed review the different algorithms available in CMS for the identification of b jets is developed, with a focus on the Jet Probability tagger. The calibration of this algorithm is explained as well as the study that has been done in order to improve its efficiency at high energy. The main analysis of this thesis is presented in the fourth chapter: the search for the Higgs boson decaying into two b quarks, produced in association with a Z boson. This search is performed using the Matrix Element Method and two different algorithms for the b jets identification, in order to compare their performance. Finally, the last chapter presents a model-independent search of signatures beyond the Standard Model, where its sensitivity is evaluated by using this approach for the Higgs search and by injecting a hypothetical signal based on a two Higgs doublet model.

Chapter 1

Theoretical background

1.1 The Standard Model

1.1.1 The components

The Standard Model (SM) [1][2][3] is a Yang and Mills theory that was created in the 60's. Based on quantum field theory, it describes the most elementary constituents of matter, represented by quantum fields, and their interactions, resulting from the gauge invariance of the $SU(2) \times SU(3) \times U(1)$ group. Adding the Brout-Englert-Higgs mechanism led to the creation of the electro-weak theory in 1967.

The SM predicts the existence of 61 particles and anti-particles. These particles are divided into two families: the fermions, of spin $\frac{1}{2}$, and the bosons, with an integer spin. Fermions obey a Fermi-Dirac statistical rule and Pauli exclusion principle: identical fermions cannot occupy the same place at the same time (they can not have the same quantum numbers). Bosons, in contrast, may be described by the same quantum numbers as they follow the statistical rules of Bose-Einstein.

The fermions compose the matter and are divided into three generations. The first one includes the electron (e^-), the first discovered fermion, revolving around the atom nucleus. The muon (μ^-) and the tau (τ^-) are replica of the electron, with higher masses. For each one of these three leptons, a neutrino is associated. Inside the atom nucleus are quarks, also distributed into three families: up (u) and down (d), charm (c) and

Particle	Mass	Electric charge
Electron (e^-)	0,511 MeV	-1
Electronic neutrino (ν_e)	$<$ 2,2 eV	0
Up (u)	1,7 - 3,3 MeV	2/3
Down (d)	4,1 - 5,8 MeV	-1/3
Muon (μ^-)	105,7 MeV	-1
Muonic neutrino (ν_μ)	$<$ 0,19 MeV	0
Charm (c)	1270 MeV	2/3
Strange (d)	101 MeV	-1/3
Tau (τ^-)	1,777 GeV	-1
Tauic neutrino (ν_τ)	$<$ 18.2 MeV	0
Top (t)	173,3 GeV	2/3
Beauty (b)	4,3 GeV	-1/3

Table 1.1: List of the Standard Model fermions, ordered by generation, with their respective mass and electric charge [4].

strange (s), then top (t) and beauty (b). Quarks have another property, with six manifestations: the color. It can be red, blue, green, anti-red, anti-blue and anti-green. The anti-colors belong, appropriately, to the anti-quarks.

Four fundamental forces are present in nature: gravitation, electromagnetism, strong force and weak force. The fermions properties have been summarized in Table 1.1. Interactions between particles are described as an exchange of mediator particles, the bosons, specific to each field: gluons for the strong field, photons for the electromagnetic field, and the $W^\pm/Z^0$ bosons for the weak field. The Higgs boson plays a unique role in the Standard Model, by explaining why the $W^\pm$ and Z are massive.

So far, this theory has been very successful: all predicted particles have been observed: the bosons Z and $W^\pm$ have been discovered at CERN in 1983 by the UA1 [5] and UA2 [6] experiments, the top quark in 1995 at the Tevatron by the CDF and D0 [7] experiments. More recently the discovery of the Higgs boson has been announced by the ATLAS and CMS experiments at CERN [8][9][10].

1.1.2 The example of the electromagnetism

A good example to get a notion of the gauge invariance principle, key property of the SM, is given when building the Lagrangian of the electromagnetic interaction, called Quantum ElectroDynamics (QED).
Let's start from the Dirac Lagrangian where $\psi(x)$ represents a free electron of mass m and charge e:

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

If a local phase transformation is applied, the Lagrangian density of 1.1 must remain unchanged under $U(1)$ transformation:

$$\psi(x) \rightarrow \psi^{'}(x) = e^{-ief}\psi(x) \qquad \bar{\psi}(x) \rightarrow \bar{\psi}^{'}(x) = e^{ief}\bar{\psi}(x) \tag{1.2}$$

where $f(x)$ is an arbitrary function. The massless term of 1.1 becomes:

$$\bar{\psi}(x)i\gamma^{\mu}\partial_{\mu}\psi(x) \rightarrow \bar{\psi}(x)i\gamma^{\mu}\partial_{\mu} + e\bar{\psi}(x)\gamma^{\mu}(\partial_{\mu}f(x))\psi(x) \tag{1.3}$$

To remove the additional term, a vectorial field $A_{\mu}(x)$ is added and transforms such as:

$$A_{\mu}(x) \rightarrow A^{'}_{\mu}(x) = A_{\mu}(x) + \frac{1}{e}\partial_{\mu}f(x) \tag{1.4}$$

A covariant derivate can then be defined by:

$$D_{\mu}\psi(x) = (\partial_{\mu} + ieA_{\mu}(x))\psi(x) \tag{1.5}$$

where $D_{\mu}\psi(x)$ transforms such as:

$$D_{\mu}\psi(x) \rightarrow e^{-ief(x)}D_{\mu}\psi(x) \tag{1.6}$$

That way, 1.1 remains invariant under $U(1)$ transformation. Now, in order to give dynamic to these gauge fields, a gauge tensor $F_{\mu\nu}$ has to be introduced:

$$F_{\mu\nu} = \partial_{\mu}A_{\nu}(x) - \partial_{\nu}A_{\mu}(x) \tag{1.7}$$

A final expression of 1.1 can then be written as:

$$\mathcal{L} = \bar{\psi}(x)(i\gamma^{\mu}\partial_{\mu} - m)\psi(x) - e\bar{\psi}(x)\gamma^{\mu}A_{\mu}(x)\psi(x) - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{1.8}$$

The second term of the Lagrangian describes the coupling of the electron and the boson of the electromagnetic field, the photon, while the last term represents the propagation of a free photon. It is interesting to notice that this last term is massless and if the vector field $A_\mu(x)$ would have been added with a mass term, the Lagrangian would not be invariant under gauge transformation. Therefore, for the theory to work, the photon has to be massless.

1.1.3 The strong interaction

The existence of particles made of quarks has been experimentally proven. These fermions have an additional quantum number with respect to leptons, carried by gluons: the color. The color property was initially introduced to allow baryons made up of three quarks with the same flavor in the fundamental state (such as the Δ^{++}) to still satisfy the Pauli exclusion principle: quarks forming the same hadron must have different colors. As a result, all three quarks in a baryon are assumed to carry different colors, and a meson must contain a colored quark and anti-quark of the corresponding anti-color.

The theory describing quarks interactions is described by a non-abelian gauge theory from the group SU(3), called Quantum ChromoDynamics (QCD). The gluon fields introduced in the Lagrangian lead to a coupling term between the gluons, which means that unlike the photon, a gluon can interact with another gluon, via the diagrams shown in Fig. 1.1.

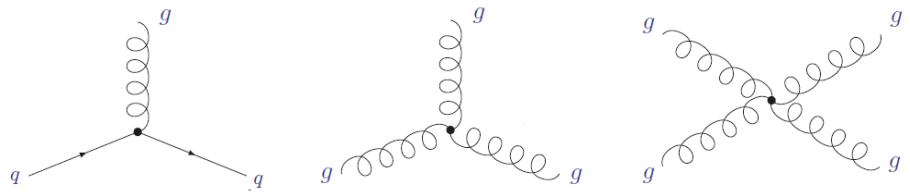

Figure 1.1: Representation of different couplings between quarks and gluons and gluons with gluons [11].

An important aspect of the strong force is its coupling constant: as the other coupling constants, it depends on the transfered energy Q; however in QCD, the dependence in energy is proportional to $\frac{1}{ln(Q^2)}$, meaning that at high energy, the coupling constant is small so quarks and gluon can be considered as free particles. On the other hand, as a quark/anti-quark pair separates, the gluon field forms a "string" of color field between

them. There is a limit to the distance that two quarks can be separated from each other, which is about the diameter of a proton: at this point it is more energetically favorable for a new quark/anti-quark pair to spontaneously appear, than to allow the string to extend further (see Fig. 1.2). This production of quarks in cascade results in the formation of hadrons out of quarks, this phenomenon being called hadronization. As a result, when quarks are produced in particle accelerators, instead of seeing the individual quarks in detectors, "jets" of many color-neutral particles (mesons and baryons), clustered together, are detected. This phenomenon is known as the color confinement: at distances comparable to the diameter of a proton, the strong interaction between quarks is about 100 times greater than the electromagnetic interaction. At smaller distances, however, the strong force between quarks becomes weaker, and the quarks begin to behave like independent particles, an effect known as asymptotic freedom.

As electro-magnetic particles, when particles contained in a jet interact with matter, a Parton Shower (PaS) is initiated: there is an emission of new particles from the initial one.

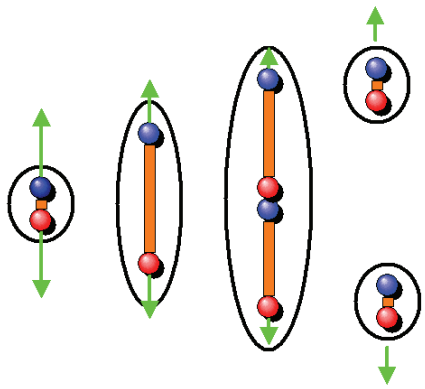

Figure 1.2: Principle of the hadronization: as the distance between two quarks increases, the "string" that binds them together starts to break, creating a new quark/anti-quark pair [12].

1.1.4 The electro-weak sector

A Dirac field ψ representing a fermion can be expressed as the sum of a left-handed part ψ_L and a right-handed part ψ_R such as ψ=ψ_L+ψ_R where:

$$\psi_L(x) = \frac{1-\gamma_5}{2}\psi(x) \qquad \psi_R(x) = \frac{1+\gamma_5}{2}\psi(x) \tag{1.9}$$

Since the weak force is known to maximally violate parity, it only involves left-handed fermions [13]. The weak interaction description is then based on the $SU(2)_L$ symmetry group, associated to the weak isospin T (and its projector T_3). Particles can be classified this way: left-handed fermions are represented by weak isospin doublets, while right-handed fermions are formed by weak isospin singlets. The $SU(2)_L \times U(1)_Y$ group has four massless gauge bosons: three coming from $SU(2)_L$, the W^i_μ, and the one from $U(1)_Y$, denoted B_μ.

For example, let's take the first generation of quarks, that can be written as:

$$Q_L = \begin{pmatrix} u \\ d \end{pmatrix}_L, u_R, d_R \tag{1.10}$$

The Lagrangian density of the electroweak sector is built to be invariant under transformations of the gauge group $SU(2)_L \times U(1)_Y$, where Y represents the hypercharge, related to the electric charge Q and the isospin T such as:

$$Y = 2(Q - T_3) \tag{1.11}$$

The Lagrangian can then be written as:

$$\mathcal{L} = \bar{Q_L}(i\gamma^\mu D_\mu)Q_L + \bar{u}_R(i\gamma^\mu D_\mu)u_R + \bar{d}_R(i\gamma^\mu D_\mu)d_R - \frac{1}{4}W^{\mu\nu}W_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \tag{1.12}$$

and the covariant derivative required to insure the local gauge invariance is given by:

$$D_\mu = \partial^\mu - igJW^\mu - ig'\frac{Y}{2}B_\mu \tag{1.13}$$

where g and g' are the coupling constants for $SU(2)_L$ and $U(1)_Y$ respectively, and J a $SU(2)_L$ group generator (represented by the Pauli matrices).

So far, the vector bosons $W^\pm$ and Z^0 are massless, to preserve the gauge invariance of $SU(2)_L \times SU(1)_Y$.

1.1.5 Spontaneous symmetry breaking

In order to fix the issue of massless $W^{\pm}$ and Z^0 bosons in the electro-weak theory, a mechanism of spontaneous symmetry breaking was proposed by Brout, Englert and Higgs [14][15][16].

A complex scalar field of type $SU(2)_L$, with no electric charge, is added in the Lagrangian:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \psi_1 + i\psi_2 \\ \psi_3 + i\psi_4 \end{pmatrix} \tag{1.14}$$

which leads to the following equation:

$$\mathcal{L} = D_\mu \phi D^\mu \phi^\dagger - V(\phi) = D_\mu \phi D^\mu \phi^\dagger + \mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2 \tag{1.15}$$

where $V(\phi)$ represents the Higgs potential. A schematic 3D representation of $V(\phi)$ can be seen on Fig. 1.3. When $\lambda < 0$ and $\mu^2 > 0$, the field represented by 1.15 is massive and reaches its fundamental state when $\phi = 0$. However if $\lambda > 0$, there is an infinite number of minima. If one minimum is chosen, the field ϕ acquires an expectation value:

$$< \phi >= \nu \equiv \frac{\mu}{\sqrt{2}} \tag{1.16}$$

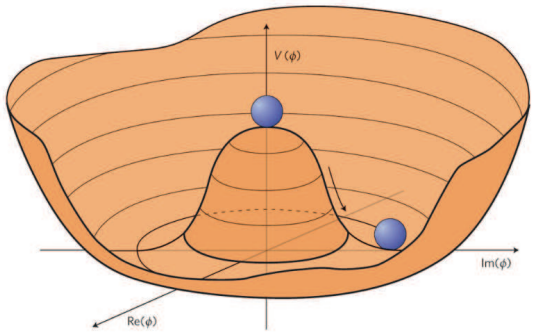

Figure 1.3: The Higgs potential energy density $V(\phi)$, when $\mu^2 < 0$. The vacuum state is the lowest-energy state. Such potential features many states of minimum energy: all the points randomly chosen along the bottom of the potential curve [17].

It is then possible to break the symmetry by choosing a fundamental state such as:

$$\phi = \begin{pmatrix} 0 \\ \nu/\sqrt{2} \end{pmatrix} \tag{1.17}$$

Once the symmetry has been broken, the vector bosons $W^{\pm}$, Z^0 and γ can be expressed from the four gauge fields previously introduced. The W^+ and W^- fields are electrically charged and can be defined as:

$$W^+ = \frac{W_\mu^1 - iW_\mu^2}{\sqrt{2}} \tag{1.18}$$

$$W^- = \frac{W_\mu^1 + iW_\mu^2}{\sqrt{2}} \tag{1.19}$$

The Z_μ corresponds to the Z boson field and is expressed by:

$$Z_\mu = \cos\theta_W W_\mu^3 + \sin\theta_W B_\mu \tag{1.20}$$

The photon field can also be defined via the fields W_μ and B_μ:

$$A_\mu = -\sin\theta_W W_\mu^3 + \cos\theta_W B_\mu \tag{1.21}$$

The θ_W angle represents the mixing angle (also called Weinberg angle). The coupling constant g and g' are related to θ_W and the electric charge e by:

$$\tan\theta_W = \frac{g'}{g} \tag{1.22}$$

$$g' cos\theta_W = e \tag{1.23}$$

The experimental value θ_W has been measured and his value can be expressed such as $\sin^2\theta_W$= 0.23122 $\pm$ 0.00015. From here, it is possible to express the masses of the $W^{\pm}$ and Z boson:

$$m_W = \frac{g\nu}{2} \qquad\qquad m_Z = \frac{\nu}{2}\sqrt{g^2 + g'^2} \tag{1.24}$$

In addition, the Higgs field gives birth to a new particle, called the Higgs boson. The mass of this particle can be expressed such as:

$$m_H = \sqrt{2\lambda}\nu \tag{1.25}$$

An unexpected consequence of the presence of the Higgs field is that, when it aquires a vacuum expectation value, it allows fermions to have a mass. Interaction terms

between the Higgs field and the fermions, the so called Yukawa interaction terms, can be added to the Lagrangian:

$$\mathcal{L} = -\lambda_\psi(\bar{\psi_L}\phi\psi_R + \bar{\psi_R}\phi\tilde{\psi_L}) \tag{1.26}$$

where $-\lambda_\psi$ are the Yukawa constants for quarks and leptons, which are free parameters of the SM.

The physical states are obtained by diagonalizing the up and down quark mass matrices by four unitary matrices. As a result, the charged current $W^\pm$ interactions couple to the physical up and down-type quarks. The resulting couplings are given by the Cabbibo-Kobayashi-Maskawa (CKM) matrix, which is expressed as:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{1.27}$$

This matrix describes the probability of a transition from one quark i to another quark j. These transitions are proportional to $|V_{ij}|^2$. Currently, the best determination of the magnitudes of the CKM matrix elements is:

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97425 \pm 0.00022 & 0.2252 \pm 0.0009 & (3.89 \pm 0.44).10^{-3} \\ 0.230 \pm 0.011 & 1.023 \pm 0.036 & (40.6 \pm 1.3).10^{-3} \\ (8.4 \pm 0.6).10^{-3} & (38.7 \pm 2.1).10^{-3} & 0.88 \pm 0.07 \end{pmatrix} \tag{1.28}$$

1.2 The Higgs boson

A consequence of the Higgs mechanism explained below is the prediction of a new particle. On July the 4th, 2012, the CMS and ATLAS collaborations announced the observation of a new particle [8][9][10], whose properties are found to be compatible with the expected SM Higgs boson.

The discovery has been claimed combining results of the searches performed in different decay channel: H$\rightarrow \gamma\gamma$, H$\rightarrow$ ZZ, H$\rightarrow W^+W^-$, H$\rightarrow \tau^+\tau^-$ and H$\rightarrow b\bar{b}$. The most significant excess comes from the two decay modes with the best mass resolution, the H$\rightarrow \gamma\gamma$, H$\rightarrow$ ZZ channels (see Fig. 1.4).

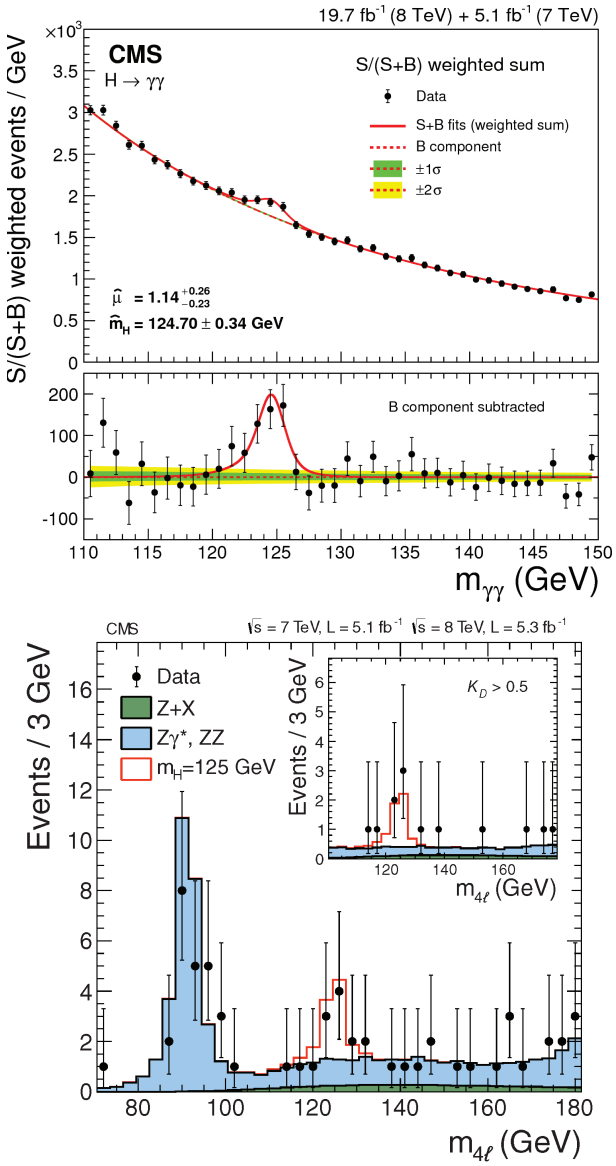

Figure 1.4: Top: di-photon mass spectrum weighted by the ratio $\frac{signal}{signal+background}$, together with the background-subtracted weighted-mass spectrum. The lines represent the fitted background and signal, and the colored bands represent the $\pm 1\sigma$ and $\pm 2\sigma$ standard deviation uncertainties in the background estimate. Bottom: distribution of the four-leptons invariant mass for the $ZZ \rightarrow 4$ leptons analysis. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass m_H = 125 GeV, added to the background expectation [8].

1.2.1 Higgs boson properties

Since this discovery, precise measurements of the Higgs mass have been performed: the latest mass measurement gives a value of $125.02^{+0.26}_{-0.27}$ (stat.)$^{+0.14}_{-0.15}$ (syst.) GeV [18] (see Fig. 1.5, top plot). In addition, its spin parity has been studied and found to be consistent with a pure scalar hypothesis [19]. The production cross section σ has also been measured and the result is compatible with the predictions of the SM (see Fig. 1.5, bottom plot).

1.2.2 Production and decay channels of the SM Higgs boson at the LHC

The Higgs boson can be produced in several ways, mainly:

- Via gluon-gluon fusion (gg$\rightarrow$ H+X), achieved through top quark loop;
- By Vector Boson Fusion (VBF) where $q\bar{q} \rightarrow q\bar{q}H$;
- In association with a $W^{\pm}/Z$ boson ($q\bar{q}$ $\rightarrow$HV +X, where V=$W^{\pm}/Z$, called the VH production);
- In association with a top quark pair ($t\bar{t}$H).

The evolution of the cross section as a function of the Higgs boson mass is shown in Fig. 1.6 (left plot) for proton-proton collision with a center of mass energy of $\sqrt{s}$ = 8 TeV.

The Higgs boson can not be directly observed in a detector, since it decays instantaneously according to several possibilities. The branching ratios for the different decay channels of a Standard Model Higgs boson depend strongly on the Higgs mass. As it can be seen on Fig. 1.6 (right plot), at a mass of 125 GeV, the Higgs will mostly decays into :

- A pair of b quarks (57%);
- A pair of τ leptons (6.3%);
- More rarely into a pair of muons (0.02%).

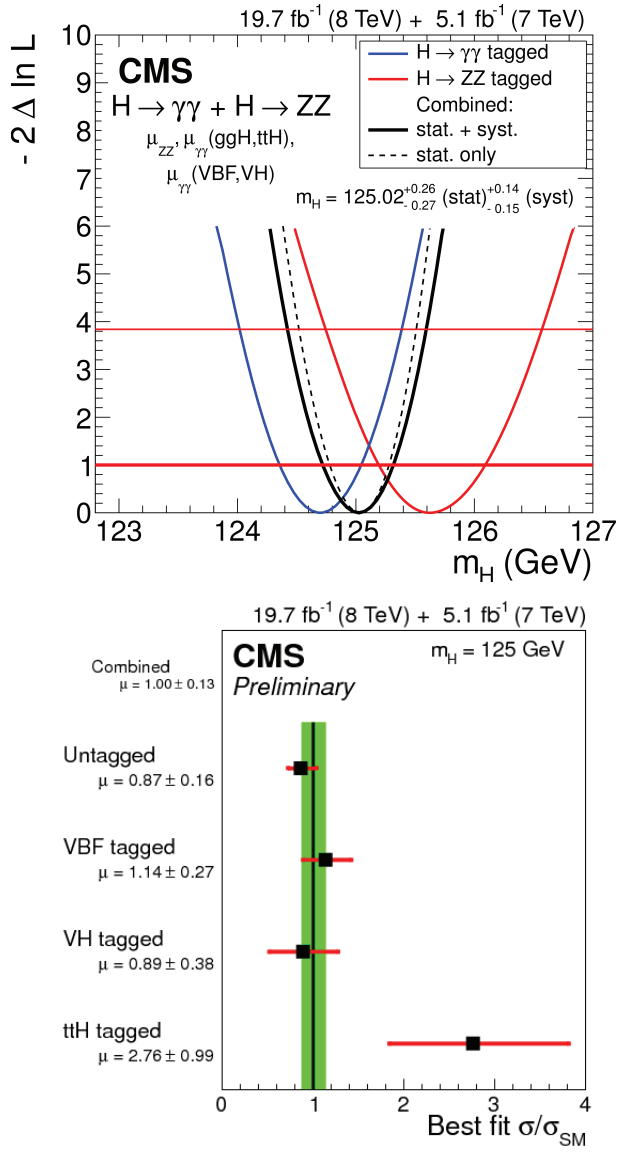

Figure 1.5: Top: Scan of the hypothesized Higgs boson mass m_H. Bottom: values of the best-fit σ/σ_{SM} for the combination (solid vertical line) and for sub-combinations by analyses targeting individual production mechanisms [18].

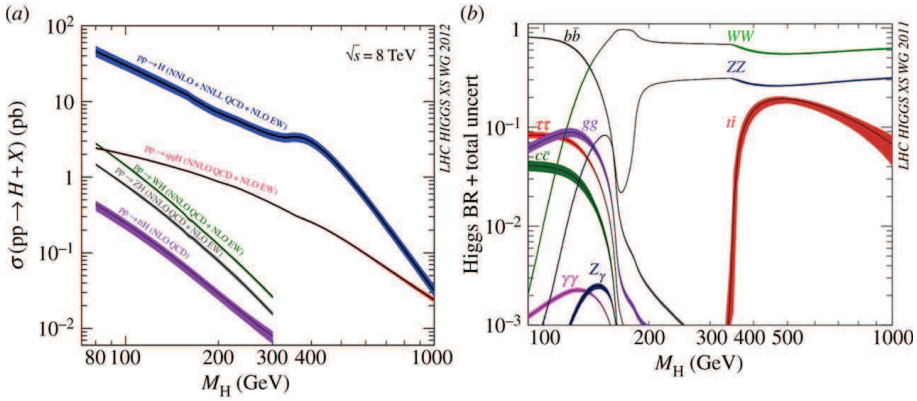

Figure 1.6: Production mode rates (left) and decay branching ratios (right) of the SM Higgs boson at $\sqrt{s}$=8 TeV [20].

Searches at the hadron colliders are particularly successful in the ZZ and $\gamma\gamma$ decay channels, due to the good separation of the signal from the irreducible backgrounds in these channels. Indeed, as it will be shown in the next chapter, the reconstruction of leptons and photons in CMS is very efficient; therefore, for these two channels it is possible to reconstruct the full energy of the decay products resulting in a high resolution mass peak, compared to backgrounds that are non-resonants. However, the $b\bar{b}$ and $\tau\tau$ decay modes remain important channels for verifying the coupling of the Higgs to fermions. They have much larger backgrounds but a considerably higher cross section times branching ratio.

1.2.3 Event generation: the production of a Higgs boson in association with a Z boson

The observation of the Higgs boson in the mass region around 125 GeV is quite challenging, especially when looking at the fermionic decay mode. Indeed, at such a mass value, the dominant decay channel is the $b\bar{b}$ production. However, this is also the mass region where the QCD background is overwhelming, drowning the H$\to b\bar{b}$ signal. An interesting research axis is then to study the associated VH production, in order to have a clearer signature. In this thesis, the channel of interest is the Higgs produced in association with a Z boson, followed by a Higgs decay in two b quarks, while the Z decays into two charged leptons. This leads to a detectable cross section and a clean signature.

This process is called the Z(ll)H(bb) process and will be referred as "ZH" in the following.

In order to confront the predictions of the SM to experimental data, an events simulation has to be performed. It is done in several steps, described below, and summarized on Fig. 1.7.

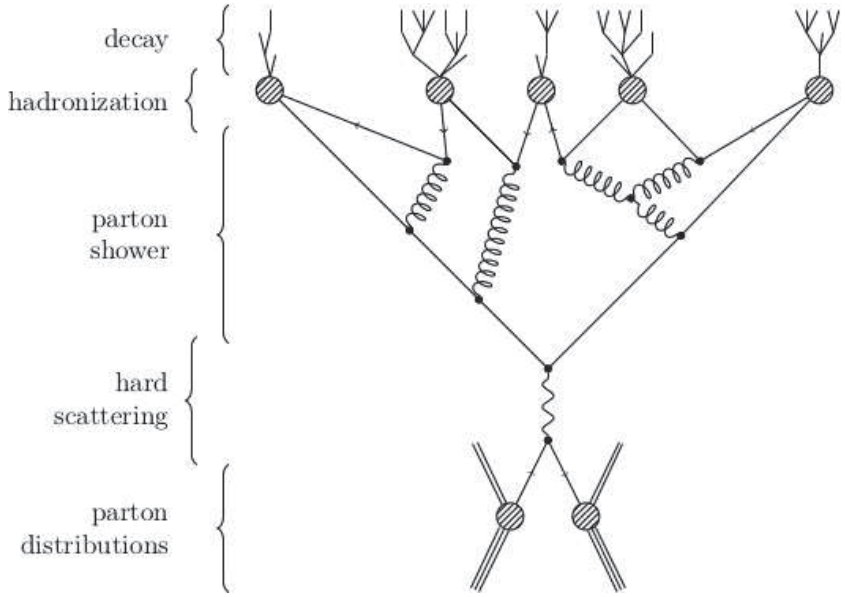

Figure 1.7: Event generation steps from the protons collision (bottom) to the final state particles decay (top).

The Proton Density Functions

Depending on the composition of the interacting protons at the moment of the collision, the production of distinctive events will be possible. Indeed, the protons are made of partons, quarks (valence quarks and sea quarks) and gluons, and each parton carries a certain amount x of the total impulsion of the proton. Functions called Proton Density Function (PDF) describe the probability that a component carries x of the proton's initial impulsion, for a specific Q.
An example of a PDF, determined by the Coordinated Theoretical-Experimental Project

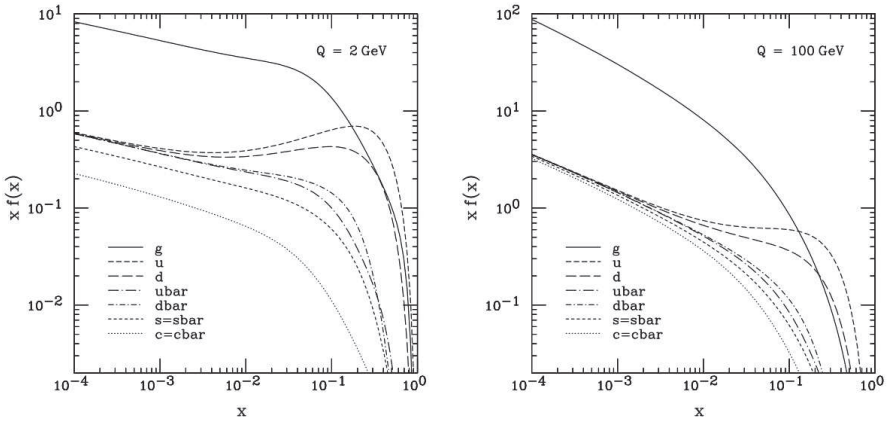

Figure 1.8: CTEQ proton's parton distribution functions for quarks, anti-quarks and gluons, for two particular protons collision energy scale Q: 2 (left) and 100 (right) GeV [21].

on QCD (CTEQ) collaboration, is shown on Fig. 1.8.

In the parton model approximation, called the QCD factorization theorem [22], where a factorization between the hard process and the free evolution of the partons can be assumed, the cross section for processes initiated by two hadrons with four momenta p_1 and p_2 is described by:

$$d\sigma_{p1p2} = \sum_{ij} \int dx1 dx2 f(x1, \mu_F^2) f(x2, \mu_F^2) \sigma_{ij}(p1, p2, \alpha_s(\mu_F^2), \frac{Q^2}{\mu_F^2}) \quad (1.29)$$

where σ_{ij} represents the cross section for hard scattering between a parton i and a parton j. The arbitrary parameter μ_F represents the factorization scale that separates the long distance interaction from the short distance. It has the same magnitude order as Q, the energy scale of the hard scattering (typically m_Z for a Z boson production). The sum is performed on all possible partons combinations.

The hard interaction and the hard process

Once the PDF have been determined, the hard process starts with a computation of the Feynman diagram's matrix element, to obtain the differential cross section. This takes into account all the possible interferences between all the alternative Feynman diagrams. Because of the perturbative development, the matrix element is actually computed using all the Feynman diagrams at various leading order (Leading Order (LO), Next to Leading Order (NLO), etc...), corresponding to a higher power of the coupling constant. An example of a hard process, corresponding to the process of interest in this thesis, the ZH process, can be seen on Fig. 1.9.

Then, once the matrix element has been determined, an integration over the phase space is computed to provide an estimation of the cross section. Finally, the event generation is performed by Monte Carlo generators, such as MadGraph [23] or Sherpa [24]. The decay of short lived particles is also handled at this point.

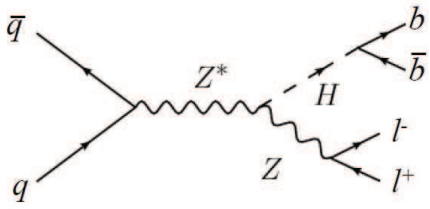

Figure 1.9: Leading order diagram illustrating the production of a Higgs boson, in association with a Z boson, and decaying into a pair of b quarks.

Hadronization

The hadronization, as it has been explained is Section 1.1.3, corresponds to the hadrons formation out of quarks and gluons produced during the hard interaction. This phenomenon is taken into account in an additional step, after the event generation. A parton-shower simulator, such as Pythia [25] or AlpGen [26], is interfaced with the matrix element generator (introduced in the previous paragraph). However, since the parton-shower is run independently from the matrix element generator, the multi-jets events generation may suffer from double counting between different jet multiplicity samples, affecting the cross section and the kinematics. A jet matching method is then applied: the phase space is split into two regions according to the QCD energy emission, Q_{cut}; soft radiations, with energies below Q_{cut}, are generated by the

parton-shower simulator, while the hard radiations (with energies higher than Q_{cut}) are simulated by the matrix element generator. It should be noticed that event generators may have a large number of free parameters, determined from experimental data, and their associated error has to be taken into account.

The pile-up events and other interactions

When two bunches of protons collide, several hard interactions may take place and the readout of the detector possibly includes information from more than one primary interaction. These multiple extra interactions are called Pile-Up (PU). These events are the price to pay when more luminosity or more time between two bunches of protons is desired. The event generation is adjusted to reflect the number of PU events seen in the data.
It is also important to model other possible interactions issued by the remnant components of the colliding protons. These events are called "underlying" events.

Initial and final state radiations

Before the collision, the partons of the protons can radiate and create extra particles in the initial state: it is called Initial State Radiation (ISR). Besides, particles in the final state, especially jets, can also radiate, and this phenomenon is called Final State Radiation (FSR). These are examples of NLO effects, and they are modeled by the parton-shower simulator.

1.3 The Matrix Element Method

In physics analyses, the standard approach consists in selecting and/or computing relevant variables for signal/background discrimination. They are then combined using a Multi Variate Analysis (MVA) tool. The Matrix Element Method (MEM) is a technique that, from the 4-vectors of the particles in final state, can directly produce a final event-by-event discriminating variable. It the matrix element described by the Feynman diagram at LO shown on Fig. 1.9 with the PDF presented in the previous section, in order to extract the maximum of theoretical information available. The MEM also exploits reconstruction level information to improve the information returned by the final variable.

This method has already been used at the Tevatron by the D0 and CDF experiments for a precise top mass measurement [27]. Within CMS, this method is used in two analysis: to test different spin hypothesis for Higgs boson (MELA [28]), and as a way

to discriminate the main background of the Higgs produced in association with a top quark pair [29].

From a specific theoretical hypothesis α and a reconstructed event characterized by p^{vis}, the MEM computes the probability $P(p^{vis}|\alpha)$ that evaluates how compatible these two inputs are. This probability is defined by the following integral:

$$P(p^{vis}|\alpha) = \frac{1}{\sigma_\alpha^{vis}} \int dx_1 dx_2 f(x_1) f(x_2) \int d\phi |M(p)_\alpha|^2 W(p^{vis}, p) \qquad (1.30)$$

This integral can be divided in three parts:

- $f(x_1)$ and $f(x_2)$ represent the PDF (see Section 1.2.3) of the two incoming protons;
- $|M(p)_\alpha|$ is the matrix element of the theoretical hypothesis at LO (one example of its representation can be seen on Fig. 1.9);
- $W(p^{vis}, p)$ are the Transfer Function (TF) introduced to translate the transition between a final state at generator level characterized by its kinematics, and the associated reconstructed final state at detector level.

The integration is done over the parton-level allowed phase space $d\phi dx1 dx2$. σ_α^{vis} is the cross section after cuts have been applied, leading to the production of events such as p^{vis}. It also consider the missing transverse energy, the momenta of the undetected particles (neutrino for example) and the momenta of the incoming partons,

Finally, the probability is normalized such that the overall expression can be interpreted as a probability density function:

$$\prod_i (\int d^3 p_i^{vis} P(p^{vis}|\alpha)) = 1 \qquad (1.31)$$

In this thesis, this normalization is not required. Instead, what will be used in the following chapters, and named "weight", is related the unnormalized probability W and defined by $-\log_{10}(W)$.

1.3.1 MadWeight

The MEM is a powerful tool, but it relies on the integration of the integral presented in Eq. 1.30, which is not trivial at all. To perform it, a tool called MadWeight [30]

is used. It is a generic and automatic software based on MadGraph that provides the leading order Feynman amplitude $|M(p)_\alpha|^2$ and the PDF, and performs the integration. The experimental inputs, the transfer functions, are provided by the user. In this work, the definition of the transfer functions is developed in Chapter 2.

The integrator used by MadWeight is VEGAS [31], that generates points in the available phase space according to the structure of a multi-dimensional grid, automatically generated. This generation depends on the Breit-Wigner present in the matrix element, and on the transfer functions: these peaked structures enhance the complexity of the numerical evaluation of the weights. To allow the integral to converge in an efficient way, these peaks must be aligned along one direction of the grid. This requires some changes of variables and a smart parametrization of the integration phase space, automatically done by MadWeight.

Besides, in events such as ZH events, the final state contains several b jets: an ambiguous parton/jet assignment can appear. In that case, for each combination of parton/jet pair, a weight is computed and MadWeight returns an averaged value.

NLO effects

As previously stated, the probability returned by the MEM exploits the LO matrix element of the theoretical hypothesis. As a consequence, if the reconstructed event presents an extra jet, produced by ISR or FSR, the event will not match the matrix element. ISR in particular can not be neglected, since those radiations are known to be important at the LHC.

- **ISR**: from a LO point of view, an additional jet produced during the initial state would violate the energy momentum conservation imposed as a constraint to the partonic state generated by $|M(p)_\alpha|^2$. The reconstructed event is then recoiling against the momentum associated to the extra radiation. In particular, the transverse momentum is not necessarily balanced among the final state particles. The MEM has been adapted to take into account these effects by computing the transverse momentum of the boost ($\vec{P}T_{boost}$) induced by the ISR, and using it as a correction in the weight computation. Two cases have to be considered: first, if there are unreconstructed particles in the final state (for example neutrinos in fully-leptonic $t\bar{t}$), the $\vec{P}_{T\,boost}$ is computed using the p_T of all the LO reconstructed particles in the final state, as well as the $\vec{E}_T^{miss}$ information, such

as:

$$\vec{P_T}_{boost} = -\sum_{extra-jets} \vec{P_T} = \sum_{LO-Particles} \vec{P_T} + \vec{E}_T^{miss} \tag{1.32}$$

However, if the final state contains only visible particles (such as ZH events), $\vec{P_T}_{boost}$ is determined using only the p_T of the final state particles:

$$\vec{P_T}_{boost} = -\sum_{extra-jets} \vec{P_T} = \sum_{LO-Particles} \vec{P_T} \tag{1.33}$$

This ISR correction is helpful when it comes to compute weights for over-constrained processes. In that case, the application of this correction allows the integral to converge in presence of narrow resonances, by relaxing the p_T conservation constraint.

- **FSR**: it is possible to handle FSR jets in two ways: either by applying another matrix element, in which an additional jet in the final state is present, or by recombining the extra jet with the particle it comes from, before applying the MEM. The first solution is highly time and CPU consuming, and the second one is not used in this analysis. However, extra information about the FSR jet can be used to build the final discriminant along with the MadWeight weight (using a MVA tool), in order to add this information in the final discrimination process.

Over-constrained systems

In the case where two narrow Breit-Wigner are present in the hard process, a high rate of failure in the integration of Eq. 1.30 is observed. This is due to the fact that there are more constrains than degrees of freedom in the system.
For example, for the ZH process, the constraints are:

- The two Breit-Wigner;
- The four particles in final state defined by their kinematics (p_T, η, ϕ);
- The conservation of the energy-momemtum between the initial and final state (four constraints);

On the other hand, the degrees of freedom are:

- The two PDF;
- The four particles in final state defined by their kinematics (p_T, η, ϕ);

In the end, there are 18 constraints against 14 degrees of freedom: the system is over-constrained. In order to allow MadWeight to compute the integral in a reasonable time, constraints have to be removed. The weights are then computed in two different ways:

1. With the correction "0" (called cor_0), where the energy-momentum (E-p) conservation is kept. In this case, if the Higgs contributes in the matrix element, its width is set wider;
2. With the correction "3" (called cor_3), where the E-p conservation is relaxed.

For each event, these two weights are kept in order to have both information in the final discriminant.

1.3.2 Advantages and critics of the method

The MEM is a complex and still not much used method, despite the various advantages of this tool:

- It maximize the amount of theoretical information for a signal/background discrimination;
- No training has to be done to get what can be the final discriminant variable, similarly to most MVA methods;
- MadWeight automatically computes the integral in a smart and efficient way, for any theoretical hypothesis;
- Many potential applications are possible for various physics analyses (top mass measurement, single top discovery, Higgs search, spin correlation measurement [32]);

However, some weaknesses of the method should be mentioned:

- The method is valid at LO order only; NLO corrections are applied but there are only approximations;
- The assignment between reconstructed jets and partons can be ambiguous if the reconstructed event is beyond LO;
- The MEM returns a weight that is fully model dependent;
- Approximations are made for the TF (see Chapter 2);

- No TF for the neutrinos is yet used;
- Depending on the model, the probability computation can be highly CPU demanding (see Table 1.2).

Table 1.2: Computing time for one weight using MadWeight 5, depending on the theoretical hypothesis.

Process	**Computation time**
ZH	$<$ 5 seconds
$t\bar{t}$ fully-leptonic	10 seconds
Z+2 b jets	18 seconds
$t\bar{t}$ semi-leptonic	41 seconds
$t\bar{t}$H fully-leptonic	1 min

Chapter 2

Experimental context

2.1 The Large Hadron Collider

To test the SM, one needs to probe the matter to see its innermost components, and how these fundamental particles interact between each other. The best way to achieve that purpose is to fragment the matter.

Two kinds of accelerators have been produced: linear and circular accelerators. The main advantage of the second type is that the beam can be recycled; by 1985, the largest circular accelerator built was the Large Electron Positron (LEP) collider synchrotron, at Centre Européen pour la Recherche Nucléaire (CERN) in Geneva, an electron/positron collider. However, the biggest disadvantage using light particles such as electrons is a significant loss of energy by synchrotron radiation. An alternative solution is to use more massive particles such as hadrons.

The LEP was dismantled in 2000 so that the underground tunnel could be used for Large Hadron Collider (LHC), a proton collider, currently the world's largest and highest-energy accelerator on Earth, which gives access to a new collision energy frontier.

The LHC is located at CERN, buried 100 meters under the ground, at the border between France and Switzerland. Its nominal purpose is to provide proton-proton col-

lisions at a center of mass energy of $\sqrt{s}$ = 14 TeV, in order to probe the edge of the SM.
Four massive detectors are placed around the accelerator ring: A Toroidal LHC ApparatuS (ATLAS) [33] and Compact Muon Solenoid (CMS) [34], which are general purpose detectors dedicated to the SM measurements and new physics searches, A Large Ion Collider Experiment (ALICE) [35], and the Large Hadron Collider beauty experiment (LHCb) [36]. ALICE uses the heavy ions collisions occurring at the LHC once a year to study the matter at very high temperature and density; the LHCb detector's goal is to study the charge parity violation in the B hadrons decay, and explain why there is more matter than anti-matter in the Universe.

2.1.1 The apparatus

The protons beams are extracted by ionization of hydrogen, then injected in the LINear particle ACcelerator (LINAC). From here, they are injected in the Proton Synchrotron (PS) and in the booster until they reach an energy of 26 GeV; the space between two bunches of protons is 25 ns at this point. Afterwards, the beams go through the Super Proton Synchrotron (SPS) to be accelerated at an energy of 450 GeV before being finally transfered to the LHC for a final acceleration. In 2012, the two beams were accelerated to collide at an energy of $\sqrt{s}$ = 8 TeV with a bunching space of 50 ns. The CERN's accelerator complex is represented on Fig.2.1.

The 27 km of the LHC ring are composed of resonant cavities, used to accelerate the two proton beams, and magnetic multi-poles: 1232 twelve meters-long supraconducting magnets produce a 8.3 Tesla magnetic field to curve and collimate the beams. More magnets are used for the beam injection, and to control the beams crossing, the beam dump and the beam stability. All of them are immersed in liquid helium at a temperature of 1.9 K.

The key parameter of the LHC is the luminosity, defining the collisions rate. The instantaneous luminosity of the accelerator is given by:

$$\mathcal{L} = f \times \frac{N_A \times N_B}{4\pi\sigma_A \times \sigma_B} \tag{2.1}$$

where f is the beam crossing frequency, N_A and N_B are the number of particles per beam; σ_A and σ_B represent the effective area of the two beams. The number of interactions per second is then:

$$\dot{N} = f \times \sigma \tag{2.2}$$

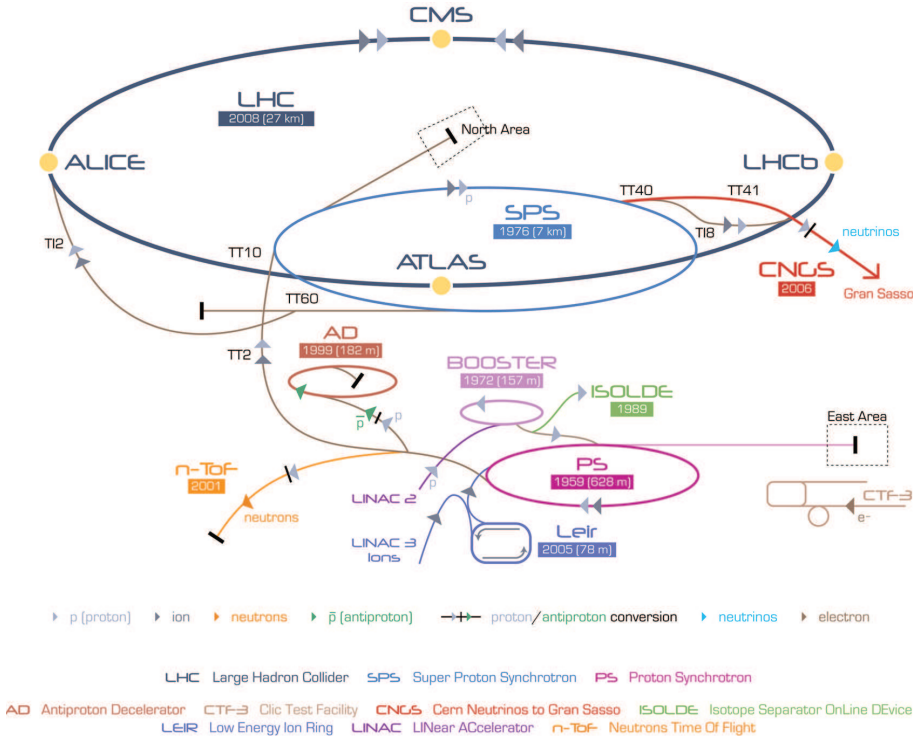

Figure 2.1: The full CERN's accelerator complex for protons acceleration [37].

where σ is the LHC proton-proton cross section. The integrated luminosity $\mathcal{L}$, corresponding to the luminosity recorded during data taking, is obtained by integration of Eq. 2.1, so that the final number of events produced during the LHC collisions is:

$$N = \mathcal{L} \times \sigma \tag{2.3}$$

Inaugurated in 2008, the LHC provided its first collisions in 2010 at $\sqrt{s}$ = 7 TeV. In 2012, the energy was increased to $\sqrt{s}$ = 8 TeV and the integrated luminosity delivered by the LHC to the CMS experiment has never stopped rising, to reach about 30 fb^{-1} of data. After the first Long Shutdown (LS) started in 2013, the collisions started to occur at $\sqrt{s}$ = 13 TeV since spring 2015; the goal of the Run II, started in summer 2015 is to reach 75-100 fb^{-1}.

The LHC was built to run with a design luminosity of $10^{34} cm^{-2} s^{-1}$ and this goal should be reached during the next run of data taking. Several upgrades are foreseen, aiming at increasing the interaction rate. After LS1, two other major LS are foreseen: LS2 should take place in 2018, the most important goal being to complete the upgrade injector chain upgrade. Then LS3 is planned for 2022, a period during which new collimators and the low-β quadrupoles will be placed. After LS3, the High Luminosity Large Hadron Collider (HL-LHC) project [38] will start.

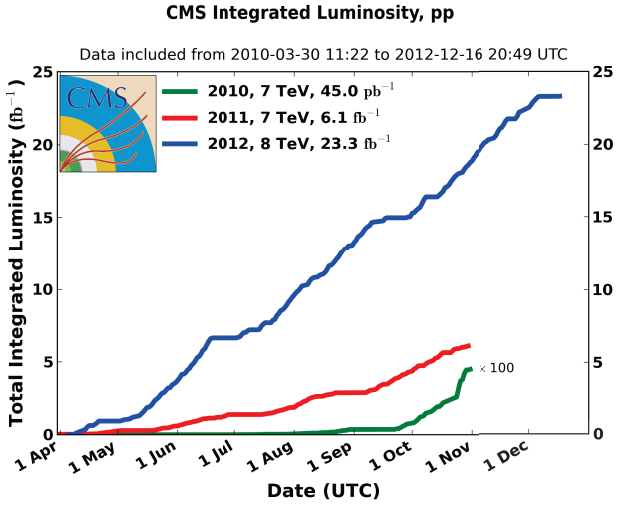

Figure 2.2: LHC integrated luminosity delivered to CMS during the 2010 (green), 2011 (red), and 2012 (blue) running periods [37].

2.1.2 Data access

The LHC Computing Grid is a global computing infrastructure that provides computing resources to store, distribute and analyze the data generated by the LHC. The data are then available to all CMS scientists, regardless of their physical location. In CMS, Cms Remote Analysis Builder (CRAB) enables users to easily perform analysis jobs on every data or Monte Carlo (MC) sets officially published by CMS, via the Grid.

Data are analyzed using the software ROOT [39], an object-oriented program and library developed by CERN.

2.2 The CMS detector

The CMS detector [40][41] is a general purpose detector designed to record the particles produced by the proton-proton interactions delivered by the LHC. It is composed of a central part and two end-caps, with a structure in layers; the collisions take place in the center of the detector, where the LHC ring lies. Around is the tracker detector, used to identify charged particles and measure their impulsion; then two calorimeters, the Electromagnetic CAlorimeter (ECAL) and the Hadronic CAlorimeter (HCAL), determine the energy of the electrons, photons and of the jets; finally, the last layer is composed of muons chambers for the muons identification and reconstruction.

Between the calorimeters and the muon chambers lies a supraconducting solenoid magnet that creates an homogeneous magnetic field of 3.8 Tesla inside the detector, used to curve the particles track. Beyond the muon chambers, iron structures (called "return yoke") are placed to close the magnetic loop. An overview of the detector with its different components can be seen on Fig. 2.3.

The coordinates system of CMS is centered at the interaction point; the x axis points in direction of the LHC center, the y points to the sky, while the z axis goes along the beam pipe. For physics analyses, several convenient angles are used to describe the particles kinematics, such as the ϕ angle, defined using the x axis in the $(x-y)$ plane, or the polar angle θ, measured from the z axis. The pseudo-rapidity η is mostly used instead of θ, and using the approximation that the mass of a particle is negligible, it is defined as:

$$\eta = -\ln(\frac{\tan\theta}{2}) \tag{2.4}$$

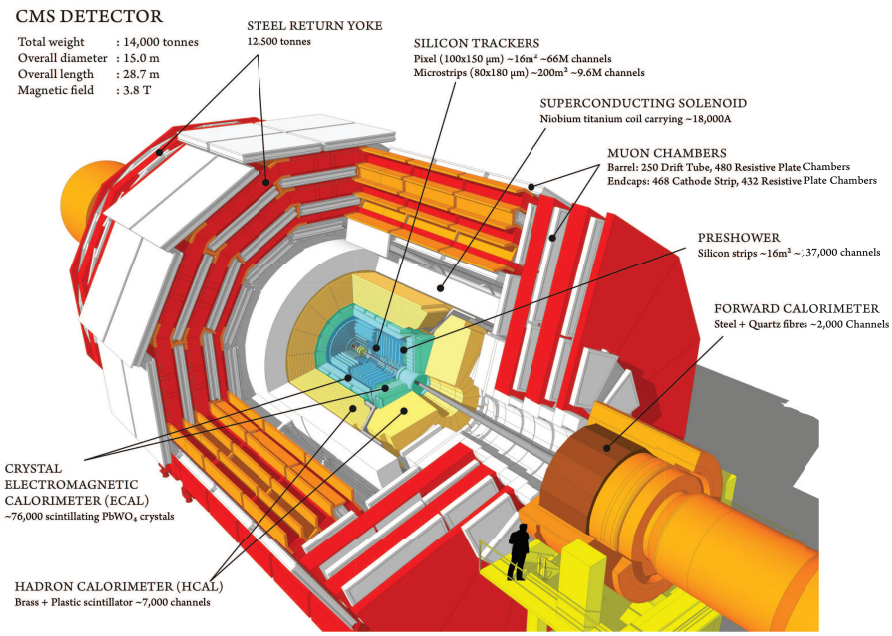

Figure 2.3: Sectional view of the CMS detector [34].

The coordinates system of CMS along with the angles definition are displayed on Fig. 2.4.

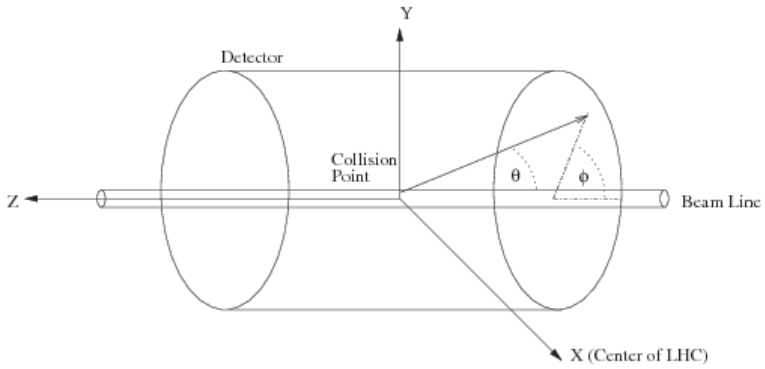

Figure 2.4: Coordinates system used in CMS and construction of the θ and ϕ angles.

In order to detect all the particles produced during the collisions and passing through the detector, and to measure with precision their properties, CMS is divided into sub-detectors, each of them dedicated to specific particles identification and properties measurement.

2.2.1 Tracking system

The CMS detector's core is composed of the largest silicon detector ever built, designed to provide a precise measurement of the charged particle trajectories, and of vertices position.

It consists of a silicon pixel tracker made of three barrel layers (BPIX) and two forward/backward disks (FPIX). Thanks to the 3-dimensional capability brought by these pixelated detectors, it is feasible to measure the properties of charged particles crossing the detector with a single hit resolution between 10-20 μm. To achieve an optimal vertex position resolution, the detector is paved with 285 μm thick, almost square shaped, silicon sensors with a pixel cell size of 100x150 μm^2. The pixel detector is composed of 66 million pixels, clustered into 1440 modules, in order to cover a pseudo-rapidity range up to $|\eta|$<2.5. The resulting signal is recorded using approximately 16000 readout chips, bump-bonded to the detector modules.

Around the pixel detector is the 5 meters-long SiStrip detector, that has a diameter of 2.5 meters. It is the first "all silicon" central tracker with about 9.6 million electronic channels. There are 10 layers in the barrel region, 4 Inner Barrel (TIB), 6 Outer Barrel (TOB) and 10+3 discs in the inner disks (TID) and end caps (TEC). The complete layout of the tracker is exposed on Fig. 2.5.

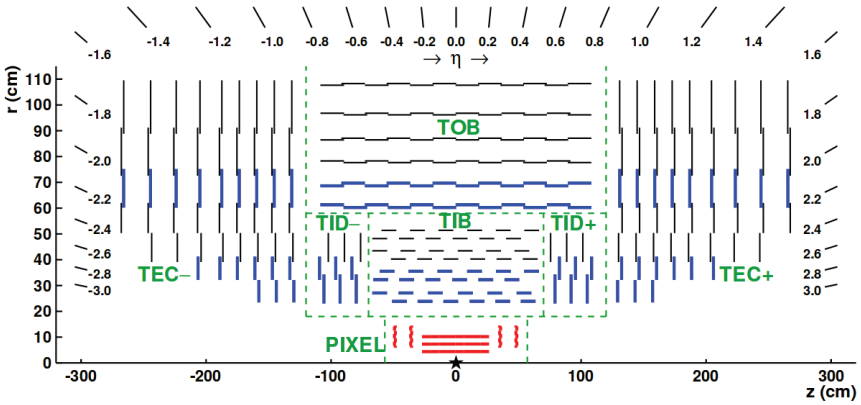

Figure 2.5: Layout of the Pixel and the SiStrip detectors in the CMS tracker [42].

Track reconstruction

The regularly used pattern recognition and track reconstruction for charged particles follows an iterative method [43]. First, the seeding starts from the pixel layers: from hit triplets or pairs compatible with the point where the two beams collide (called the beam-spot), seeds are created while the others are discarded. Then, the estimation of the trajectory is performed: each seed is propagated to the successive layers, using a Kalman filter technique [44]. It allows the reconstruction of the trajectory even if there is a missing hit in a layer. The propagation continues until there are no more layers or there is more than one missing hit. Finally, the track is fitted: further hits are added and the track parameters estimation is updated for each new hit obtained. A final fit is performed to obtain the track parameters at the interaction point. Afterward, vertices reconstruction can be achieved.

Vertex reconstruction

It is essential to properly determine the primary vertex, where tracks from the main interaction originate. The candidates are selected by clustering reconstructed tracks, based on the z coordinate of their closest approach to the beam line. An adaptive vertex fit is then used to estimate the vertex position with a sample of tracks that are compatible with an origin close to the interaction region. Among the primary vertices found, the one with the highest $\sum_{tracks} p_T$ is selected as a candidate for the origin of the hard interaction. The primary vertex reconstruction efficiency is close to 100%. The vertex resolution measurement is performed with a data driven method: the vertex is split into two, then an independent fit is done on the two resulting vertices and their difference in position gives the resolution. This resolution strongly depends on the number of tracks, as it is shown in Fig. 2.6.

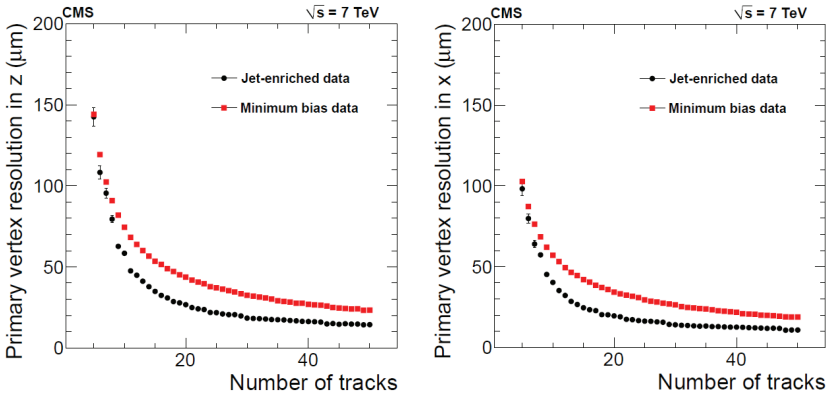

Figure 2.6: Primary vertex resolution along the z-axis (left) and the x-axis (right) as a function of the number of tracks coming from the vertex, for jet-enriched data and for minimum bias data events ([43]).

Secondary vertex reconstruction is a rather challenging process in comparison with primary vertex reconstruction. The discrimination between a primary and a secondary vertex is based on the distance between the vertex and the beam line in the transverse plan (or a primary vertex if one has already been reconstructed). Since most vertex finder algorithms are sensitive to both primary and secondary vertices, a vertex filter is necessary to ensure the selection of, and uniquely, secondary vertex candidates.

Performance

Since its start-up in 2008, more than 95% of the pixel channels of the CMS detector are active during the data taking, allowing good detector performance. Thanks to its fine granularity, the pixel detector can provide high quality seeds for off-line track reconstruction algorithms. The hit efficiency is determined by the quantification of missing hits on reconstructed tracks during LHC runs. It is estimated to be above 99%, as it is shown on Fig. 2.7, left plot. The pixel thresholds increase with the integrated luminosity, reducing the cluster size and affecting the resolution. Recalibrations are then performed during LHC technical stops to partially recover from this degradation (see Fig. 2.7, right plot).

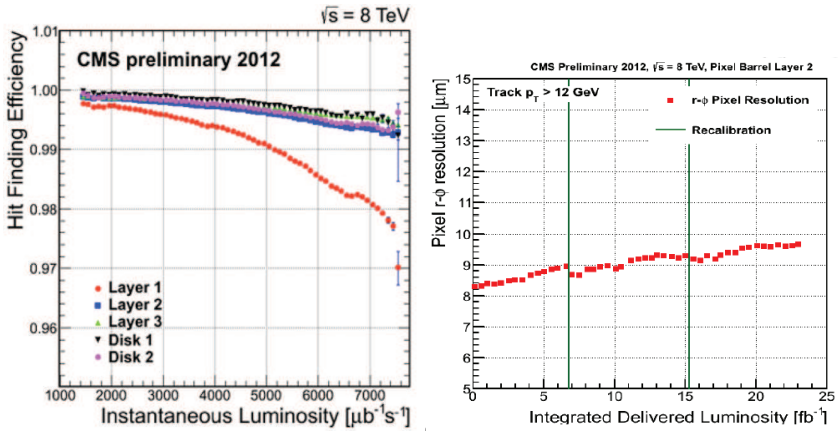

Figure 2.7: Left: hit finding efficiency as a function of the instantaneous luminosity, for every component of the pixel detector. Right: pixel resolution in the $(r - \phi)$ plane as a function of the integrated delivered luminosity. The recalibration phases are indicated in green [45].

2.2.2 The calorimetry

The role of a calorimeter is to measure the energy of the particle passing through by stopping it: a loss of energy is induced by the interaction between the particle and the calorimeter material, then recorded. If the particle is completely stopped, since the detector is almost hermetic, the deposed energy should correspond to its initial energy.

The calorimetry system of CMS is divided into two parts: one dedicated to the electromagnetic particles, the electrons and the photons, and one designed for the hadronic objects.

The electromagnetic calorimeter

The ECAL is the closest to the center of the detector. It's an hermetic scintillator detector made of lead tungstate $PbWO_4$ crystals, offering the best performance for energy resolution; this material also has the advantage to be resistant to the radiation environment of the LHC. Besides, the scintillation decay time of these crystals is of the same order of magnitude than the LHC bunch crossing time.

The ECAL is composed of a barrel section and two end-caps. The cylindrical barrel consists of 61200 22x22 mm^2 and 230 mm-long crystals, clustered into 36 three tonnes super-modules, oriented in the direction of the interaction point. 15000 more 28.6x28.6 mm^2 and 220 mm-long crystals compose the end-caps: this high segmentation of the detector allows a good spatial resolution. At the end-caps the ECAL inner surface is covered by the pre-shower sub-detector, consisting of two layers of lead interleaved with two layers of silicon strip detectors. Its purpose is to aid in the pion-photon discrimination in the $1.6 < |\eta| < 2.6$ region, and to improve the electrons and photons position determination. Avalanche Photo Diodes (APD) are used to detect the scintillation light in the barrel region while Vacuum Photo-Triodes (VPT) are used in the end-cap region.

ECAL covers a pseudo-rapidity region up to $|\eta| < 3$ ($|\eta| < 1.479$ for the barrel), and precise energy measurement for photons and electrons can be performed until $|\eta| <$ 2.6 This limit has been determined by considering the radiation dose received and the amount of pile-up energy deposits; it also matches the geometric acceptance of the inner tracking system. A geometric view of the ECAL can be seen on Fig. 2.8.

Electrons and photons reconstruction

At high energy, an electron looses its energy by emitting photons, that emit pairs of electron-positron, and so on, until the produced particles reach a critical energy E_c, too low to allow another pair emission. This phenomena is called Bremsstrahlung or deceleration radiation. At this point, the final particles interact with the ECAL material creating light, detected by the APD. Since the detector is almost hermetic, it is possible to measure all the components of this electromagnetic shower, in order to go back to the initial electron's energy. However, some low energy electrons and photons may already start the Bremsstrahlung inside the tracker material, before reaching the ECAL. For this reason, the electrons are reconstructed combining tracking information and energy deposits in ECAL.

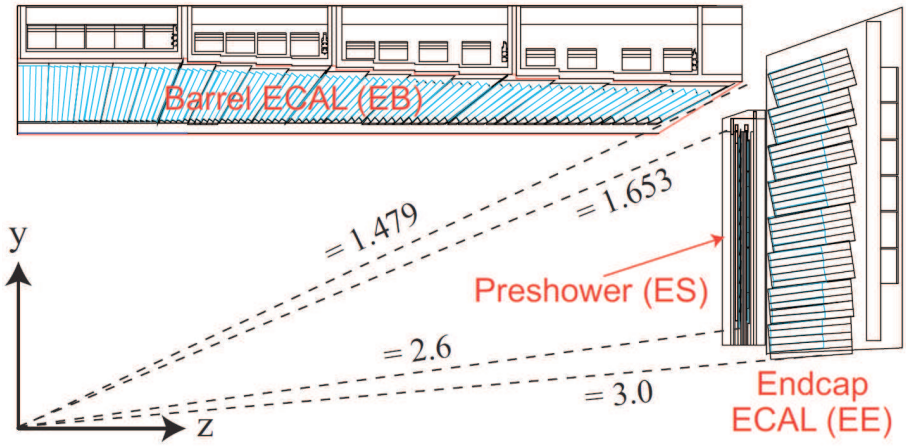

Figure 2.8: Geometric view of one quarter of the ECAL.

Because of the 3.8 Tesla magnetic field presence, the particles trajectory is bent; as a consequence, the energy deposit clusters are spread in η and ϕ. Super-clusters of energy are then built to gather all the energy deposits coming from the electron decay: the crystal with the highest transverse energy deposit, above 4 GeV, is selected and a narrow η-larger ϕ window is created around this seed; finally, super-clusters are built collecting all the crystals in the way. The electrons energy is then computed as the sum of the energies of all the crystals in the super-cluster and its position is determined as the mean position of the crystals in the super-cluster, weighed by their energy.
After that, super-clusters are matched to track seeds. Here, two complementary algorithms are used: one starts the seeding from the tracker and is more dedicated to low p_T electrons. The second one begins in the ECAL: from the position of the super-cluster, a propagation to the pixel layers through the magnetic field is done to search for compatible hits; for the first pixel hit found, loose requirements are applied; a tighter selection is made on the second hit.

The electron track reconstruction [46] starts from the seeds found in the pixel and a Gaussian Sum Filter (GSF) algorithm [47] is used for the full track reconstruction. This algorithm is an extension of the Kalman filter, developed to take into account the Bremsstrahlung effects, since it strongly affects the low energy electrons tracks. Once the track reconstruction is achieved, a super-cluster association is performed and a first selection is applied, based on kinematic and geometric properties, to keep only good electrons candidates, interesting for physics analyses.

The photon reconstruction [48] follows the same procedure with different super-cluster requirement: the energy sum of the 3x3 crystals centered on the most energetic crystal in the super-cluster divided by the energy of the super-cluster is used. This allows a good discrimination between photons that convert before reaching the calorimeter and unconverted photons.

The ECAL resolution on the energy measurement is parametrized as follow:

$$\frac{\sigma(E)}{E} = \frac{S}{\sqrt{E}} + \frac{N}{E} + C \tag{2.5}$$

where C is a constant taking into account the non-uniformity of longitudinal light detection, the calibration uncertainty and the energy leakage coming from the back of the crystal; S is a stochastic term and N is a noise term that includes the electronic noise, digitization and pile-up related effects. For central photons with energies in the range of interest for physics analyses (100 GeV), typical values for S, N and C are respectively 2.8%, 0.128 and 0.3.

ECAL performance

The ECAL performance has been measured for the 8 TeV dataset and found to be very good: as it can be seen on the plots on Fig. 2.9, for photons with a medium working point selection applied, the identification efficiency is above 80%. For the electrons with the corresponding selection applied, the selection efficiency is above 80% in the barrel region, depending on the electron transverse momentum, as it is shown on Fig. 2.10.

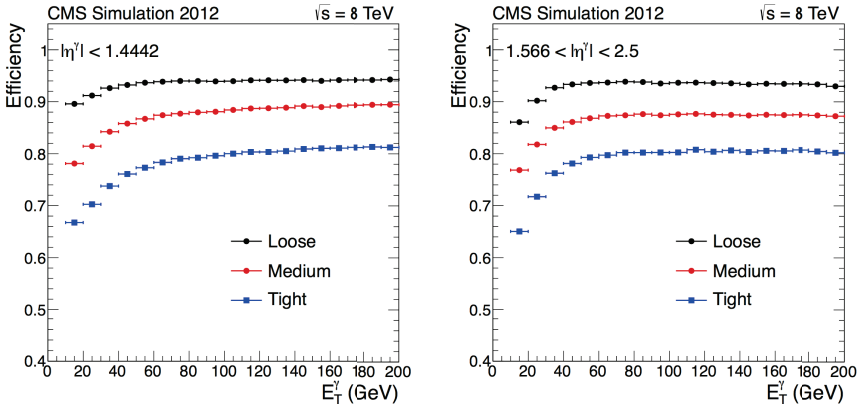

Figure 2.9: Efficiency in simulation of the Tight (blue), Medium (red), and Loose (black) stringency selection of the cut-based photon identification, as a function of photon transverse energy in the ECAL barrel (left) and in the end-caps (right). The signal events are γ + jets events ([48]).

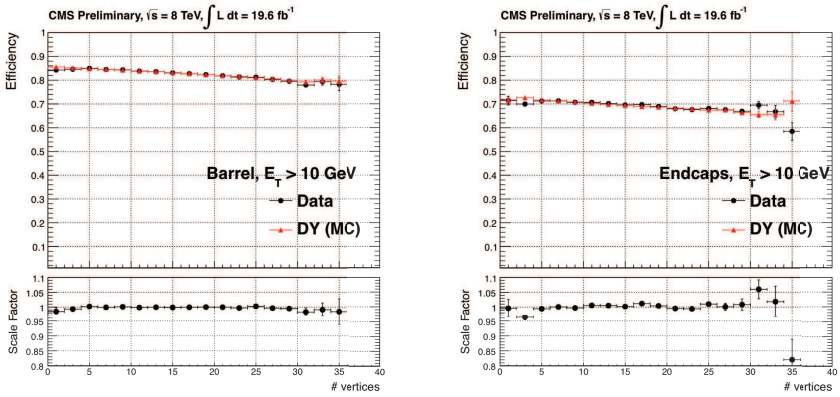

Figure 2.10: Left: electron selection efficiency for a medium stringency selection, on data and on a Drell-Yan Monte Carlo sample, as a function of the number of reconstructed vertices, in the ECAL barrel (left) and end-cap (right) regions. Both statistical and systematic errors are included [49].

The hadronic calorimeter

Surrounding the ECAL and enveloped in the CMS solenoid is the HCAL, whose role is to measure the energy of particles made of quarks and gluons. Additionally, using the tracker and ECAL information, it allows an indirect measurement related to the presence of non-interacting uncharged particles.

The HCAL is a sampling calorimeter: it finds a particle's position, energy and arrival time, using alternating layers of absorber and fluorescent scintillator materials that produce a rapid light pulse when the particle passes through the detector. When a hadronic particle hits a plate of brass or steel, an interaction can occur producing secondary particles. These particles fly through successive layers of absorber and also interact, resulting in a particle shower. As this shower develops, the particles pass through the alternating layers of active scintillation material, emitting blue-violet light. Within each tiny optical fibers, with a diameter of less than 1 mm, the light is absorbed. They shift the blue-violet light into the green region of the spectrum, and optic cables carry the green light away to readout boxes, located at strategic locations within the HCAL volume. When the amount of light in a given region is summed up over many layers of tiles in depth, called a tower and oriented to the interaction point, this total amount of light corresponds to the particle's energy.

The HCAL is organized into barrel (HB and HO), end-cap (HE) and forward (HF) sections, as it can be seen on Fig. 2.11. The HB region ($0 < |\eta| < 1.4$) has a granularity of $\Delta\eta \times \Delta\phi$ = 0.087×0.087; the outer part, HO, is outside the solenoid magnet; this is due to the fact that in this region, the ECAL and HCAL materials combined are not enough to completely stop the hadron showers. The end-caps HE are placed in the pseudo-rapidity region $1.4 < |\eta| < 3$, and they have a granularity of $\Delta\eta \times \Delta\phi$ = 0.17×0.17. The high pseudo-rapidity region ($3.0 < |\eta| < 5.0$) is covered by the HF detector. Located 11 meters on either side of the interaction point, it uses a slightly different technology of steel absorbers and quartz fibers for readout, designed to allow better separation of particles in the forward region. The HF is also used to measure the relative on-line luminosity system in CMS [34].

Given the difference of detector geometry in the forward and end-cap/barrel regions, the energy resolution is defined accordingly. For the forward region:

$$\frac{\sigma(E)}{E} = \frac{0.9}{\sqrt{E}} + 0.045 \tag{2.6}$$

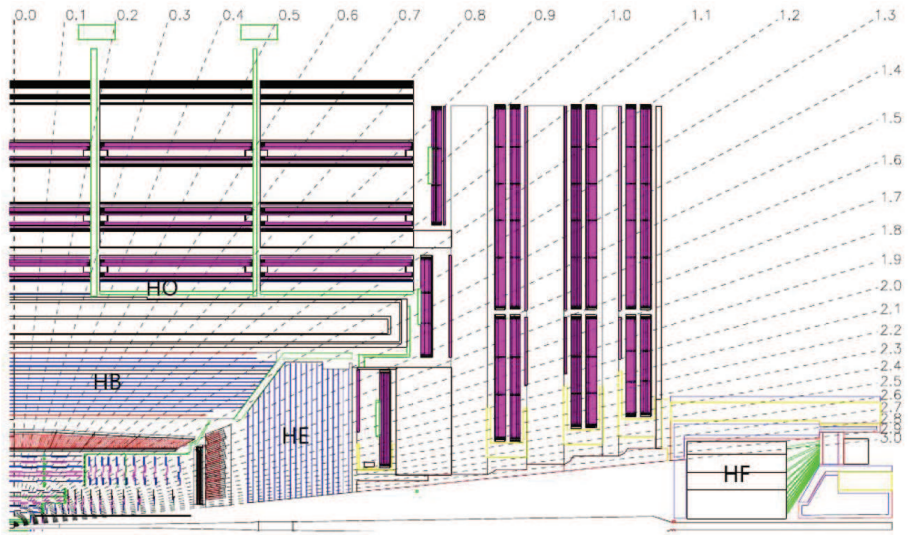

Figure 2.11: Longitudinal view of one quarter of the detector showing the positions of the HCAL sections: hadron barrel (HB), hadron outer (HO), hadron end-cap (HE) and hadron forward (HF).

and for the barrel/end-cap region:

$$\frac{\sigma(E)}{E} = \frac{1.72}{\sqrt{E}} + 0.09 \tag{2.7}$$

Jet reconstruction

Jets have a complex substructure, reflecting the properties of the quark/gluon they originate from. The simplest way to reconstruct a jet follows the same principle as the electron/photon reconstruction: it is to sum the calorimetric energy deposits in a cone of angular size ΔR around the incoming quark/gluon, in order to get back to its initial energy. Several algorithms for jet reconstruction are available in CMS. For the analyses presented in this thesis, the anti$-k_t$ algorithm is used [50].

This algorithm is a cluster algorithm, that starts from all the elementary objects available and performs an iterative pair-wise clustering to build larger objects, using geometric and kinematic properties of the objects. It starts by defining a distance d_{ij} between two objects i and j:

$$d_{ij} = min(k_{ti}^{-2}, k_{tj}^{-2}) \frac{{\Delta R_{ij}}^2}{D^2} \tag{2.8}$$

based on their angular separation in the $(\eta-\phi)$ plane ${\Delta R_{ij}}^2 = (\eta_i-\eta_j)^2+(\phi_i-\phi_j)^2$, where $d_{iB} = k_{ti}^{-2}$ represents the distance between the particle i and the beam (B). The parameter D scales d_{ij} with respect to d_{iB} such that any pair of final jets a and b are at least separated by ΔR_{ab}=D^2.

The algorithm computes the distances d_{ij} and d_{iB}, and finds the smallest one: if it is d_{ij}, it sums the four momenta of the two particles i and j, updates the distances and proceeds in finding the next smallest distance; if the smallest one is d_{iB}, the particle i is removed and called a jet. The procedure is repeated until all the particles are clustered into jets. The anti-k_t algorithm is robust with respect to hadronization effects or underlying events contamination: this is due to the $1/k_t^2$ dependence of d_{ij}, implying that soft particles will tend to cluster with the harder particles instead of clustering with other soft particles. It produces circular cone-shaped jets, improving the momentum resolution of the jets. In this thesis, this algorithm is applied with the parameter D set at 0.5.

After reconstruction, corrections on the jets are applied to account for the differences between the jets reconstructed by the algorithm and the jets at MC generation level. These differences are mainly due to the non linear/uniform response of the CMS calorimeter to the jet shower.

Jet energy corrections

The calorimeter response is not linear with respect to the detected particles and therefore it is not straightforward to translate the measured jet energy to the true particle energy. Besides, jets are not always fully reconstructed. Therefore, CMS analyses apply several factorized energy correction factors, where each level of correction takes care of a different effect. Each level is essentially a scaling of the jet four momentum with a corrective scale factor which depends on various jet related quantities (jet p_T, η, flavor, etc.). The corrections are applied sequentially, the output of each step being the input to the next, with a fixed order:

- The level 1 (L1) correction removes the energy coming from pile-up events. In principle this will remove any dataset dependence on luminosity so that the following corrections are applied upon a luminosity-independent sample;
- The level 2 (L2) Relative correction makes the jet response uniform in pseudo-rapidity. This is achieved by correcting any jet using a jet in the central region ($|\eta|$ <1.3). The derivation of the relative correction is done either by using MC information or by employing a data driven method (di-jet balance method [51]);

- The level 3 (L3) absolute correction assures the jet response to be flat with respect to the jet p_T. Once L2 correction has been applied to a jet, it is corrected back to particle level. The derivation of the absolute correction is done either by using MC information or by employing data driven techniques (for example using the Z/γ+jet balance method [52]);
- After the first collisions at $\sqrt{s}$=7 TeV, it appeared that the CMS jet energy response simulation was very successful. However, the comparison between data and MC was not perfect, with some small differences, up to 10%, depending on the η region. Therefore, after applying L2 and L3 corrections, a small residual calibration, η and p_T dependent, is applied in order to fix the differences between data and MC. By definition, this correction is applied to MC only.

Jet energy resolution

Measurements show that the Jet Energy Resolution (JER) in data is worse than in the simulation: the jets in MC need to be smeared to describe the data. CMS provides scale factors to correct the jet energy resolution directly on the raw anti-k_t jets, for different η and p_T ranges.

Missing transverse energy

The Missing Transverse Energy (MET)[53] ($\vec{E}_t^{miss}$) is defined as the measured energy imbalance in the transverse plane to the colliding protons beams. This imbalance can be caused by several phenomena, such as particles escaping from the detector without any interaction (neutrinos, very forward particles), detector effects (noise, dead cells) or unaccounted physics processes (pile-up events, new physics). Some new physics signatures may contain weakly interacting particles, making the $\vec{E}_t^{miss}$ an object of first importance: it has to be well understood and measured.

A global definition of the $\vec{E}_t^{miss}$ is the negative vector sum of the transverse momenta of all final state particles reconstructed in the detector: for an event containing N particles in the final state, the E_t^{miss} is defined as the magnitude of a 2D vector:

$$E_t^{miss} = \sqrt{(\sum_i^N E_x^i)^2 + (\sum_i^N E_y^i)^2} \tag{2.9}$$

where E_x^i is the energy component of the i^{th} particle along the x axis. CMS has developed three distinct algorithms to reconstruct the $\vec{E}_t^{miss}$:

- The "Calo MET" is based on the calorimeter energy deposits and the reconstruction algorithm uses the calorimeter towers geometry. It is calculated using the energy contained in calorimeter towers to define pseudo-particles. The sum excludes energy deposits below noise thresholds;
- The previous MET can be corrected by including tracks reconstructed in the inner tracker after correcting for the tracks expected energy deposits in the calorimeter ("TC MET").
- The Particle Flow (PF) MET is calculated using a complete particle-flow technique, from the reconstructed PF particles (see Section 2.3);

As for jets, corrections are applied on the $\vec{E}_t^{miss}$, in order to have a better MC/data agreement:

- The type I correction is based on the energy response of the reconstructed jets in the event, and uses the jet energy scale correction;
- A correction is applied to take into account the presence of muons, leaving little energy in the calorimeter and creating an imbalance of energy. Since their momentum is very well determined by the tracking system, it can easily be removed from the $\vec{E}_t^{miss}$ computation;
- The electron correction follows the same principle as the muon correction; this correction is expected to be small due to the excellent energy resolution and coverage of the ECAL;
- Tau corrections are also applied using PF taus: it removes the energy towers in a cone of ΔR=0.5 around the τ;
- A type II correction takes into account underlying events, pile-up effects and double counting of unclustered energy.

In most of CMS analyses, the PF MET is used since this variable benefits from the best resolution. In this thesis, the MET significance [54], based on the PF MET, is used: on an event-by-event basis, it tests the probability that an observed MET is consistent with a fluctuation from zero due to finite detector resolution. It is constructed from a likelihood ratio, assuming Gaussian resolutions for the objects, such as:

$$S \equiv 2 \ln \frac{\mathcal{L}(\vec{\epsilon} = \sum \vec{\epsilon_i})}{\mathcal{L}(\vec{\epsilon} = 0)} \tag{2.10}$$

where $\vec{\epsilon}$ is the true $\vec{E}_t^{miss}$ and $\sum \vec{\epsilon_i}$ is the observed $\vec{E}_t^{miss}$, computed by summing over all reconstructed objects in the event. In the numerator, the likelihood evaluates

the probability that the true value of $\vec{E}_t^{miss}$ equals the observed value while the denominator corresponds to the null hypothesis. The objects resolution is propagated into the denominator.

2.2.3 The muon system

The detection of muons is one of CMS's most important task. However, muons can penetrate several meters of iron without interacting: unlike most particles, they are not stopped by any of CMS's calorimeters. Therefore, chambers to detect muons are placed at the very edge of the experiment, where they are the only particles likely to register a signal. The muon stations sit outside the magnet coil and are interleaved with iron "return yoke" plates, that allow to fully exploit the 1.8 Tesla return flux of the magnetic field. By tracking the muon's position through the multiple layers of each station, combined with tracker measurements, the detector precisely traces the muon's trajectory. Since the trajectory is bent by the magnetic field, a measurement of the muon's impulsion can be precisely determined: the p_T resolution in the central region for muons with p_T up to 1 TeV is better than 10%.

In total there are 1400 muon chambers using three different technologies: 250 Drift Tube (DT), 540 Cathode Strip Chamber (CSC) and 610 Resistive Plate Chamber (RPC). The DTs are located in the barrel region ($|\eta| < 1.2$) where radially four detection stations are placed in 5 wheels of 12 sectors. 60 chambers compose the three first layers while there are 70 chambers in the fourth one. The first three stations are made of 8 chambers providing a measurement of the coordinates in the (r-ϕ) bending plane and a measurement of the z coordinate.
The CSCs are placed in the end-cap region ($0.9 < |\eta| < 2.4$). They provide a fast response, are radiation resilient, and have a fine segmentation. There are 4 stations of CSC in each end-cap, where the chambers are perpendicularly positioned with respect to the beam line and interspersed between the return yoke plates.
A complementary system composed of RPC, more dedicated to the trigger system, helps to improve the p_T resolution. A total of 6 layers of RPCs are embedded in the barrel muon system, 2 in each of the first 2 stations and one in each of the last 2 stations. The whole muon system is shown on Fig. 2.12.

Muon reconstruction

In CMS, the muon reconstruction can be done by using only muon chambers information ("StandAlone" muon), only tracker information ("Tracker" muon), or both ("Global" muon).

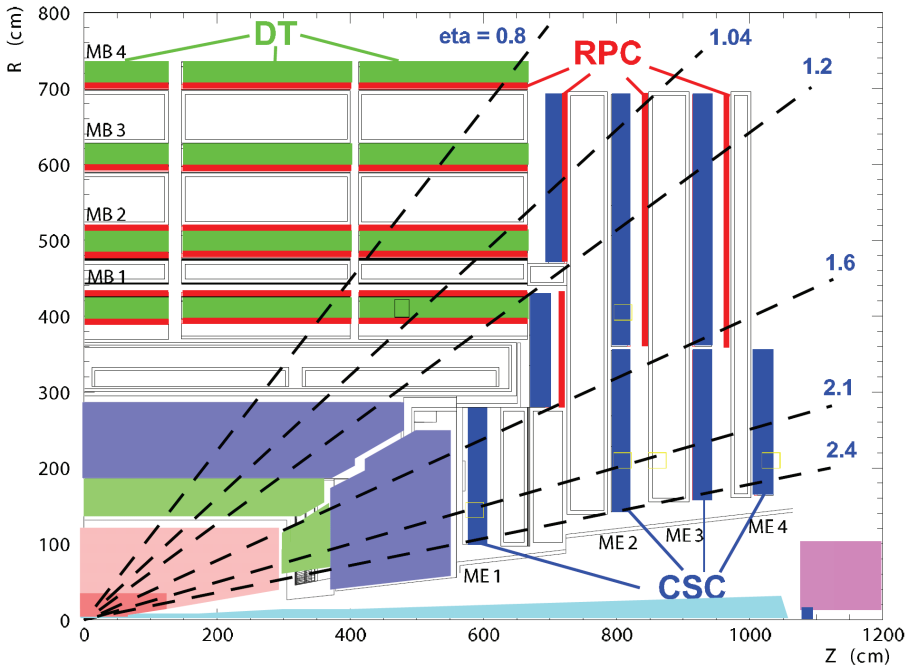

Figure 2.12: Layout of one quadrant of CMS. The four DT stations in the barrel (MB1-MB4, green), the four CSC stations in the end-cap (ME1-ME4, blue), and the RPC stations (red) are shown.

- The ”StandAlone“ muon is reconstructed using the muon track segment inside the muon system;
- The ”Tracker“ muon is reconstructed using tracker tracks with $p_T > 0.5$ GeV. Their trajectory is extrapolated to the muon system taking into account the bending of the magnetic field as well as the energy loss and scattering with the detector material. If a compatible hit in the muon chamber is found, the track is qualified as a tracker muon;
- The ”Global“ muon is reconstructed by taking the opposite approach of the tracker muon method: the initial seed for the reconstruction is one track segment in the DTs or three in the RPC. Then, a Kalman filter technique is applied to reconstruct the track in the direction of the tracker. The trajectory parameters are updated at each step and once the extrapolation of the trajectory is done, the final parameters are compared to the ones of the tracker muon candidate. If the match is satisfactory, all the hits from the tracker and the muon system are combined to perform a final fit on the Global muon's track.

The muon identification is then performed by requiring additional cuts depending on the desired working point:

- A ”Loose“ muon, which is a global or a tracker muon identified by the particle flow algorithm (see Section 2.3).
- A ”Soft“ muon: the tracker muon is matched with muon segments not used by other reconstructed muons. Additional cuts are applied on the number of tracker (pixel) layers presenting hits (> 5 (0)), the transverse and longitudinal impact parameter ($d_{xy} < 0.3$ cm and $d_z < 20$ cm with respect to the primary vertex). Besides, the muon's track has to be considered as very well reconstructed; Soft muons are mainly used for analyses using muons with $p_T < 10$ GeV;
- A ”Tight“ muon is a global muon reconstructed by the particle flow algorithm, matching the following quality criteria: normalized χ^2 of the track < 10, at least one muon chamber hit included in the muon track fit, presence of muon segments in at least two muon stations, muon tracker track with a transverse impact parameter $d_{xy} < 2$ mm and longitudinal distance of the tracker track $d_z < 5$ mm with respect to the primary vertex, number of hits in the pixel detector > 0 and number of tracker layers presenting hits > 5. Tight muons are mostly used in CMS analyses;

Performance

The determination of the reconstruction and identification efficiency is done by means of a so called "Tag and Probe" method [55], which uses $J/\psi \to \mu\mu$ or $Z \to \mu\mu$ events, known to contain at least two muons with a certain invariant mass. This method allows to obtain an almost unbiased estimation of the efficiencies for the different stages of the reconstruction. The "tag" is a muon that passes a very tight selection criteria (being a reconstruction, identification, trigger, or isolation criteria) and a very low fake rate, while the "probe" has a looser criteria, loose enough to not bias the efficiency estimation. The efficiency is then:

$$Eff = \frac{N_{passing-probe}}{N_{all-probes}} \tag{2.11}$$

This tool is used to compute the reconstruction, identification, trigger and isolation efficiency separately, for different η regions and p_T ranges; the obtained SF are then convoluted together, to obtain a "per-event" scale factor.

The muon reconstruction is robust and efficient at 99% within the detector acceptance [56]. As it is shown on Fig. 2.13, it allows to reproduce the different di-muon resonances.

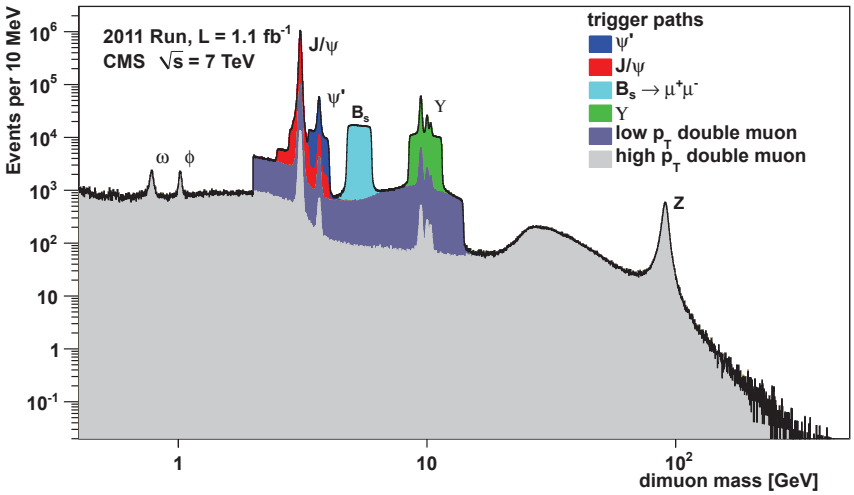

Figure 2.13: Invariant-mass spectra of opposite-sign muon pairs, for 2011 data corresponding to an integrated luminosity of 1.1 fb^{-1} [57].

2.2.4 Trigger system and storage

At the LHC, collisions occur every 50 ns (40 MHz frequency), but only a small fraction of interesting events for physics analyses are produced. In order to select only these events, a trigger system has been developed. The CMS trigger system has two levels:

- The level 1 trigger level (L1) [58]: it is designed to reduce the 40 MHz input rate to 100 kHz, in order to make the data acquisition possible. The selection is achieved by a hardware system, using information from the muon chambers and from the calorimeters. The four best reconstructed muons are kept and transfered to the high level trigger; the HCAL towers information is combined with the ECAL crystals response and if a threshold in p_T and E_T is reached, the event is kept. No information from the tracking system is used here, in order not to exceed the 3.2 μs of decision allowed time.
- The High Level Trigger (HLT) [59]: it reduces the data rate to 1 kHz by analyzing the events kept by the L1 trigger and reconstructing more complex objects. Several selections are applied, corresponding to different trigger paths, created to cover the needs of a large set of physics analyses. Unselected events are lost.

Data storage

Finally, the event is fully reconstructed, using the official framework of the CMS collaboration, CMS SoftWare (CMSSW) [40]. They are then stored to different computation centers, called ”Tier”. The Tier-2 centers in CMS are the only locations, besides the specialized analysis facility at CERN (Tier-0), where users are able to obtain guaranteed access to CMS data samples. Tier 0 distributes the raw data and the reconstructed output to Tier-1s, and reprocesses data when the LHC is not running. The Tier-1 centers (13 computer centers) are used primarily for organized processing and storage. The Tier-2s are specified with data export and network capacity to allow the centers to refresh the data in disk storage regularly for analysis. A nominal Tier-2 of 810 TB of storage for CMS was deployed in 2012. There are around 155 Tier-2 sites around the world.

2.2.5 Detector simulation

The detector response simulation is essential to simulate the behavior of the detector when particles are passing through. This is determined with GEANT4[60], a framework that contains a detailed description of the detector: material budget, areas with sensor readout/dead material, geometry, alignment, etc... This allows an accurate detector response simulation to determine acceptance, nuclear interactions and detector reconstruction effects.

First, each sub-detector geometry is modeled and the simulated particles go through each of them, starting from the interaction point. At this level, particle can decay and interact with the detector material. At the end of this step, particles and hits have been simulated.
Then the detector response is simulated: the energy deposits are converted and digitalized in order to be used by the reconstruction algorithms, the calorimetric deposits are converted in photo-electrons, and hits in the muons chambers are collected. Finally, the event reconstruction is performed, using the same reconstruction algorithms as for data event reconstruction.
With this framework, it is also possible to artificially degrade the hit position or energy deposits determination, in order to simulate a wrong detector alignment or bad detector calibration.

2.3 Particle-flow reconstruction

The Particle Flow (PF) algorithm [61][62] is an advanced method, more powerful than the reconstruction procedures described in the previous sections. The aim of the PF algorithm is to provide a single list of reconstructed particles (photons, charged hadrons, neutral hadrons, muons and electrons), combining information from all the CMS subdetectors. This list constitutes a complete description of the event, easy to handle, which is then used as input to higher level reconstruction algorithms: reconstruction of jets, calculation of the MET, and identification of τ and b jets. A schematic representation of the algorithm can be seen on Fig. 2.14.

The particle flow algorithm consists of the following steps:

- The first step gathers the fundamental ingredients coming from all the subdetectors: calorimeter clustering, tracking information and extrapolation to the calorimeters, muons and electrons identification;

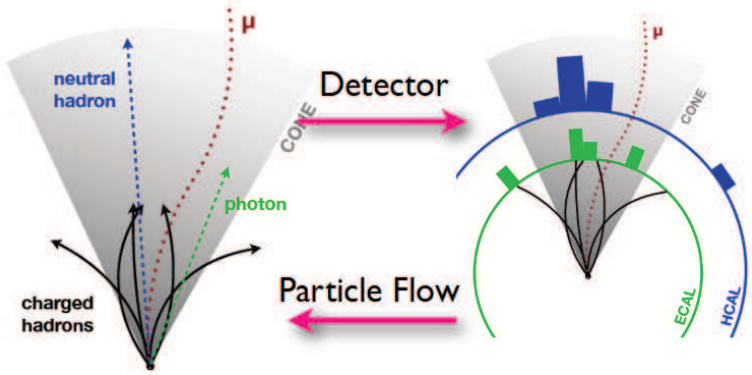

Figure 2.14: Schematic representation of the PF objects in comparison to detector level objects: blocks are reconstructed for each sub-detector and then combined for particle reconstruction and identification [63].

- These elements are combined by producing ”blocks“. Indeed, a given particle is expected to give rise to several PF elements in various CMS sub-detectors (for instance a muon can have one track block and one muon block). These elements must then be connected to each other by a link algorithm to fully reconstruct the particle, while avoiding double counting from different detectors.
  For example, a link between a track block and a calorimeter block is done by extrapolating the trajectory of the track block from its last measured hit in the tracker to ECAL and to HCAL, taking into account the particle shower profile. The track block is then linked to any given cluster whose position is matching the extrapolation;

- The particle identification and reconstruction is performed. For each block, the algorithm proceeds as follows:

  1. First, each global muon gives rise to a ”particle-flow muon”; the corresponding track is removed from the block list;
  2. The electron reconstruction and identification follows. Since electrons tend to produce short tracks and to loose energy by Bremsstrahlung in the tracker layers on their way to the calorimeter, they are first identified then their tracks are refitted with a GSF to follow their trajectories all the way to the ECAL. A final identification is performed using tracking and calorimetric variables. Each identified electron gives rise to a “particle-

flow electron". The corresponding track and ECAL clusters are removed from further processing of the block;

3. Tighter quality criteria are applied to the remaining tracks; while about 90% of them are fake tracks, the other 10% come from charged hadrons, photons or neutral hadrons. The neutral particles calorimetric clusters are well separated from the extrapolated position of the tracks, which constitutes a clear signature for photons (for ECAL deposits) and neutral hadrons (for HCAL deposits). Neutral particles overlapping with charged particles in the calorimeters can be detected as an excess of calorimetric energy with respect to the sum of the associated track momenta;
4. The PF jets are reconstructed using the jet algorithm described in Section 2.2.2, taking as input the PF collection, and the PF $\vec{E}_t^{miss}$ is calculated using the reconstructed PF particles.

Performance of the PF algorithm

The performance of several algorithms taking jets as input is found to be significantly improved when using PF jets instead of Calo-jets, reconstructed using only calorimetric information. On Fig. 2.15, the jet matching efficiency using PF jets with a given p_T is shown and found to be 95-97% for jets with $p_T < 200$ GeV. A gain of factor 2-3 in angular resolution is observed when using PF jets instead of Calo-jets, and a reduction of the dependency on the jet parton flavour is appreciated: it is less than 2% for jets with $p_T > 20$ GeV, instead of 10%.

The performance on the $\vec{E}_t^{miss}$ is also better when using the PF algorithm: the resolution is improved by almost a factor 2 with respect to the calorimeter-based $\vec{E}_t^{miss}$, at low E_T sum (see Fig. 2.16).

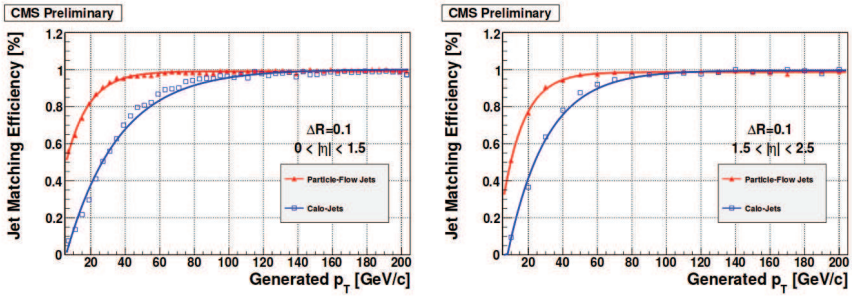

Figure 2.15: Jet matching efficiency as a function of the jet p_T, as obtained for Calo-jets (open squares) and PF jets (triangles) pointing to the barrel ($|\eta| < 1.5$, left plot), and to the end-caps ($1.5 < |\eta| < 2.5$), 1.5, with a matching distance (distance in the (η-ϕ) plane between the reconstructed jet and the matched generated jet) of 0.1. Efficiencies are fitted with an exponential functions of p_T ([62]).

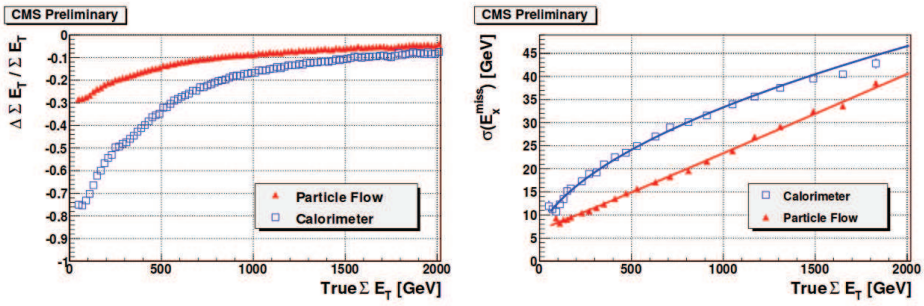

Figure 2.16: Left: response, for QCD multi-jets events, of the total visible transverse energy of the event, defined to be $(\Sigma E_T^{reco} - E_T^{true})/\Sigma E_T^{true}$, as a function of the true total visible transverse energy of the event. Right: resolution of the x-projection of $\vec{E}_t^{miss}$, obtained from a Gaussian fit, versus the total true visible transverse energy of the event. The solid triangles represent quantities based on PF reconstruction; the open squares represent quantities based on calorimeter reconstruction only ([62]).

2.4 Transfer functions for the MEM

The Transfer Functions (TF) $W(p^{vis}, p)$ have been introduced in Chapter 1 to consider the detector effects such as the resolution. Indeed, the detector is not perfectly hermetic and the event reconstruction is not 100% efficient. The TF are extracted from reconstructed events and defined as the conditional probabilities that translate the evolution between the final state at generator level p, characterized by its kinematics, and its associated reconstructed final state at detector level p^{vis}. For instance, colored particles produced by the hard interaction radiate and hadronize, producing jets. The evolution from quarks to jets has to be considered.

In principle, the TF depend on the topology of the reconstructed event, as well as the E, η and ϕ of the final state objects.The $W(p^{vis}, p)$ can thus end up as very complex functions. For simplification, particles are assumed to be independent from each other. From there, several approximations can be drawn: the TF are identical for any physics processes, they can be factorized for the N particles in the final state, and for their different kinematic parameters. Under these assumptions, the TF can be written as:

$$W(p^{vis}, p) = \prod_{i}^{N} W(p_i^{vis}, p_i) = \prod_{i}^{N} W_i(E_i^{vis}, E_i) W_i(\eta_i^{vis}, \eta_i) W_i(\phi_i^{vis}, \phi_i) \quad (2.12)$$

If the leptons direction is supposed to be well reconstructed, the corresponding TF become three dimensional δ-functions. The same assumption is made for the jet direction. Finally, only TF in energy must be determined.

The number n of generator-detector level particle pairs associated to reconstructed object having an energy between E^{vis} and $E^{vis} + \delta E^{vis}$, and a generated object with an energy between E and $E + \delta E$, is represented by $n(E^{vis}, E)\delta E^{vis}\delta E$, that can be written as:

$$n(E^{vis}, E)\delta E^{vis}\delta E = n(E)\delta E \times W(E, E^{vis}) \quad (2.13)$$

where $W(E, E^{vis})$ is the TF for energy.

2.4.1 Parametrized transfer functions

A first set of TF has been produced. The difference in energy between the partonic and reconstructed particle (jet or lepton) can be represented by a double Gaussian function, to take into account the non trivial tails and possible bias, especially at high

energy. As a result, the TF have been parametrized as the sum of two Gaussians not centered on the same value:

$$W(E, E^{vis}) = \frac{1}{2\pi(a_2 + a_3 \times a_5)}[\exp(-\frac{(E - E^{vis} - a_1)^2)}{2 \times a_2^2}) + a_3 \times \exp(\frac{(E - E^{vis} - a_4)^2)}{2 \times a_5^2})] \tag{2.14}$$

The parameters a_1 and a_2 represent respectively the mean and the width of the first Gaussian, while parameters a_4 and a_5 stand for the second Gaussian mean and width. The remaining a_3 parameter gives the ratio factor between the two Gaussians. All five parameters depend on the energy, such as:

$$a_i = a_{i,0} + a_{i,1} \times E + a_{i,2} \times \sqrt{E} \tag{2.15}$$

where i=1,...,5. This choice is motivated by the parametrization of the calorimeter energy resolution. This results in fifteen parameters, extracted by maximizing an unbinned maximum likelihood, using a significant number n of partonic-reconstructed pairs of a particle j, for the dedicated TF. The likelihood is build as:

$$-ln(L) = -\sum_j \ln(n(E_j^{vis}, E_j)) = -\sum_j n(E_j) - \sum_j E \times W_j^E(E_j, E_j^{vis}) \tag{2.16}$$

The first term does not depend on the a_i parameters and can be ignored. The transfer function parameters are extracted from the second term using MINUIT (a ROOT package), by performing a minimization of:

$$-ln(L) = -\sum_j W(E_j, E_j^{vis}) \tag{2.17}$$

A significant downside of this TF production is that the fit must be done several times in order to optimize the parameters, which can be very time consuming and complex. On the other hand, there is a significant advantage in using this set: since these TF are analytic functions, the tails are very precisely modeled.

Parametrized TF have been determined for the 2011 analysis using a high statistic sample of fully-leptonic $t\bar{t}$ and DY events. The b quark – jet pairs are selected within the detector and trigger acceptance and such that the angular distance in the $(\eta - \phi)$ plane between the parton and the reconstructed jet is less then 0.3. Besides, the jet energy reconstruction is not uniform on all the rapidity range covered by the CMS detector: forward/backward jets (with $1.6 < |\eta| < 2.4$) are poorly reconstructed in comparison with central jets (with $0.0 < |\eta| < 1.6$). Therefore, two TF are extracted,

for each η region. A quality check of the TF and its parameters is done by comparing the TF with the projection of $\Delta E = E - E^{vis}$ for four different range of partonic energy.

The results of this check for the electrons and jets TF can be seen in Appendix A.

I was in charge to build the muon transfer function, that has been parametrized differently with respect to the jets and electrons TF: instead of using the energy, the variable $\frac{1}{p_T}$ has been used. It is motivated by the fact that the muon reconstruction is achieved using tracker and muons chambers information, whereas calorimetric deposits seed the jets and electrons reconstruction. Therefore, the main resolution effect comes from the uncertainty on the Sagitta of the muon track. Beside, unlike the jets and the electrons, the muons TF has not been divided into different η regions. The parameters extracted by the fit are presented in Table 2.1, and plots on Fig. 2.17 show the TF projection for different p_T windows, in good agreement with respect to $\Delta E = E - E^{vis}$. The extracted parameters for the electrons and jets TF can be seen in Appendice A.

Table 2.1: Parameters of the muon TF extracted by maximizing an unbinned likelihood fit for muons in the detector acceptance. A double Gaussian parametrization (Eq. 2.14) depending on $\frac{1}{p_T}$ is chosen here, with $a_i = a_{i,0} + a_{i,1} \times \frac{1}{p_T} + a_{i,2} \times \sqrt{\frac{1}{p_T}}$.

	Independent term	$\frac{1}{p_T}$ **term**	$\sqrt{\frac{1}{p_T}}$ **term**
a_1	$a_{10} = -1,89.10^{-04} \pm 1,00.10^{-06}$	$a_{11} = 1,87.10^{-06} \pm 8,30.10^{-07}$	a_{12} = 0,00
a_2	a_{20} = 0,00	a_{21} = 0,00	$a_{22} = -2,99.10^{-03} \pm 7,52.10^{-08}$
a_3	a_{30} = 0,00	$a_{31} = 4,90.10^{-02} \pm 1,38.10^{-03}$	a_{32} = 0,00
a_4	$a_{40} = -1,09.10^{-03} \pm 1,49.10^{-04}$	a_{41} = 0,00	$a_{42} = 4,98.10^{-01} \pm 1,02.10^{-04}$
a_5	$a_{50} = 1,69.10^{-03} \pm 1,49.10^{-04}$	a_{51} = 0,00	a_{52} = 0,00

The parametrized TF have also been used for the analysis performing a search for a SM Higgs boson decaying to bottom quarks and produced in association with a Z boson, using the CSV tagger and the 2012 dataset [64]. In this thesis, a new set of TF has been created, the main goal being the improvement of TF production procedure.

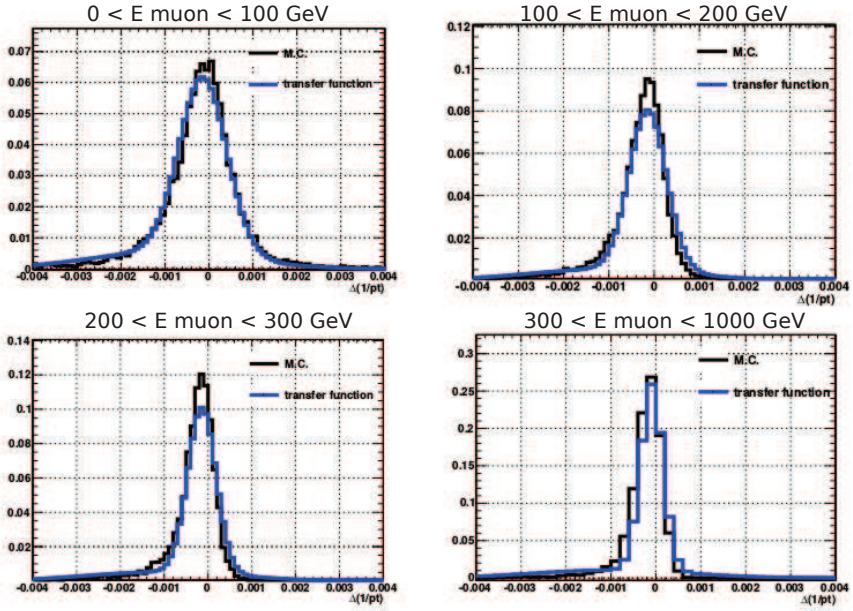

Figure 2.17: Comparison of the TF obtained for muons, with the expected $\Delta\frac{1}{p_T} = (\frac{1}{p_T}) - (\frac{1}{p_T}^{vis})$ distribution, for different generated muon energy ranges.

2.4.2 Binned TF

The binned TF have been produced following a procedure already set up for producing Delphes [65] TF; however, small adaptations had to be performed to create CMS binned TF. This new procedure is rather straightforward: a 2D-histogram containing the information about $\Delta E = E - E^{vis}$ as a function of E is filled using a high statistic sample of fully-leptonic $t\bar{t}$ for leptons/jets, within the detector and trigger acceptance. Then this histogram is smoothed and normalized to one for each bin in E. The content of this final histogram is extracted to get the TF. The main advantage of this TF set, in addition to its production simplicity, is the direct determination of the TF (no fit is performed).

As for the previous TF set, electrons and jets TF have been estimated for two η regions. The TF for muons remains undivided and for this set, and it is parametrized in energy as well, for simplicity reasons.

An important parameter in this procedure is the binning of the histograms: it has to be chosen with care, to uniformly fill the histograms as the energy increases. The remaining spikes are removed by a smoothing procedure, using a ROOT function based on the algorithm 353QH [66], to guarantee a good integration of Eq. 1.30. An example of such histogram can be seen on Fig. 2.18 for the central jets TF.

The new framework directly produces the control plots to check the quality of the TF, and they can be seen for jets on Fig. 2.19. For electrons and for muons, the plots are available in Appendix A. A good agreement can be seen between the MC distribution of ΔE and the projected TF for a specific range in E. Small discrepancies between the two curves can appear when the tails are truncated on the plot since both curves are renormalized to one once the x range has been set.

A comparison of the performance of the parametrized fitted TF and the new binned TF has been done, in order to compare the discrimination power between signal (ZH events) and one of its main background (fully-leptonic $t\bar{t}$ events). To do so, the distributions of the weights, computed with the fitted TF and with the binned TF, are used: a scan is performed to compute the signal efficiency as a function of the background efficiency. Similar performance are obtained in both cases. This result can be appreciated for the muon case on Fig. 2.20. For the electrons and jets TF, the plots are available in Appendix A.

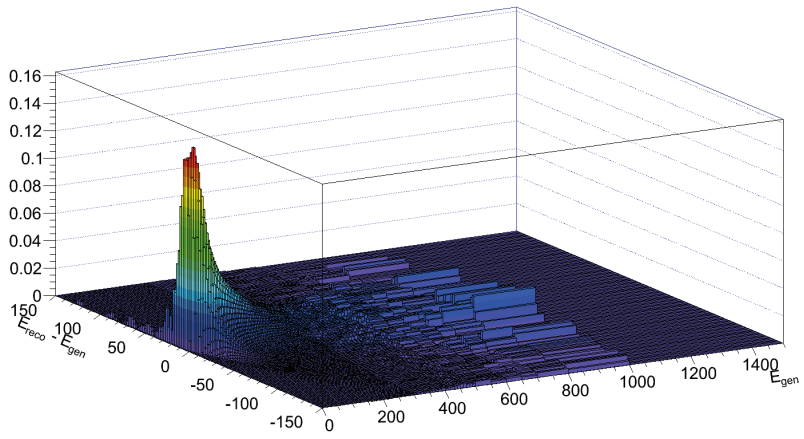

Figure 2.18: 2D histogram representing $\Delta E = E - E^{vis}$ as a function of E, where E represents the energy of a central jet ($0.0 < |\eta| < 1.6$).

A table summarizing the changes between the production of parametrized TF and new binned TF can be found in Table 2.2.

Table 2.2: Comparison between the parametrized fitted TF and the new binned TF.

	Fitted TF	**Binned TF**
Production	Not user friendly, long and iterative	Very user friendly, direct
Adapted for 2012 analysis	No	Yes
Performance	Similar	
Advantages	Determination of the tails for high E_{gen}	TF determination more accurate

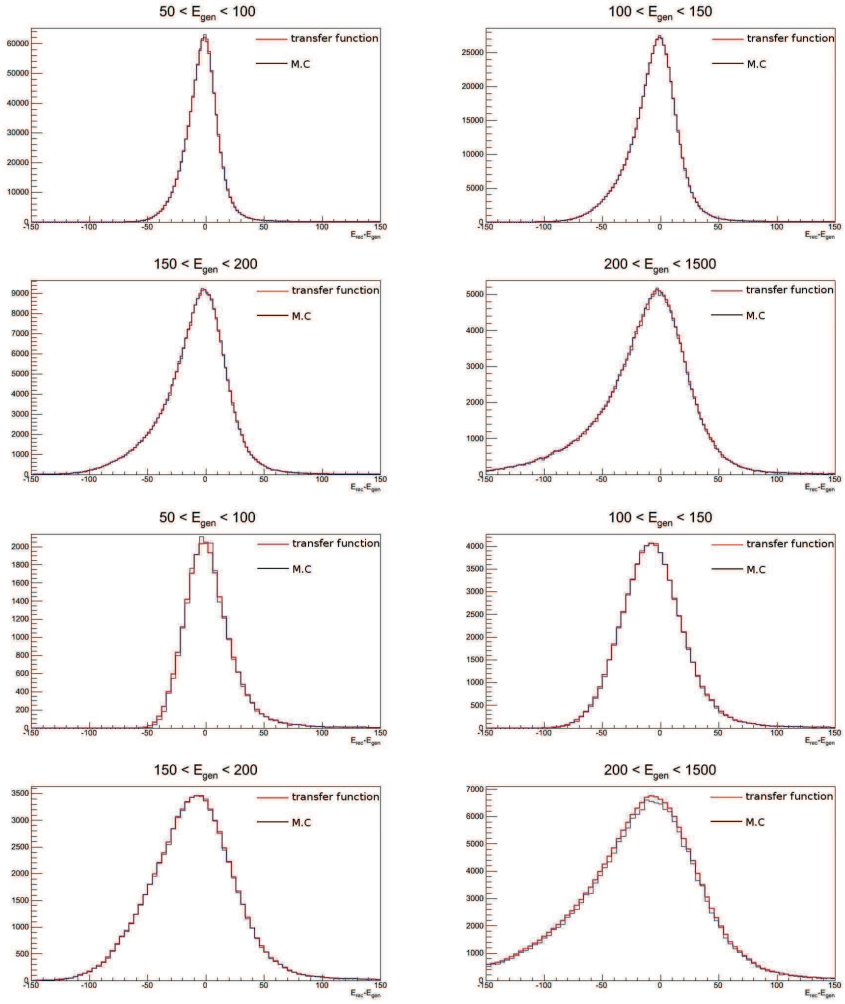

Figure 2.19: Comparison of the TF function obtained for jets in the central region (top) and in the forward-backward regions (bottom), with the projection of $\Delta E = E - E^{vis}$, for different generated b quark E ranges.

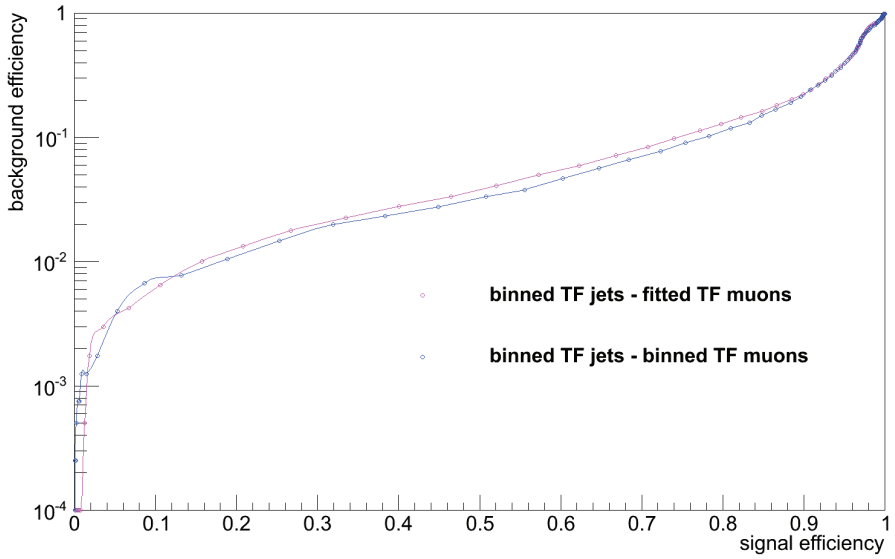

Figure 2.20: Comparison of performance for ZH (signal) versus $t\bar{t}$ events, using parametrized fitted TF (pink curve) and binned TF (blue curve) for the muons.

Chapter 3

b jets identification in CMS

3.1 Algorithms for b jet identification

Many physics analyses performed in CMS, from SM measurements to searches for new physics, rely on the identification of jets originating from the hadronization of a b quark, called b-tagging. These jets have special properties, induced by the long lifetime of the B hadron produced during the hadronization, such as the presence of an inner Secondary Vertex (SV) (as seen on Fig. 3.1, left plot). Tracks originating from this vertex have a large Impact Parameter (IP), whereas tracks coming from the Primary Vertex (PV) have an IP compatible with zero, reflecting the tracking resolution. Besides, due to the semi-leptonic decay of the B hadron, $\sim$20% of b jets contain a muon or an electron.

These properties are used to build taggers, algorithms capable of distinguishing b jets from light jets, defined as jets arising from the hadronization of u, d and s quarks, as well as jets from gluons.

On Fig. 3.2, the b-tagging efficiency versus the probability to misidentify a light jet as a b jet (mistag) is shown, for the taggers used in analyses performed with the $\sqrt{s}$ = 7 TeV dataset [67]. Among them, Jet Probability (JP) and Combined Secondary Vertex (CSV) are the two taggers that give the best performance: a 80% probability to tag a b jet corresponds to an expected mis-identification rate (or mistag) under 10%. Along with the Track Counting High Purity (TCHP), JP and CSV are the only

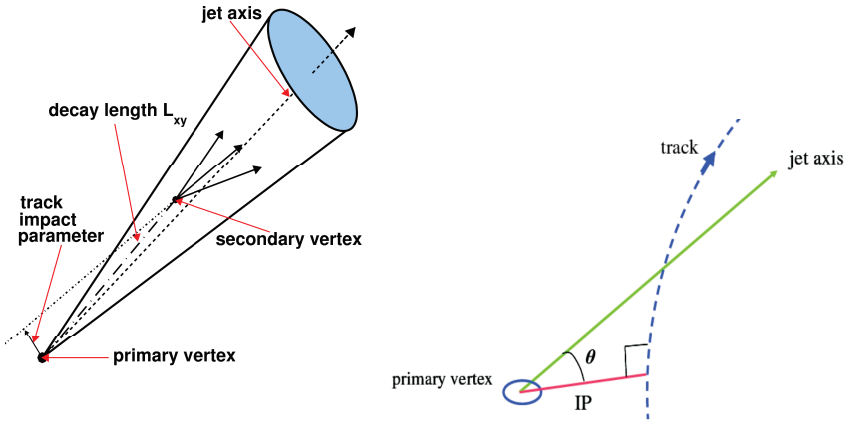

Figure 3.1: Left: schematic view of a B hadron decay inside a b jet [33]. Right: construction of the IP of a track in 2 dimensions [21].

Tagger	Variables used in the algorithm	Energy supported	b-tagging efficiency at 1% mistag rate
TCHE	Track IP	7 TeV	60%
TCHP	Track IP	7 and 8 TeV	55%
JP	Track IP	7 and 8 TeV	63%
JBP	Track IP	7 TeV	67%
SSVHE	SV flight distance from PV	7 TeV	55%
SSVHP	SV flight distance from PV	7 TeV	-
CSV	All variables from SV + Track IP	7 and 8 TeV	67%

Table 3.1: List of the different algorithms for b jet identification in CMS, along with their distinctive characteristics, the energy of collision for which the tagger was used, and the b-tagging efficiency corresponding to a mistag rate of 1%.

three taggers supported by the b-tagging group at $\sqrt{s}$ = 8 TeV [68][69]. For the $\sqrt{s}$ = 7 TeV dataset, other taggers were available, such as Simple Secondary Vertex High Efficiency (SSVHE) and Simple Secondary Vertex High Purity (SSVHP), the TC tagger in the high efficiency mode, Track Counting High Efficiency (TCHE), and a version of the JP tagger in which the four tracks with the highest IP weight more in the algorithm, Jet B-Probability (JBP). Table 3.1 summarizes all the taggers available in CMS for the 7 and 8 TeV datasets, along with their main properties.

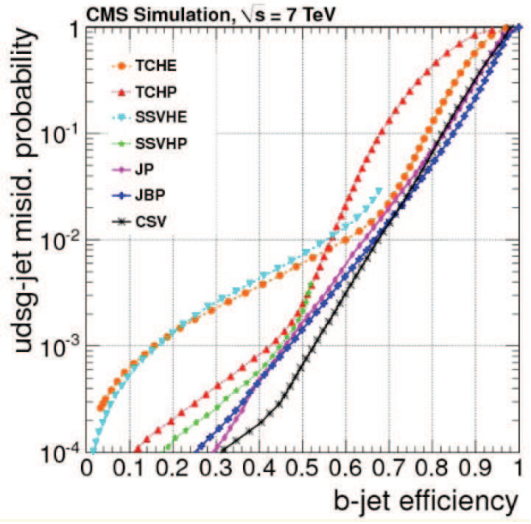

Figure 3.2: b-tagging performance as a function of the mis-identification probability, for all the b-tagging algorithms used at $\sqrt{s}$ = 7 TeV [67].

3.1.1 Track based b-tagging

TC(HE-HP) and JP are based on a single discriminating variable: the IP of a charged track. It represents the minimal distance between the track and the PV, and is calculated in three dimensions by taking advantage of the excellent resolution of the pixel detector along the z axis. It is determined by the linearization of the track trajectory from the closest point of approach to the PV, and signed accordingly to the scalar product of the vector pointing from the PV to the point of closest approach with the jet direction (see Fig. 3.1 plot for its geometric construction). Tracks originating from the decay of particles traveling along the jet axis, like a B hadron, tend to have positive IP values. In contrast, the IP of tracks coming from the PV can be equally positive or negative, reflecting the detector resolution. Tracks coming from c jets can also have high IP values resulting from the long life-time of charm hadrons, but with a less noticeable impact on the track IP distribution.

The IP significance, IP/σ, defined as the ratio of the IP to its estimated uncertainty, is used as an observable in order to take into account the effect of the resolution. For light jets, the IP/σ distribution is expected to be symmetric around zero, whereas for b jets (and c jets), it is more populated at the positive and large values. The distributions of IP/σ for tracks coming from light jets, c jets and b jets are shown on Fig. 3.3. A small

asymmetry is seen for the light jets, coming from K_s^0 and Λ, light strange particles with a long life-time, called V_0 particles.

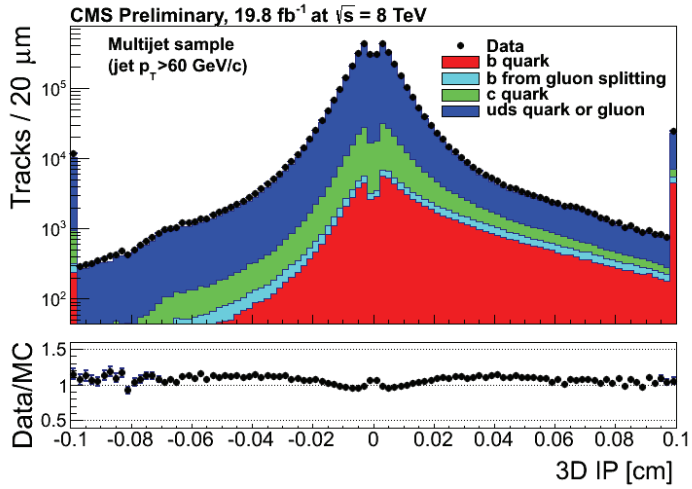

Figure 3.3: Distribution of the IP for light jets (dark blue), c jets (green) and b jets (red). The filled circles correspond to the data sample recorded at $\sqrt{s}$ = 8 TeV [68].

The TCHP(HE) algorithm is based on the third (second) track with the highest IP/σ: if this tracks passes a given threshold value, the jet is tagged as a b jet. This tagger is simple and robust, and the distribution of the returned discriminator value for the 2012 $\sqrt{s}$ = 8 TeV dataset is shown on Fig. 3.4.

JP is a more sophisticated algorithm, computing the compatibility for a set of tracks associated to a jet to come from the PV: if this probability is low, the jet is likely to be a b jet.

Jet Probability

The JP algorithm takes as inputs tracks with negative IP, used to build resolution functions $R(x)$, from which the probability $P_{tr}(S)$ that a track with a given IP/σ is coming from the PV is extracted:

$$P_{tr}(S) = sign(S).\int_{|S|}^{\infty} R(x)dx \tag{3.1}$$

Different $R(x)$ are used, depending on the quality of the track. These resolution functions are not perfect Gaussians and are difficult to model. Therefore, the $R(x)$ are

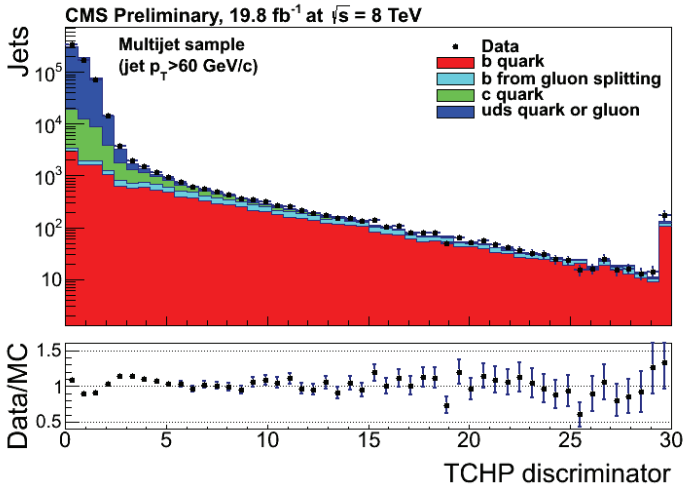

Figure 3.4: Distribution of the TCHP discriminator for light jets (blue), c jets (green) and b jets (red). The filled circles correspond to the data sample recorded at $\sqrt{s}$ = 8 TeV [68].

determined directly from the distributions of track IP/σ when $IP/\sigma < 0$. For these tracks, by construction, $P_{tr}(S)$ is flat between -1 and 0, as they are mostly light tracks for which the distribution of IP/σ is symmetric.

The probability P_{jet} that a jet containing N tracks is coming from the PV is computed by combining the probabilities $P_{tr}(S)$ of the jet's N tracks with a Poissonian function:

$$P_{jet} = \prod \times \sum_{j=0}^{N-1} \frac{(-ln \prod)^j}{j!} \tag{3.2}$$

where:

$$\prod = \prod_{i=1}^{N} \widetilde{P}_{tr}(i) \tag{3.3}$$

$\widetilde{P}_{tr}$ is the redefined track probability $\widetilde{P}_{tr} = P_{tr}/2$ for $P_{tr} > 0$ and $\widetilde{P}_{tr} = 1 + P_{tr}/2$ for $P_{tr} < 0$, is introduced to keep the track probability always positive. It is also required that if $P > 5.10^{-3}$, P is set at $P > 5.10^{-3}$, to avoid P_{jet} to accept zero values.

As mentioned previously, the $R(x)$ functions are known to be different depending on the track quality reconstruction. This means that the algorithm should be calibrated to

be sensitive to all kind of tracks. As a consequence, several categories of tracks are defined and for each track, the $P_{tr}(S)$ is computed according to the category the track belongs to. The $R(x)$ functions can be built from data events, allowing a data driven calibration of JP.

Calibration of JP

The calibration of JP is done in several steps. Firstly, only tracks fulfilling the following track selection criteria are used:

- Number of hits associated to the track in the Pixel detector NPix ≥ 2;
- Number of hits associated to the track in the tracker (Pixel+SiStrip)≥ 8;
- Transverse IP with respect to the PV $|IP_{2D}| < 0.2$ cm;
- Transverse momentum $p_T > 1$ GeV;
- Normalized χ^2 (χ^2/nod, the number of degrees of freedom) of the track reconstruction's quality < 5;
- Distance to jet axis < 0.07 cm;
- Decay Length (DL) < 5 cm.

Then these tracks are sorted in categories. The different categories are defined using the following variables:

- Number of hits associated to the track in the Pixel detector;
- Number of hits associated to the track in the tracker (Pixel+SiStrip);
- Momentum p of the track;
- Pseudo-rapidity η of the track;
- Normalized χ^2 of the track.

The categories used for the official calibration of JP are the nine following categories:

- 1 category for tracks with $\chi^2 > 2.5$ (category 1);

For tracks with $\chi^2 < 2.5$:

- 3 categories for tracks with $|\eta|$ in the ranges [0, 0.8], [0.8, 1.6] and [1.6, 2.5] with NPix $>= 3$ and $p < 8$ GeV (respectively category 2, 3 and 4);
- 1 category for tracks with NPix = 2 and $p < 8$ GeV (category 5);
- 3 categories for tracks with $|\eta|$ in the ranges [0, 0.8], [0.8, 1.6] and [1.6, 2.5] with NPix $>= 3$ and $p > 8$ GeV (respectively category 6, 7 and 8);
- 1 category for tracks with NPix = 2 and $p > 8$ GeV (category 9).

The distributions of the resolution functions for all the track categories can be seen on Fig. 3.5. It is clear on this plot that $R(x)$ differs from one category to another, and it highlights the relevance of having different categories for the calibration of JP.

The calibration of the algorithm and its validation are achieved using the following procedure:

- The $R(x)$ functions are constructed for each category of tracks, using a specific sample of data events, by filling binned histograms. Afterwards, they are stored in a calibration file;
- This new calibration is applied to the same sample of events used to build the $R(x)$ functions;
- The distribution of P_{tr} for tracks with $IP/\sigma < 0$ is analyzed: since the calibration is applied on the same events used for the $R(x)$ construction, by construction the distribution should be flat. An example of the distribution used to check the calibration, using the $R(x)$ shown on Fig. 3.5 (top plot), can be found on Fig. 3.5 (bottom plot).

The JP discriminator is constructed to be proportional to $-\ln(P_{jet})$ and its distribution, using the calibration for the $\sqrt{s}$ = 8 TeV dataset, can be seen on Fig. 3.6. Some structures in peaks are visible: they are due to an artifact of the JP algorithm : tracks with probabilities < 0.005 are accounted as tracks with probabilities strictly equal to 0.005 (stated in Section 3.1.1). This means that multiple tracks can have the exact same track probability. The first (second) peak thus means that there is only one (two) track(s) with a probability of 0.005.

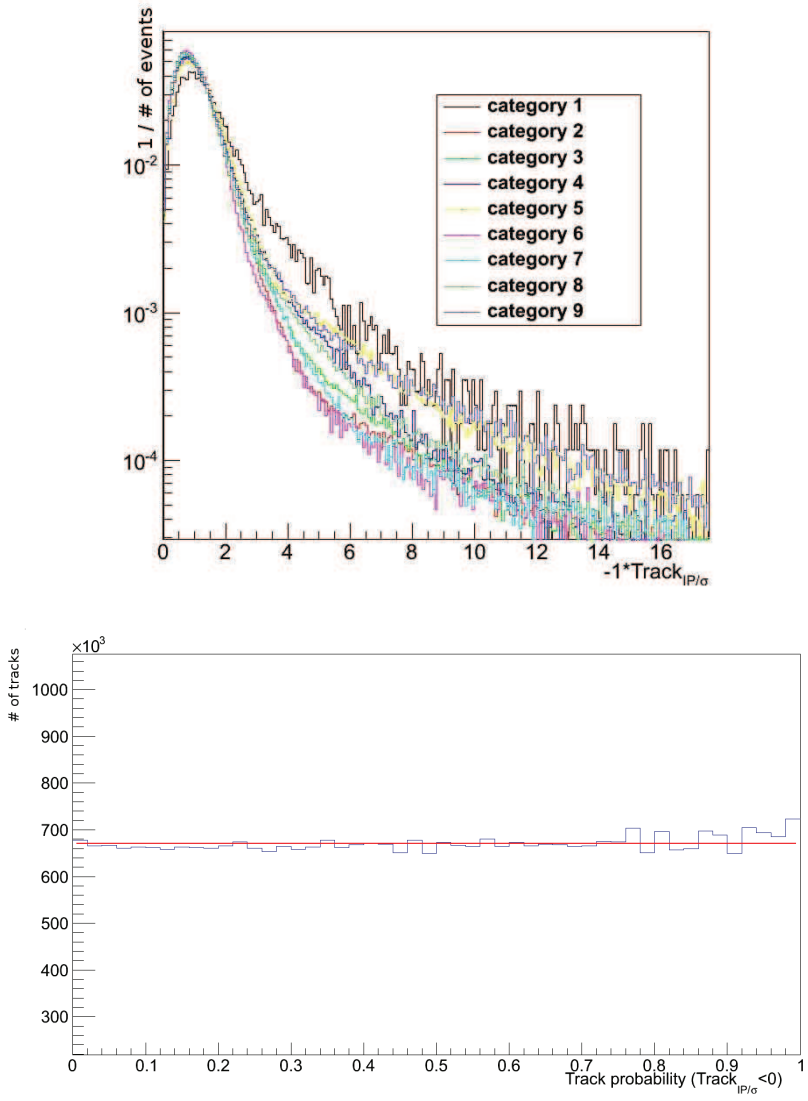

Figure 3.5: Top: positive resolution functions for the nine categories of tracks defined above, using tracks coming from multi-jets events with jets with $80 < p_T < 120$ GeV. The histograms have been normalized to unity. Bottom: distribution of the track probability (P_{tr}) for tracks with $IP/\sigma < 0$, from all the nine categories defined below, for jets with $80 < p_T < 120$ GeV. This distribution has been done with the same jets used for building the $R(x)$ functions seen on the top plot. The red ligne represents a straight line.

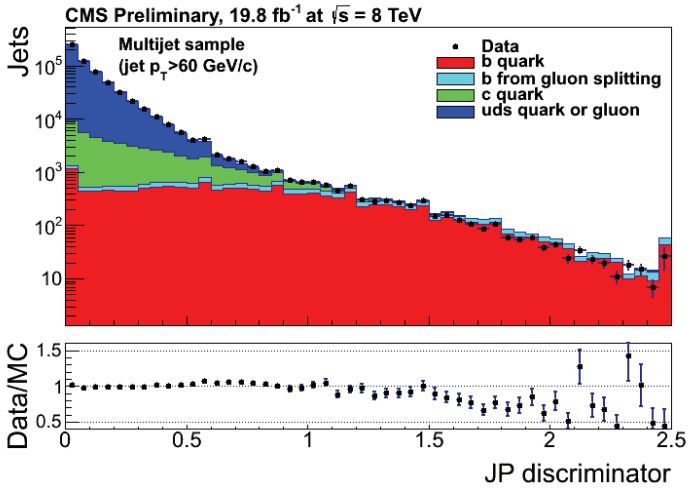

Figure 3.6: Distribution of the JP discriminator for light jets (dark blue), c jets (green) and b jets (red). The filled circles correspond to the data sample recorded at $\sqrt{s}$ = 8 TeV [68].

JP presents several specific properties:

- The main advantage is its robustness since its calibration is performed directly using data. This can be a significant advantage with respect to the CSV tagger, since the latter first needs to be trained on MC, but JP can be directly calibrated and used, even if the MC shows inconsistencies, or is not available;
- Since the IP/σ is used for the calibration of the algorithm, the calibration already takes into account the detector resolution effects;

The variables entering the categories for the calibration are highly tracking-dependent, which means that for every new detector and data taking conditions, the calibration has to be changed. Calibrations need to be done and validated for each data-taking period, and in that context, the framework used to perform these two tasks should be robust and fast.

The first existing calibration framework uses two different and independent packages: the $ImpactParamerLearning$ package, which performs the calibration, and the $BTagAnalyzer$ package, to test and validate it. However, it is not straightforward that the track selection, for example, is the same in both packages. Besides, the

$BTagAnalyzer$ package is also used by the group in charge of the b-tagging commissioning and the b-tagging efficiency measurement: if the $ImpactParamerLearning$ package is not synchronized, inconsistent results inside the b-tagging group might be produced. Thus, the new framework has been based on the $BTagAnalyzer$ package.

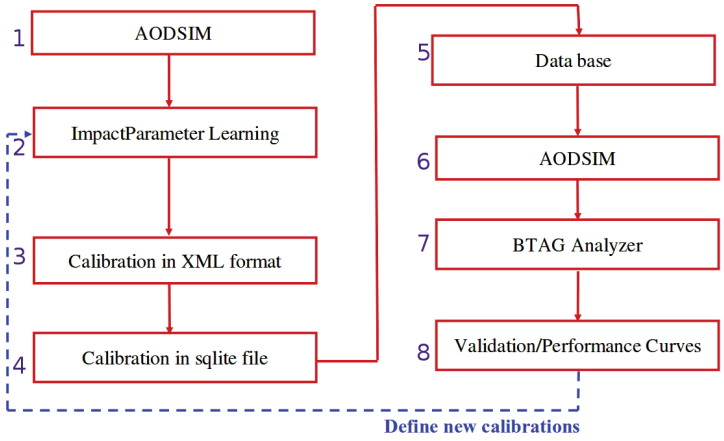

Figure 3.7: Previous framework to perform the calibration of JP, based on the $ImpactParameterLearning$ package.

As it shown on Fig. 3.7, the calibration performed using the original framework is done in several steps. Steps 1 to 4 are dedicated to the calibration production, and steps 5 to 7 are for its validation:

1. First, from the light format samples (called AODSIM samples), the $ImpactParamerLearning$ package is run on the CMS computing grid with CRAB [70];
2. The calibration is directly written in a raw format (xml format);
3. This file is converted into a file readable by the CMS database (sqlite file);
4. The calibration is transfered to the CMS database;

Then, the calibration is tested:

5. New AODSIM events are produced, including the new calibration in the data-taking scenario (CRAB has to be run again);

6. The $BTagAnalyzer$ code is run on the previous events to create files with specific selected information, called "NTuples";

7. From these NTuples, the performance of the new calibration is tested: the histograms of track probability for tracks with $IP/\sigma < 0$ are analyzed for every category of tracks and for all the tracks together. If these histograms are flat, the calibration is validated.

If the calibration is not satisfactory, all the steps from 1 to 7 must be redone. Otherwise, the calibration is included in a new data-taking scenario.

The main weaknesses of this procedure are that several grid jobs have to be run twice, which can be time consuming, and that inconsistencies can appear between the $ImpactParamterLearning$ package and the $BTagAnalyzer$, since two different codes have to be maintained.

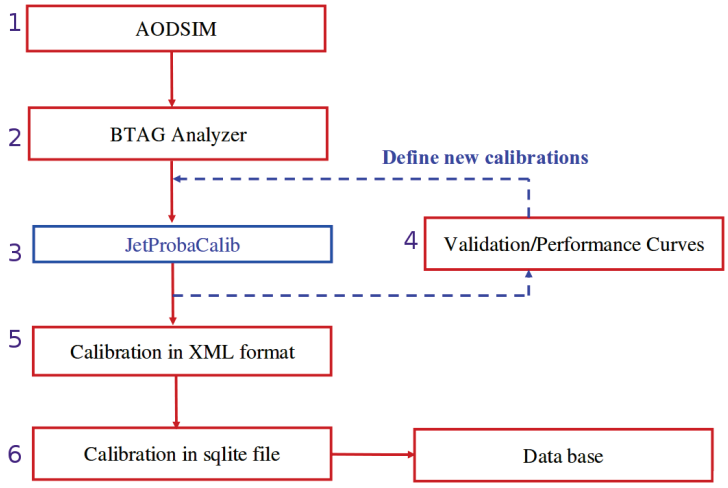

Figure 3.8: New framework for performing the calibration of JP, based on the $BTagAnalyzer$ package.

Hence a new calibration framework has been introduced in order to make the calibration process faster and easier to use with only one package: the $BTagAnalyzer$ package. Fig. 3.8 shows all the steps of this new procedure. From step 1 to step 3, the calibration is created, and then it is tested during step 4:

1. First, AODSIM events are produced similarly to the previous case, using CRAB;

2. Then the $BTagAnalyzer$ is run to produce the NTuples. The NTuples production is necessary for other b-tagging tasks (mainly performance measurements and commissioning, as mentioned before), so the Ntuples might have already been produced, and this step could be skipped;

3. These NTuples are used as inputs for a ROOT-based code, $JetProbaCalib.C$, which is CMSSW environment free. This code produces a ROOT file containing the categories in histogram format and the new calibration in xml format;

Then the calibration is directly tested:

4. The histograms contained in the output ROOT file are used to check the calibration with another independent ROOT code, $JetProbaValidation.C$: it produces the histograms of P_{tr} for tracks with $IP/\sigma < 0$, for all the different categories of tracks, so they can be directly analyzed;

5. If the calibration is satisfactory, it is added to the CMS database.

If the calibration does not give good performance, it can be easily redone by repeating step 3 only, and then tested by re-running step 4. More information about how to calibrate JP with this new framework can be found on a dedicated Twiki page [71].

There are various advantages to perform the calibration with this new framework:

- CRAB is only run once, implying a large gain of time and CPU occupation (especially if the NTuple production has already been done);
- The $JetProbaCalib$ code is CMSSW independent, so it can be run with less incompatibilities and more flexibility (if the working release changes from one calibration to another for example);
- The code creates $CategoryDef$ objects, inheriting from $TObject$ to store the entire calibration in a ROOT file, including the category definition. It is then straightforward to add and to edit categories;
- A book keeping can be done to easily and quickly check the old calibrations;
- The previous point is of main interest for studying the impact of the calibration in the high p_T jets region (see Section 3.4.3): all the calibration histograms are stored in one ROOT file and can be retrieved by another code to study their impact on the JP performance;

- The definition of the categories is not anymore hard-coded in multiple codes, making the procedure more robust;

Validation of the procedure

In order to validate the new calibration procedure, several calibrations have been produced with the same sample of events (QCD simulated events with a generated p_T in the range [30, 1000] GeV):

- A calibration using the previous procedure with the $ImpactParameterLearning$ package, named "old calibration";
- Another calibration using the $ImpactParameterLearning$ package, in which the track selection has been modified to correspond to the track selection of the $BtagAnalyzer$, named "old calibration modified". This calibration is done to perform a closure test, to make sure the new procedure is behaving similarly to the previous one;
- A new calibration, following the new procedure, named "new calibration".

Then, JP is computed for these three different calibrations, and the distributions of the track probability, for tracks with $IP/\sigma < 0$, are displayed on Fig. 3.9. The "old calibration modified" and the "new calibration" curves are on top of each other, meaning that the closure test is successful. Besides, the "new calibration" curve is flat, implying that the calibration produced with the new procedure is valid.

A summary of all the changes between the two procedures can be found in Tab. 3.2.

Table 3.2: Comparison between the previous and the new framework to perform the JP calibration.

	Previous procedure	**New procedure**
Required time for production	At least 3 days	Less than a day
Run with CRAB	2 times	1 time or not necessary
b-tagging group sync.	Not direct	Direct
Framework	CMSSW dependent	CMSSW independent, ROOT-based
Output	Calibration in xml	Calibration in xml, ROOT file with categories

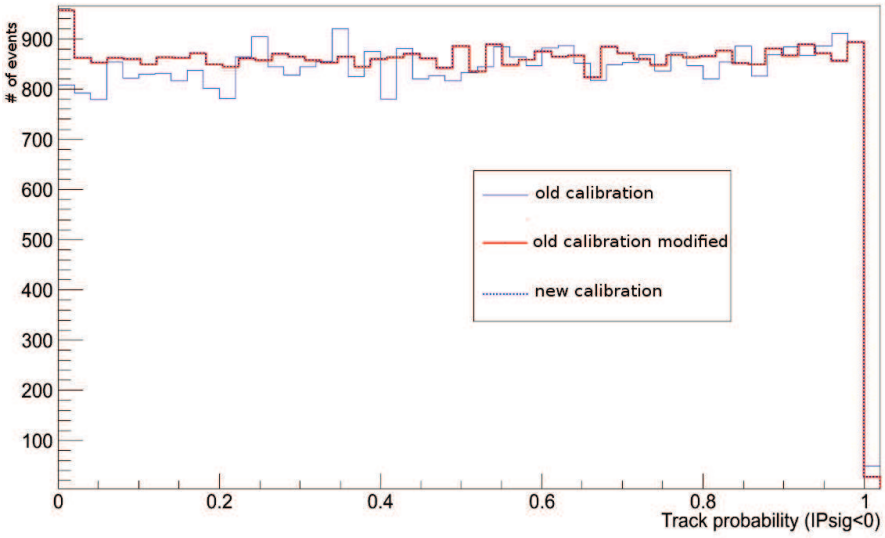

Figure 3.9: Comparison of the track probability distribution, for tracks with $IP/\sigma < 0$, using the old calibration (solid blue line), the old calibration with the same track selection than the one in the $BtagAnalyzer$ (red line), and with the new calibration (dashed blue line). The curves of the old modified calibration and the new one are on top of each other.

3.1.2 Combined Secondary Vertex

The CSV [72] algorithm is the tagger that currently gives the best b-tagging performance and as a consequence, is the one mainly used in CMS analyses, as in those presented in Chapter 4 and 5. It is a sophisticated and complex tagger that exploits all known discriminating variables between b jets and non b jets. It combines different topological and kinematic variables linked to the SV with the track IP/σ, using a likelihood ratio technique to compute the b-tagging discriminator. A variant of this tagger uses a Multi-variate Analysis (MVA) tool. This algorithm has the advantage to cover the cases when no SV is found, using three categories:

- "Real vertex": in this case the algorithm exploits the full SV information;
- "Pseudo vertex": in this scenario, a pseudo vertex is created from two tracks with $IP/\sigma > 2$ and the properties of this pseudo vertex are used;
- "No vertex": here, the algorithm solely accounts for the tracks information.

The distribution of the discriminator value is shown on Fig. 3.10. A visible disagreement between MC and data can be seen; however, this distribution will be truncated (b jets are required to pass a given discriminator threshold value), and scale factors are applied afterwards: this will correct for the MC/data discrepancy.

3.2 Performance measurements

The b-tagging efficiency and the mis-identification rate are extracted from simulated events. These numbers are then corrected by the efficiencies measured in data, and Scale Factor (SF) are extracted.

3.2.1 b-tagging efficiency measurements

The b-tagging efficiency is measured in data using several methods applied to QCD multi-jets and $t\bar{t}$ events, in different ranges in p_T and η. The CMS b-tagging group has designed several methods to compute this efficiency, summarized in the table shown on Table 3.3.

The efficiency ϵ measured in data is compared with the identification efficiency for b jets in the simulation, resulting in data/MC scale factor: $SF_b = \epsilon_b^{data}/\epsilon_b^{MC}$.

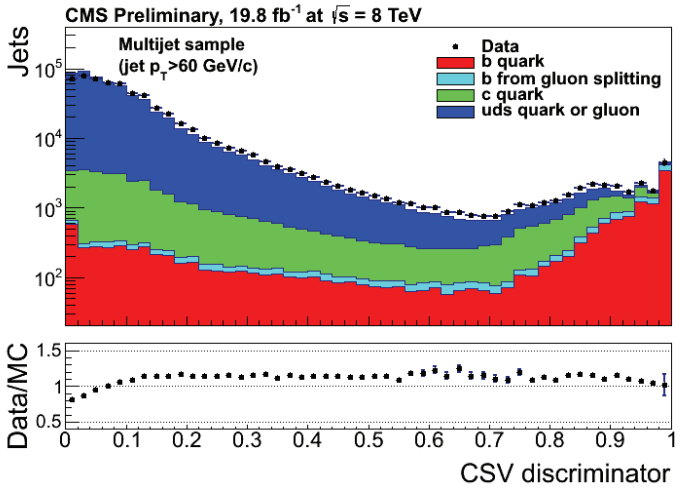

Figure 3.10: Distribution of the CSV discriminator for light jets (dark blue), c jets (green) and b jets (red). The filled circles correspond to the data sample recorded at $\sqrt{s}$ = 8 TeV [68].

Several systematic uncertainties affect the measurement of the b jet tagging efficiency. Some are common to all the methods, while others affect a specific method. The common systematic uncertainties for all methods are:

- The number of pile-up (PU) events;
- The difference of gluon splitting rates between data and simulation;
- For the methods using muon-jets, the central value of the b-tagging efficiency is extracted from data with muon $p_T > 5$ GeV, which affects the shape of the template distributions used in fits, and also the number of events used to measure the tagging efficiencies.

For the efficiency measurements using $t\bar{t}$ events, the uncertainties related to the PDF (see Chapter 1) is one of the most important systematic sources, as well as the jet-parton matching uncertainty. Other important uncertainties, such as the energy scales of the jets and, to a lesser extent, of the leptons, are taken into account, as they shift the momenta of the reconstructed objects.

Method	Variables used	Efficiency measurement technique
Using QCD multi-jets events		
IP3D/p_T^{Rel}	Muon IP3D/p_T^{Rel}	The fraction of each jet flavour in the sample is extracted using a likelihood fit. The efficiency is: $\epsilon_b^{tag} = \frac{f_b^{tag} \times N_{data}^{tag}}{f_b^{tag} \times N_{data}^{tag} + f_b^{untag} \times N_{data}^{untag}}$ where f_b^{tag} (f_b^{untag}) is the fraction of b jets tagged (un-tagged) extracted from the data and N_{data}^{tag} (N_{data}^{untag}) the total yields of tagged (un-tagged) jets.
LT	JP tagger (or CSV for JP)	A fit on the JP distribution is done to get the b, c and light contributions with the constrain $f_b + f_c + f_{light}$ =1. The efficiency is: $\epsilon_b^{tag} = \frac{C_b \times f_b^{tag} \times N_{data}^{tag}}{f_b^{before-tag} \times N_{data}^{before-tag}}$ where C_b is the fraction of jets with a JP information.
System 8	2 independent LT taggers, 2 different datasets, 2 categories of jets (b and non b jets)	8 unknowns, all depending on the number of passing or failing tags, are correlated by a set of 8 equations related to the tagging efficiencies. This equation is solved numerically to get the b-tagging efficiency.
Using $t\bar{t}$ events		
FTC ($t\bar{t}$ di-leptonic)	Number of tagged jets (4-7 jets)	The method requires consistency between the observed and the expected number of tagged jets, using a likelihood fit. The $t\bar{t}$ cross section and b-tagging efficiency are free parameters of the fit.
FTM ($t\bar{t}$ semi-leptonic)	Number of tagged jets (2-3 jets)	The method requires consistency between the observed and the expected number of tagged jets, using a likelihood fit.
bSample ($t\bar{t}$ semi-leptonic)	Reconstructed top pair event	A kinematic fit is used to associate jets to the quarks from the top decay; b jets from the leptonic decaying top quarks are used to define a b jets-enriched sample. This sample is further sub-divided into two sub-samples: one enriched in b jets, the other depleted (using the invariant mass of the charged lepton + candidate b jet). The efficiency is derived by the discriminator distribution in the enriched subsammple, subtracted by the same distribution in the b-depleted sub-sample.

Table 3.3: Methods used for the measurement of the b-tagging efficiency within CMS at $\sqrt{s}$ = 8 TeV [68], along with their specific features.

tagger	SF_b in muon-jets	SF_b in $t\bar{t}$ events
JP (mistag rate 10%)	0.982 ± 0.020	0.966 ± 0.015
CSV (mistag rate 10%)	0.983 ± 0.017	0.987 ± 0.018
JP (mistag rate 1%)	0.947 ± 0.034	0.961 ± 0.012
CSV (mistag rate 1%)	0.951 ± 0.024	0.953 ± 0.012
TCHP (mistag rate 0.1%)	0.896 ± 0.035	0.921 ± 0.010
JP (mistag rate 0.1%)	0.866 ± 0.036	0.922 ± 0.017
CSV (mistag rate 0.1%)	0.916 ± 0.032	0.926 ± 0.036

Table 3.4: Scale factors SF_b obtained in muon-jet data and $t\bar{t}$ data, averaged over the p_T spectrum of jets from top decays [68], for the JP, CSV and TCHP taggers at different mistag rates, for the 8 TeV dataset. The overall uncertainties are given.

Combination of efficiency measurements

The combination is based on a weighted mean of the different SF measurements, taking into account correlated and uncorrelated uncertainties and evaluating the shared fraction of events between the different methods. Table 3.4 compares the combined scale factors SF_b measured in multi-jets and $t\bar{t}$ events, averaged over the p_T spectrum of jets from top decays.

3.2.2 Misidentification probability

The measurement of the misidentification probability for light jets relies on the definition of inverted tagging algorithms, selecting non-b jets instead of b jets, using the same variables and techniques as the standard versions. These "negative" taggers can be used in the same way as the regular b jet tagging algorithms both in data and in the simulation. As the negative-tagged jets are enriched in light flavours, the misidentification probability can be measured from data, and the value obtained for simulated events is used to extract a correction factor.

The discriminator values for negative and positive taggers are expected to be almost symmetric for light jets by resolution effect. Therefore the misidentification probability ϵ^{misid} can be derived from the rate of negative-tagged jets ϵ^{-} in inclusive jets data samples. The negative taggers are built from tracks with a negative impact parameter or from secondary vertices with a negative decay length. When a negative tagger is applied to jets of any flavour, the corresponding tagging efficiency is denoted "negative tag rate".

A correction factor, $R_{light} = \epsilon_{MC}^{misid}/\epsilon_{MC}^{-}$, is evaluated from the simulation in order to correct for second-order asymmetries in the negative and positive tag rates of light-flavour quark and gluon jets, and for the heavy flavour contribution to the negative tags:

$$\epsilon_{data}^{misid} = \epsilon_{data}^{-} \times R_{light} \tag{3.4}$$

The data/MC scale factors for the misidentification probabilities are then defined:

$$SF_{light} = \epsilon_{data}^{misid}/\epsilon_{MC}^{misid} \tag{3.5}$$

and their values are given in Table 3.5.
There are several systematic effects on the misidentification probability based on negative tags:

- The fraction of b jets that has been measured in CMS to agree with the simulation within a 20% uncertainty;
- The average fraction of gluon jets that depends on the details of the parton density and hadronization functions used in the simulation. An uncertainty of 20% is applied;
- The amount of reconstructed K_s^0 and λ particles; these light particles have a long lifetime and create an asymmetry in the light contribution of discriminator distribution;
- The PU model used in the simulation;
- The mis-measured tracks, coming from jets with a reconstructed track not associated with a genuine charged particle;
- The rate of secondary interactions in the detector, leading to photon conversions and nuclear interactions in the pixel detector layers;
- Small differences in the angle between a track and the jet axis can lead to a change of the impact parameter sign ("sign flip") and therefore modify the negative tag rate;
- Physics analyses use jets from different event topologies. For a given jet p_T, the misidentification probability is different for the leading jet or in case there are other jets with higher p_T values in the same event. Measured misidentification scale factors for leading and sub-leading jets have a dispersion of about 7%. In addition, misidentification SF vary by 2-7%, depending on the tagger, and for different running periods.

tagger	SF_{light}
JPL	$1.03 \pm 0.01 \pm 0.07$
CSVL	$1.10 \pm 0.01 \pm 0.05$
JPM	$1.10 \pm 0.02 \pm 0.20$
CSVM	$1.17 \pm 0.02 \pm 0.15$
TCHPT	$1.27 \pm 0.06 \pm 0.27$
JPT	$1.11 \pm 0.07 \pm 0.31$
CSVT	$1.26 \pm 0.07 \pm 0.28$

Table 3.5: Data/MC scale factors SF_{light} for different algorithms and operating points for jet p_T in the range [80-120] GeV [68]. Both statistical and systematic uncertainties are quoted.

3.3 The commissioning

The goal of the b-tagging commissioning group [73] is to validate the b-tagging algorithms for a specific dataset: all the b-tagging related variables are checked and compared with data to see if the agreement is satisfactory enough. In particular, one needs to validate:

- The MC generation;
- The trigger effects;
- The tracks selection;
- The discriminator distributions;
- The detector alignment;
- The algorithms calibration/training;
- The SV-related distributions;

During the first period of my thesis, I produced the first plots of comparison between the $\sqrt{s}$ = 7 and the $\sqrt{s}$ = 8 TeV dataset: a comparison between the MC at different energies, and a MC/data comparison at $\sqrt{s}$ = 8 TeV. For example, as it can be seen on Fig. 3.12, the PU distribution for the $\sqrt{s}$ = 8 TeV dataset is shifted to significantly higher values than for the $\sqrt{s}$ = 7 TeV dataset.

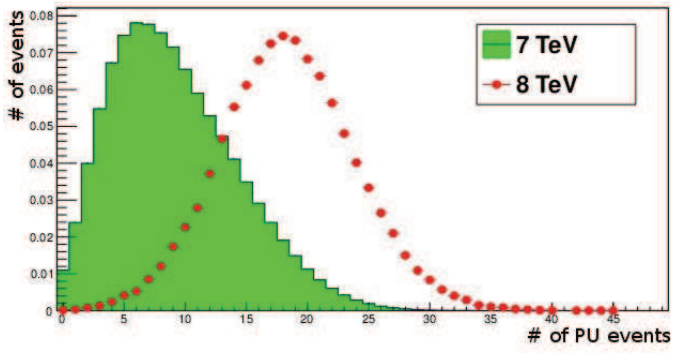

Figure 3.12: Comparison of the distributions of the PU for MC at $\sqrt{s}$ = 7 TeV (green histogram) and at $\sqrt{s}$ = 8 TeV (red dots).

3.3.1 New code and procedure

This first plot production revealed several issues with respect to the framework that was used back then: the code was not efficient enough and rather complex. Beside, all the plots were produced in a row. To get a single plot, one had to re-run the code on all events, artificially extending the procedure duration. The goal was then to create a new commissioning code using the $BTagAnalyzer$ package once again, and to make the plots production as simple and as user-friendly as possible. The use of the $BTagAnalyzer$ package is of main interest since this code is already used by the group in charge of the performance measurement. A common NTuples production is possible, resulting in a significant gain of time.

The content of the NTuples can be chosen by the user, depending on the plots to be produced, to lighten the procedure. Once the NTuples have been produced, one can start the plot production with many different configuration options, listed hereafter:

- Choice of the PU treatment, with the possibility to include the official CMS PU reweighing recipe [74]. It is also possible to choose a PU scenario (depending on the used samples), including a personal one;
- Choice of the MC generator used for the MC production (Pythia or Herwig);
- Various QCD samples with different generated p_T are available. However, since their cross section differ, each event is reweighted to take into account its impact among all the other QCD events. This procedure is done automatically;

- Possibility to use samples enriched in muons;
- A wide set of trigger selection is available.

Finally, a new code to produce the plots allows to make a selection on the list of plots. Several categories of plots are set and for every category, a boolean variable indicates if the plots should be produced:

- Track-related variables;
- Track-related variables at N-1 cut (to check the effect of the cut of interest);
- Secondary vertex-related variables;
- Muons-related variables;
- Taggers distributions;
- Tag rates;
- 2D plots.

Comparison between the old and new commissioning codes has been performed through the production of several plots. A good compatibility was observed. All the b-tagging related plots shown in this thesis, comparing data and MC, have been produced within this new framework, such as Fig.3.3, Fig.3.6 and Fig.3.10. Tab. 3.6 lists the advantages associated to the use of the new code.

Table 3.6: Comparison between the previous and the new framework to produce commissioning plots.

	Previous code	**New code**
Run the code	Not intuitive	User friendly
Run with CRAB	Necessary	May not be necessary
Plot production	Several hours	Several minutes
Sync. with b-tag. group	Not direct	Direct
PU scenario	Hard-coded	Several choices
MC generator	Pythia	Pythia and Herwig
Generated p_T-reweighing	Hard-coded	Automatic
Muon enriched samples	Hard-coded	Automatic configuration
Plot production	All plots produced at once	Plots selection

3.4 Study of JP at high p_T

As previously stated, JP is one of the tagger that gives the best performance: for a mistag rate of 10%, more than 80% of signal efficiency can be reached for b jets and almost 60% for c jets. However, this concerns only jets with a $p_T < 200$ GeV. Above that limit, as it is shown on Fig. 3.13, a clear degradation of the JP performance is observed between jets with $80 < p_T < 120$ GeV and jets with $200 < p_T < 300$ GeV. For a 1% probability to mistag a light jet, the b-tagging efficiency drops from 67% down to 59%. In addition, for a similar probability to mistag a c jet as a b jet, the b-tagging efficiency drops from 30% to 27%.

Recovering from this degradation is a goal to achieve in order to improve future physics analyses.

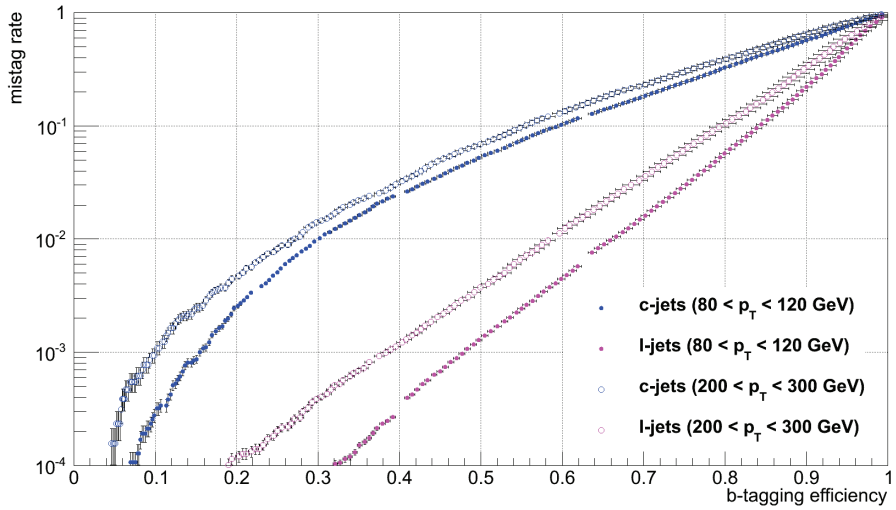

Figure 3.13: b-tagging efficiency as a function of the mistag rate, when a c jet is tagged as a b jet (blue) and when a light jet is tagged as a b jet (pink), for jets with $80 < p_T < 120$ GeV (plain circles) and for jets with $200 < p_T < 300$ GeV (empty circles). The gaps are directly related to the peak structures present in the distribution on Fig. 3.6.

3.4.1 Degradation of the performance

For this study, the $TrackHistory$ [75] class is used for MC events: it returns the origin of a specific track, and reveals if the track is coming from the decay of the B hadron (real B track).

First of all, in order to understand the behavior of the tracks inside high p_T jets, the b jet's track multiplicity as a function of the jet p_T is analyzed. Fig.3.14 shows the multiplicity of reconstructed tracks inside a b jet, as a function of the jet p_T. For jets with $p_T > 200$ GeV, the track multiplicity appears roughly constant. However, when tracks are sorted between real/non-real B tracks, a different behavior is observed: the multiplicity of reconstructed real B tracks inside a b jet decreases with the p_T of the jet (as it is visible on Fig. 3.15), highlighting a tracking inefficiency. On the other hand, the number of light components of the b jet, the non-real B tracks, logarithmically increases, as it is shown on Fig. 3.16.

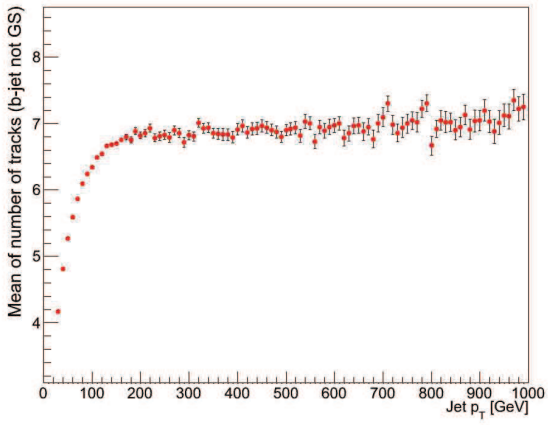

Figure 3.14: Multiplicity of reconstructed tracks inside a b jet, as a function of the jet p_T.

A direct consequence on the JP algorithm can be deduced by looking at the distribution of the track IP/σ inside a b jet, depending on the tracks origin: if only the real B tracks are taken into account, the distribution of IP/σ does not vary much when the b jet p_T increases, which is expected since IP/σ is Lorentz invariant (see Fig. 3.17, top plot).

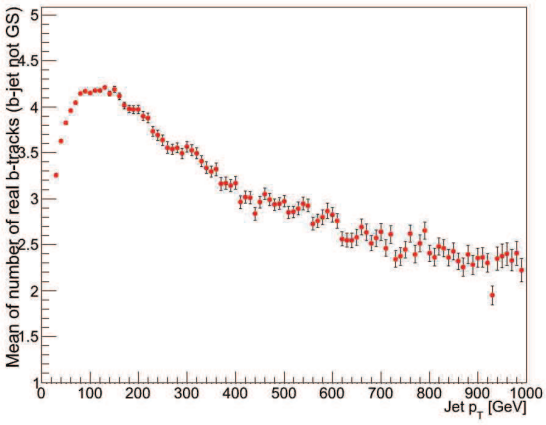

Figure 3.15: Multiplicity of reconstructed real B tracks inside a b jet, as a function of the jet p_T.

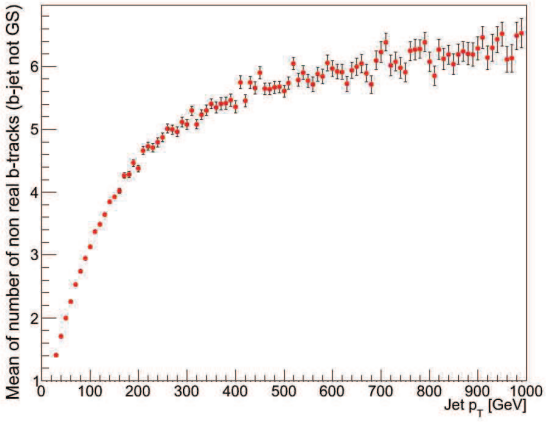

Figure 3.16: Multiplicity of reconstructed non-real B tracks inside a b jet, as a function of the jet p_T.

However, when all the tracks in the b jet enter the distribution, a clear dependence in jet p_T is visible (Fig. 3.17, bottom plot). This means that the light contribution of tracks in the b jet, which increases with the jet p_T, degrades the JP algorithm.

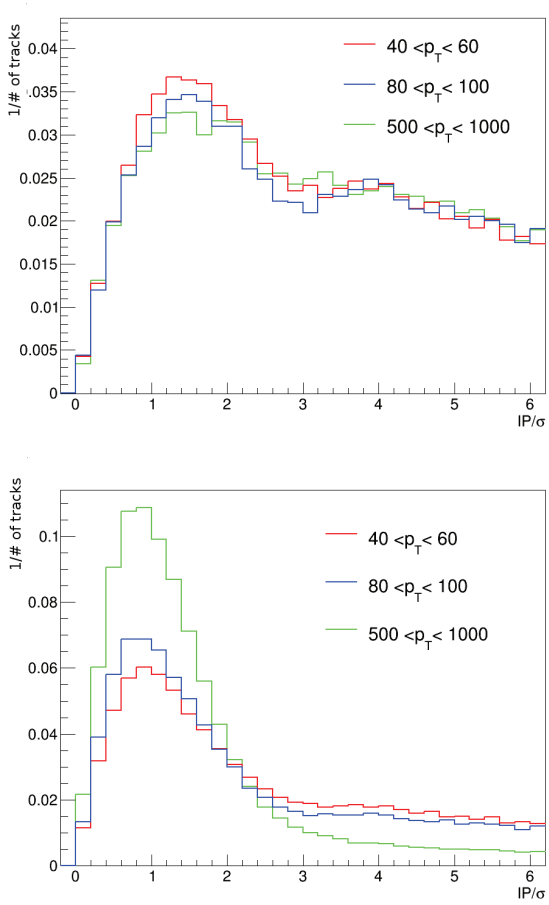

Figure 3.17: Distribution of the track IP/σ as a function of the jet p_T, for the real B tracks in the b jet (top) and for all the tracks in the b jet (bottom). The distributions have been renormalized to unity.

Gluon splitting

In the following, a distinction is made to separate b jets arising from Gluon Splitting (GS) from the other b jets. It can be seen on Fig. 3.18 that the fraction of b jets from

GS becomes significant as the jet p_T increases. However, these jets show a different real B tracks multiplicity distribution, as a function of the b jet p_T (see Fig.3.19). This can be explained by the fact that jets arising from GS can be very collimated if the initial gluon has a large enough momentum. In that case, the two arising b jets may be contained in a single reconstructed b jet, with tracks from both jets. Indeed, from Fig. 3.20, it can be deduced that in jets coming from GS, tracks tend to be closer to each other than tracks from non GS b jets.

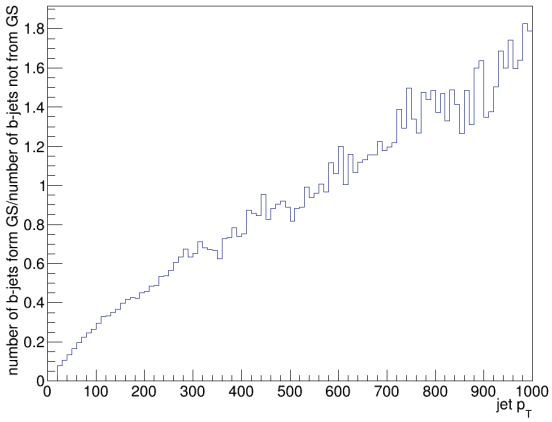

Figure 3.18: Ratio of the number of b jets coming from GS over the number of b jets not coming from GS, as a function of the jet p_T.

Therefore, for the rest of this study, in order to analyze the behavior of tracks within b jets in interest for physics analyses only, b jets coming from GS are discarded.

The amount of fake tracks (tracks not associated to a reconstructed track) inside b jets is also checked. However, the fake track rate only slightly increases with the p_T of the jet, as seen on Fig. 3.21, meaning that the non-B tracks, whose number increases as it is seen on Fig.3.19, are mostly real tracks coming from hadronization.

At this point, a first statement can be drawn: the degradation of JP for high p_T jets is coming from the loss of real B tracks inside the b jet, but also from the increasing

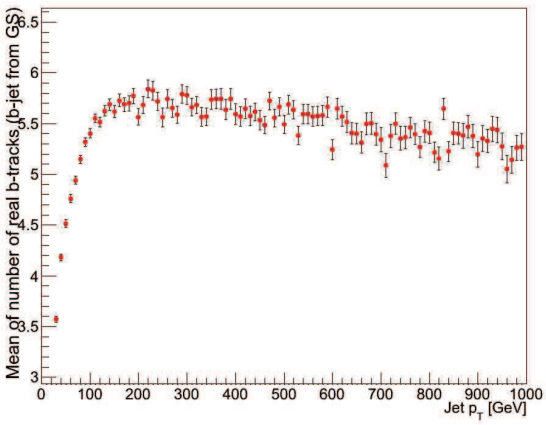

Figure 3.19: Multiplicity of tracks real B tracks from GS jets, as a function of the jet p_T, inside a b jet.

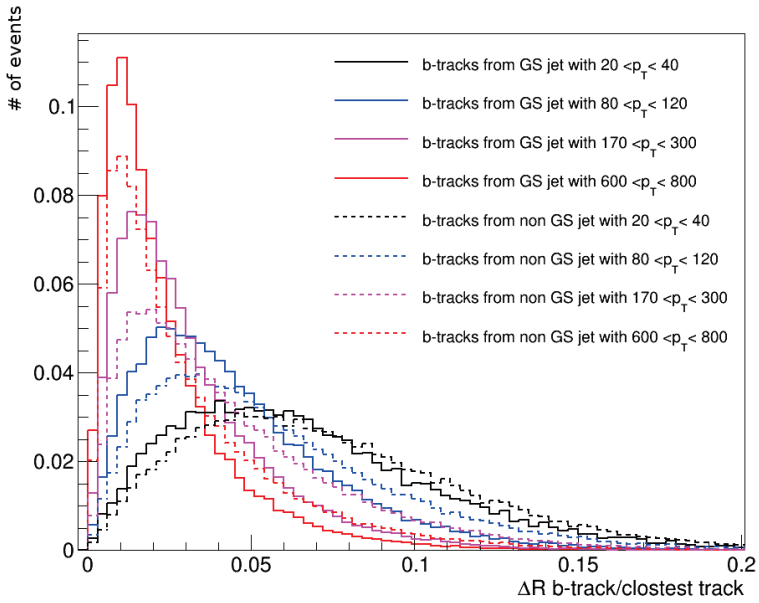

Figure 3.20: Angular distance between one track and its closest track inside a b jet, for non GS jets (dashed lines) and GS jets (plain lines). The distributions have been renormalized to unity.

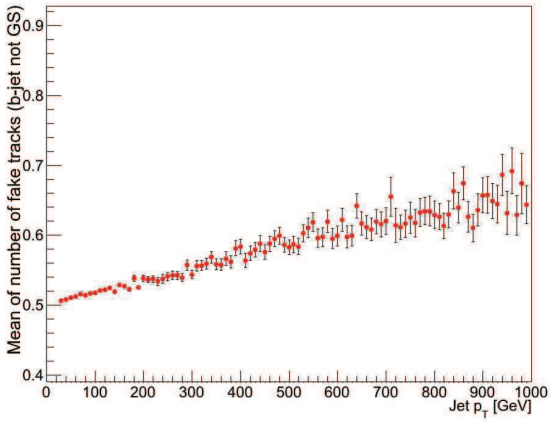

Figure 3.21: Multiplicity of fake tracks inside a b jet, as a function of the jet p_T.

number of non real B tracks inside the same jet, coming from the light components of the b jet. The association of these two phenomena would dilute the B tracks information in the computation of the jet probability performed in equation 3.2, decreasing the algorithm efficiency. Two approaches can be considered to improve JP at high p_T: either to recover the lost real B tracks, or to remove the non real B tracks from the algorithm computation.

Loss of B tracks

The loss of B tracks inside a b jet at high p_T was highlighted. This could mean that the track selection is not optimal for high p_T regions. The decay length of the B hadron in the (x-y) plane (rho, ρ) is then studied: the left plot on Fig. 3.22 shows that as the jet p_T increases, the B hadron flies a greater distance. On the right plot of Fig. 3.22, it can be seen that the fraction of B hadrons with $\rho > 4$ cm (at the edge the first layer of the pixel detector) is getting more and more important for higher jet p_T (5-8% for $p_T > 300$ GeV). As previously stated in Section 3.1.1, a cut on the track decay length (DL) is applied such that tracks with DL < 5 cm are not kept (these tracks most likely come from a hard interaction with the tracker material). However, this selection criterion may not be suitable for high p_T jets; effects of this cut are investigated in Section 3.4.3.

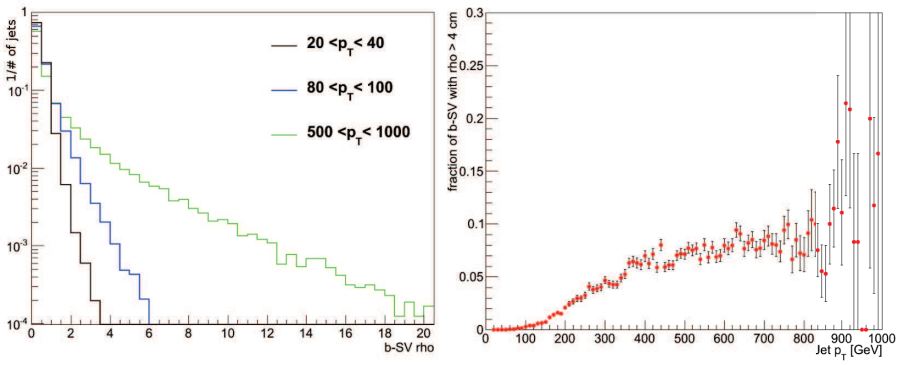

Figure 3.22: Left: distribution of the flying distance of the B hadron in the ρ plane for different jet p_T. Right: fraction of B hadrons when $\rho > 4$ cm, as a function of jet p_T.

Discriminating variables

Several variables are found to give a good discrimination between the real B tracks and the non real B tracks.

The p_T of the track is an obvious variable to look at: on Fig. 3.23 (top plot), it is possible to see that the mean p_T is significantly higher for real B tracks than for the other tracks. Another relevant variable is the angular distance ΔR between the track and the b jet: Fig. 3.23 (bottom plot) shows that a clear difference is observed between the B tracks and the non B tracks distributions, as the jet p_T increases. The real B tracks tend to be closer to the jet axis while the non B tracks are more spread out. An optimization of the jet-track association cone could be necessary at high p_T to remove the contamination.

The number of hits in the pixel detector is displayed on Fig. 3.24 (top plot). For this plot, the track selection (see Section 3.1.1) has been loosen to NPix ≥ 1. This reveals that real B tracks in high p_T jets tend to populate the bin 1, which is not filled by tracks fulfilling the nominal track selection. This implies that the track selection might not be optimal for the high p_T region. The p_T^{rel} distribution (shown on Fig. 3.24, bottom plot), corresponding to the transverse momentum of the muon relative to the direction of the total muon-jet momentum vector, shows higher values for real B tracks.

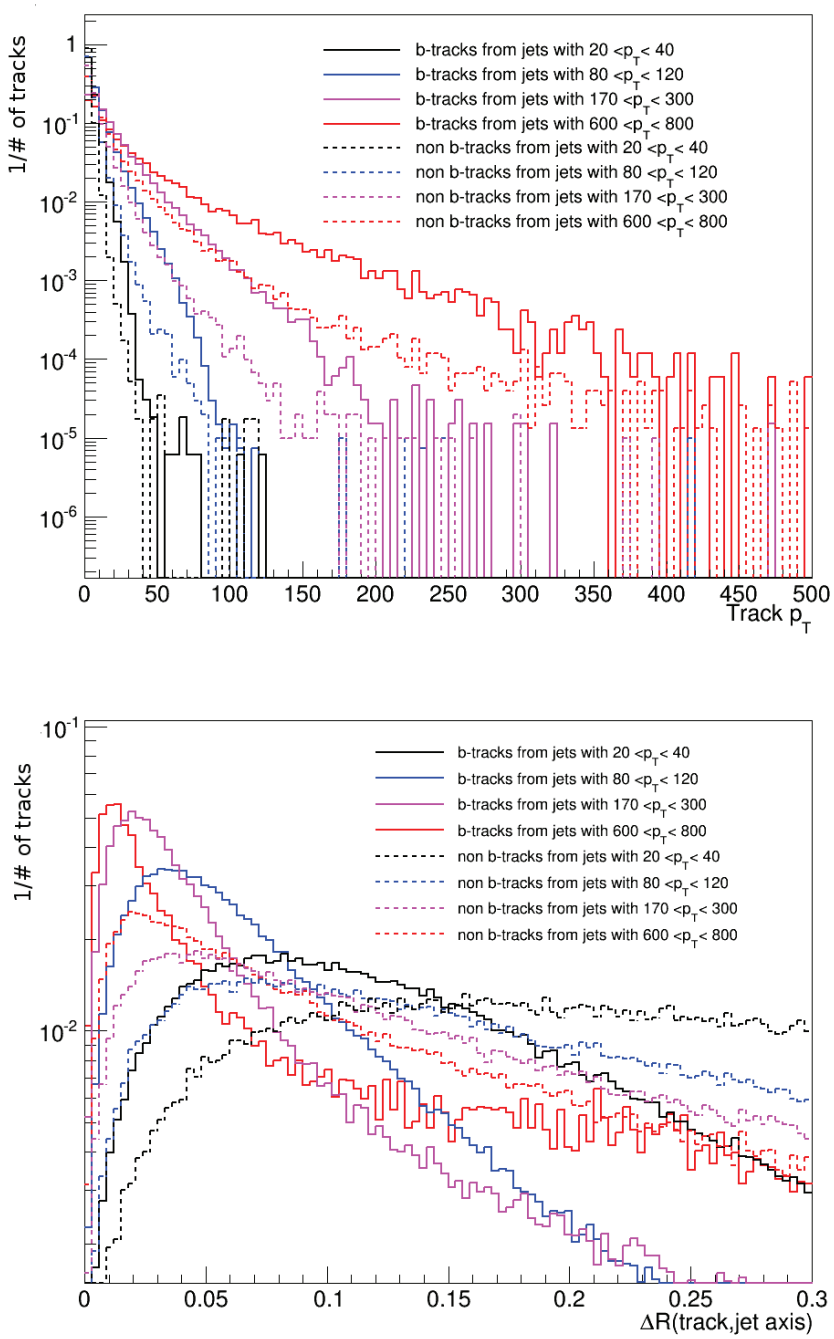

Figure 3.23: Distribution the of the track p_T (top) and of the ΔR between the track and the jet (bottom), for real B tracks (plain lines) and for non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

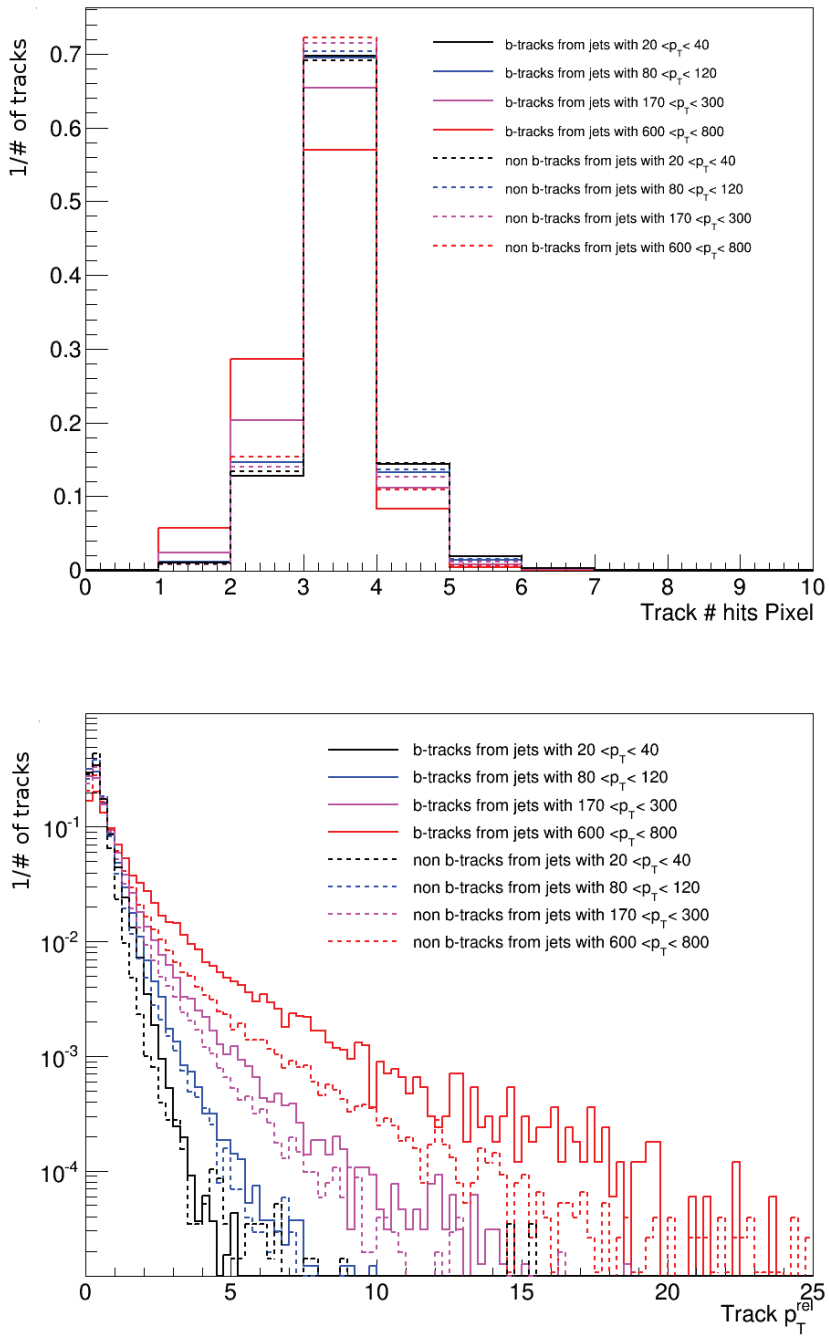

Figure 3.24: Distribution of the number of hits in the Pixel detector (top) and of the p_T^{rel} (bottom), for B tracks (plain lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

The decay length is displayed on Fig. 3.25 (top plot) and it is found again that B tracks mostly populate the high value bins. The same conclusion can be drawn from Fig. 3.25 (bottom plot), where the distance of the track to the jet axis is shown.

Finally, the distance of the track to the $(x - y)$ plane (d_{xy}, d_0 being the track 2D impact parameter) is displayed on Fig. 3.26. Again, B tracks have significantly larger values than non B tracks.

More plots about the discriminating variables can be seen in Appendix B.

This study revealed a brand new set of discriminating variables between real B tracks and other tracks within b jets. Combining them and adding this information to the JP algorithm can be, in principle, a viable solution to improve the algorithm. This will be tested in the following section.

3.4.2 Use of a Boosted Decision Tree

A Boosted Decision Tree (BDT) is one of the MVA tool available in ROOT: it combines several discriminating variables and returns a value that evaluates how much the input is signal-like compared to a background hypothesis. An extended description of BDTs can be found in Appendix D. The nominal BDT parameters set by ROOT have been applied here.

In this study, the goal is to estimate how likely a track is a real B track, and to apply an additional selection cut based on this criteria to remove non B tracks from the JP algorithm. The signal is composed of real B tracks coming from non GS b jets, while the background is defined as the set of non real B tracks.

Optimization in Jet Probability

A training is performed using the following variables:

- Track p_T;
- ΔR between the track and the jet;
- Number of hits in the pixel detector associated to the B track candidate;
- Number of hits in the SiStrip detector associated to the B track candidate;

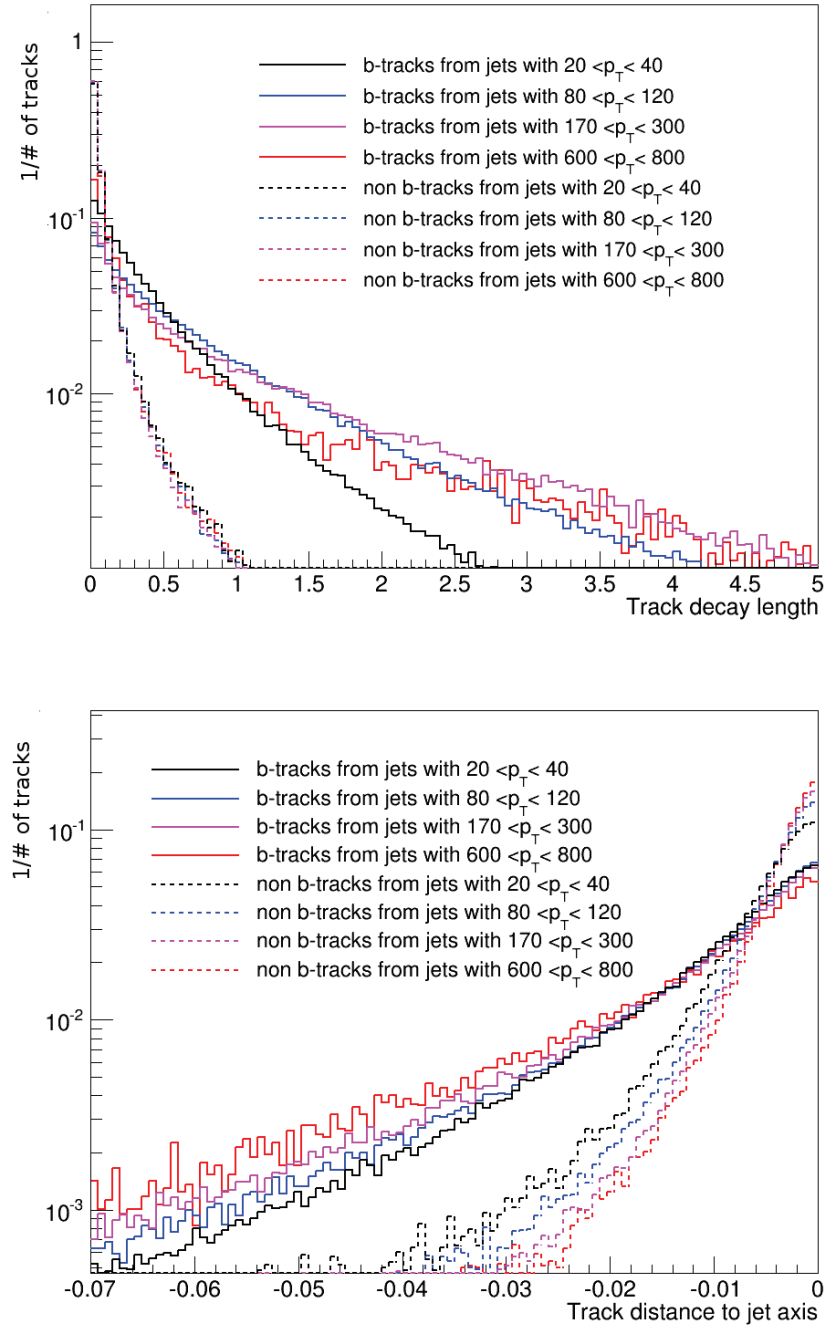

Figure 3.25: Distribution of the track DL (top) and of the distance to the jet axis (bottom), for B tracks (plain lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

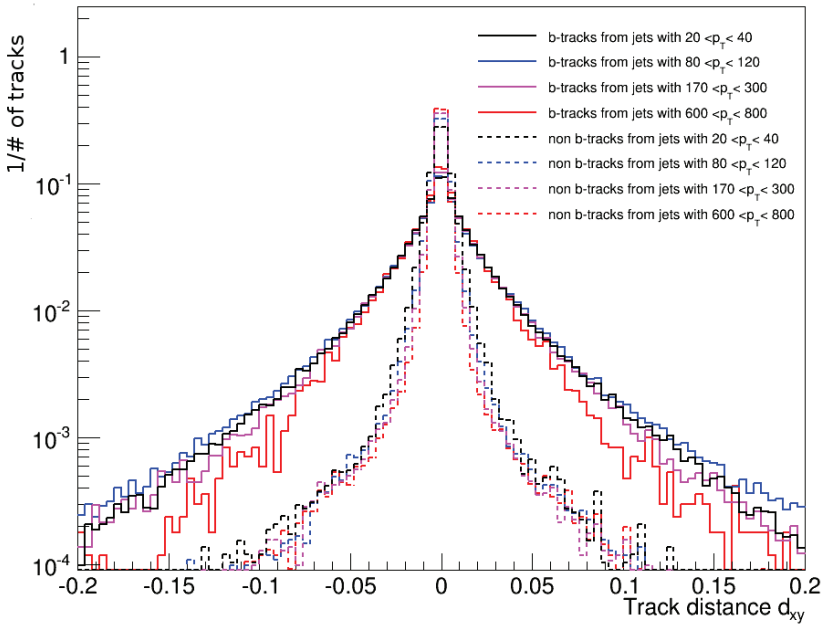

Figure 3.26: Distribution of the track distance to the $(x - y)$ plane, d_{xy}, for B tracks (plain lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

- Relative p_T between the track and the jet axis p_T^{rel};
- Normalized χ^2;
- Distance of the track to the $(x-y)$ plane, d_{xy};
- Distance of the track to the z plane, d_z;
- Track decay length;
- Track distance to jet axis;
- Track-pair invariant mass.

No significant correlation is found between any two variables, as it can be seen on Fig. 3.27).

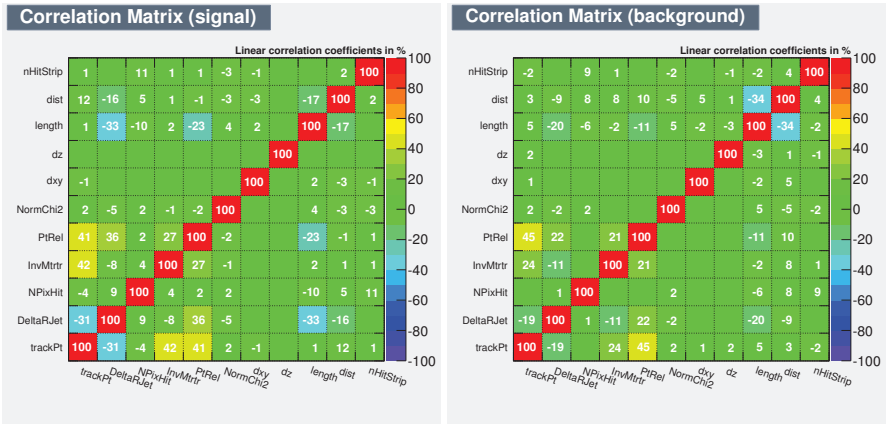

Figure 3.27: BDT correlation matrices for signal (left) and background (right).

The BDT output presents a good discriminating power between signal and background (see Fig. 3.28). The shape of the training curves are similar to the shapes of the test curves, implying a similar discrimination power between the training and the test samples (no over-training).

An optimization of the BDT output cut is performed for jets with $250 < p_T < 300$ GeV: 11 cuts are applied (from -0.5 to 0.5 with a step of 0.1) for which the b-tagging efficiency is recomputed. Fig. 3.29 shows the b-tagging efficiency as a function of the BDT cut applied. The cut at -1.0 being equivalent to no cut at all (c.f. Fig. 3.28), the b-tagging efficiency for this value can be used as the reference one.

According to Fig. 3.30, a significant gain of b-tagging efficiency can be achieved when applying a high BDT cut.

In particular, an absolute gain of 4-5% efficiency is observed for a cut at 0.3: for the Loose Working Point (WP) (corresponding to 10% of mistag rate) the gain in efficiency is 4.5%, for the Medium WP (corresponding to 1% mistag rate) it is 3.8% and finally, a 5.0% gain is observed for the Tight WP (meaning 0.1% mistag rate). Since the additional BDT cut affects the track multiplicity, the gaps in the curves, explained in Section 3.4, are more pronounced on Fig. 3.30 than on Fig. 3.13.

Another aspect to discuss is the loss of efficiency observed at high BDT cut: since this cut is directly applied on the tracks and not on the jet, when a tight BDT cut is applied, fewer tracks enter the JP algorithm, affecting the performance of the algorithm: there are not enough inputs, even though the number of jets remains the same.

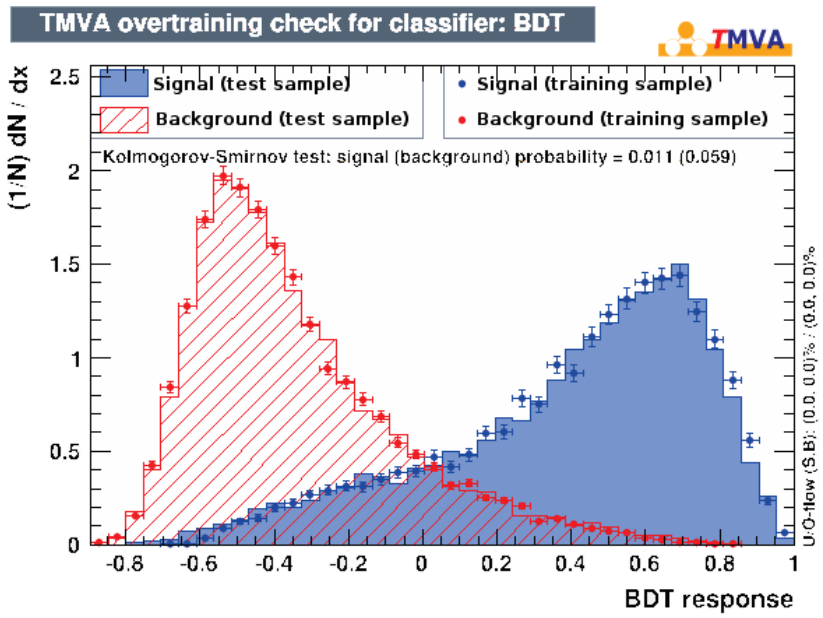

Figure 3.28: Distribution of the BDT output for signal (blue) and background (red). The full distributions are for the test sample while the dots represent the training sample.

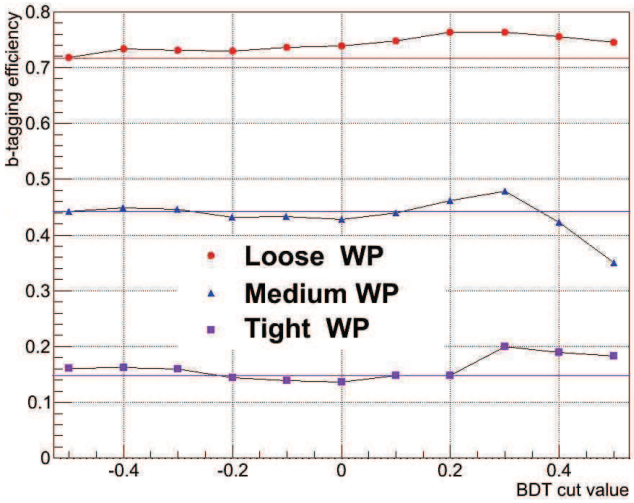

Figure 3.29: Evolution of the b-tagging efficiency for JP as a function of the BDT cut applied, for the three WP: Loose (red), Medium (blue) and Tight (purple), for jets with $250 < p_T < 300$ GeV.

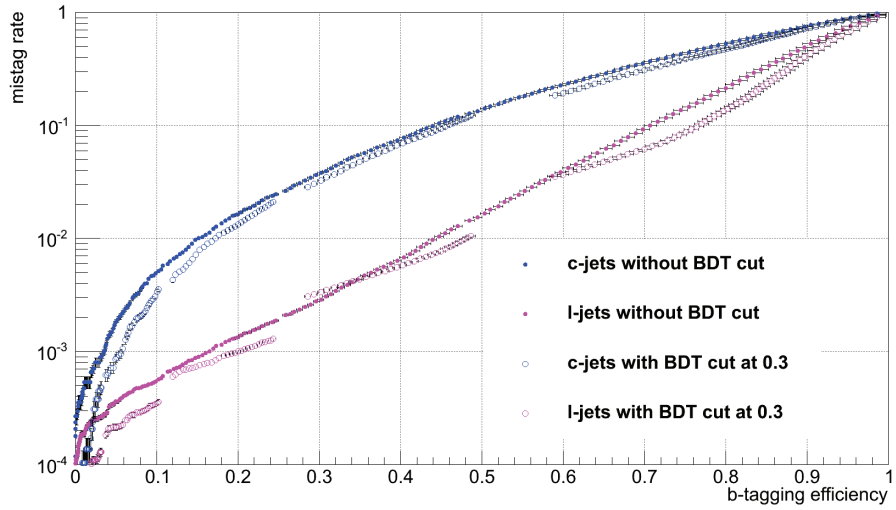

Figure 3.30: JP performance curves for c jets (blue) and light-jets (pink) with $250 < p_T < 300$ GeV, using nominal tracks (solid markers) and tracks passing the additional BDT cut of 0.3 (empty markers).

Optimization in Jet B-Probability

The Jet B-Probability (JBP) is the same algorithm as JP but in which the four tracks with the highest IP/σ give more weight in the jet probability computation. The same study is repeated for this algorithm, and a significant gain of efficiency can be reached, as it is shown on Fig. 3.31, for jets with $250 < p_T < 300$ GeV.

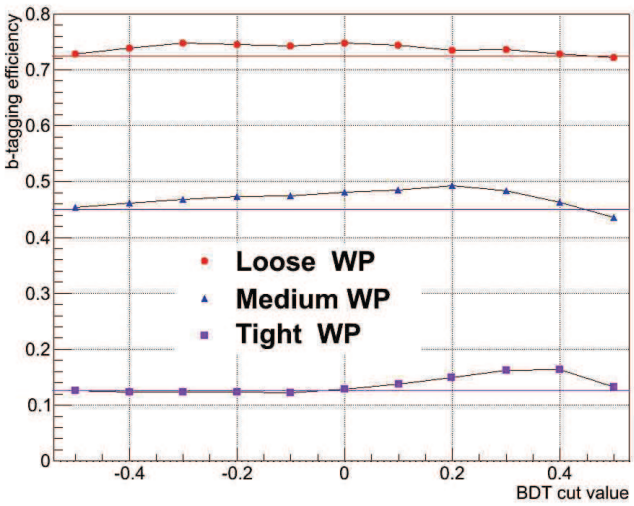

Figure 3.31: Evolution of the b-tagging efficiency for JBP as a function of the BDT cut applied, for the three WP: Loose (red), Medium (blue) and Tight (purple), for jets with $250 < p_T < 300$ GeV.

The final results have been summarized in the Tables 3.7, listing the gain of b-tagging efficiency for the two algorithms when applying different BDT cuts, for jets with $250 < p_T < 300$ GeV.

The same study, with similar results, has been performed with jets with $450 < p_T < 550$ GeV, and related plots and tables are presented in Appendix B.

Table 3.7: Absolute gain of b-tagging efficiency (in %) for JP and JBP for the different WP, applying a specific BDT cut and using jets with $250 < p_T < 300$ GeV.

	JP			JBP		
BDT cut	**Loose WP**	**Medium WP**	**Tight WP**	**Loose WP**	**Medium WP**	**Tight WP**
-0.5	0	0	1	0.06	0.4	0
-0.4	1.6	0.6	1.2	1	1	-0.3
-0.3	1.2	0.4	1	2	1.8	-0.3
-0.2	1.2	-0.1	-0.3	2	2.4	-0.3
-0.1	1.9	-1	-1.1	1.9	2.4	-0.3
0	2.2	-1.3	-1.3	2	3	0.2
0.1	3.1	-0.3	-0.2	2	3.5	1.2
0.2	4.4	1.8	-0.2	1	4.2	2.4
0.3	4.5	3.8	5	1	3.3	3.7
0.4	3.8	-2	4	0.06	1.3	3.8
0.5	2.6	-9.2	3.3	-0.3	-1.4	0.6

BDT with sorted tracks

Another test is performed by changing the JBP algorithm: instead of assigning a bigger weight to the four tracks with the highest IP/σ, more weight is given to the four tracks with the highest BDT values. The method is tested on jets with $250 < p_T < 300$ GeV (see Fig. 3.32), and a measurable gain in b-tagging efficiency is observed.

As a conclusion of this study, it appears that the best way to achieve a significant gain of b-tagging efficiency is to use a BDT and to apply a high cut on the output to reject non B tracks.

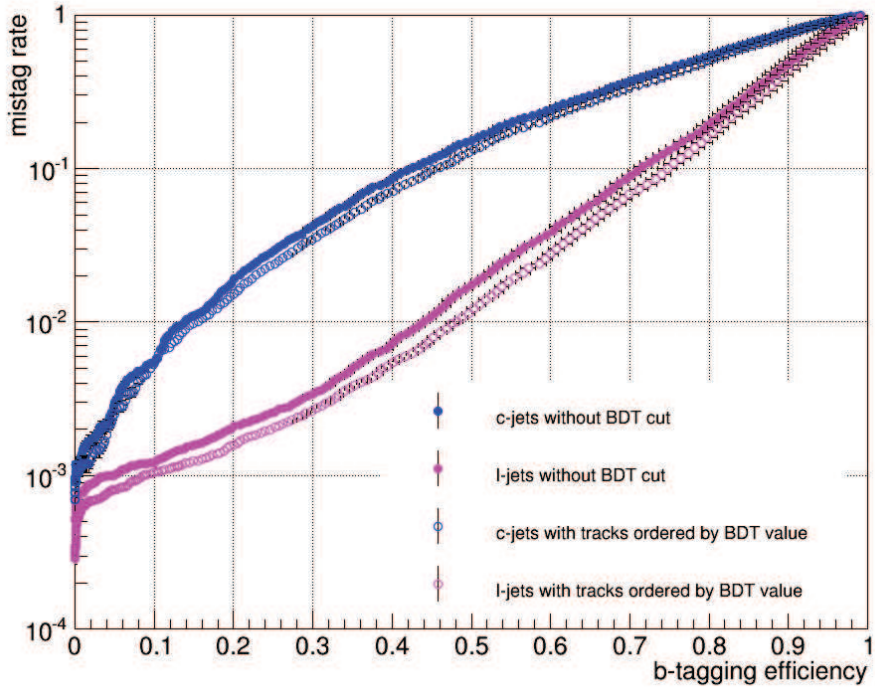

Figure 3.32: JBP performance curves for c jets (blue) and light-jets (pink), using the reference algorithm (solid markers) and the new JBP with the BDT outputs sorted (empty markers), for jets with 250$<$ p_T $<$300 GeV.

3.4.3 New categories for the calibration

As it was mentioned before, one of the new calibration procedure advantages is that categories can easily be added/removed/modified (see Section 3.1.1). It is now much quicker to estimate the effect of new categories on the calibration. The following study consists in adding new categories that are sensitive to the high p_T jets region.

As a first attempt, one can perform a new calibration with the same categories as the one presented in Section 3.1.1, but using only high p_T jets (jets with $p_T > 470$ GeV). The performance curves are compared with the reference calibration, achieved with events with a jet p_T in the range [80-120 GeV], but no visible improvement appears (see on Fig. 3.33).

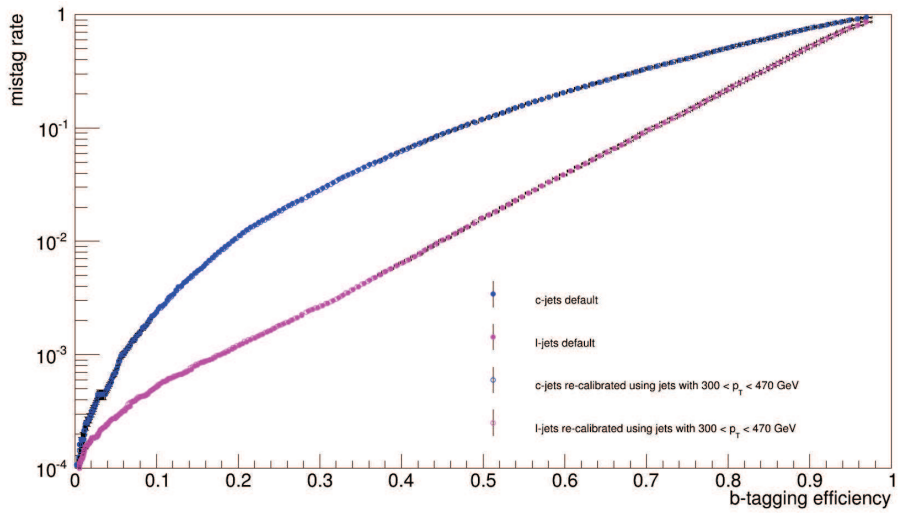

Figure 3.33: JP performance curves for c jets (blue) and light-jets (pink) with $80 < p_T < 120$ GeV, using the reference calibration (solid markers) and the new calibration, achieved with only high p_T jets (empty markers).

Various variables have been found to be sensitive to the track IP, such as the number of hits in the pixel detector, the track p_T (Fig. 3.34 top plot), and the track decay length (Fig. 3.34, bottom plot). They also depend on the jet p_T, as it was shown in Section 3.4.1. This means that the resolution functions (mentioned in Equation 3.1) are known to depend on these variables, as the p_T of the jets increase. Therefore, they could be

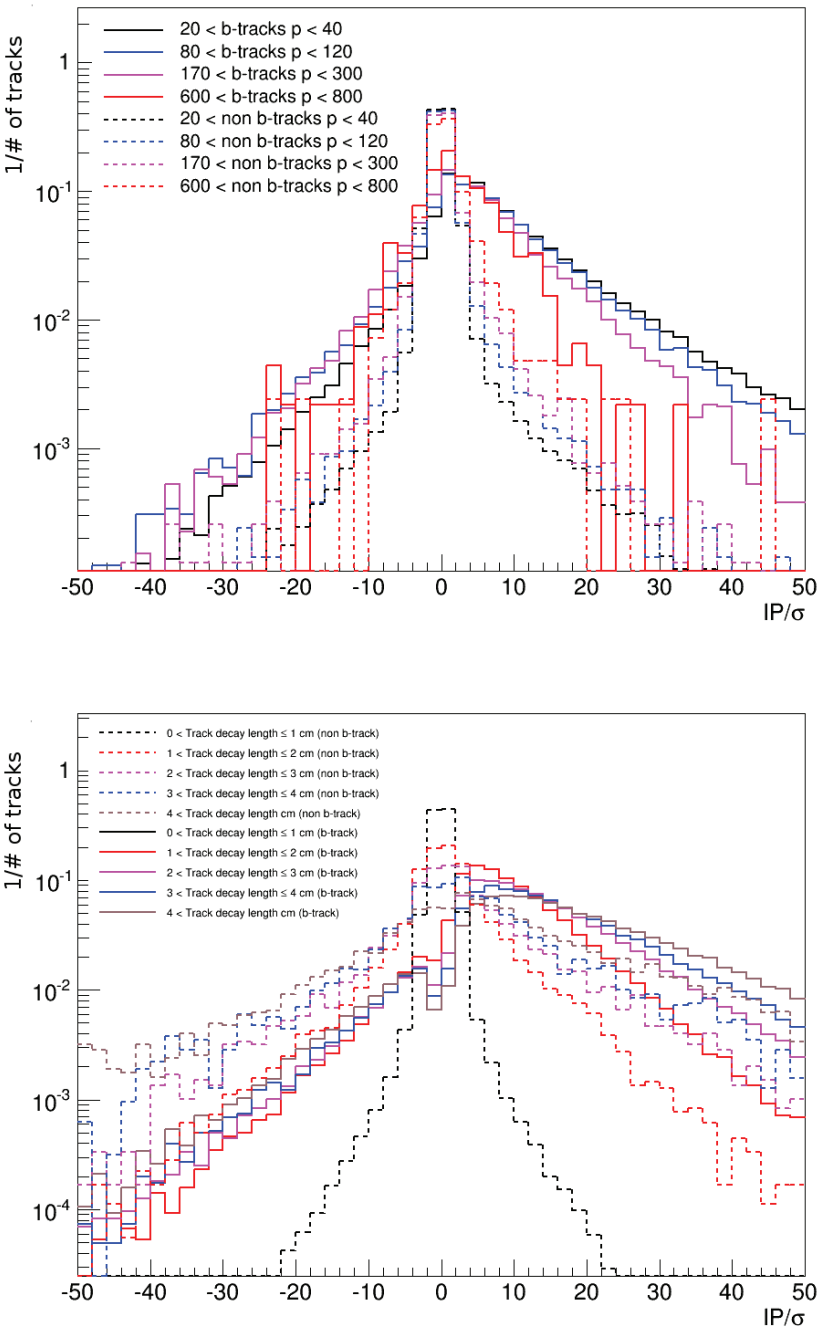

Figure 3.34: Distribution of the IP/σ for tracks with different track impulsion ranges (top) and different track decay length ranges (bottom), for real B tracks (solid lines) and non B tracks (dashed lines).

incorporated in new categories in order to make the calibration sensitive to high p_T jets.

The following calibrations using new categories are then tested:

- A new category for tracks with one hit in the pixel detector. The goal of this category is to recover from the tracking inefficiency observed at high p_T. The track selection is changed accordingly, but only a slight improvement of the performance is seen (see Fig. 3.35).
- New categories with decay length ranges. Since the IP/σ is sensitive to the decay length, it can be used as a parameter in the categories definition: new categories for different decay length have been created. New categories based on decay length are added with the following ranges: [0-1, 1-1.5, 1.5-2, 2-3, 3-5, 5-10 cm]. The track selection has been loosen accordingly. Only a very slight improvement of the performance has been noticed.
- Refinement of the track-momentum based categories. The categories have been defined at the beginning of the data taking but never re-examined since then. A refinement is done by improving the splitting of the track momentum: new categories for track p in the ranges [8-20, 20-40, 40-80, 80-150, $>$150 GeV] are added. Again, no significant improvement can be seen.

More plots about the results related to this section can be found in Appendix B. Adding categories for the calibration to tune JP for high p_T jets region gives, so far, moderate results.

3.4.4 Conclusion

The JP tagger is one the main taggers currently used in the CMS analyses, performing very well up to 200 GeV. After this limit, a loss of b-tagging efficiency is observed, ensued from a loss of B tracks inside the b jet, in association with a rise of non B track contamination. New high p_T jets based categories have been created, aiming to recover the degradation of performance observed in this region. A slight improvement of a few % is reached. Another study has been done to improve the B track purity inside the b jets using a BDT, and the first results are promising with a gain 4-7% of efficiency, depending on the WP and the jet p_T. These changes have not implemented and used in the following of this thesis, but this work paves the way for further investigations and potential improvements, such as the optimization of the list of variables used for the BDT, or the performance of the training using data events. All these results have been summarized in a CMS analysis note [76].

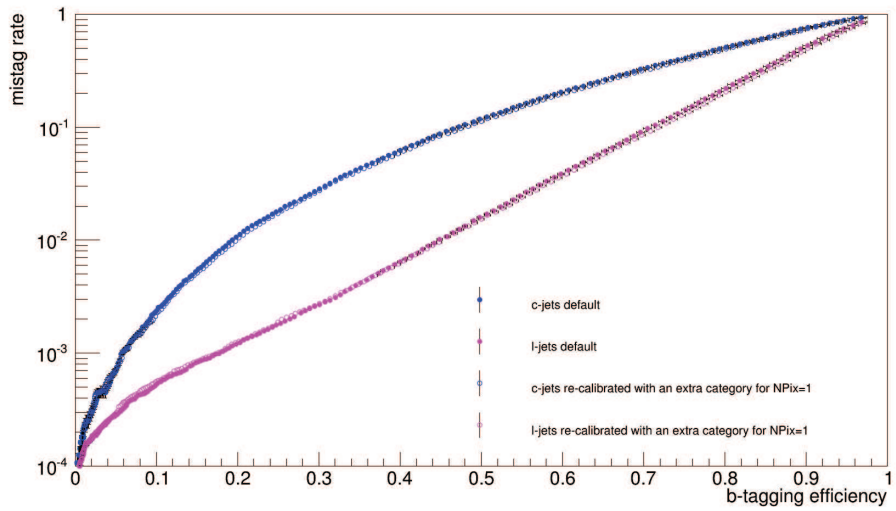

Figure 3.35: JP performance curves for c jets (blue) and light-jets (pink), using the reference calibration (solid markers) and the new calibration with a new category for tracks with NPix = 1 (empty markers).

Chapter 4

Search for the associated production of Higgs and Z bosons with the Matrix Element Method

The Higgs boson and its properties have been presented in Chapter 1. One missing part of the puzzle, to claim that the newly discovered particle is the SM Higgs boson, is the observe its coupling to the fermions, a more challenging task. Such a discovery would not only strengthen the consistency of the Brout Englert Higgs (BEH) mechanism, but would also be a confirmation of the hypothesis that this new particle is responsible for the generation of the fermion's mass. The Higgs boson decay into two b quarks is therefore of main interest.

Unfortunately, when the Higgs is produced by gluon-gluon fusion, this channel is considered as nearly impossible to exploit due to the overwhelming di-jet events from the QCD background. For this reason, a preferred Higgs production mode is the VH mode, when the Higgs is produced in association with a Z boson. This channel has a significantly lower cross section but shows a very clean signature.

The purpose of this study is to perform an independent cross check of the main CMS analysis on the dedicated subject, while comparing the performance of two algorithms of b jets identification. The search is based on the Z(ll)+H(bb) analysis (with

l = e, μ and using the CSV discriminant [64][77]), derived itself on the Z+1/2b jets cross section measurement [78]. The reprocessed data collected during 2012 by the CMS experiment at a center-of-mass energy of 8 TeV, corresponding to a luminosity $\mathcal{L}$ = 19.7 fb^{-1}, are exploited here.

4.1 Phenomenology of $llbb$ topology

Several SM processes present a final state containing two leptons, b jets and no Missing Transverse Energy (MET) (called the $llbb$ topology):

1. The signal, the ZH process, presented in Chapter 1. For this process, the cross section combined with the branching ratio leads to a small production rate (0.0249 pb at 8 TeV); considering the luminosity $\mathcal{L}$ = 19.7 fb^{-1} recorded at 8 TeV, around 500 ZH events are expected in this data sample. However, the fiducial region of the detector is restricted to $|\eta| < 2.4$, and in addition, the detector reconstruction and identification efficiencies are less than 100%. Moreover, as it was discussed in the previous chapter, the identification of b jets also leads to a loss of signal events. This leads to a selection of only $\sim 4\%$ of the produced signal events. As conseqa uence, it will be difficult to see a 3σ excess in this channel;

2. The Drell-Yann (DY)+jets events, corresponding to the production of a Z boson (decaying in two leptons) in association with jets, is the most significant background (its production rate including the branching ratio $Z \rightarrow l^+l^-$ is 3531.9 pb). These events are categorized, based on the number of b partons matched with b tagged jets:

   - "Zbb": the two b-tagged jets are associated to real b partons;
   - "Zbx": one of the two b jets matches a real b parton while the other matches a c or a light parton (mis-tagged), and is therefore referenced as 'x';
   - "Zxx": the two selected b jets are mis-tagged jets.

   The main irreducible background is then the production of a Z boson in association with two b jets, called the Zbb process. Zbb events can be produced via $q\bar{q}$ annihilation (10%) or via gluon-gluon fusion (90%). The related Feynman diagrams can be seen on Fig.4.1;

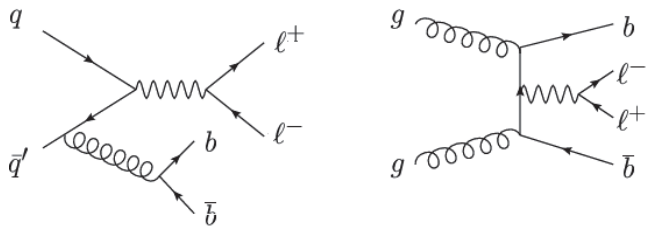

Figure 4.1: Production of Zbb events at the LHC: by $q\bar{q}$ annihilation (left) or by gluon-gluon fusion (right).

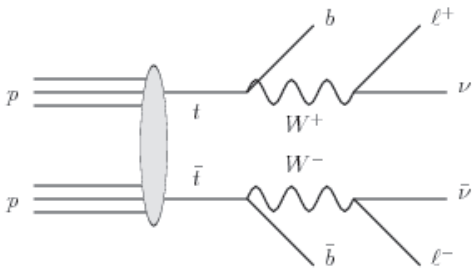

Figure 4.2: Production of $t\bar{t}$ events, followed by a leptonic decay of both W bosons.

3. The second most important background is the top quark pair production, followed by a leptonic decay of both W bosons. The associated production rate including the branching ratio $W \to l\nu$ is 26.6 pb. However, the kinematics of the corresponding final state is significantly different from the signal ones. Indeed, for $t\bar{t}$ events the two leptons are not issued from the same particle, and the presence of neutrino implies the production of real $\vec{E}_t^{miss}$. The related Feynman diagram is shown on Fig. 4.2).

4. The final process to take into account is the di-boson production ZZ where two Z bosons fake the signal signature: one decays into two leptons, the second into two b jets, as it can be seen on Fig. 4.3. This process has a very small production rate (0.168 pb, including the branching ratios $Z \to b\bar{b}$ and $Z \to l^+l^-$) but since the Z and the Higgs masses are very close, this background is difficult to reduce.

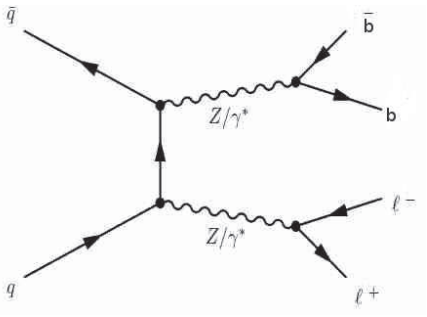

Figure 4.3: Production of ZZ events, followed by a leptonic decay of one Z boson while the second Z decays into $b\bar{b}$ [33].

Several processes are not taken into account in this analysis, since their contributions are found to be negligible:

- The fully-hadronic $t\bar{t}$ decay process;
- The semi-leptonic $t\bar{t}$ final state;
- Single top events: s-channel, t-channel and tW events;
- Di-boson production: WW and WZ processes.

For the fully-hadronic $t\bar{t}$, since a small contribution is expected, no events have been processed; the yields of the unused contributions, in the signal region of this analysis, are displayed in Table 4.1.

Table 4.1: Expected yields after selection in the signal region (displayed in Table 4.3 and Table 4.4), for the processes not taking into account in the background fit and background rejection procedures. Statistical errors are shown. The algorithm of b-tagging used here is JP.

	tW/$\bar{t}$W	$t\bar{t}$ semi lept.	WW	WZ
Event yields	7.6 ± 2.8	2.4 ± 1.5	2.8 ± 1.7	0.5 ± 0.7
Total contribution	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$

4.1.1 Samples used

The full 8 TeV dataset recorded during 2012, by the un-prescaled double lepton triggers[1] (see Table 4.2, "Data" section), is used in this analysis. It corresponds to a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}; the data are compared with generated and simulated samples listed on Table 4.2. The assumed cross sections for the different background processes have been chosen following the CMS recommendations. For the signal, the cross section is computed to take into account until NNLO effects [20], and the decay of the Higgs into a pair of b quarks.

Data and MC samples are processed via official Physics Analysis Tools (PAT), using the CMSSW software version CMSSW_5_3_14 patch1.

The Zbb process is extracted from an inclusive DY +jets sample; in addition to this inclusive sample, four additional DY+jets samples are used, enriching the content of events with a high boost of the Z boson. Each of the samples presents a cut at generator level on the $p_T(Z)$, and based on this cut the following 5 bins are defined:

- [0 - 50] GeV: for which only the DY inclusive sample contributes;
- [50 - 70] GeV: in addition to the DY inclusive sample, a DY sample with a p_T(Z) generated between 50 and 70 GeV contributes;
- [70 - 100] GeV: a DY sample generated with a p_T(Z) between 70 and 100 GeV is used in addition to the DY inclusive sample;
- [100 - 180] GeV: the DY sample with generated p_T(Z)$>$ 100 GeV for this bin is used along with the DY inclusive sample;
- [180 - ∞[GeV: both DY samples with generated p_T(Z) $>$ 100 GeV and p_T(Z) $>$ 180 GeV are used in addition to the DY inclusive sample.

On top of that, DY samples produced with a high Hadronic Activity (H_T) (defined as the scalar sum the p_T of the jets) are added to the initial sample:

- [0 - 200] GeV: a DY sample with a H_T between 0 and 200 GeV;
- [200 - 400] GeV: a DY sample with a H_T between 200 and 400 GeV.

[1] HLT_Mu17_Mu8 or HLT_Mu17_TkMu8 for muons and HLT_Ele17_Ele8 for electrons

The cross sections of the different DY samples are taken from the production and reprocessing management tool for CMS to compute the corresponding effective luminosities, which are used in order to reweight the events in a merging step so that the effective luminosity of the DY combined sample matches the luminosity of the inclusive DY sample. As a consequence, 10 weights are computed and assigned according to the p_T and the HT of the DY event. This merging procedure is explained in Appendix C.

Table 4.2: Data and MC samples used in this analysis, with the corresponding cross section for simulated events. All samples are taken from AOD (Data) and AODSIM (MC) format files.

Dataset	Data	
Electrons Run A	/DoubleElectron/Run2012A-22Jan2013/	
Electrons Run B	/DoubleElectron/Run2012B-22Jan2013/	
Electrons Run C	/DoubleElectron/Run2012C-22Jan2013/	
Electrons Run D	/DoubleElectron/Run2012D-22Jan2013/	
Muons Run A	/DoubleMu/Run2012A-22Jan2013/	19.7 fb^{-1}
Muons Run B	/DoubleMuParked/Run2012B-22Jan2013/	
Muons Run C	/DoubleMuParked/Run2012C-22Jan2013/	
Muons Run D	/DoubleMuParked/Run2012D-22Jan2013/	
Dataset	MC	$\sigma(Pb)$
$t\bar{t}$ fully-leptonic	/TTJets_FullLeptMGDecays_8TeV-madgraph-tauola/Summer12_DR53X-PU_S10_START53_V7C-v2/	27.3 [79]
ZZ	/ZZ_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53_V7A-v1/	8.2 [80]
ZH_{125}	/ZH_ZToLL_HToBB_M-125_8TeV-powheg-herwigpp/Summer12_DR53X-PU_S10_START53_V7A-v1/	0.0249 [81]
DY inclusive	/DYJetsToLL M-50 TuneZ2Star 8TeV-madgraph-tarball/*_ START53 V7A-v1/	2950 (LO)
DY $p_T(Z) \in$ [50-70] GeV	/DYJetsToLL_PtZ-50To70_TuneZ2star_8TeV-madgraph-tarball/Summer12_DR53X-PU_S10_START53_V7A-v1/	93.8
DY $p_T(Z) \in$ [70-100] GeV	/DYJetsToLL_PtZ-70To100_TuneZ2star_8TeV-madgraph-tarball/Summer12_DR53X-PU_S10_START53_V7A-v2/	50.31
DY $p_T(Z) >$ 100 GeV	/DYJetsToLL_PtZ-100_TuneZ2star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v2/	34.1
DY $p_T(Z) >$ 180 GeV	/DYJetsToLL_PtZ-180_TuneZ2star_8TeV-madgraph-tarball/Summer12_DR53X-PU_S10_START53_V7C-v1/	4.5

4.2 Event selection

The selection applied aims to favor signal events, and the following criteria, summarized in Table 4.3, must be fulfilled.

General requirements

- As previously stated, the triggers used in this analysis are the double muons and electrons triggers: the trigger is fired when there are at least two muons or two electrons in the events, with $p_T > 8$ GeV, including one with $p_T > 17$ GeV;
- Jets and the leptons fulfilling the trigger requirement must come from the same primary vertex PV, considered as the vertex of the hard interaction, with at least four associated tracks. To reduce the risk of selecting a pile-up (PU) vertex, the longitudinal and radial distances of the vertex from the center of the detector must be smaller than 24 cm and 2 cm, respectively. For events with more than one selected PV, the PV containing N tracks with the largest $\sum_i^N p_T^i$ is chosen;
- Leptons and jets had to be reconstructed using the CMS particle flow algorithm, described in Chapter 2.

Leptons requirements

- The p_T and η of the leptons are set to fit the trigger and detector acceptance: only muons (electrons) with $p_T > 20$ GeV and $|\eta| < 2.4$ (2.5) are kept. In addition, electrons lying between 1.442 $|\eta| < 1.566$ are discarded;
- An isolation criterion is applied to remove leptons coming from heavy flavor hadron decays: the relative isolation of a lepton within a cone of size $\Delta R = \sqrt{(\eta)^2 + (\phi)^2}$ = 0.4 (0.3) for muons (electrons) is set to be less than 0.2 (0.15);
- The di-lepton pair must be composed of the two same-flavor and opposite-charged leptons with the highest p_T. Their invariant mass should be compatible with the one of the Z boson: $76 < M_{ll} < 106$ GeV for a "tight" selection, and $60 < M_{ll} < 120$ GeV for a looser selection. This criterion depends on the studied region (see Table 4.4).

Jets requirements

- Jets must to fulfill the following kinematic criteria: $|\eta| < 2.4$ and $p_T > 20$ GeV. An extra requirement is made on the p_T of the two leading jets: $p_T(j1), p_T(j2) > 30$ GeV. This cut is driven by the b-tagging SF, provided only for jets with $p_T > 30$ GeV;
- The leading jets must be identified as b jets. Two algorithms are tested: Jet Probability (JP) and Combined Secondary Vertex (CSV). The two b jets have to pass a given discriminator threshold, set at 0.545 for JP and 0.675 for CSV, in order to be tagged by the "Medium" Working Point (WP). This WP corresponds to the mis-identification rate of 1% (based on QCD events). A more detailed discussion on the subject is held in the next paragraph;
- A categorization on the number of jets in the event is done. Indeed, for events in the category where exactly two jets are required (called the "2-jets" category), a better background rejection is observed, as well as a better resolution on the di-jet mass M_{bb}. The other category with at least one extra jet in the event (called the "3-jets" category) shows a degradation of the M_{bb} resolution. Such behavior is expected since in the 3-jets category, some extra jets are emitted from the b jets (FSR jets), compromising the M_{bb} reconstruction;
- Since the mass of the Higgs boson measured by CMS and ATLAS is 125 GeV, a cut on the M_{bb} is applied such that it corresponds to the Higgs mass: for the 2-jets (3-jets) category 90 (70) $< M_{bb} <$ 150. This cut is set to keep 90% of the signal is both categories. It is thus loosen in the 3-jets category, in order to cover the 20% of resolution degradation expected in M_{bb}.

Missing transverse energy requirement

The reconstruction of the MET ($\vec{E}_t^{miss}$) is based on PF reconstruction (explained in Chapter 2). The MET significance is used in the selection. As it was explained in Chapter 2, this variable estimates the compatibility of the reconstructed missing transverse energy with zero and leads to smaller systematic uncertainties. It is used to suppress background originating from $t\bar{t}$, by keeping only events with MET significance < 10.

b-tagging

As previously stated, two algorithms dedicated to the identification of b jets are used: JP and CSV, presented in Chapter 3. A comparison of the performance for the two

algorithms can be seen on Fig. 4.4, for QCD events with a p_T in the range [30-120] GeV. For the Medium WP, the b-tagging efficiency is around 58% while it ∼64% for CSV. As a consequence, the probability to tag two b jets using the Medium WP is around 34% for JP while it is roughly 41% when using CSV, leading to an expected difference of 22% in the raw events yields. Besides, the b-tagging scale factors (SF) are not the same for these two taggers, as it can been seen on Fig. 4.5: the CSV SF are about 9% higher than the JP SF, inducing a total difference of around 30% in the final yields for processes containing two real b jets.

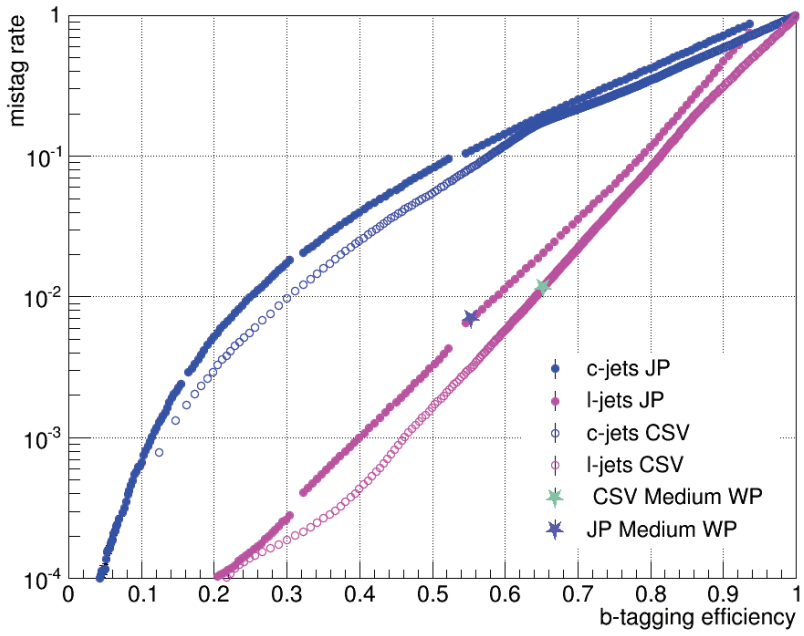

Figure 4.4: Comparison of the performance for JP (plain markers) and CSV (empty markers), for QCD events with p_T in the range [30-120] GeV. The pink dots represent the probability to tag a light jet as a b jet, while the blue dots represent the probability to tag a c jet as a b jet.

Several regions are defined for the next steps of the analysis: the Full Region (FR) is a region where the M_{ll} window is set at "loose" and no M_{bb} cut is applied. This region is the first region used for a data/MC yields comparison. The background fit is

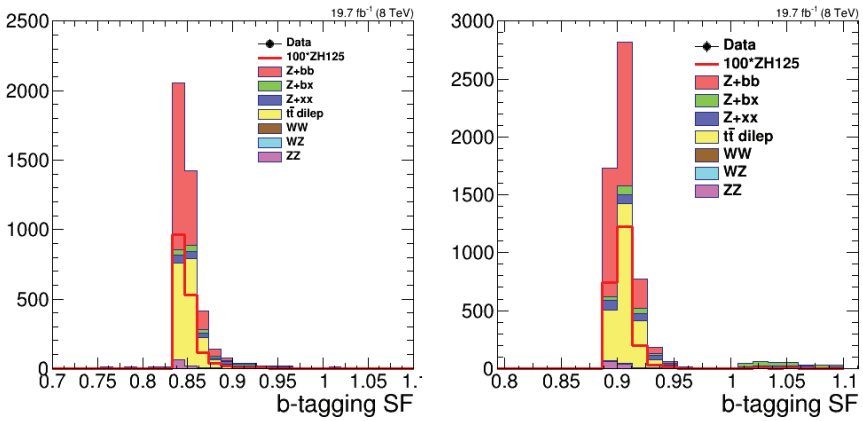

Figure 4.5: Comparison of the b-tagging SF applied for the JP (left) and CSV (right) tagged events, for the different MC contributions. These SF have been computed for the case where two "Medium" tagged b jets have been requested.

performed in the Control Region (CR): here, the cut on M_{bb} is reversed in order to cut most of the signal events. Finally, the Signal Region (SR) is defined in order to select most of the signal and the least background events.
These regions along with their specific cuts are summarized in Table 4.4.

4.2.1 Data-simulation efficiency scale factors

It is fundamental to correct for the efficiencies relative to the selection criteria and triggers applied in the analysis and consequently rescale the MC in order to match data distributions. This step allows to rely on simulation in the computation of the final results. Simulated samples have therefore been rescaled according to following extracted correction factors (SF):

- The number of PU events is an important aspect to take into account since this contamination is unavoidable; a correction is applied as a function of the expected number of PU events that presumably took place while data taking. Depending on the data taking period, the PU scenario can be very different and not match the one expected during the event generation. A reweighting of each

Table 4.3: Selection criteria applied in this analysis.

Trigger	DoubleMuon/DoubleElectron
Leptons	$p_T > 20$ GeV $\|\eta\| < 2.4$ for muons (2.5 electrons) Veto 1.442 $\|\eta\| < 1.566$ for electrons Isolation criteria for muons (electrons) using ΔR=0.4 (0.3): < 0.2 (0.15)
Jets	Leading jet: $p_T > 30$ GeV Sub-leading jet: $p_T > 30$ GeV Extra jet: $p_T > 20$ GeV $\|\eta\| < 2.4$ b-tagging: two Medium tagged b jets
Missing transverse energy	$\vec{E}_t^{miss}$ Sig. < 10

Table 4.4: Definition of the regions of interest for the search, with the corresponding cuts in leptons/jets invariant mass.

	Full Region	Control Region	Signal Region
M_{ll}	$60 < M_{ll} < 120$ GeV	$60 < M_{ll} < 120$ GeV	$76 < M_{ll} < 106$ GeV
M_{bb}	no cut	2-jets cat.: $M_{bb} < 90$ GeV or $M_{bb} > 150$ GeV 3-jets cat.: $M_{bb} < 70$ GeV or $M_{bb} > 150$ GeV	2-jets cat.: $90 < M_{bb} < 150$ GeV 3-jets cat.: $70 < M_{bb} < 150$ GeV

simulated event is directly applied by requiring that the simulated PU distribution matches exactly the one observed in data. The data distribution is obtained using the proton-proton inelastic scattering cross section in association with the instantaneous luminosity per bunch crossing for each run of data taking. As it can be seen on Fig. 4.6 (left plot), a SF of 50% can be applied for some events;

- As previously stated when discussing the lepton reconstruction and identification in Chapter 2, these procedures are not 100% efficient. Besides, trigger and isolation efficiencies should also be considered. The efficiency to select a lepton according to the selection shown in Table 4.3 has been estimated for simulated events and 2012 data, using the "Tag and Probe" method presented in Chapter 2. This method returns a per-event SF, reflecting the probability that a selected lepton fired the trigger. On Fig. 4.6 (right plot), the distribution of this SF is shown: a reweighting of at most 15% is applied;

- The b-tagging group provides per-jet SF (presented in Chapter 3), that are combined to give a per-event SF, depending on number of b, c and l jets in the event, as well as the requirement on the number of tags in the event and the mis-identification probabilities. Methods to do the perform the combination are

also provided by the CMS b-tagging working group [82]. The distributions of these SF have been shown on Fig. 4.5.

All these correction factors are applied for the following plots and event yields shown in this chapter.

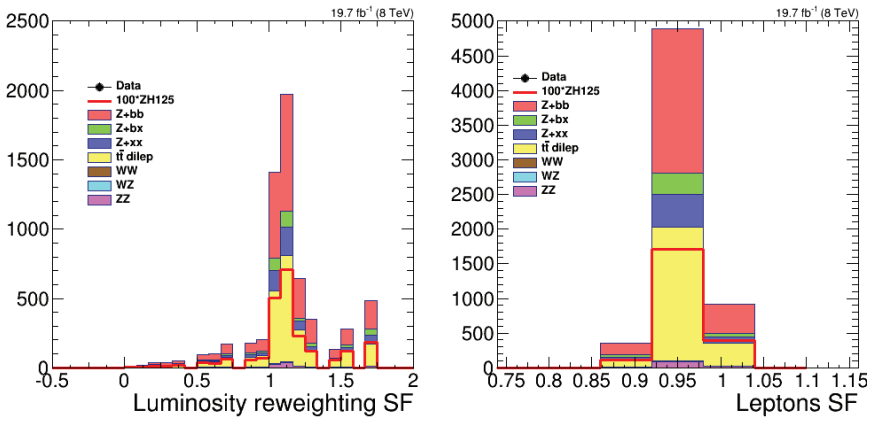

Figure 4.6: Distribution of the SF applied to take into account the PU events (left) and to correct for the lepton reconstruction and identification efficiencies (right).

4.2.2 Yields and data/MC comparison

The selection presented in Section 4.2 is applied to the samples listed in Table 4.2, with the corresponding cross section normalization. The resulting number of events for data and MC are shown in Table 4.5 and Table 4.6 for the FR, using the JP and CSV tagger respectively.

Table 4.5: Data yields for the FR (displayed in Table 4.3 and Table 4.4), compared with the expectation from the different main MC processes, normalized to their theoretical cross section. The data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}. The tagger used is JP.

JP	Zbb	Zbx	Zxx	$t\bar{t}$	ZZ	Tot. MC	Data	Data/MC
Z($\mu^+\mu^-$)	1135.1 $\pm$ 33.7	102.4 $\pm$ 10.1	170.6 $\pm$ 13.1	1030.9 $\pm$ 32.1	47.6 $\pm$ 6.9	2486.6 $\pm$ 49.9	3001	1.21
Z(e^+e^-)	821.4 $\pm$ 28.7	100.6 $\pm$ 10.0	121.2 $\pm$ 11.0	773.6 $\pm$ 27.8	36.2 $\pm$ 6.0	1853.0 $\pm$ 43.0	2082	1.12
Total	1956.5 $\pm$ 44.2	203.0 $\pm$ 14.2	211.8 $\pm$ 14.6	1804.5 $\pm$ 42.5	83.8 $\pm$ 9.2	4339.6 $\pm$ 65.9	5083	1.17

Table 4.6: Data yields for the FR displayed in Table 4.3 and Table 4.4, compared with the expectation from the different main MC processes, normalized to their theoretical cross section. The data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}. The tagger used is CSV.

CSV	Zbb	Zbx	Zxx	$t\bar{t}$	ZZ	Tot. MC	Data	Data/MC
Z($\mu^+\mu^-$)	1540.5 $\pm$ 39.2	220.0 $\pm$ 14.8	373.1 $\pm$ 19.3	1340.4 $\pm$ 36.6	67.8 $\pm$ 8.2	3541.8 $\pm$ 59.5	4214	1.19
Z(e^+e^-)	1133.0 $\pm$ 33.7	163.2 $\pm$ 12.8	254.3 $\pm$ 15.9	1005.5 $\pm$ 31.7	49.7 $\pm$ 7.0	2605.7 $\pm$51.0$\pm$	2880	1.11
Total	2673.5 $\pm$ 51.7	383.2 $\pm$ 19.6	627.4 $\pm$ 25.0	2345.9 $\pm$ 48.4	117.5 $\pm$ 10.8	6147.5$\pm$ 78.4$\pm$	7094	1.15

A global 35% difference in the event yields is measured for events containing two real b jets (Zbb, ZZ and $t\bar{t}$ events) between the two taggers. However, one can also see that JP better rejects events with at least one mis-tagged jet (Zxx events are for instance rejected three times more when using JP instead of CSV).

Most importantly, an overall 16% of data/MC discrepancy is observed, with both taggers. The source of this disagreement is not known, although it seems to mainly come from a deficit of events in the Z peak, as it was shown on Fig. 4.7, left plot. The discrepancy is also located at low MET (Fig. 4.7, right plot), and for events with jets having a low b-tagging discriminator value: on the plots of Fig. 4.8, it is even more striking in the JP case. It also come from the b jets at low p_T, as it is visible on Fig. 4.9, left plot. However,, the discrepancy does not seem to appear in a specific η region (see Fig. 4.9, right plot). All these plots tend to suggest a lack of DY events, especially the contributions with mis-tagged jets. More plots in the Full region are available in Appendix C. This MC/data disagreement is also seen in other CMS analyses [83].

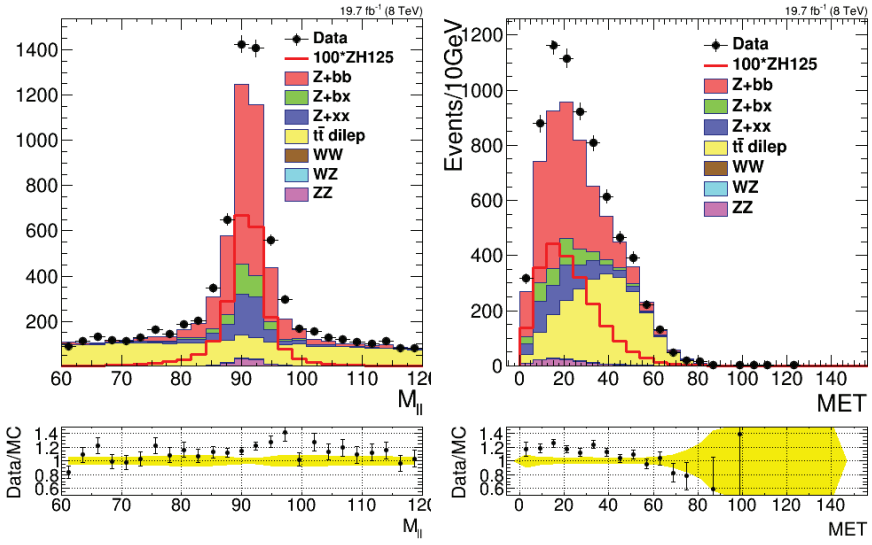

Figure 4.7: Distributions of the di-lepton invariant mass (left), and of the MET (right), in the FR, for the CSV selection.

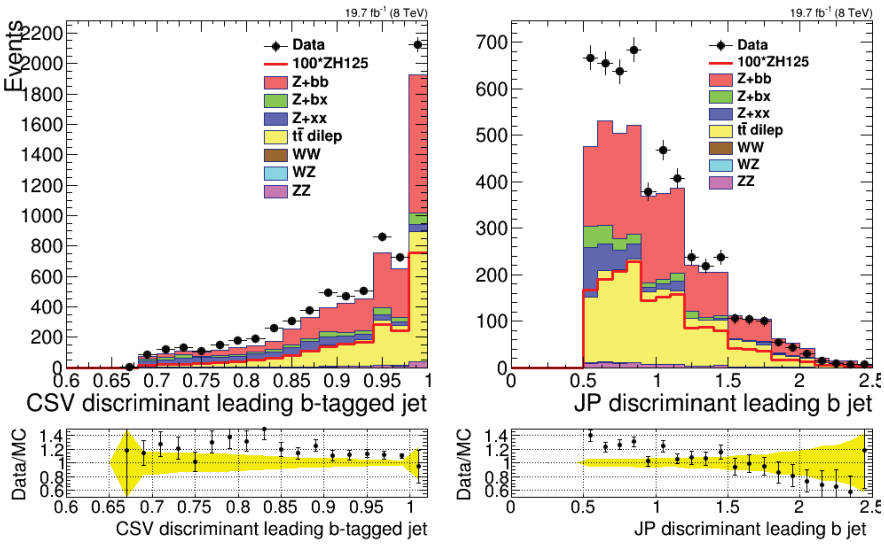

Figure 4.8: Distribution of the leading b jet CSV (JP) discriminator value on the left (right) plot, in the FR.

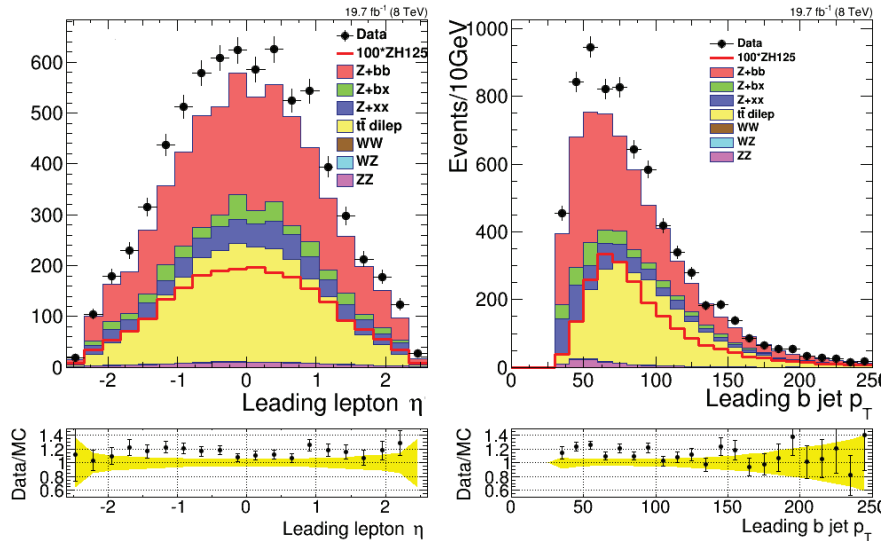

Figure 4.9: Distributions of the leading lepton η (left) and on the leading b jet p_T (right) in the FR, for the CSV selection.

Since such a discrepancy is not expected to come from new physics phenomena, but from a bad modeling of the MC or the corrections applied, a background fit is performed to improve the data/MC agreement in the control region.

4.3 Background fit

As previously stated, the main background contributions are coming from the $t\bar{t}$ dileptonic process and DY events, categorized based on the number of real b jets selected. Thus, several SF can be defined:

- "SF_{Zbb}", used to describe the Zbb contribution, in the 2-jets category only;
- "SF_{Zbx}", used to describe the Zbx events, in the 2-jets category but also in the 3-jets category. Besides, it also describes the Zbb events in the 3-jets category. Indeed, these events can be by illustrated the same Feynman diagram where the Zbb event has been produced by $q\bar{q}$ annihilation, and the additional jet is a FSR/ISR jet;
- "SF_{Zxx}", used to describe the Zxx contributions in both 2 and 3-jets categories;
- "$SF_{t\bar{t}}$", used to describe the $t\bar{t}$ contribution in both 2 and 3-jets categories.

These contributions are renormalized to the data, in the CR previously defined, using a single two-dimensional fit to the following variables distributions:

1. The product of the JP/CSV discriminator values of both b-tagged jets. This product is sensitive to the non-b jets contamination, and therefore helps to discriminate Zbb events from Zbx and Zxx events. The plots representing the product of the two b jets JP (CSV) discriminator values are displayed on Fig. 4.10 (Fig. 4.11), for the 2-jets category. The plots for the 3-jets category can be seen in Appendix C. As the data/MC discrepancy is located at low discriminator value, it can be expected that the Zxx and Zbx SF provided by the fit will be higher than the Zbb SF;
2. (1): For the JP selection, the second variable used for the fit is the output of a Neural Network (NN), discriminating the $t\bar{t}$ and the DY processes, based on the MEM weights (whose definition was given in Chapter 1). The description of NNs can be found in Appendix D.
   A single training for both jet categories is performed for simplicity reason, and the training plots can be seen in Appendix C. Plots of the NN output can be seen on Fig. 4.12 for the 2-jets category, before the background fit. A good

discrimination is observed and it can be predicted that the SF associated to the $t\bar{t}$ process will be close to one.

2. (2): For the CSV selection, the second variable used in the 2D-fit is the di-lepton invariant mass. This last variable allows to distinguish the $t\bar{t}$ from the DY events and does not require a training.

The ZZ cross section is measured in CMS with using an event selection compatible with this analysis. Therefore, ZZ events are directly renormalized to this value [80].

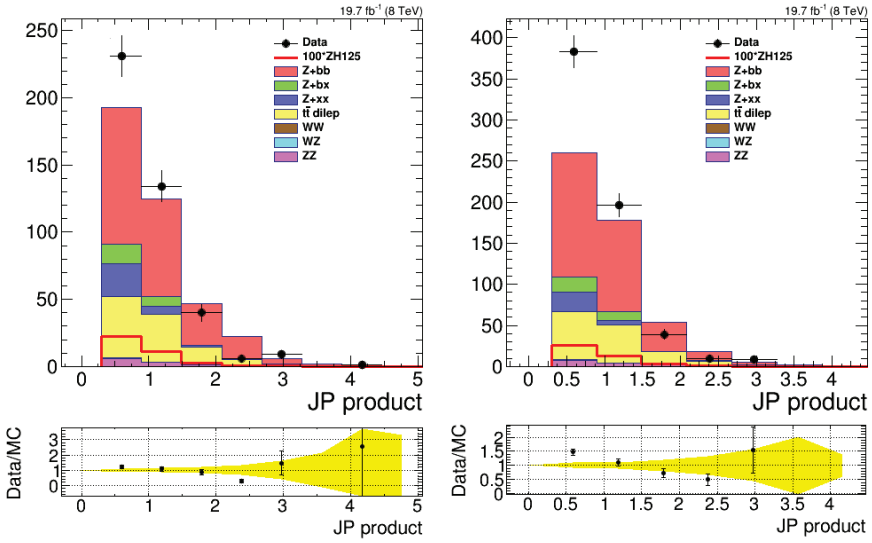

Figure 4.10: Product of the two leading b jets JP discriminator values, used to discriminate b jets from light jets contributions, in the CR. The left plot is in the di-electron channel and the right plot in the di-muon channel, for the 2-jets category.

Finally, the fit result can be seen on Fig. 4.13 for the JP selection. The values of the SF are displayed in Table 4.7: for $t\bar{t}$, it is compatible with one, as expected. For the different DY contributions, the SF are higher, especially when one or two b jets have been mis-tagged. Given the characteristics of the observed discrepancy (previously discussed), such results are meaningful.

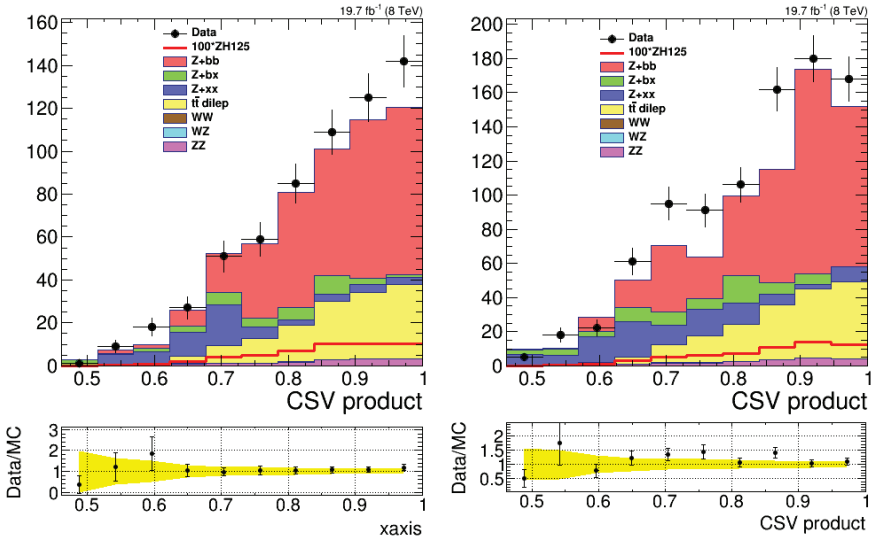

Figure 4.11: Product of the two leading b jets CSV discriminator values, used to discriminate b jets from light jets contributions, in the CR. The left plot is in the di-electron channel and the right plot in the di-muon channel, for the 2-jets category.

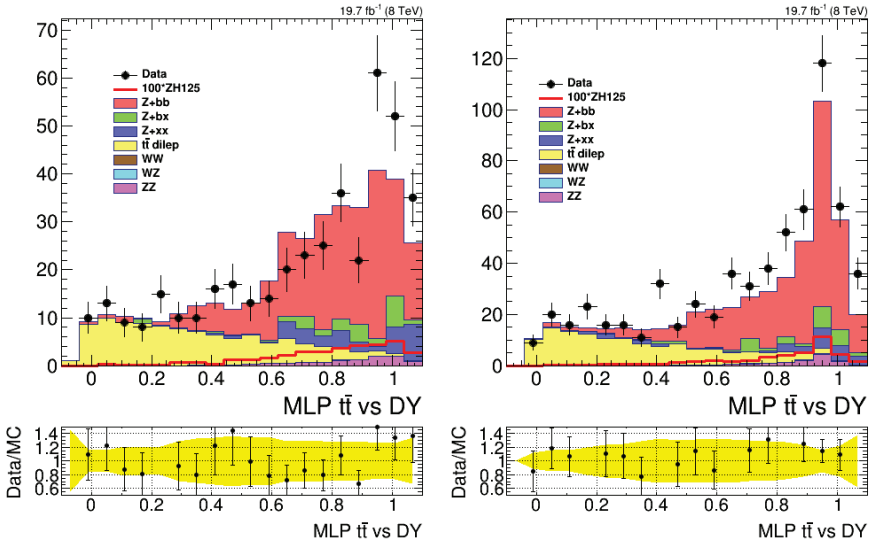

Figure 4.12: Neural-Net output used to discriminate DY events from $t\bar{t}$ events, in the CR. Left plots represent the di-electron channel, right plots the di-muon channel, for the 2-jets category.

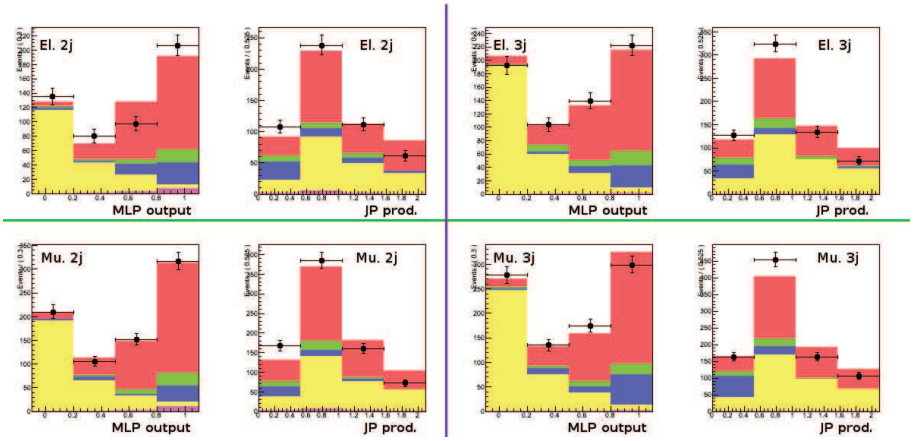

Figure 4.13: Post-fit distributions for electrons (top plots) and muons (bottom plots), in the 2-jets (left plots) and 3-jets (right plots) categories. The represented distributions are the Neural Net output used to discriminate DY events from $t\bar{t}$ events and the product of the two b jets JP discriminator values.

The SF are found to be similar for both taggers.

The correlation matrix for the JP SF is:

$$\begin{array}{c} \\ t\bar{t} \\ Zbx \\ Zxx \\ t\bar{t} \end{array} \begin{pmatrix} 1.000 & -0.135 & -0.191 & -0.018 \\ -0.135 & 1.000 & 0.093 & -0.324 \\ -0.191 & 0.093 & 1.000 & -0.438 \\ -0.018 & -0.324 & -0.438 & 1.000 \end{pmatrix} \quad (4.1)$$

Table 4.7: Scale factors obtained by the 2D simultaneous fit, for the events selected with the JP/CSV tagger, using the data sample that corresponds to the one recorded at 8 TeV, and representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}..

SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
JP			
1.036 ± 0.06	1.273 ± 0.06	1.637 ± 0.24	1.027 ± 0.04
CSV			
1.140 ± 0.06	1.348 ± 0.06	1.359 ± 0.14	1.006 ± 0.04

while for the SF obtained using CSV, the correlation matrix is:

$$\begin{array}{c} \\ t\bar{t} \\ Zbx \\ Zxx \\ t\bar{t} \end{array} \begin{pmatrix} 1.000 & -0.316 & -0.424 & 0.055 \\ -0.316 & 1.000 & 0.194 & -0.320 \\ -0.424 & 0.194 & 1.000 & -0.461 \\ 0.055 & -0.320 & -0.461 & 1.000 \end{pmatrix} \quad (4.2)$$

As expected, the most correlated SF are the Zbb and Zbx ones, since these events are very similar. One can also notice that the correlation between the various background sources are smaller for the JP tagger; this is due to the better purity of obtained with the JP tagger, with respect to the CSV.

In order to see if the value of SF_{Zxx} is compatible with the expectations, it is compared with the latest SF for the misidentification probability provided by the b-tagging group of CMS, listed in table 4.8 [68]. In this analysis, as two Medium tagged b jets are required, this leads a global mis-tagging SF of 1.21 ± 0.31 for JP and 1.37 ± 0.25 for CSV. These numbers are almost compatible with the SF_{Zxx} value (taking into account the fact that they already contained the SF in Table 4.8). This means that the observed discrepancy seems to come mainly from an under-estimation of the SF_{light} by CMS, probably amplified by NLO effects.

Table 4.8: Data/MC scale factors SF_{light} (related to the mis-identification probability), provided by the CMS b-tagging working group [68], for the CSV and JP taggers at the Medium WP, for jet p_T in the range [80-120] GeV. Both statistical and the systematic uncertainties are given.

b-tagger	SF_{light}
JP Medium WP	$1.10 \pm 0.02 \pm 0.20$
CSV Medium WP	$1.17 \pm 0.02 \pm 0.15$

4.3.1 Cross-checks

The data/MC discrepancy observed is quite important, as it was discussed in the previous section. Several hypotheses can be drawn: the trigger efficiency can be badly reproduced by the simulation. However, the leptons η distributions show a flat data/MC ratio (as it is shown on Fig. 4.9), which tends to discard the trigger as the source of the discrepancy. The preferred assumption is thus the mis-modeling of the b-tagging

SF, since at the stage "ll+jj+X" (when two leptons and two jets are selected, and no cut is applied on the MET), the data/MC agreement is within 5%, and a good reproduction of the main kinematic variables shapes is observed (see Fig. 4.14 and Fig. 4.15). The discrepancy appears when b-tagging is required. More plots on the FR and at "ll+jj+X" stage can be found in Appendix C.

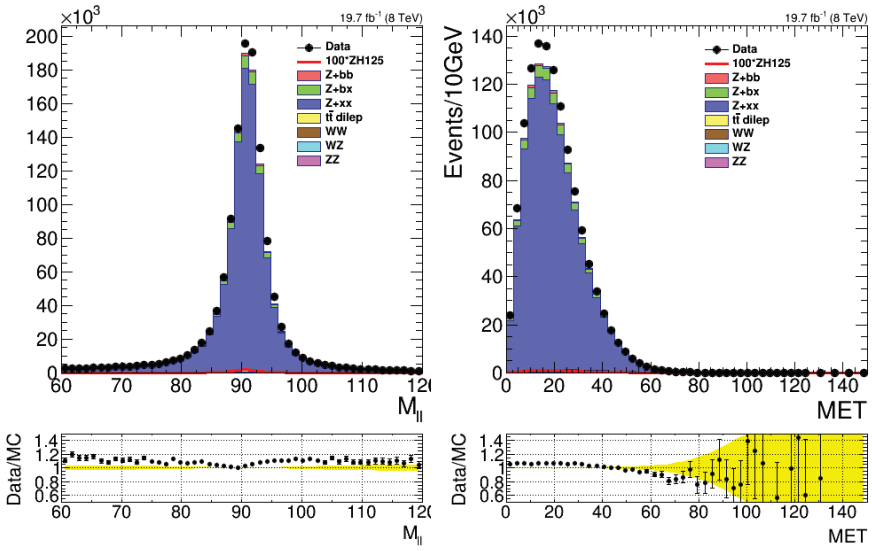

Figure 4.14: Distributions of the invariant mass of the di-lepton system (left), and of the MET (right), at "ll+jj+X" stage. The FR selection has been applied, except for the b-tagging requirement and the MET cut.

Two tests are performed: a reweighting of the di-lepton invariant mass is done, in two different ways, to cancel the data/MC disagreement. Scale factors are extracted then propagated further.
The goal of these tests is to show that whenever the discrepancy comes from a trigger effect or a mis-modeling of the b-tagging SF, it is absorbed during the background fit procedure and therefore, there is no impact on the final result. The tagger used for these test is JP.

M_{ll} reweighting I:

A first bin-by-bin reweighting is done by adjusting the all the DY contributions together to the data at the stage “ll+bj+X” (two leptons, one b jet and one jet), independently for muons and electrons.

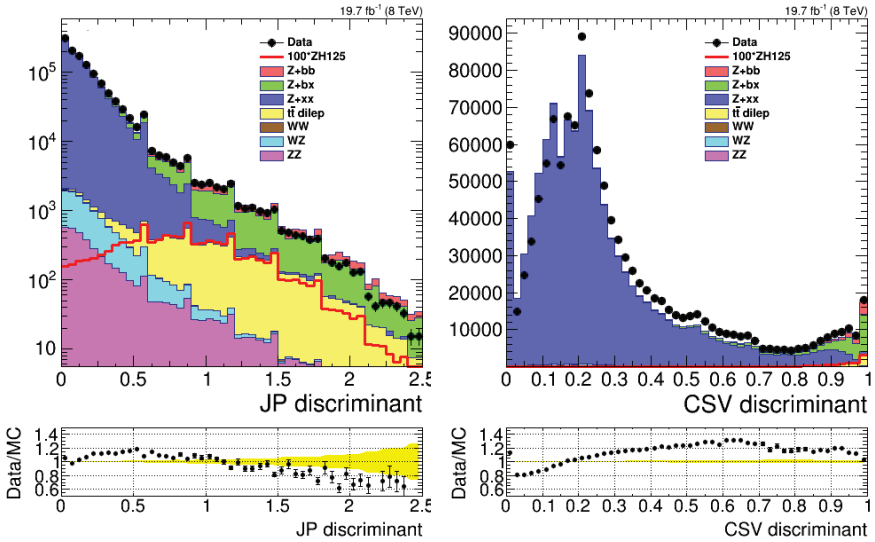

Figure 4.15: Distributions of the JP discriminator (left), and of the CSV discriminator (right), at "ll+jj+X" stage. The FR selection has been applied, except for the b-tagging requirement and the MET cut.

The use of “ll+bj+X” events is relevant since at this stage, a discrepancy is already seen, and the requirement of only one b jet improves the statistics for the reweighting. In this case, the DY categorization is done using the b-tagged jets with the highest p_T and discriminator value.

A closure test is performed and the result can be seen on Fig. 4.16. A global scale factor of 1.16 is applied to all the DY contributions in the di-muon channel, while the SF for the di-electron channel is of 1.10 (see Table 4.9).

Table 4.9: Scale factors obtained by the M_{ll} reweighting I, performed independently for muons and electrons, using the JP tagger.

SF($\mu\mu$)	SF(ee)
1.16	1.10

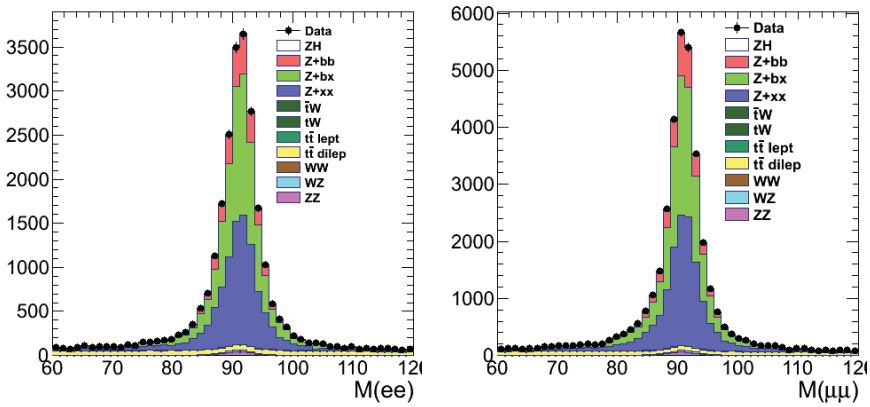

Figure 4.16: Invariant mass of the two electrons (left) or muons (right) after a bin-by-bin reweighting of all the DY contribution to match the data, at "llbj+X" stage. The tagger used is JP.

Table 4.10: Scale factors obtained by the 2-D simultaneous fit, for the events selected with the JP tagger. The SF found using the M_{ll} reweighting I have been propagated before the fit procedure.

SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
0.981 ± 0.053	1.126 ± 0.056	1.455 ± 0.218	1.022 ± 0.037

The background fit is then performed, taking into account the global SF obtained for the different DY contributions. The new obtained SF can be seen in Table 4.10.

When using these new SF in the following instead of the nominal ones, the deviation on the final result is found to be negligible at the 1% level.

M_{ll} reweighting II:

The second reweighting performed is a bin-by-bin reweighting of the Zbx contribution to the data in the 2-jets category and of the Zxx contribution in the 3-jets category, for electrons and muons together. This choice is motivated by the fact that these two contributions are dominant in the dedicated categories. This reweighting aims at roughly correct for the mis-modeling of the b-tagging SF, as it is suspected to be the origin of the observed data/MC discrepancy. The plots of the closure test can be seen

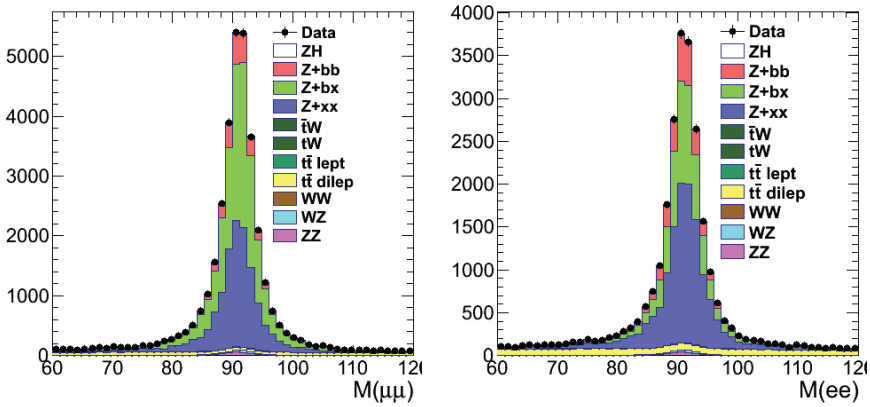

Figure 4.17: Invariant mass of the two leptons after a bin-by-bin reweighting of the Zbx contribution for the 2-jets category (left) and of the Zxx contribution for the 3-jets category (right), in order to match the data, at "llbj+X" stage. The tagger used is JP.

on Fig. 4.17. A SF of 1.14 is applied to the Zbx events in the 2-jets category while Zxx events are reweighted by a SF of 1.56 in the 3-jets category (see Table 4.11).

Table 4.11: Scale factors obtained by the M_{ll} reweighting II, performed for muons and electrons together. The Zbx contribution is reweighted in the 2-jets category while the SF for the 3-jets category is obtained using the Zxx events. The tagger used is JP.

SF(Zbx/2j)	SF(Zxx/3j)
1.14	1.56

In this case again, the background fit is performed taking into account the new SF for the Zbx and the Zxx events. The final SF can be seen in Table 4.12.

Again, the new SF of Table 4.12 are used to redo the complete analysis, and the final result agrees better than 2% with the nominal one.

As a conclusion, both of the reweightings lead to final result similar to the nominal one, meaning that the background fit absorbs the source of the discrepancy, whenever it is coming from a bad trigger efficiency in the simulation or a mis-modeling of the b-tagging SF.

Table 4.12: Scale factors obtained by the 2D simultaneous, for the events selected with the JP tagger. The SF found using the M_{ll} reweighting II have been propagated before the fit procedure.

SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
1.141 ± 0.059	1.227 ± 0.061	0.979 ± 0.167	1.025 ± 0.037

4.4 Yields and Control Plots

The SF obtained by the background fit are applied for both JP and CSV selection, and the yields in the signal region can be found respectively in Table 4.13 and Table 4.14. The $Signal/\sqrt{Background}$ is very similar for both taggers: it is 0.36 for JP and 0.40 for CSV.

Control plots in the signal region of various kinematic variables are shown in this section, for the selection using JP in both jet categories: the di-lepton and di-jet invariant mass (Fig. 4.18, the MET significance and the leading b jet p_T on Fig. 4.19, the sub-leading b jet p_T and leadind JP discriminator value on Fig. 4.20. The sub-leadind JP discriminator value and the leadind CSV discriminator value are displayed on Fig. 4.21 while the sub-leadind CSV discriminator value is shown on Fig. 4.22 (using events selected by the CSV tagger). All of them show a reasonable agreement between MC and data, and a good reproduction of the kinematic shapes.

More plots can be found in Appendix C.

Table 4.13: Data yields for the SR displayed in Table 4.3 and Table 4.4, compared with the expectation from the different main processes, normalized to the theoretical cross section, for the electron and muon channels, in both jets categories. The tagger used is JP and the data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}.

JP	Zbb	Zbx	Zxx	$t\bar{t}$	ZZ	Signal	Tot. MC (no signal)	Data	Data/MC
$\mu\mu$ 2j	182.2 ± 13.5	16.6 ± 4.1	39.8 ± 6.3	70.3 ± 8.4	11.4 ± 3.4	4.9 ± 2.2	320.3 ± 17.9	378	1.18
ee 2j	147.2 ± 12.1	15.9 ± 4.0	46.5 ± 6.8	50.2 ± 7.1	8.0 ± 2.8	3.5 ± 1.9	267.8 ± 16.4	251	0.94
$\mu\mu$ 3j	361.2 ± 19.0	28.4 ± 5.3	80.0 ± 8.9	135.6 ± 11.6	17.5 ± 4.2	3.6 ± 1.9	622.7 ± 25.0	649	1.04
ee 3j	251.5 ± 15.9	35.6 ± 6.0	41.9 ± 6.5	106.2 ± 10.3	13.6 ± 3.7	2.8 ± 1.7	448.8 ± 21.2	425	0.95
Total	942.1 ± 30.7	96.5 ± 9.8	208.2 ± 14.4	362.3 ± 19.0	50.5 ± 7.1	14.8 ± 3.8	1659.6 ± 40.7	1703	1.03

Table 4.14: Data yields for the SR displayed in Table 4.3 and Table 4.4, compared with the expectation from the different main processes, normalized to the theoretical cross section, for the electron and muon channels, in both jets categories. The tagger used is CSV and the data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}.

CSV	Zbb	Zbx	Zxx	$t\bar{t}$	ZZ	Signal	Tot. MC (without signal)	Data	Data/MC
Z($\mu\mu$) 2j	297.6 ± 17.3	38.3 ± 6.2	76.7 ± 8.8	90.3 ± 9.5	16.3 ± 4.0	6.3 ± 2.5	519.2 ± 22.8	530	1.02
Z(ee) 2j	213.6 ± 14.6	31.6 ± 5.6	62.8 ± 7.9	64.7 ± 8.0	11.3 ± 3.4	4.8 ± 2.2	384.0 ± 19.6	338	0.88
Z($\mu\mu$) 3j	499.0 ± 22.3	71.5 ± 8.5	123.0 ± 11.1	173.3 ± 13.2	25.0 ± 5.0	4.8 ± 2.2	891.8 ± 29.9	921	1.03
Z(ee) 3j	382.5 ± 19.6	56.2 ± 7.5	97.4 ± 9.9	134.7 ± 11.6	18.2 ± 4.3	3.8 ± 1.9	689.0 ± 26.2	605	0.88
Total	1392.7 ± 37.3	197.6 ± 14.1	359.9 ± 19.0	463.0 ± 21.5	70.8 ± 8.4	19.7 ± 4.4	2484.0 ± 49.8	2394	0.96

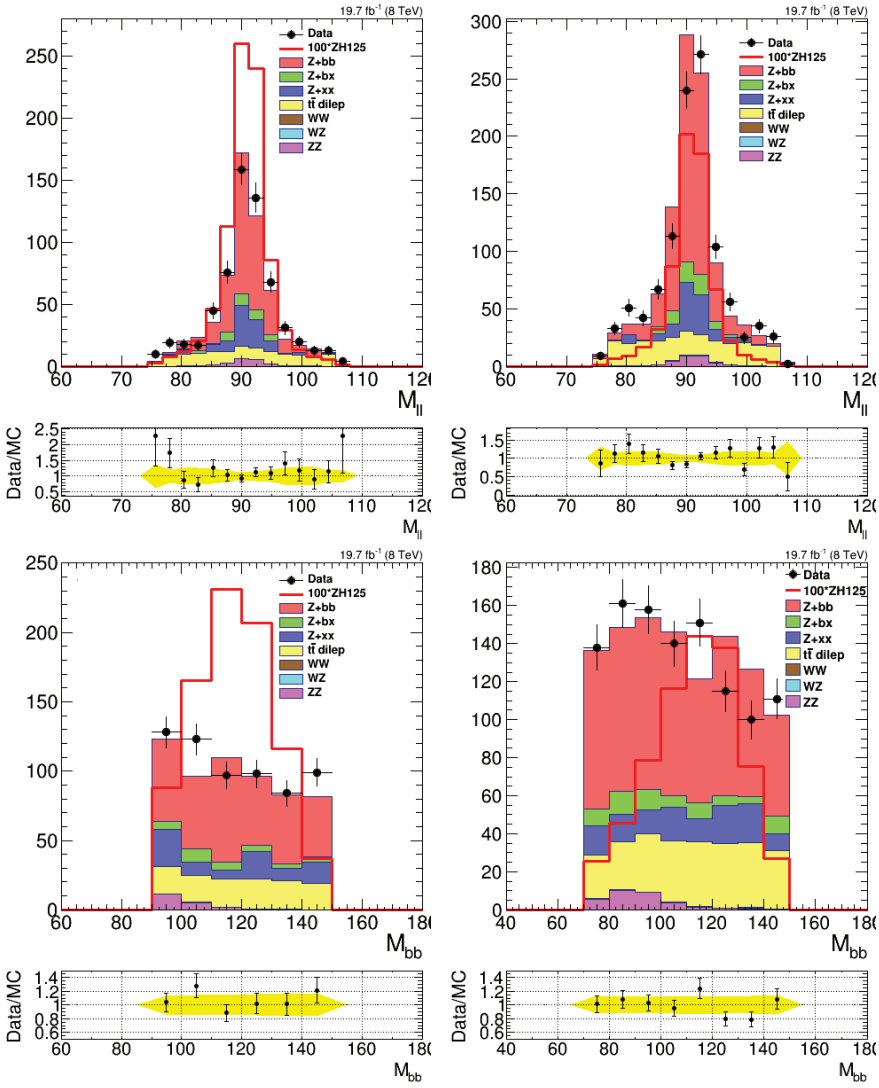

Figure 4.18: Top: di-lepton invariant mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories Bottom: di-jet invariant mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories. The JP tagger is used, and events have been renormalized according to their cross section. All the correction scale factors are applied.

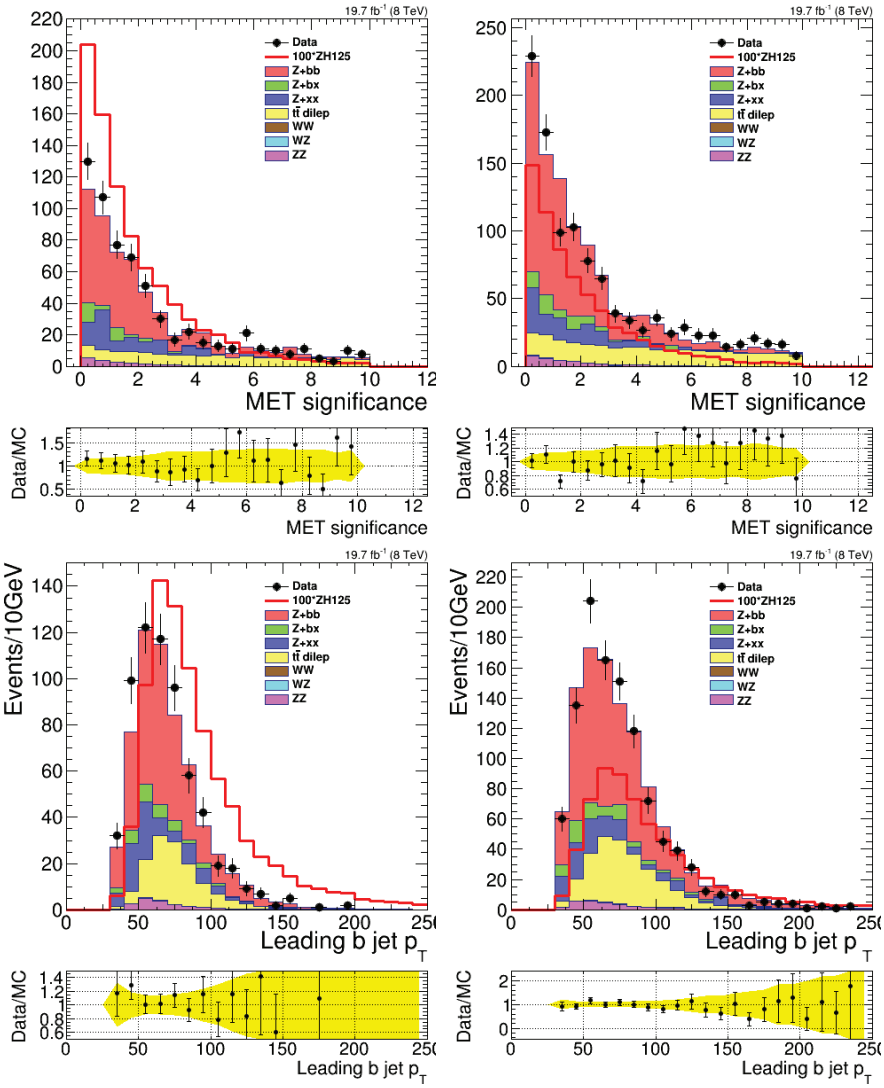

Figure 4.19: Top: MET significance mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories. Bottom: leading b jet p_T in the signal region, for the 2-jets (left) and the 3-jets (right) categories. The JP tagger is used, and events have been renormalized according to their cross section. All the correction scale factors are applied.

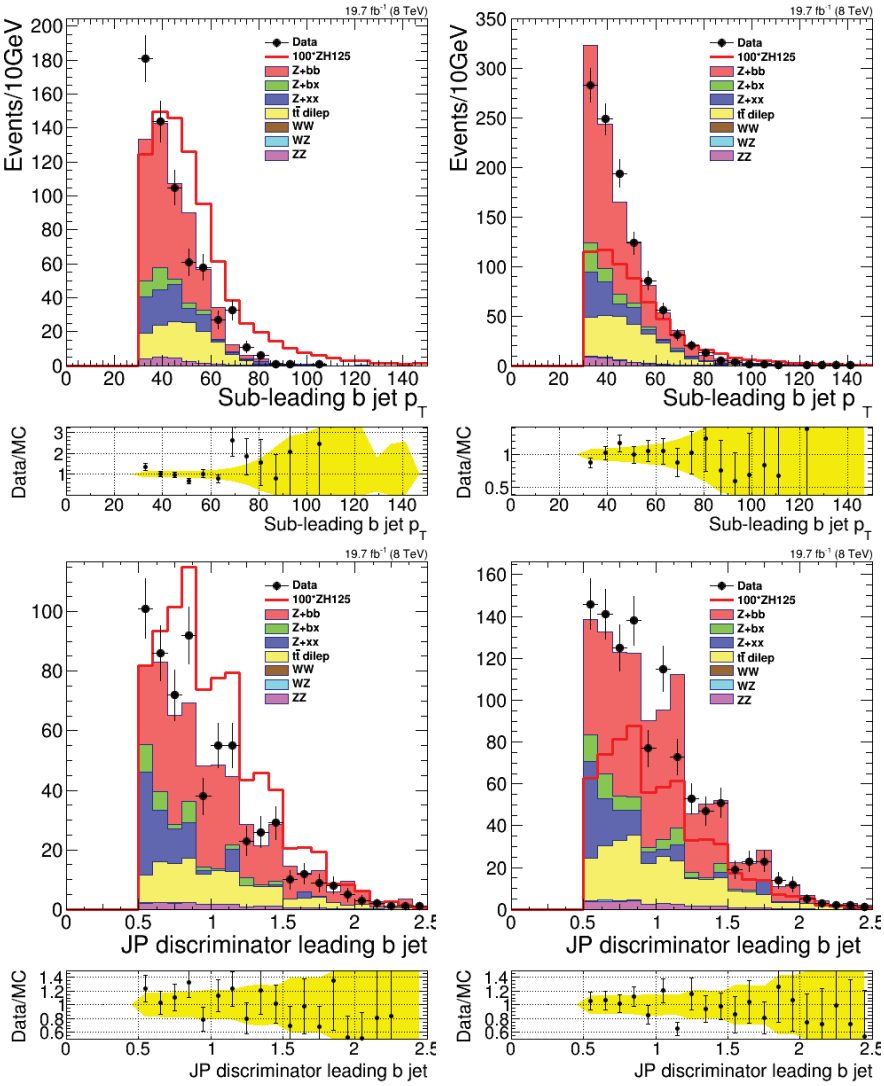

Figure 4.20: Top :sub-leading b jet p_T in the signal region, for the 2-jets (left) and the 3-jets (right) categories. Bottom: leading b jet JP discriminator value in the signal region, for the 2-jets (left) and the 3-jets (right) categories. The JP tagger is used and events have been renormalized according to their cross section. All the correction scale factors are applied.

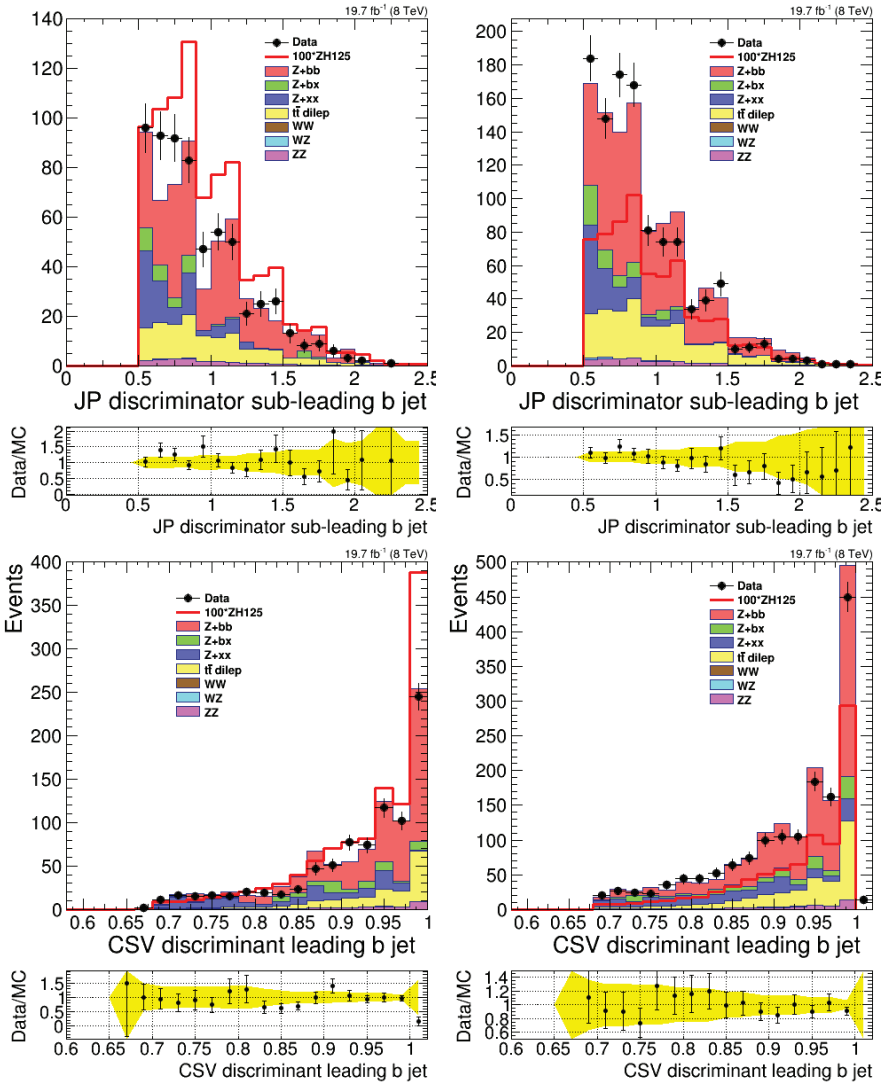

Figure 4.21: Top :sub-leading b jet JP discriminator value in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the JP selection. Bottom: leading b jet CSV discriminator value in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV selection. Events have been renormalized according to their cross section. All the correction scale factors are applied.

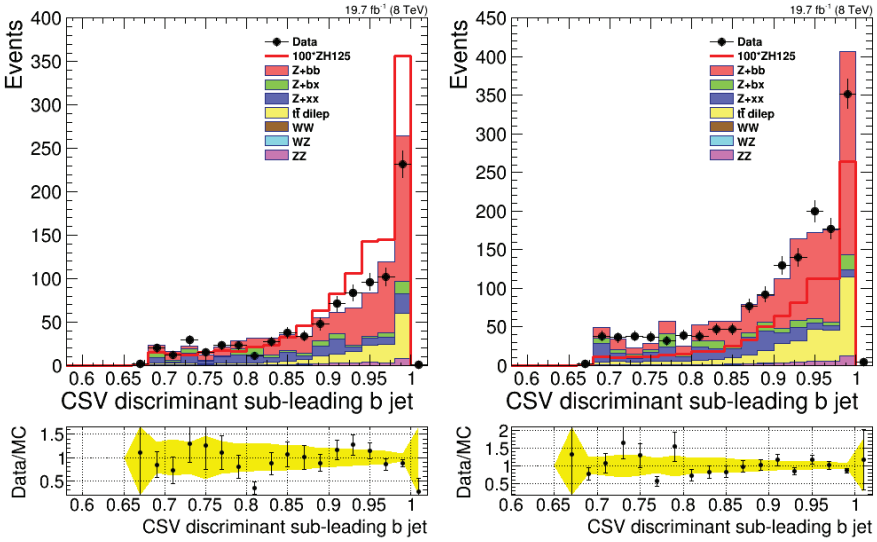

Figure 4.22: Sub-leading b jet CSV discriminator value in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

4.4.1 Validation of the MEM weights

The weights returned by the MEM, presented in Chapter 1, have already been used for the background fit procedure, in Section 4.3. These weights also allow to distinguish the signal events from the main background processes.

Five weights corresponding to different background hypotheses, and two weights corresponding to the signal hypothesis are evaluated:

- Weights for two Zbb hypothesis: Zbb induced by $q\bar{q}$ (Zbb_{qq}) and Zbb induced by gg (Zbb_{gg}). An approximation is made at this point: the same Zbb weights are applied for the Zbx and Zxx events, meaning that no specific hypothesis was created for the Zbx and Zxx topology. This is justified by the fact that Zxx events come from the same Feynman diagram than Zbb events (see Fig. 4.1). For Zbx, the issue will be solved in the next Chapter.
- Weights for the $t\bar{t}$ di-leptonic hypothesis;
- Weights for the ZZ hypothesis, with (without) imposing E-p conservation (see Chapter 1): ZZ_{cor0} (ZZ_{cor3});

- Weights for the ZH hypothesis, with (without) imposing E-p conservation: ZH_{cor0} (ZH_{cor3});

The validation of the MEM weights can be seen in this section, for the JP tagger for both 2-jets and 3-jets categories: the $t\bar{t}$ weights on Fig. 4.23, Zbb induced by $q\bar{q}$ (Zbb_{qq}) and Zbb induced by gg (Zbb_{gg}) on Fig. 4.24 and 4.25; The ZZ weights are shown respectively for the correction 0 (Fig. 4.26) and 3 (Fig. 4.27), and finally the ZH weights for both correction 0 (Fig. 4.28) and correction 3 (Fig. 4.29).

Given the available statistics, the agreement between MC and data is satisfactory. The plots for the CSV selection are available in Appendix C.

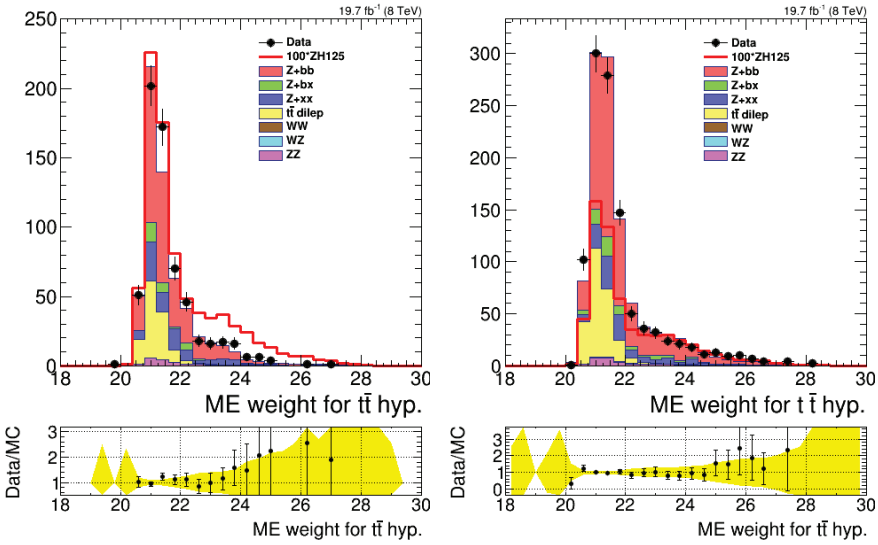

Figure 4.23: ME weights for the $t\bar{t}$ hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

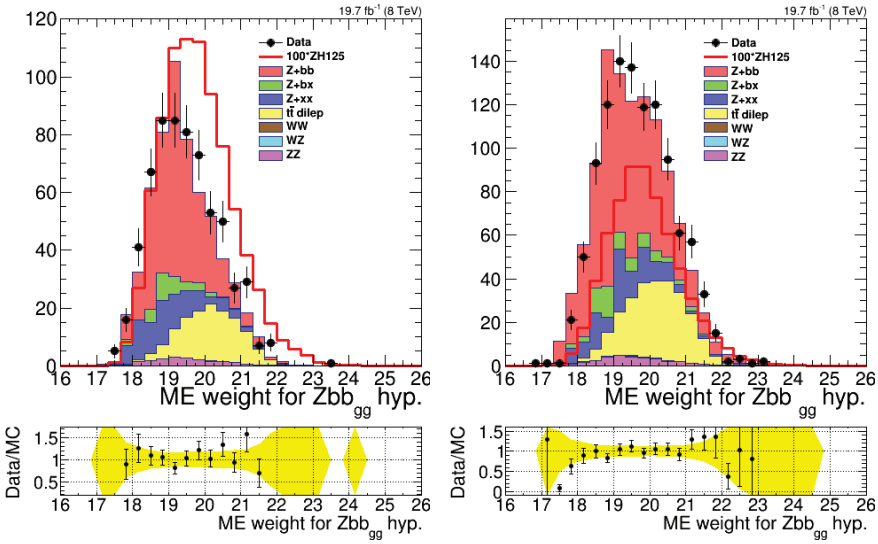

Figure 4.24: ME weights for the Zbb induced by gluon-gluon hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

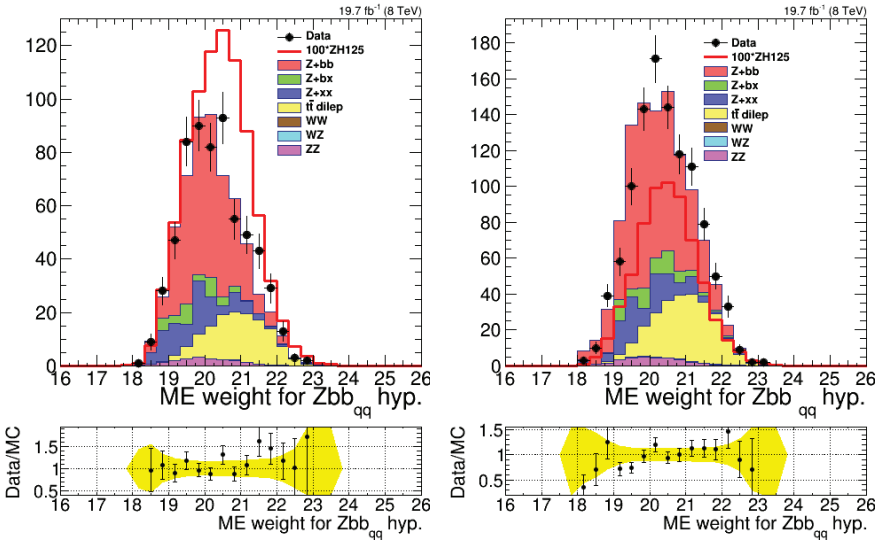

Figure 4.25: ME weights for the Zbb induced by $q\bar{q}$ hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

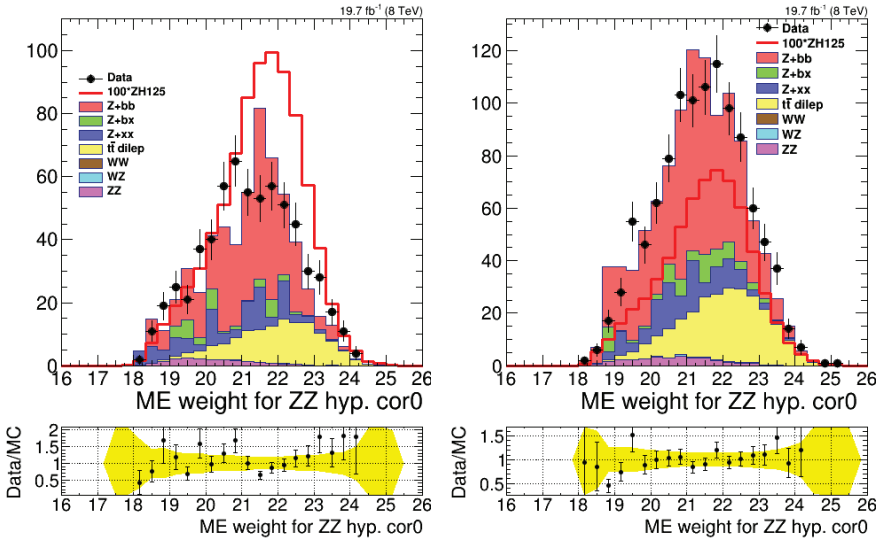

Figure 4.26: ME weights for the ZZ hypothesis with E-p conservation (cor_0), for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

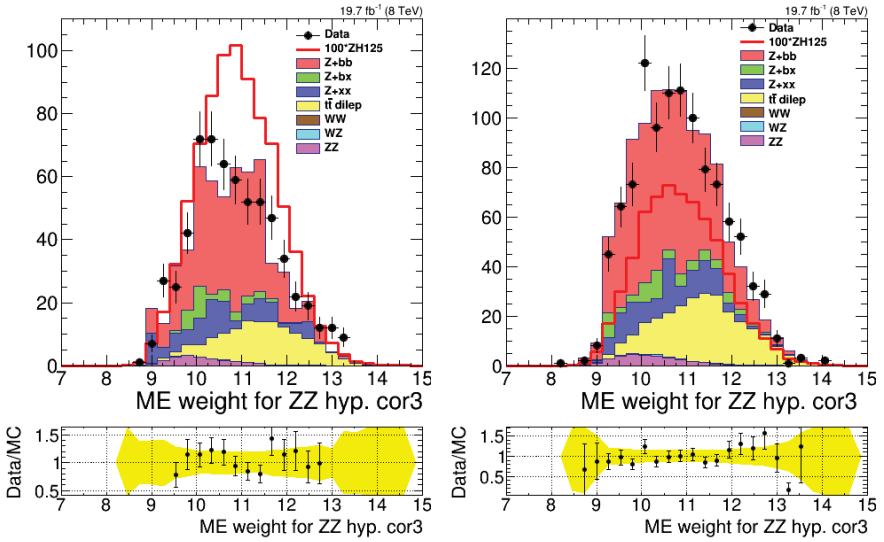

Figure 4.27: ME weights for the ZZ hypothesis without E-p conservation (cor_3), for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

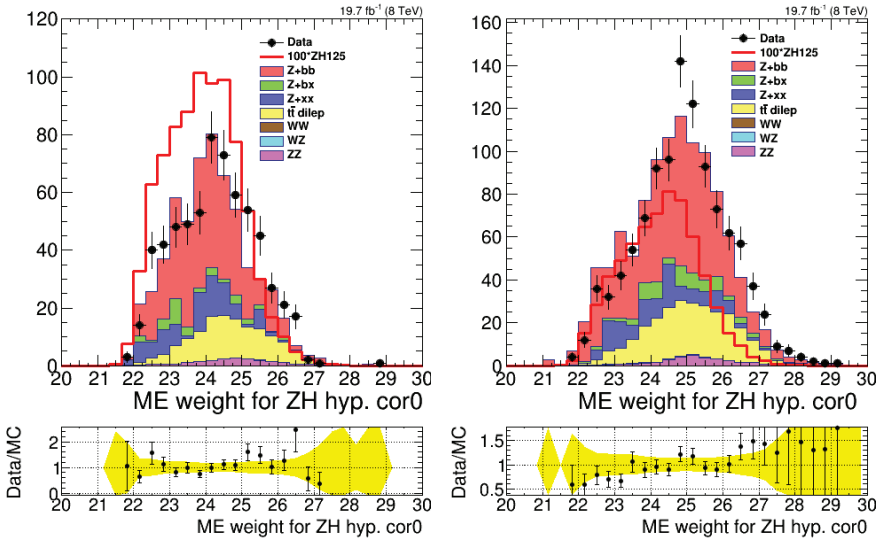

Figure 4.28: ME weights for the ZH hypothesis with E-p conservation (cor_0), for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

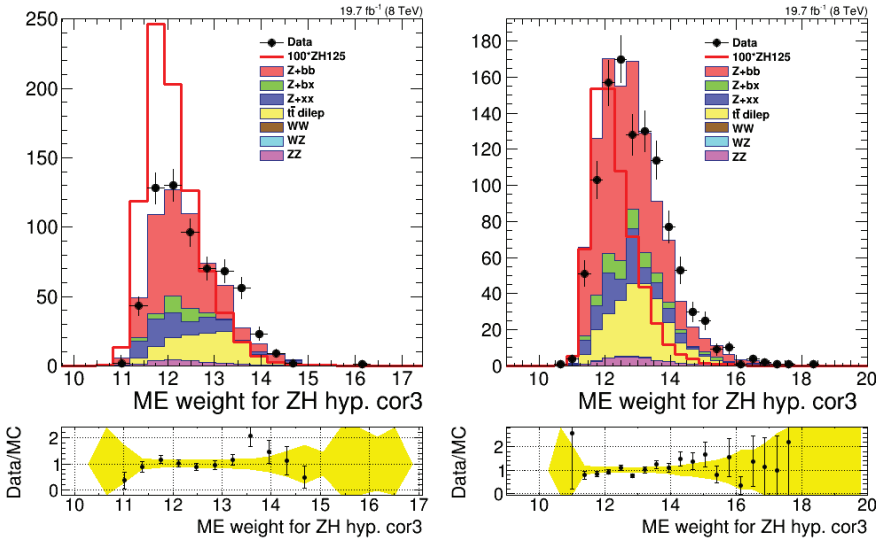

Figure 4.29: ME weights for the ZH hypothesis without E-p conservation (cor_3), for the 2-jets (left) and the 3-jets (right) categories, for the JP tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

4.5 Background estimation

The different weights described in the previous section are combined using a MVA tool. Two methods have been tested: the use of a BDT, or a complex MLP (both tools are described in Appendix D). The final result, the exclusion limit on the ZH signal strength, has been computed using both tools, and the best limit was obtained using a BDT. Therefore, only the procedure using the BDT is presented in this section. More details about the tests performed with a MLP can be found in Appendix C.

The BDT takes as input the seven weights presented in the previous section, and separate trainings are performed for the 2-jets and 3-jets categories. On the other hand, test and training samples are composed by events in the both di-muon and the di-electron channels. Indeed, the different kinematic distributions do not differ significantly between the two channels and a a good statistic is required for the training step. Only 50% of the available samples are used for the training; the remaining 50% are used to perform over-training tests. The selection criteria corresponding to the SR are imposed to the simulated events entering in the training. Since the contribution of each of the background processes to the total expected background varies significantly, a weight is assigned to each event, based on the process contribution. The weight for the DY, $t\bar{t}$, and ZZ samples are respectively 0.9, 0.085, 0.015.

In the 3-jet category, in addition to the Matrix Element weights, extra variables are added to amplify the discrimination: the invariant mass of the two b jets with the FSR jet, $M(bb, fsr)$, the angular distance between the FSR jet and its closest b jet, $\Delta R(fsr, b)$, and finally the angular distance between the two b jets, $\Delta R(bb)$. The plots of the first two variables can be seen respectively on Fig. 4.30. It should be noticed here that for technical reasons, the Zbb 4-flavour sample used for the training of the Neural Net in the previous analysis was not available for this study. As a consequence, a lack of statistics is clearly affecting the BDT training.

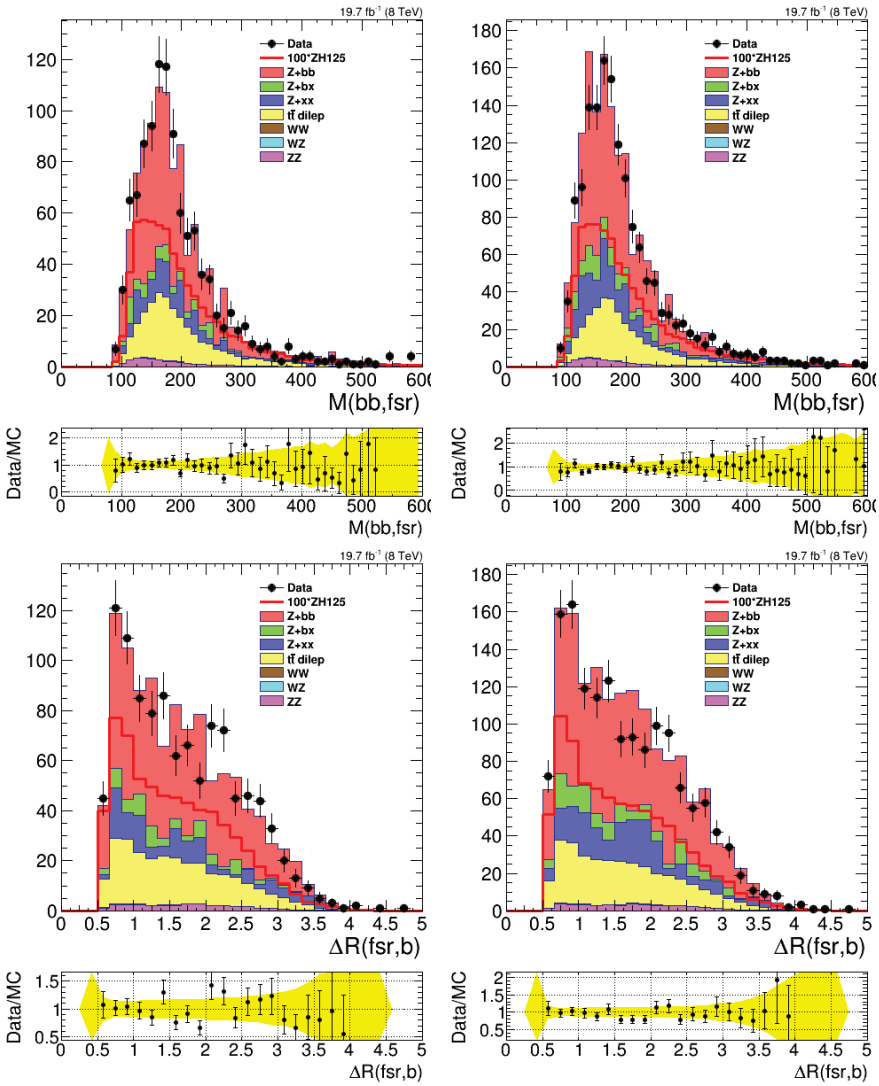

Figure 4.30: Top: distribution of the angular distance between the FSR jet and the closest b jet, $M(bb, fsr)$, used in the 3-jets category, for the JP (left) and the CSV (right) selection. Bottom: distribution of the invariant mass between the two b jets and the FSR jet, $\Delta R(bj)$, used in the 3-jets category, for the JP (left) and the CSV (right) selection. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

4.5.1 Final discriminant plots

The plot of the BDT output can be seen for JP (CSV) on Fig.4.31 (Fig. 4.32), for both jet categories and both muon and electron channels, in the signal region. Again, a quite reasonable agreement is found between MC/data, given the few available statistics. Plots for the MLP output are available in Appendix C.

4.6 Systematics

Systematic uncertainties affecting the estimated rates of signal and background processes, as well as the shape of the final discriminator, can bias the outcome of this search. First each independent source of uncertainty is identified then its effect on the event yields of the different processes is described. An individual source of uncertainty can affect multiple processes (also across different channels) and all these effects will be correlated.

The most common model used for systematical uncertainties is the log-normal. The event distribution is characterized by a parameter κ, and affects the expected event yields in a multiplicative way: a positive deviation of +1σ corresponds to a yield scaled by a factor κ compared to the nominal one, while a negative deviation of -1σ corresponds to a scaling by a factor 1/κ. For small uncertainties, the log-normal is approximately a Gaussian. This model is used to determine the effect implied by the following systematics: the luminosity, the lepton scale factor, the ZH cross section and the background fit.

In this analysis, the limit depends on the distribution of the BDT output, built in Section 4.5; the shape of this discriminating variable can be affected by several effects, such as the jet energy scale/resolution, the uncertainty on the b-tagging scale factors, and the limitation in MC statistics. To determine the impact of these systematics, the difference in the discriminant shape is evaluated: the observed distribution of the expectations from the signal and all backgrounds affected by the systematic, are provided as histograms, all with the same binning.

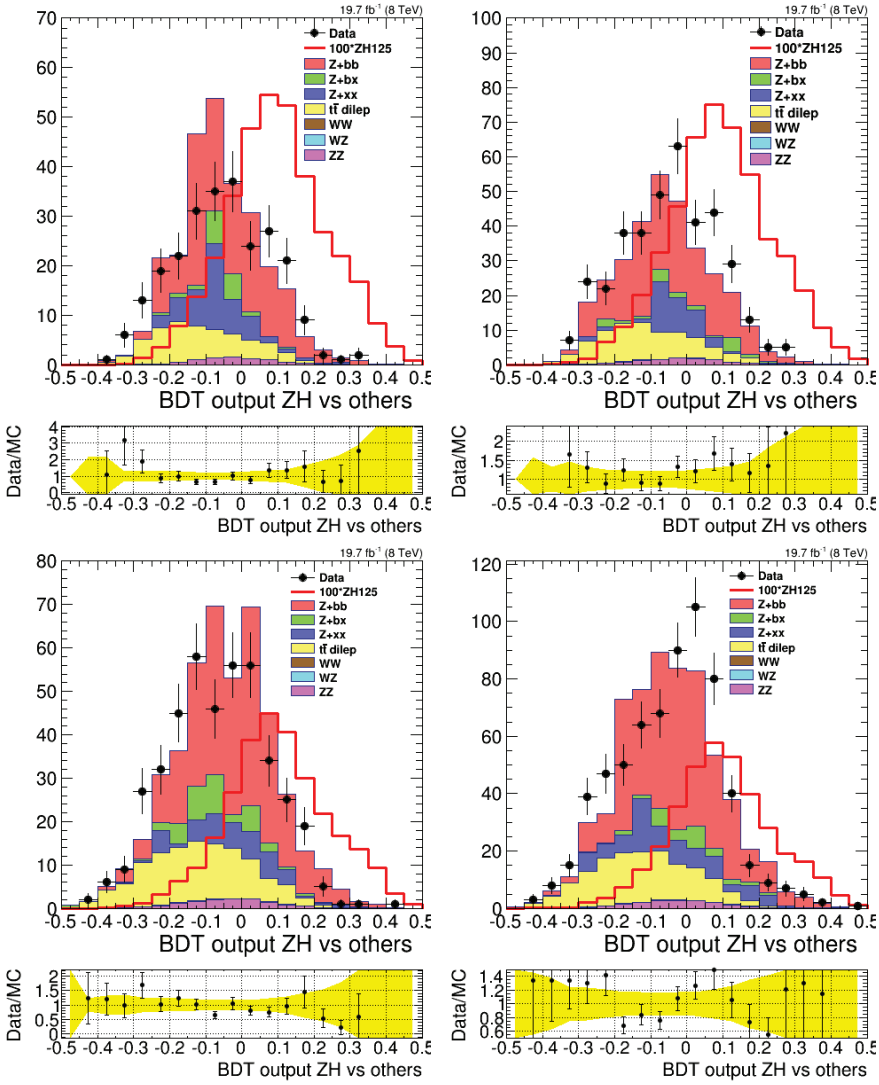

Figure 4.31: Top: BDT output in the SR for the electrons (left plots) and muons (right plots) in the 2-jets category. Bottom: BDT output in the SR for the electrons (left plots) and muons (right plots) in the 3-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is JP.

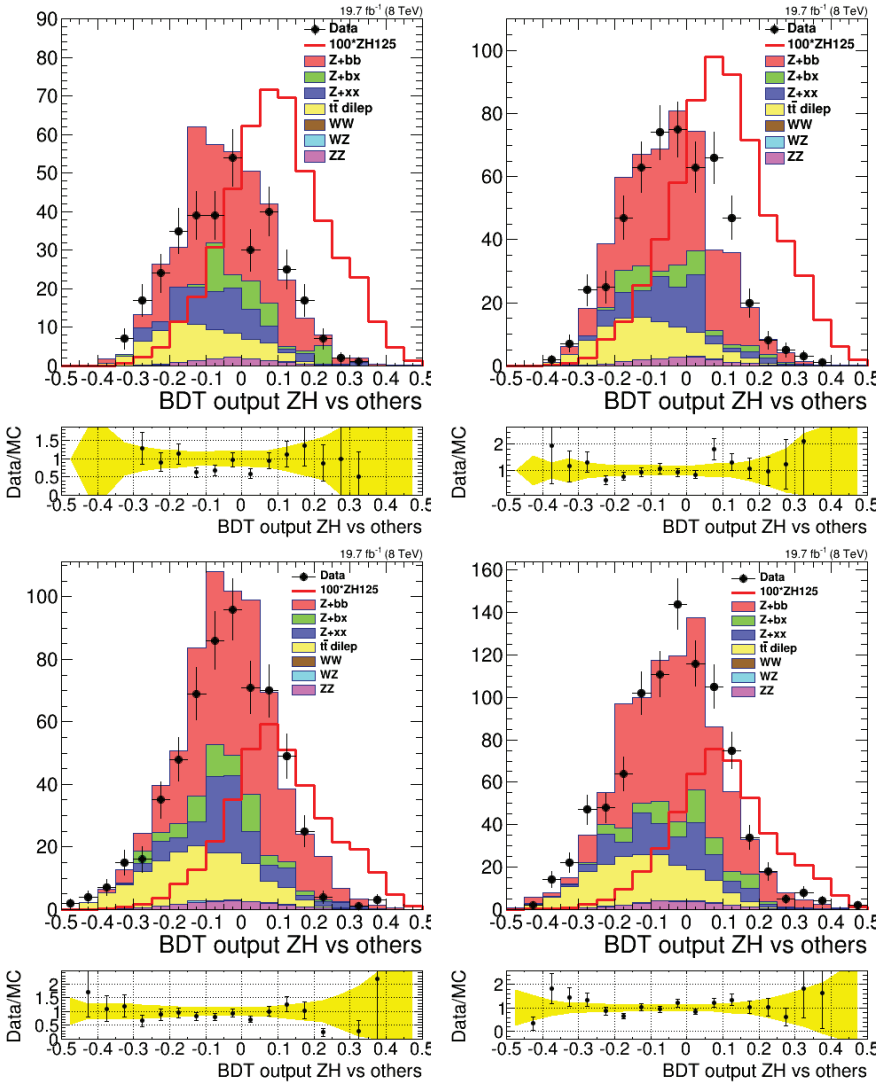

Figure 4.32: Top: BDT output in the SR for the electrons (left plots) and muons (right plots) in the 2-jets category. Bottom: BDT output in the SR for the electrons (left plots) and muons (right plots) in the 3-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is CSV.

A detailed list of all the systematic uncertainties considered in this search is given below:

- **Luminosity**: the uncertainty on the luminosity affects the normalization of the signal and the di-boson background. Indeed, for the other background processes, this uncertainty is absorbed in the background fit procedure. It is assumed to be 4.4% for the 8 TeV dataset;

- **Signal cross section**: the uncertainty on the total signal cross section is 4% [81]. This number accounts as well for the PDF uncertainty;

- ZZ **normalization**: An uncertainty of 15% is assigned to the ZZ normalization. This value corresponds to the uncertainty on the cross-section measured by CMS [80].

- **Lepton reconstruction and trigger efficiencies**: these efficiencies are measured using the "Tag and Probe" technique (explained in Chapter 2). A flat uncertainty of 2% is assigned to the total trigger and lepton reconstruction efficiency for both electrons and muons. Uncertainties between electrons and muons are assumed to be uncorrelated. It is only applied to the signal process since it is already taken into account for the ZZ measurement;

- **b-tagging and mis-tagging efficiencies**: the scale factors associated to the b-tagging are variated up and down according to their uncertainties, following the recommendations of the b-tagging working group. The variations are performed separately for heavy flavor jets for the b-tagging uncertainty, and for light jets for the mis-tagging uncertainty. In order to assess the effect of these uncertainties on the normalization of the different backgrounds, the background fit procedure has been repeated using the up and down variations. For all the background processes the impact has been found to be small compared to the corresponding statistical uncertainty of the fit;

- **Jet energy scale/resolution**: the Jet Energy Scale (JES) uncertainty is evaluated by applying jet-energy corrections that describe one standard deviation variations with respect to the default correction factors (up and down). The JER effect is estimated by increasing (decreasing) the default smearing by a factor of two for the up (down). For these uncertainties, two effects are measured:

    1. The effect on the event selection: the whole analysis is redone with events affected by the change in energy scale. These events have to pass the selection described in Section 4.2. The yields of events passing the selection for each systematic variation considered can be seen in Appendix C, along with the associated SF computed by the background fit procedure.

2. The ME weights are recomputed with events for which the jet energy resolution/scale have been variated. A special patch has been created to run Madweight on the same event with five different configurations: one nominal, two for JES$\pm$ and two for JER $\pm$. An approximation is made at this stage: the ISR correction computed for the nominal event is applied to the four other configurations. However, this approximation is estimated to be small with respect to the resolution on the di-jet mass used for the ISR evaluation;

- **Background Fit**: the statistical uncertainties associated to the four scale factors extracted by the background fit has been considered for each process. In order to obtain uncorrelated systematic uncertainties, the correlation and covariance matrices returned by the fit are used in a procedure explained in Appendix C;

- **Monte Carlo statistics**: the limited size of the generated Monte Carlo samples represents an additional source of uncertainty. To account for this effect, alternative shapes that vary exclusively the contents of one of the bins of the discriminator are introduced for each process. The considered bin is multiplied by factors representing $\pm$ one standard deviation of a Poisson distribution whose parameter is the number of MC events populating the bin. This means that the up and down fluctuations are not symmetrical for the bin, especially if it is populated only by a small number of events. The statistical uncertainties corresponding to the different bins are included only for the last 10 bins, the ones with the most sensitivity, and are assumed to be uncorrelated.

4.7 Results

4.7.1 The CL_s tool

The Confidence Levels (CL) method [84] is used to claim a discovery or establish an exclusion. It is based on a frequentist significance test using a likelihood ratio. The purpose of this analysis is to perform the exclusion of the signal hypothesis. The p-value, that quantifies the statistical significance of an observed signal, is set at 0.05, which corresponds to a confidence level of 95%. The limits are computed from the shape of the multivariate discriminators described in Section 4.5. The likelihood function is then the product of Poisson probabilities on all the N bins of the distribution:

$$\mathcal{L}(\mu,\theta) = \prod_{j=1}^{N} \frac{(\mu \times s_j + b_j)^{n_j}}{n_j!} e^{-\mu \times s_j - b_j} p(\tilde{\theta}|\theta) \tag{4.3}$$

where in the bin j n_j, s_j and b_j represent respectively the number of observed candidates, the signal and background expected rates. μ is the signal strength modifier (σ/σ_{SM}) and θ is the set of nuisance parameters, corresponding to the systematic uncertainties, their expected values being represented by $\tilde{\theta}$. The method computes the probability P_μ defined by:

$$P_\mu = -2\ln\frac{\mathcal{L}(X|\mu,\hat{\theta}_\mu)}{\mathcal{L}(X|\hat{\mu},\hat{\theta})} \tag{4.4}$$

where $\hat{\mu}$ and $\hat{\theta}$ maximize the likelihood given the observed data X, while $\hat{\theta}_\mu$ maximize the likelihood value for a given value $\hat{\mu}$ such as $0 \leq \hat{\mu} \leq \mu$. P_μ represents the incompatibility between the data and the hypothesis for a value of μ, and when P_μ is high, so is the inconsistency between the data and the hypothesis. The p-value for a given $P_{\mu,obs}$ is:

$$p_\mu = \int_{P_{\mu,obs}}^{\infty} f(P_\mu|\mu)dP_\mu \tag{4.5}$$

where $f(P_\mu|\mu)$ is the probability density function of P_μ assuming the hypothesis μ, obtained using toy experiments. The confidence level is then built as:

$$CL_s(\mu) = \frac{p_\mu}{p_0} \tag{4.6}$$

If μ=1, CL_s worths less than 0.05 and it can be claimed that the signal is excluded for a nominal production rate, at 95% of confidence level.

4.7.2 Limits

The CL_s prescription is employed to set upper limits on the SM Higgs boson associated production with a Z boson and decaying in a pair of b quarks. The different sources of systematic uncertainties described in Section 4.6 are taken into account in the computation of the limits. In order to improve the signal sensitivity, the limit is evaluated for each dataset in the di-electron and di-muon final states separately, as well as in the 2-jets and 3-jets categories independently.

First, the expected blinded upper limits are presented in Table 4.15, for JP and CSV, with and without systematics. These limits are computed using the MC distributions to generate pseudo-data in order to perform a MC/data fit. These limits are used to tune the analysis sensitivity. For JP, the expected limit is $3.52\times\sigma_{SM}$, and for CSV, the expected limit is $3.39\times\sigma_{SM}$: both taggers return similar results.

Then, an un-blinding is performed: the data are used in the data/MC fit performed to compute the limits. For JP, the observed limit is $4.89\times\sigma_{SM}$, and for CSV, it is $5.46\times\sigma_{SM}$ (see Table 4.16).
The results are summarized on Fig. 4.33.

A significant degradation of the limit is observed when going from the blinded to the unblinded limits, and this is due to the fact that the observed MC/data discrepancies can be explained by a fluctuation induced by the systematics. CMS provides a tool to check the impact of the systematics on the unblinded limits. If the systematics have a significant impact, the limit on the signal strength can vary a lot because an excess in data might be explained by a variation of a systematic within its uncertainty.
This behavior is observed for the JP approach, as it can be seen on the pool plots of Fig. 4.34 and Fig. 4.35: a much better agreement between data and MC is found after the fit including the systematics, explaining the degradation of the limit observed after unblinding.
A similar impact is visible when using the CSV tagger, and the related plots are shown in Appendix C. This effect is due to the systematic linked to the MC statistics that affect the discriminator shape, since each bin can vary independently.

Table 4.15: Blinded limits on the SM Higgs boson associated production with a Z boson cross section times the $b\bar{b}$ branching ratio, for both JP and CSV tagger. The limits are normalized to the SM predictions.

Blinded			
	Median expected value (no syst.)	Median expected value (with syst.)	$+1\sigma$ band
JP	2.73	3.52	5.01
CSV	2.59	3.39	4.89

Table 4.16: Unblinded limits on the SM Higgs boson associated production with a Z boson cross section times the $b\bar{b}$ branching ratio, for both JP and CSV tagger. The limits are normalized to the SM predictions.

Unblinded			
	Median expected value (with syst.)	$+1\sigma$ band	Observed value
JP	4.82	6.64	4.89
CSV	3.73	5.33	5.46

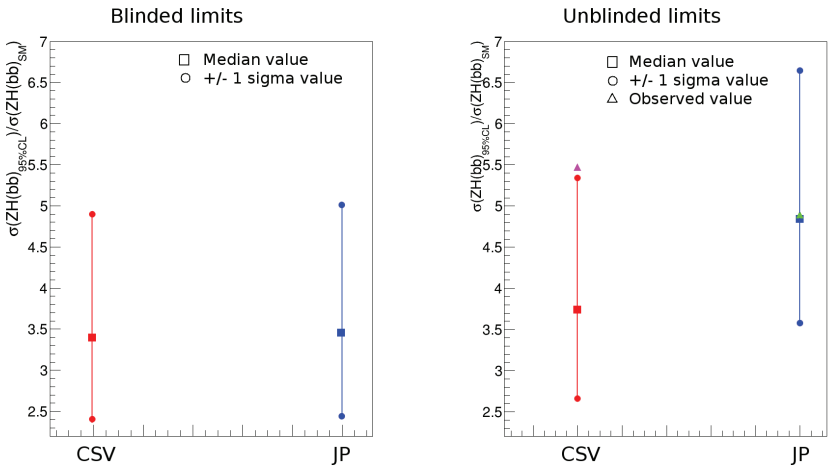

Figure 4.33: Blinded and unblinded limits on the SM Higgs boson associated production with a Z boson cross section times the $b\bar{b}$ branching ratio, for both JP (blue) and CSV (red) tagger. The median expected values are indicated by the squared markers, while the $\pm\sigma$ expected values are indicated by the round markers. The observed limits are displayed using triangle markers.

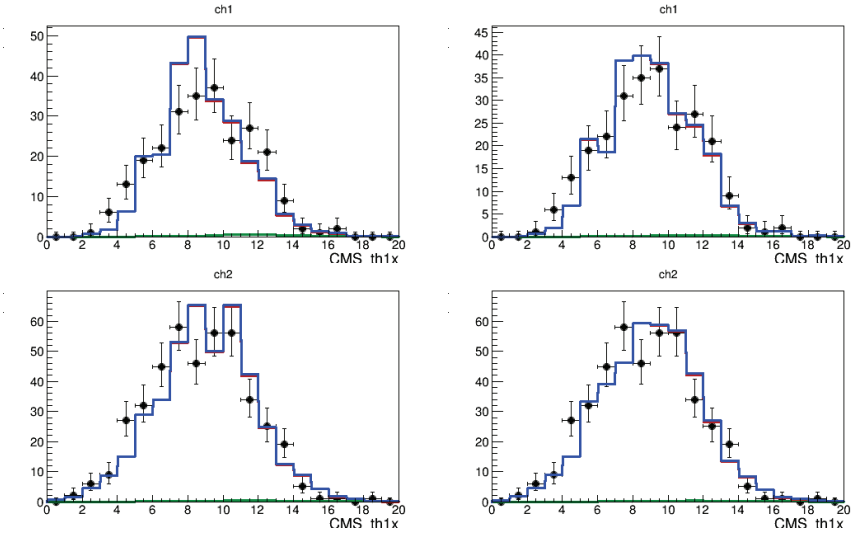

Figure 4.34: Distribution of the final discriminant before (left) and after (right) fit, for the electron channel in the 2-jets (top) and 3-jets category, using the JP tagger.

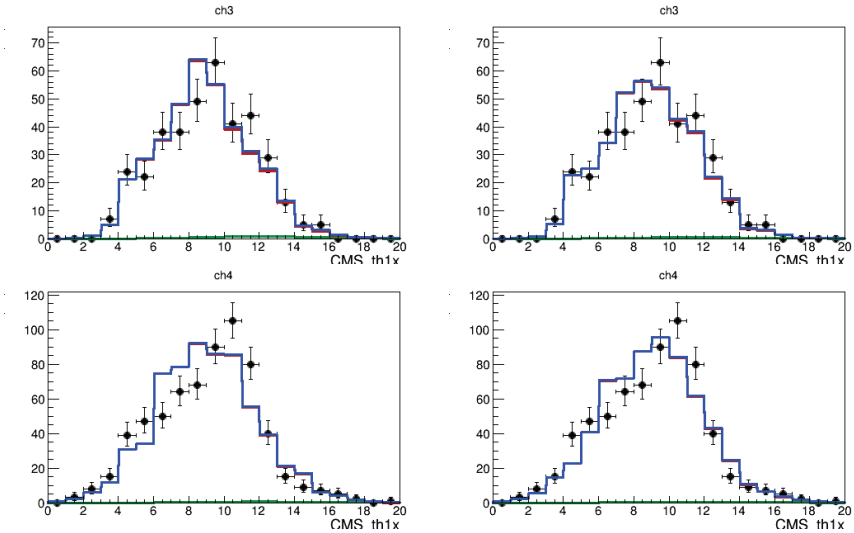

Figure 4.35: Distribution of the final discriminant before (left) and after (right) fit, for the muon channel in the 2-jets (top) and 3-jets category, using the JP tagger.

4.7.3 Comparison with other results

These results can be compared to the ones found by similar analyses. First of all, the CMS VH analysis has measured an observed 95% CL exclusion limit on the ZH process (with m_H = 125 GeV) of 1.89$\times\sigma_{SM}$ [85], by combining the W($l\nu$,$\tau\nu$)H($b\bar{b}$), Z(ll)H($b\bar{b}$) and Z($\nu\nu$)H($b\bar{b}$) channels. This corresponds to a small excess, as it can be seen on Fig. 4.36. It should be noticed here that most of the sensitivity is coming from the W($l\nu$,$\tau\nu$)H($b\bar{b}$) and Z($\nu\nu$)H($b\bar{b}$) channels, and that the limits are obtained combining the searches performed at 7 and 8 TeV.

The previous ME-based ZH analysis [64], using the CSV tagger and the combination of the 7 and 8 TeV data (before reprocessing), has observed a 95% CL exclusion limit on the ZH process of 1.6$\times\sigma_{SM}$, corresponding to a deficit of data events in the signal region. This result is shown on Fig. 4.37. The expected limits are better than the ones computed in this chapter, and this degradation can easily be explained by the clear lack of statistics the BDT training suffers from.

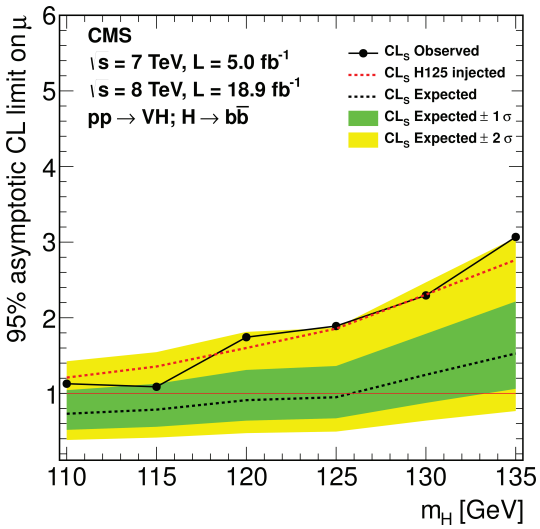

Figure 4.36: The expected and observed 95% CL upper limits on the product of the VH production cross section times the $H \to b\bar{b}$ branching ratio, with respect to the expectations for the SM Higgs boson. The limits are obtained combining the results of the searches using the 2011 (7 TeV) and 2012 (8 TeV) data. The red dashed line represents the expected limit obtained from the sum of expected backgrounds and the SM Higgs boson signal with a mass of 125 GeV.

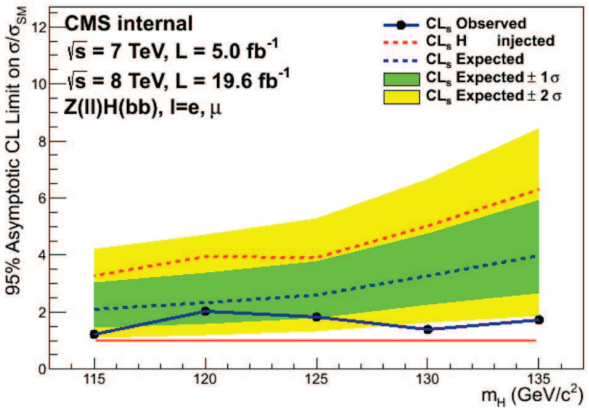

Figure 4.37: Expected and observed 95% CL upper limits on the ZH signal strength, with respect to the expectations for the standard model Higgs boson. The limits are obtained combining the results of the searches using the 2011 (7 TeV) and 2012 (8 TeV) data. The red dashed line represents the expected limit obtained from the sum of expected backgrounds and the SM Higgs boson signal with a mass of 125 GeV.

4.7.4 Impact of systematic uncertainties

The impact on the results of each systematic uncertainty is studied by removing them separately from the set of systematic uncertainties: each time, only the considered uncertainty is dropped. The results of this study are summarized in Table 4.17 (for JP) and Table 4.18 (for CSV). The main degradation on the expected limit comes from the limited statistics available for the main background, the Zbb process. The second most important source of degradation is the systematic associated to the background fit. One can also see that the JP selection is a little less sensitive to the systematics, due to the better rejection of Zbx and Zxx events, leading to a smaller impact of the MC statistics.

Table 4.17: Breakdown of the systematics on the final limits, for the **JP** tagger.

Systematic	Value	Degradation
No syst.	2.73	-
All syst.	3.52	28.9 %
-lumi.	3.52	-
-lepton SF	3.52	-
-ZH cross section + ZZ normalization	3.52	-
-JES	3.52	-
-JER	3.52	-
-b-tag. SF bc	3.48	0.9 %
-b-tag. SF light	3.52	-
-Background fit	3.39	3.7 %
MC statistics		
-All	3.20	9.8 %
-Zbb	3.27	7.7 %
-Zbx	3.52	-
-Zxx	3.48	0.9 %
-$t\bar{t}$	3.52	-
-ZZ	3.52	-
-ZH	3.52	-
MadWeight		
-JES	3.48	0.9 %
-JER	3.52	-

Table 4.18: Breakdown of the systematics on the final limits, for the **CSV** tagger.

Systematic	Value	Degradation
No syst.	2.59	-
All syst.	3.39	30.9 %
-lumi.	3.39	-
-lepton SF	3.39	-
-ZH cross section + ZZ normalization	3.39	-
-JES	3.39	-
-JER	3.39	-
-b-tag. SF bc	3.39	-
-b-tag. SF light	3.39	-
-Background fit	3.27	3.7 %
MC statistics		
-All	2.93	15.7 %
-Zbb	3.10	9.4 %
-Zbx	3.33	1.8 %
-Zxx	3.33	1.8 %
-$t\bar{t}$	3.39	-
-ZZ	3.39	-
-ZH	3.39	-
MadWeight		
-JES	3.39	- %
-JER	3.39	-

4.8 Conclusion

A search for the SM Higgs boson in the Z(ll)+H(bb) final state has been described, where two algorithms dedicated to the identification of b jets have been tested, in order to compare their performance. The method used is a multivariate analysis based on a Matrix Element Method, that produces a set of discriminating observables, the weights, sensitive to the Higgs boson signal. These quantities are then combined using a multivariate technique to produce final discriminators. The event selection has been optimized in order to improve the signal/background ratio, and events have been categorized according to the number of jets in the final state to include effects from final state radiation.

A significance data/MC discrepancy has been observed, most likely coming from a bad modeling of the mis-tagging rate. To correct for this effect, background normalization scale factors have been estimated from two-dimensional fits in a region of control.

The main systematic uncertainties affecting the estimation of the upper limit on the ZH production arise from the normalization of the backgrounds and the limited size of the Monte Carlo samples. These and other uncertainties have been estimated and taken into account in the evaluation of the limits. The blinded 95% CL expected limits for a SM Higgs boson mass of 125 GeV, based on the CL_s method, is $3.52\times\sigma_{SM}$ when using the JP tagger, and $3.39\times\sigma_{SM}$ when using the CSV tagger.

After unblinding, the observed limit obtained when using the CSV tagger is $5.46\times\sigma_{SM}$, while it is $4.89\times\sigma_{SM}$ when using the JP algorithm. For JP, the result is in agreement with the MC predictions while a small excess is observed for CSV. When comparing these results with the ones obtained by the previous ME-based analysis, a degradation of the expected limits is observed, coming from a significant lack of statistics this analysis suffers from.

As a conclusion, this analysis shows that the performance of the JP and CSV tagger are comparable: the JP selection is little bit less sensitive to the systematics while the CSV selection gives a better background rejection. This means that both taggers could be used for the model-independent search; however, since CSV gives the best performance, it will be used in the next chapter. For this analysis, many tools have been tuned and many ME weights have been computed: they will be directly re-used in the following chapter.

Chapter 5

Model Independent search of new physics phenomena with a Z boson and two b jets in the final state

The goal of the model-independent search is to design an analysis, using the MEM, to discriminate all the SM processes between them in order to categorize the $llbb$ phase space. To do so, a brand new method is applied, based on a recursive approach. It creates "boxes" enriched with a specific SM hypothesis at each step and as a result leads to a decomposition of the data through a tree defined from a purity criterion.

5.1 The Zbb final state

The $llbb$ topology has been defined in Chapter 4. This analysis focus this topology, restrained to Zbb final state: the two leptons are required to come from a Z boson. Several SM processes populate this phase space, among them the ones already presented in the previous chapter:

- The DY events, for which the same events categorization than to one used in the previous chapter can be done:
- "Zbb" events: the two b-tagged jets correspond to real b partons;

- "Zbx" events: one of the two b jets corresponds to a real b parton while the other selected jet matches a c or a light parton, and is therefore referenced as 'x';
- "Zxx" events: the two selected b jets have been mis-tagged.
- The top quark pair production, followed by a leptonic decay of the two W;
- The di-boson production where two Z bosons fake the signal signature: one decays into two leptons, the second into two b jets;
- The Higgs production in association with a Z boson, followed by the Higgs decay into two b quarks: the ZH process.

In addition to these events, new processes are taken into account for this study:

- The di-boson production WW, where the two leptons come from the leptonic decay of each W. The contribution of this process is very limited in this study, since the cut on the MET is maintained;
- The di-boson production WZ. In this case, the two leptons arise from the Z boson decay while the two b jets are two mis-tagged jets coming from the W boson decay. Similarly to the WW process, WZ events are not expected to contribute much.

Samples used

On top of the samples used in the previous chapter, the ones listed in Table 5.1 have been added for this analysis.

Table 5.1: Additional samples used in this analysis, with the corresponding cross section for simulated events. All samples are taken AODSIM format files.

Dataset	MC	$\sigma(Pb) \times BR$
WW	/WW_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53 V7A-v1/AODSIM	60.1 [86]
WZ	/WZ_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53 V7A-v1/AODSIM	24.6 [87]

5.1.1 Event selection and simulation

The event selection used aims at selecting SM processed having a Zbb final state. Extra jets are accepted but a cut on the MET is applied. This selection is based on the

one presented in Chapter 4, corresponding to the Full Region (FR), and is displayed in Table 5.2. The CSV tagger is used in this analysis.

Table 5.2: Selection criteria used in this analysis

Trigger	DoubleMuon/DoubleElectron
Leptons	$p_T > 20$ GeV $\|\eta\| < 2.4$ (2.5) for muons (electrons) Veto 1.442 $\|\eta\| < 1.566$ for electrons Isolation criteria for muons (electrons) using ΔR=0.4 (0.3): < 0.2 (0.15) $60 < M_{ll} < 120$
Jets	Leading jet: $p_T > 30$ GeV Sub-leading jet: $p_T > 30$ GeV Extra jet: $p_T > 20$ GeV $\|\eta\| < 2.4$ b-tagging: two CSV Medium tagged b jets
Missing transverse energy	$\vec{E}_t^{miss}$ significance < 10

All the selected events are renormalized according to the corresponding cross section; in addition, the correction SF presented in Chapter 4 (b-tagging, lepton and luminosity reweighting) are applied as well.

Finally, the scale factor found for the CSV tagger, using the background fit procedure explained in Chapter 4, are applied. A reminder of the SF values can be found in Table 5.3.

Table 5.3: Scale factors obtained by the 2D simultaneous fit, for the events selected with the CSV tagger.

SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
1.140 ± 0.06	1.348 ± 0.06	1.359 ± 0.14	1.006 ± 0.04

Event yields

The yields corresponding to the selection described in Table 5.2, using the SF shown in Table 5.3, can be found in Table 5.4.

Table 5.4: Data yields for the FR displayed in Table 5.2, compared with the expectation from the different main processes, normalized to the theoretical cross section. The data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}.

	Zbb	Zbx	Zxx	$t\bar{t}$	ZZ	ZH	WW	WZ	Tot. MC	Data	Data/MC
Z($\mu^+\mu^-$)	1930.0 $\pm$ 43.9	296.6 $\pm$17.2	507.0 $\pm$22.5	1348.4 $\pm$ 36.7	67.8 $\pm$ 8.2	12.6 $\pm$ 3.5	0.5 $\pm$ 0.7	7.8 $\pm$ 2.8	4170.7 $\pm$ 64.6	4214	1.01
Z(e^+e^-)	1415.5 $\pm$ 37.6	220.0 $\pm$14.8	345.6 $\pm$ 18.6	1011.6 $\pm$ 31.8	49.7 $\pm$ 7.0	9.6 $\pm$ 3.1	0.3 $\pm$ 0.5	4.5 $\pm$ 2.1	3056.8 $\pm$ 55.3	2880	0.94
Total	3345.5 $\pm$ 57.8	516.6 $\pm$22.7	852.6 $\pm$ 29.2	2360.0 $\pm$ 48.6	117.5 $\pm$10.8	22.2 $\pm$ 4.7	0.8 $\pm$ 0.9	12.3 $\pm$ 3.5	7227.5 $\pm$ 85.0	7094	0.98

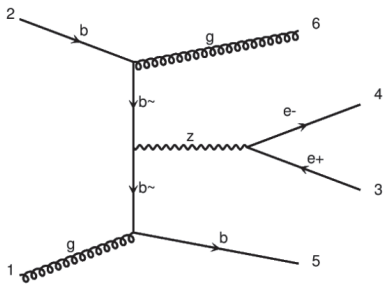

Figure 5.1: Zbj hypothesis associated to Zbx events, induced by the gluon-quark interaction.

5.1.2 Control plots and discriminating variable

The discrimination of the SM processes between them is done using the weights produced by the MEM. Most of these weights have already been computed for the analysis presented in Chapter 4 and for this study, additional ME weights have been calculated:

- Weights for the WW hypothesis;
- Weights for the DY events induced by the quark-gluon interaction, shown on Fig. 5.1, since the Zbx contribution is known to mainly come from this diagram. Therefore, Zbx events are associated to this new weight, in order to increase the discrimination power among the DY contributions. The Zbb and Zxx events are affected with the Zbb_{gg} weight.

The weights for these two new hypotheses are displayed on Fig. 5.2. Here it should be mentioned that for each event, a single weight can be assigned. Therefore, for the ZZ and ZH hypothesis, only the weights computed with the energy-momentum conservation (cor_0) are used. Control plots of the weights distributions are shown on Fig. 5.3 and Fig. 5.4.

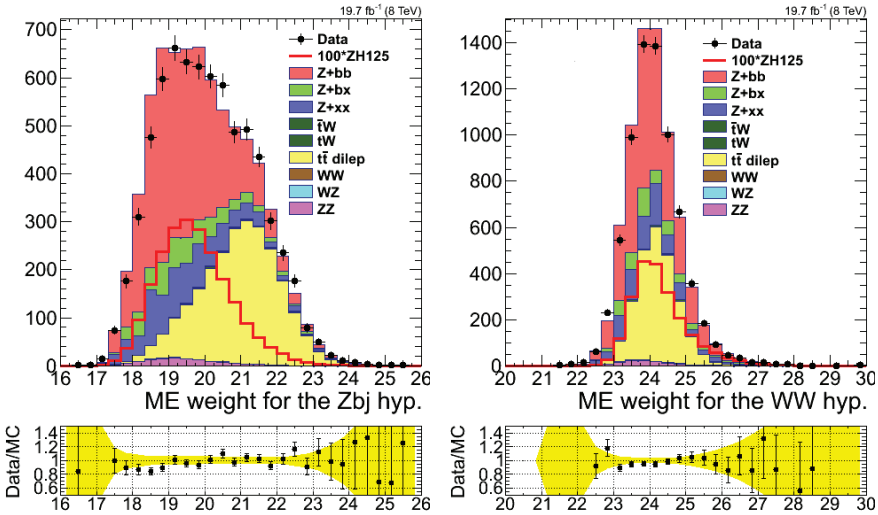

Figure 5.2: ME weights for the Zbx hypothesis (left) and for the WW hypothesis. Events have been renormalized according to their cross section, and the correction scale factor have been applied. The data sample corresponds to the 8 TeV sample ($\mathcal{L}$ = 19.7 fb^{-1}).

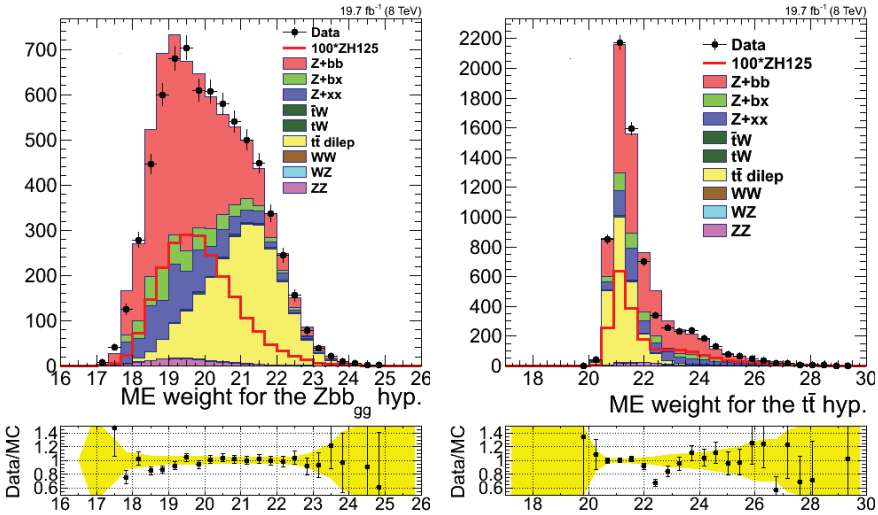

Figure 5.3: ME weights for the Zbb_{gg} hypothesis (left), for the $t\bar{t}$ hypothesis (right). Events have been renormalized according to their cross section, and the correction scale factor have been applied. The data sample corresponds to the 8 TeV sample ($\mathcal{L}$ = 19.7 fb^{-1}).

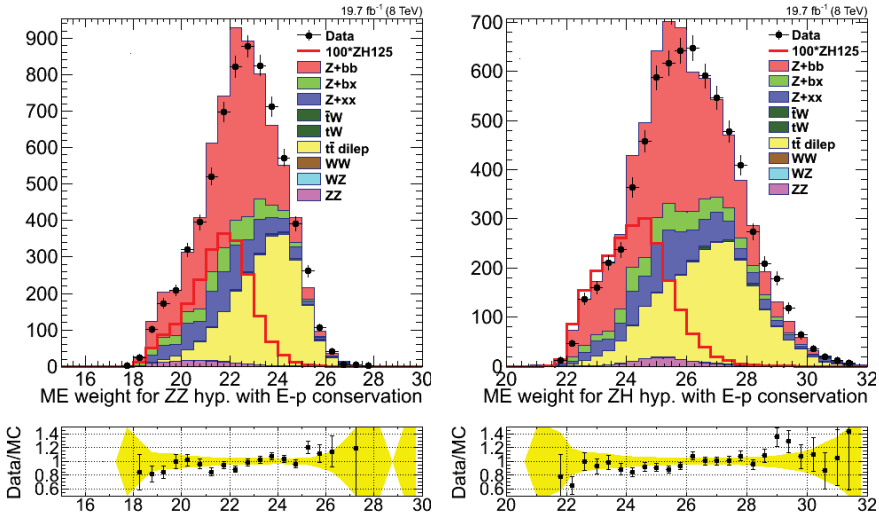

Figure 5.4: ME weights for the ZZ_{cor0} (left) and for the ZH_{cor0} (right) hypotheses. Events have been renormalized according to their cross section, and the correction scale factor have been applied. The data sample corresponds to the 8 TeV sample ($\mathcal{L}$ = 19.7 fb^{-1}).

Discriminating variable

From the MEM weights, it is possible to separate a process a from a process b, without the need of an MVA tool, by using a discriminating variable D defined as:

$$D = \frac{ArcTan(W_a - W_b) + \frac{\pi}{2}}{\pi} \tag{5.1}$$

This definition allows to have a final discriminant between 0 and 1. Examples of distributions of D are shown on ig. 5.5: to discriminate Zbb from $t\bar{t}$ events (left plot), and to separate the Zbb from the Zbx events (right plot).

The variable $D(Zbb, t\bar{t}$ shows a discrimination power as good as the NN use in Chapter 4 to separate the $t\bar{t}$ from the DY for the background fit procedure. The excess of data in the middle of the distribution corresponds to the zero results return by the MEM, more important in the data than in the MC. Events with a weight of 0 are indeed neither Zbb, neither $t\bar{t}$ like.
Unlike $D(Zbb, t\bar{t})$, the variable $D(Zbb, Zbx)$ does not show a real discrimination power between the Zbb and Zbx events.

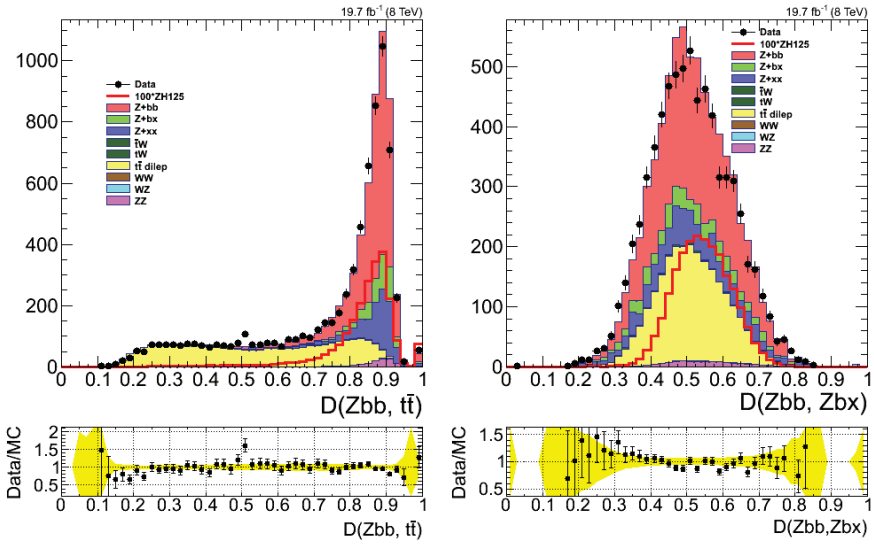

Figure 5.5: Distribution of $\frac{ArcTan(W_a - W_b) + \frac{\pi}{2}}{\pi}$, where a is the Zbb process and b the dileptonic $t\bar{t}$ process (left) or the Zbx process (right). Events have been renormalized according to their cross section, and the correction scale factors have been applied. The data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}.

5.2 Construction of a final discriminant

The method is based on a recursive separation of the SM processes, using the ME weights. Based on a purity criterion and the available MC statistics, the procedure will continue or stop. The procedure separates in priority the processes with the smallest yields.

5.2.1 Phase space decomposition

Let's take an example where three processes are known to populate the phase space of interest: a, b and c, ordered by yields. N_a, N_b and N_c represent the number of generated MC events respectively available for each process;

1. The procedure starts by looking the two processes with the smallest yields: a and b will then be discriminated first, in the "mother box"; besides, since a has the smallest yield, it will be referred as the "signal" process;

2. The distributions of the ME weights W_a and W_b corresponding to the two processes are renormalized to unity such that $n_a = a*N_a = 1$ and $n_b = b*N_b = 1$;

3. From these renormalized weight distributions, D (defined in Section 5.1) is built;

4. If a purity criterion is fulfilled, two "daughter boxes" are created, the "signal"-like daughter box containing $f_{signal}\%$ of the "signal". This purity criterion is defined as:

$$\frac{P_2 - P_1}{P_1} > cut_{Pur} \tag{5.2}$$

where P_1 and P_2 are the purities computed respectively in the mother containing n_{a1} events, and in the daughter "signal"-like box that contains n_{a2} events:

$$P_1 = \frac{n_{a1}}{n_{a1} + n_{b1}} \qquad and \qquad P_2 = \frac{n_{a2}}{n_{a2} + n_{b2}} \tag{5.3}$$

This step can be visualized on Fig. 5.6. Since in the mother box, a renormalization has been performed such as $n_a = n_b = 1$, this leads to a purity criterion of:

$$2 \times P_2 - 1 > cut_{Pur} \tag{5.4}$$

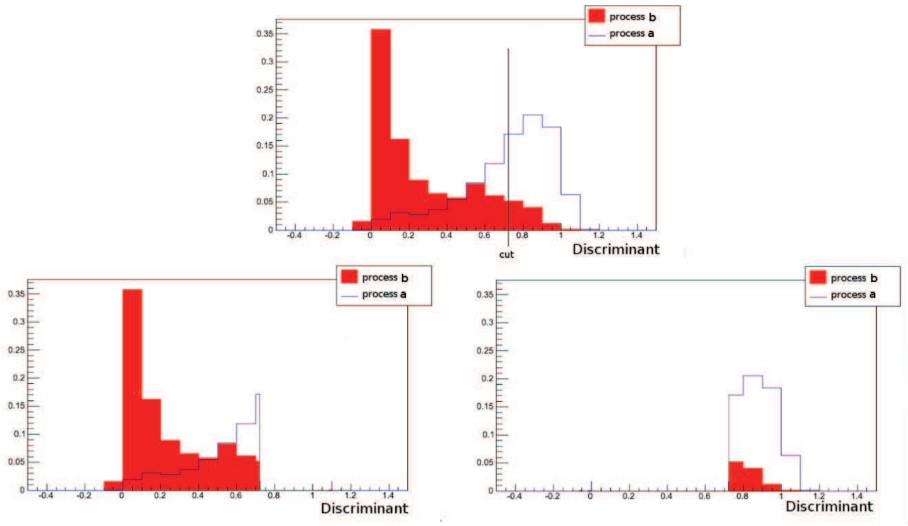

Figure 5.6: Creation of two daughter boxes (bottom), from a mother box (top) containing the process a and b, normalized to unity. The process a has the smallest yields and is therefore considered as the "signal", and therefore the bottom right box is called the daughter "signal"-like box. The cut applied separate a such that the daughter "signal"-like box contains $f_{signal}\%$ of the "signal". On this plot, f_{signal}=50%.

5. In order to prevent for the creation of daughter boxes when not enough MC statistics is available (and therefore induce an over-training of the method), an additional criterion is applied: N_a and N_b have to be greater than N_{MC} in mother and daughter boxes.

If the purity criterion is not fulfilled or if N_a and N_b are smaller than N_{MC}, the method will try to discriminate a and c, then b and c. If none of these combinations succeed, the procedure stops.

Therefore, the free parameters of this method that have to be optimized are:

- The fraction of signal f_{signal} going to the "signal"-like daughter box, when creating the daughter boxes;
- The improvement on the purity criterion required to allow the procedure to continue, cut_{Pur};
- The minimum number of generated MC events N_{MC} the mother and daughter boxes should contain.

These parameters are set to nominal values, leading to a reasonable number of final daughter boxes (47):

- f_{signal} = 50%;
- cut_{Pur} = 0.4;
- N_{MC} = 100.

A tuning of these parameters is performed in the following, taking as reference the 95% CL exclusion limit on the ZH signal strength.

Final tree

The method is applied, using the nominal values of the free parameters. The final can be seen in Fig. 5.7. It has been divided into two parts for a good visualization: the left part of the tree is shown on the top plot and the right part on the bottom plot. This tree starts by separating the ZH and the ZZ processes, as they have smallest yields. It leads to the creation of two daughter boxes, the red one being the "ZZ-box" while the blue one is the "ZH-box". In the blue box, the two processes considered for the discrimination are the ZH and the $t\bar{t}$. Since the ZZ process must still have a smaller yield than the $t\bar{t}$ process, it means that not enough discrimination was found between ZZ and ZH. The same thing occurs in the red box.

5.2.2 Additional event categorization

The boxes have been created using information from the ME weights; however, these weights do not take into account some interesting discriminating information, such as the b-tagging or the number of extra-jets in the events. An event categorization is then performed on all the final daughter boxes.

- **b-tagging:** once all the boxes have been produced, they are further sub-divided depending on the b purity. If the product of the two b jets CSV discriminator value passes a threshold value, the event is tagged as "b-like". The threshold value is arbitrarily chosen to equally populate the b-like and the light-like daughter boxes. If the MC statistics very limited, this value is set at 0.75. This separation could be further optimized by adding a fourth free parameter;

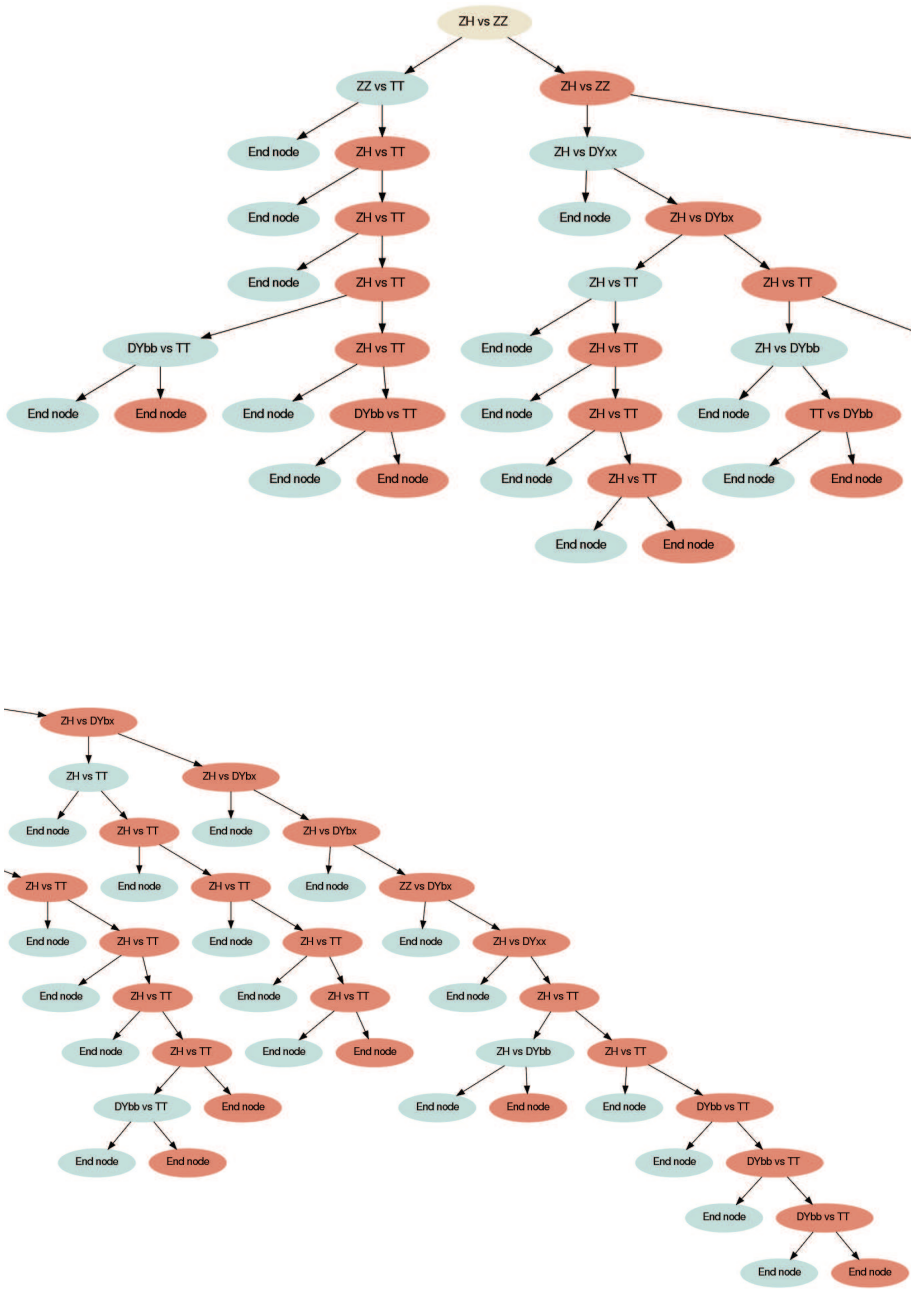

Figure 5.7: First (top) and second (bottom) part of the discrimination tree built with the method exposed in Section 5.2.1. The red boxes correspond to the daughter ”signal“-like boxes.

- **Extra jets:** after the b-purity-based categorization, events are sorted depending on the number of extra jets they contain: if no extra jet is found, the event goes in the 2-jets category. In the opposite case, it goes in the 3-jets category.

In the end, if the tree contains X final daughter boxes, there are now $N = X \times 4$ boxes. An histogram of N bins is created, merging the four histograms for the b/light-like events in the 2-jets/3-jets category. The bin i is filled with the expected yield in the i^{th} box; the distribution of this histogram is used as final discriminant. An example of such distribution, built from the final daughter of the tree represented on Fig. 5.7, can be seen on Fig. 5.8.

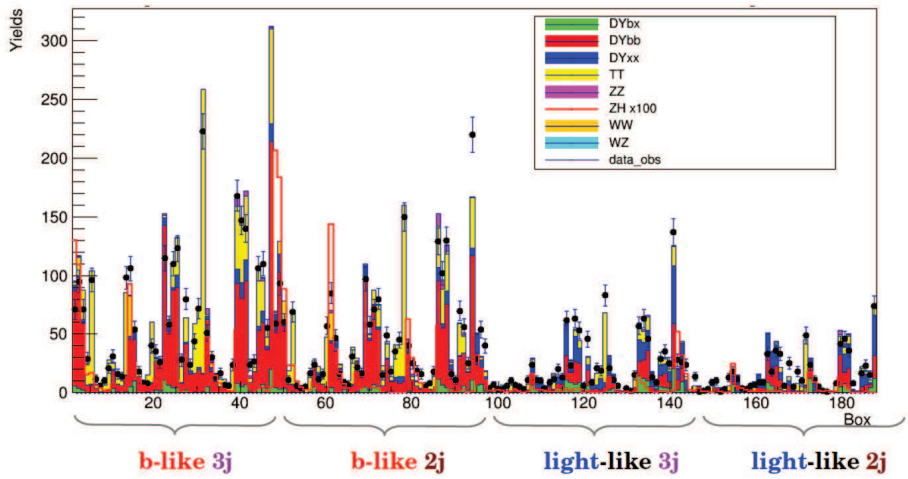

Figure 5.8: Distribution of the yields of each box produced by the tree shown on Fig. 5.7, further sub-divided to create four categories of boxes. Each MC process has been renormalized to the related expected cross section. The x axis represents the number of box: the bin i is filled with the expected yield in the i^{th} box. The data sample corresponds to the one recorded at 8 TeV, representing a luminosity of $\mathcal{L}$ = 19.7 fb^{-1}.

5.3 Sensitivity of the method

5.3.1 Template fit scale factors

A toy MC is used to test the sensitivity of the method: pseudo-data events are generated from the distribution of the total MC contributions, renormalized to their expected

yields. These events are generated such as for each bin of the distribution, the data point is within the statistical MC error. The experiment is repeated 1000 time and for each experiment, a fit is performed to evaluate the data/MC agreement. The scale factors (SF) for each process contribution are extracted. For the WW and WZ processes, the SF values are set at 1.

This method is tested on the tree shown on Fig. 5.7. The value of the χ^2 of the fit is displayed on Fig. 5.9. As the number of degrees of freedom equals 180, the quality of the fit is on average satisfactory. The SF distributions obtained for every SM process are shown on Fig. 5.10, Fig. 5.11 and Fig. 5.12, along with their mean and RMS values.
As expected, all mean values are found to be compatible with 1, and the RMS gives the statistical precision one can be expected when measuring each process: 3.7% for the Zbb, 18.8% for Zbx, 8.9% for Zxx, 3.1% for $t\bar{t}$, 53.7% for the ZZ and 256.5% for the ZH.

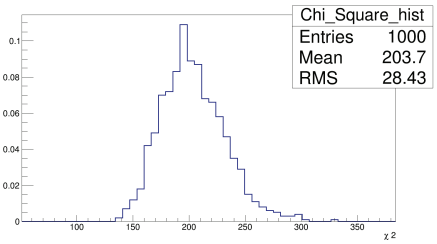

Figure 5.9: Value of the χ^2 obtained when fitting the pseudo-data distribution to the MC distribution. The mean and RMS of the distributions are displayed. The number of degrees of freedom of the system is 180, which means that on average, the quality of the fit is good.

5.3.2 Fit to data

The fit is now performed using the real data distribution. The obtained χ^2 is 435.9, corresponding to a value in the tail of the previous χ^2 distribution. This means that the observed disagreement is not compatible with the SM predictions. However, the fit has been performed without taking into account the systematic uncertainties, and therefore the agreement is worse than it should be.

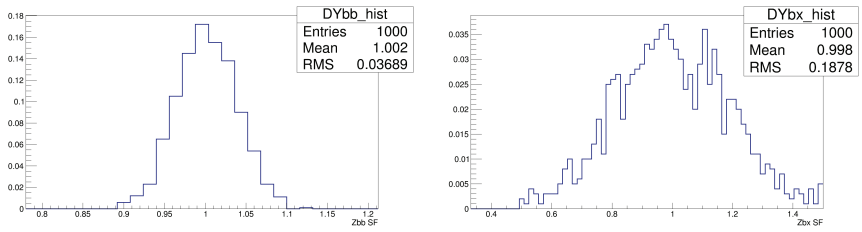

Figure 5.10: Distribution of the SF values extracted from the template fit for the Zbb (left) and Zbx (right) contributions, along with the mean and RMS values.

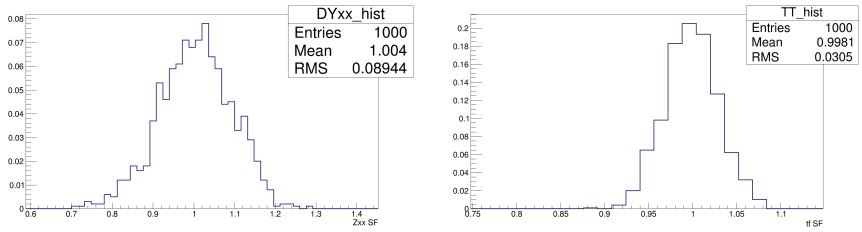

Figure 5.11: Distribution of the SF values extracted from the template fit for the Zxx (left) and $t\bar{t}$ (right) contributions, along with the mean and RMS values.

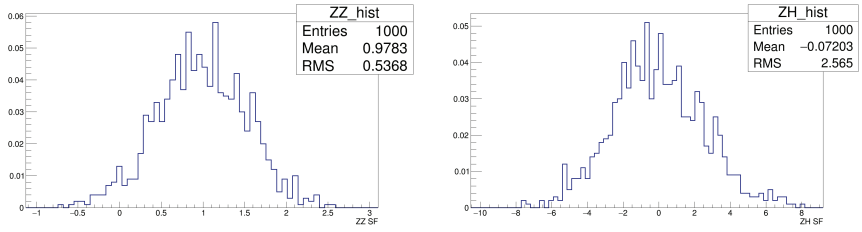

Figure 5.12: Distribution of the SF values extracted from the template fit for the ZZ (left) and ZH (right) contributions, along with the mean and RMS values.

New SF are extracted for each contribution; for the DY and $t\bar{t}$ processes, these SF have to be scaled to the ones applied before the fit (presented in Table 5.3). Thus, the final values are the following:

- Zbb: $0.90 \pm 0.04 \rightarrow \times 1.14 = 1.03 \pm 0.05$;
- Zbx: $1.48 \pm 0.76 \rightarrow \times 1.35 = 2.00 \pm 1.03$;
- Zxx: $0.76 \pm 0.09 \rightarrow \times 1.36 = 1.03 \pm 0.12$;
- $t\bar{t}$: $1.05 \pm 0.03 \rightarrow \times 1.01 = 1.06 \pm 0.03$;
- ZZ: 1.39 ± 0.51;
- ZH: 0.99 ± 2.37.

All these SF are compatible with 1. A precision of around 37% is measured for the ZZ process, and a small deficit of data in the ZH signal region is observed. In addition, as it can be seen on Fig. 5.13, none of these SF are highly correlated.

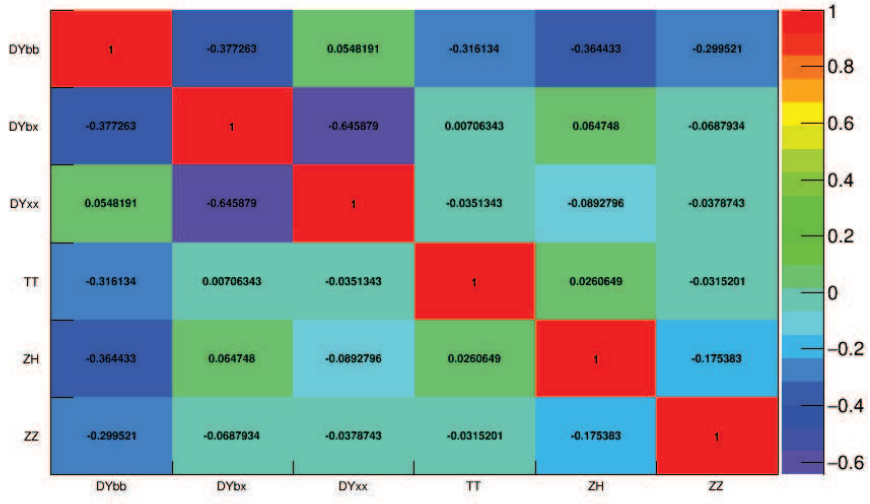

Figure 5.13: Correlation matrix obtained when extracting the SF of the different SF processes, using the real data distribution.

Therefore, when looking at the final discriminant distribution, no significant excess of data in several bins is seen. If when injecting new data (13 TeV data for instance), an excess appears in several bins (less than 15) with only one σ deviation, because of the bin multiplicity of this distribution nothing could be conclude. The distribution should be redone with another set of data, to see if the excesses are still here. If it is the case, the look-elsewhere effect can be considered as small and a study of these excesses should be foreseen. This is the goal of the model-independent search. However, the method first needs to be checked and to do so, signals have to be injected.

5.3.3 Exclusion limits for the ZH search

As previously mentioned in Section 5.2.1, the method possesses free parameters that can be tuned using different figures of merit. An interesting one to choose is the 95% CL_s upper limits on the ZH signal strength since the optimization of this parametrization will give the best discrimination between the smallest $llbb$ process, the ZH, and the other SM processes. An other advantage is the direct comparison of the obtained limit with the results from the previous Chapter. The expected limit is computed from histograms similar to the one shown on Fig. 5.8.

Therefore, different trees have been tested, changing each time the value of one parameter:

- Several values of cut_{Pur} are tested, going from 0.1 to 0.9 with a step of 0.1. Going to high values means drastically increase the purity at each step;
- The f_{signal}, fraction of "signal" events in the mother box going to the "signal"-like daughter box, is tested at 60 % and 70%;
- Two values of N_{MC}, the minimum of generated MC events mother and daughter boxes must contain, are tested: 75 and 50;
- The weights ZZ_{cor0} and ZH_{cor0} are replaced by the ZZ_{cor3} and ZH_{cor3} weights. Indeed, an arbitrary choice was made on the used correction for the ZZ and ZH hypothesis, since events can only be associated to one weight;
- The Zbx events are associated to the same MEM hypothesis than the Zbb and Zxx events, the Zbb hypothesis induced by gluon-gluon. This test is done to measure the improvement related to the new Zbx hypothesis.

The resulting limits can be seen in Table 5.5 for the tests performed on the cut_{Pur} parameter, Table 5.6 for ones on the N_{MC} parameter and Table 5.7 for tests with

different f_{signal} values. The limits computed using the ZZ_{cor3} and ZH_{cor3} weights is shown in Table 5.8, along with the limit obtained when the Zbx hypothesis has been removed.

The values are displayed for two scenarios: when the limit computation takes into account the systematics linked to the statistics (assumed to be the dominant one) and when it does not. This systematic is evaluated from the histogram used for the limit calculation: for each bin, the related number of generated MC event N is taken; the fluctuation "up" of a given bin is defined as $N + \sqrt{N}$ and the fluctuation down as $N - \sqrt{N}$.

Table 5.5: 95% CL_s upper exclusion limits obtained for a given value of cut_{Pur}, when a categorization of events has been made using the CSV product or both the CSV product and the number of extra jets in the event. The systematics related to the MC statistics have been added in the columns labeled "+ MC stat.". The values of the other free parameters are the nominal ones: f_{signal} = 50% and N_{MC} = 100.

cut_{Pur} value	# of boxes	Limit CSV	Limit CSV + MC stat.	Limit CSV+ # jets	Limit CSV+ # jets + MC stat.
0.1	90	3.61	3.80	3.23	3.42
0.2	72	4.23	4.51	3.89	4.23
0.3	73	4.02	4.36	3.77	4.20
0.4	47	4.23	4.61	4.02	4.52
0.5	40	4.27	4.61	4.05	4.52
0.6	28	4.33	4.73	4.11	4.66
0.7	18	4.61	5.02	4.39	4.92
0.8	11	4.92	5.30	4.64	5.14
0.9	4	5.30	5.70	5.02	5.55

Table 5.6: 95% CL_s upper exclusion limits obtained for a given value of N_{MC}, when a categorization of events has been made using the CSV product of both the CSV product and the number of extra jets in the event. The systematics related to the statistics MC have been added in the columns labeled with "+ MC stat.". The values of the other free parameters are the nominal ones: f_{signal} = 50% and cut_{Pur} = 0.4.

N_{MC} value	# of boxes	Limit CSV	Limit CSV + MC stat.	Limit CSV+ # jets	Limit CSV+ # jets + MC stat.
50	87	4.14	4.45	3.86	4.30
75	75	4.14	4.45	3.89	4.36

Table 5.7: 95% CL_s upper exclusion limits obtained for a given value of f_{signal}, when a categorization of events has been made using the CSV product of both the CSV product and the number of extra jets in the event. The systematics related to the statistics MC have been added in the columns labeled with "+ MC stat.". The values of the other free parameters are the nominal ones: N_{MC} = 100 and cut_{Pur} = 0.4.

f_{signal} value	# of boxes	Limit CSV	Limit CSV + MC stat.	Limit CSV+ # jets	Limit CSV+ # jets + MC stat.
60%	57	4.05	4.39	3.80	4.33
70%	46	4.20	4.61	3.95	4.48

Table 5.8: 95% CL_s upper exclusion limits obtained when the ME correction 3 is used instead of the correction 0, and when the Zbx hypothesis is not used. The categorization of events has been made using the CSV product of both the CSV product and the number of extra jets in the event. The systematics related to the statistics MC have been added in the columns labeled with "+ MC stat.". The values of the free parameters are the nominal ones: f_{signal} = 50%, N_{MC} = 100 and cut_{Pur} = 0.4.

	# of boxes	Limit CSV	Limit CSV + MC stat.	Limit CSV+ # jets	Limit CSV+ # jets + MC stat.
cor3	58	3.83	4.05	3.58	3.86
no Zbx	37	4.30	4.64	4.08	4.55

The best limit, including the systematic error linked to the statistics uncertainty, is found for the configuration where cut_{Pur}=0.1, f_{signal}=60% and N_{MC}=75, using the correction 3 for the ZZ and ZH hypothesis, and keep the Zbx weight. Besides, splitting the final boxes using the CSV product of the b jets and the number of extra jets in the events give better results. In the end, it seems that the more boxes there are, the best the limit gets. However, the limit obtained using the systematic related to the MC statistics should get worse at some point, when too many boxes do not have enough statistics.

For this configuration, the expected blinded limit obtained is 3.05 $\times\sigma_{SM}$, and taking into account all the systematics listed in Chapter 4 and recomputed for this analysis, it is 3.58 $\times\sigma_{SM}$. This limits can be compared to the one obtained with the dedicated search exposed in Chapter 4, using the CSV tagger, that equals 3.36 (with all the systematics included): they are similar. Besides, the last results present a smaller degradation due to the systematics than in the previous analysis (17.4% instead of 31%): this is induced by the smaller impact of the statistics in this analysis. Results are summarized in Table 5.9.

Table 5.9: Comparison of the limits obtained for the ZH search when using the dedicated analysis presented in Chapter 4, and the method exposed in this chapter, with and without the systematics.

Dedicated search (with syst.)	Model-independent search (no syst.)	Model-independent search (with syst.)
3.36	3.05	3.58

The final discriminant distributions for the four categories with the different MC contributions and the data, and for the sum of all the MC processes versus the data, can be seen on Fig. 5.14, for the b-like 2-jets category. Similar plots are available for the b-like 3-jets category (Fig. 5.15), the light-like 2-jets (Fig. 5.16) and 3-jets categories (Fig. 5.17).

The plots in the b-like categories present a reasonable agreement between data and MC and some bins show a good signal sensitivity. The data/MC agreement is worse in the light-like categories, as expected.

5.3.4 Exclusion limits for the ZA search

A 2-Higgs-Doublet Model (2HDM)[83] signal, characterized by the decay channel $H \rightarrow Z(ll)A(bb)$ and therefore called the "ZA" process, is a new physics contribution that enters the Zbb topology. A is a pseudo-scalar light Higgs with a mass at 142 GeV and H a heavy Higgs boson with a mass at 329 GeV (as it can be seen on Fig. 5.18); the cross section of this process, including the branching ratios, is 0.076 pb.

Using the same configuration as the one giving the best limit for the ZH search, in which the ZA process does not participate in the tree elaboration but only populate the final boxes, the upper blinded expected limit on the ZA signal strength obtained is 0.36 $\times\sigma_{BSM}$. Taking into account the systematics related to the MC statistics, it is 0.38 $\times\sigma_{BSM}$.
Then, if the systematics degradation is assumed to be the same than for the ZH case (17.4%), the final limit including all the systematics is 0.42 $\times\sigma_{BSM}$.

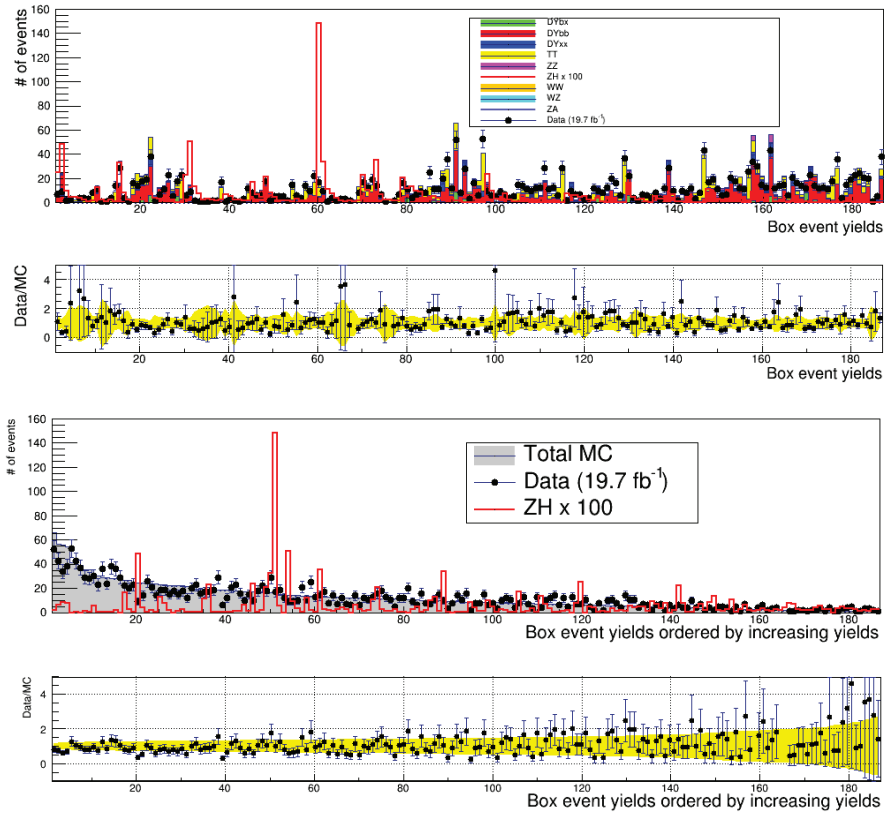

Figure 5.14: Distribution of the final discriminant used for the exclusion limit oh the ZH strength determination, for the b-like 2-jets boxes, for all the different processes and the data (top), and for the sum of all the MC processes versus the data ordered by expected event yields (bottom). The yellow bands represent the statistical uncertainty on the MC distributions.

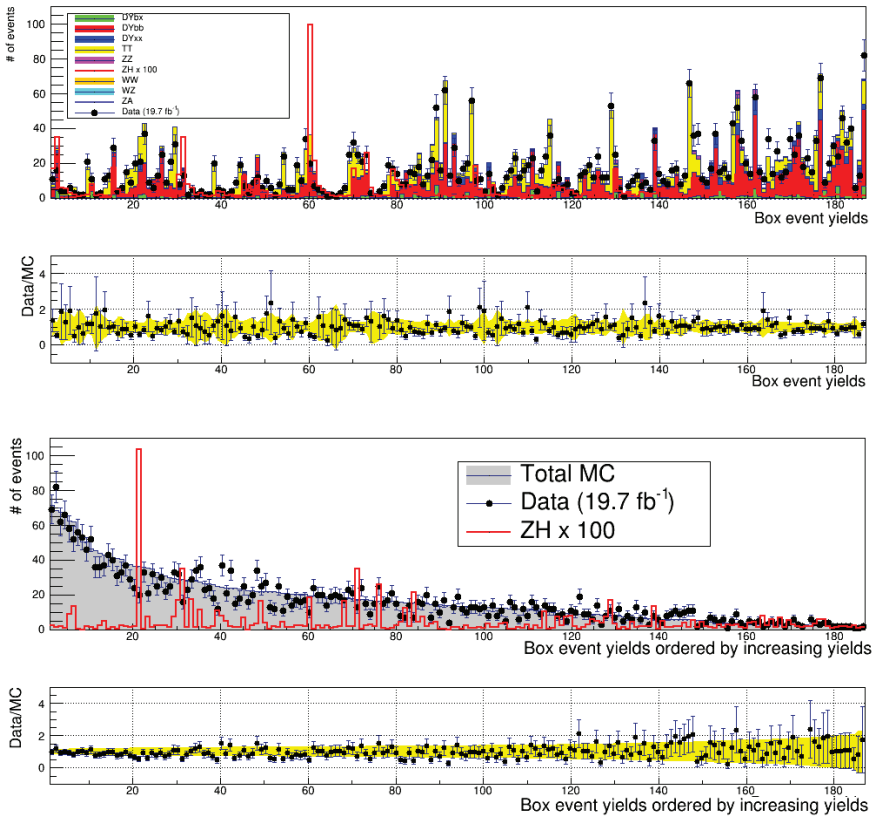

Figure 5.15: Distribution of the final discriminant used for the exclusion limit oh the ZH strength determination, for the b-like 3-jets boxes, for all the different processes and the data (top), and for the sum of all the MC processes versus the data ordered by expected event yields (bottom). The yellow bands represent the statistical uncertainty on the MC distributions.

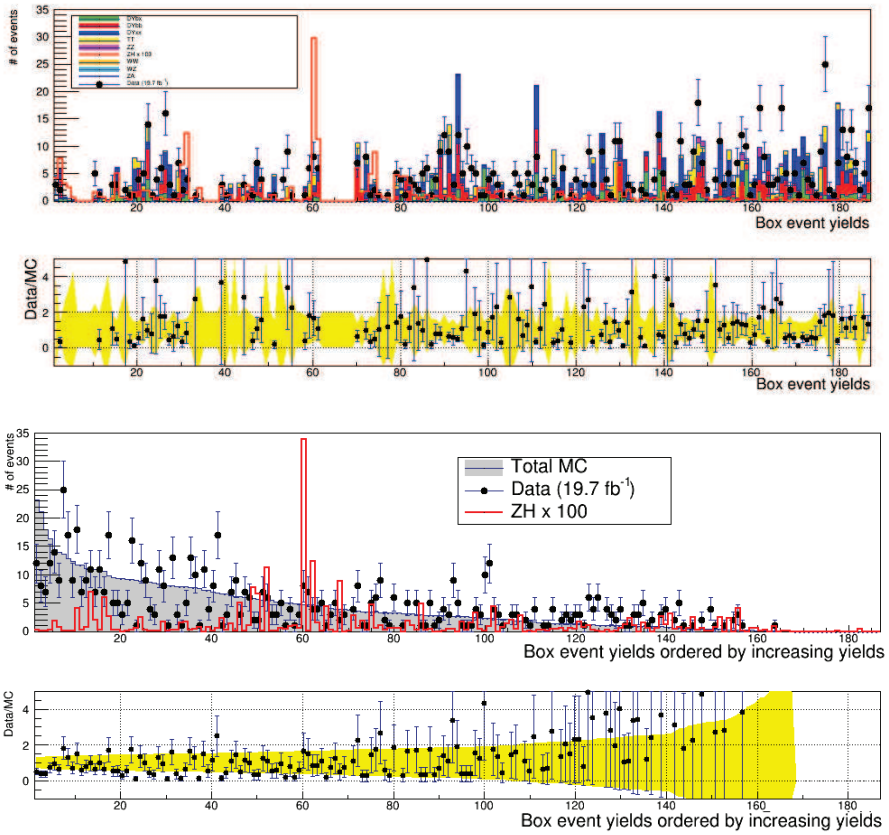

Figure 5.16: Distribution of the final discriminant used for the exclusion limit oh the ZH strength determination, for the light-like 2-jets boxes, for all the different processes and the data (top), and for the sum of all the MC processes versus the data ordered by expected event yields (bottom). The yellow bands represent the statistical uncertainty on the MC distributions.

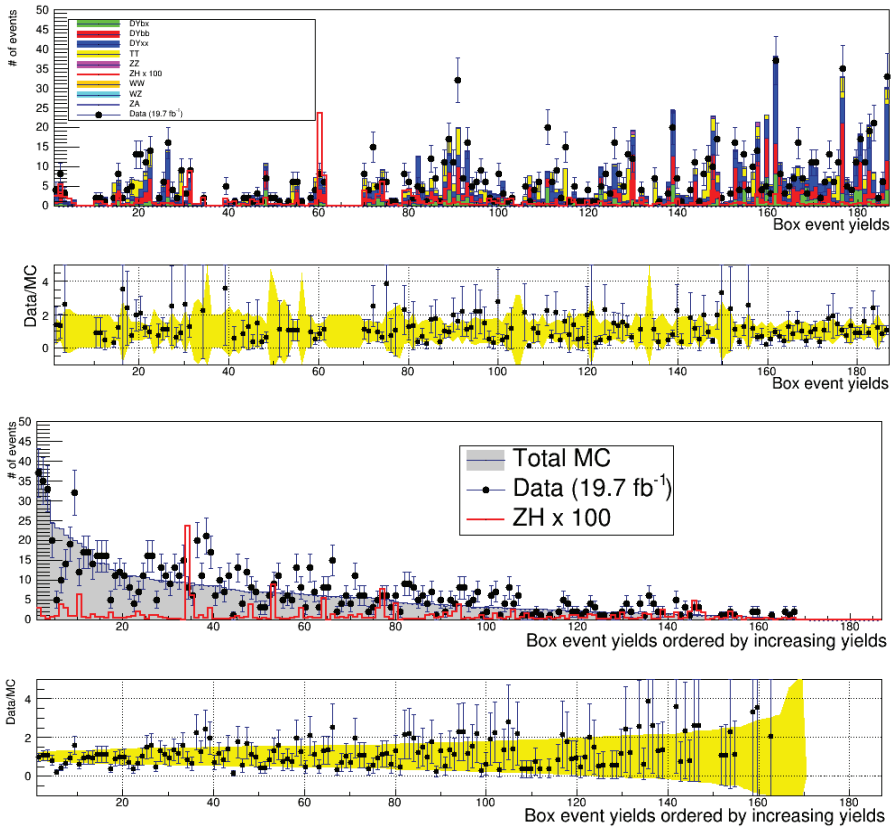

Figure 5.17: Distribution of the final discriminant used for the exclusion limit oh the ZH strength determination, for the light-like 3-jets boxes, for all the different processes and the data (top), and for the sum of all the MC processes versus the data ordered by expected event yields (bottom). The yellow bands represent the statistical uncertainty on the MC distributions.

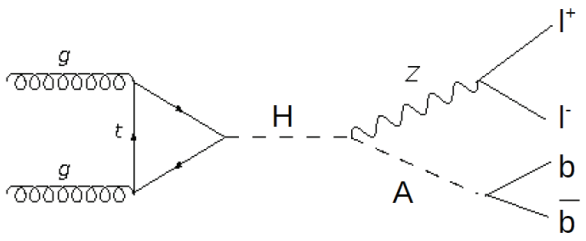

Figure 5.18

The limit obtained when using a dedicated search [83] is $0.83\times\sigma_{BSM}$: this means that even if the two results can not be exactly compared, this method returns a limit with the same order of magnitude than the one obtained using a dedicated search, highlighting the potential of the method. These results are summarized in Table 5.10.

Table 5.10: Comparison of the limits on the ZA signal strength when doing an exclusive search [83], and the method exposed in this chapter, using the parameters tuned for the ZH search.

Dedicated search	Model-independent search (no syst.)	Model-independent search (with syst.)
0.83	0.36	0.42

An interesting thing to look at is the distribution of the final discriminant when the bins have been ordered by $(\sqrt{S+B}-\sqrt{B}$, in order to observed the obtained significance between the ZH and ZA processes. Both processes have been renormalized to be compared. These plots are shown on Fig. 5.19 and Fig. 5.20, and it reveals the two processes do not similarly appear in the same bins: even if the ZA has not been used to build the tree and is very similar to the ZH process, the method allows to have some discrimination power for this process as well.

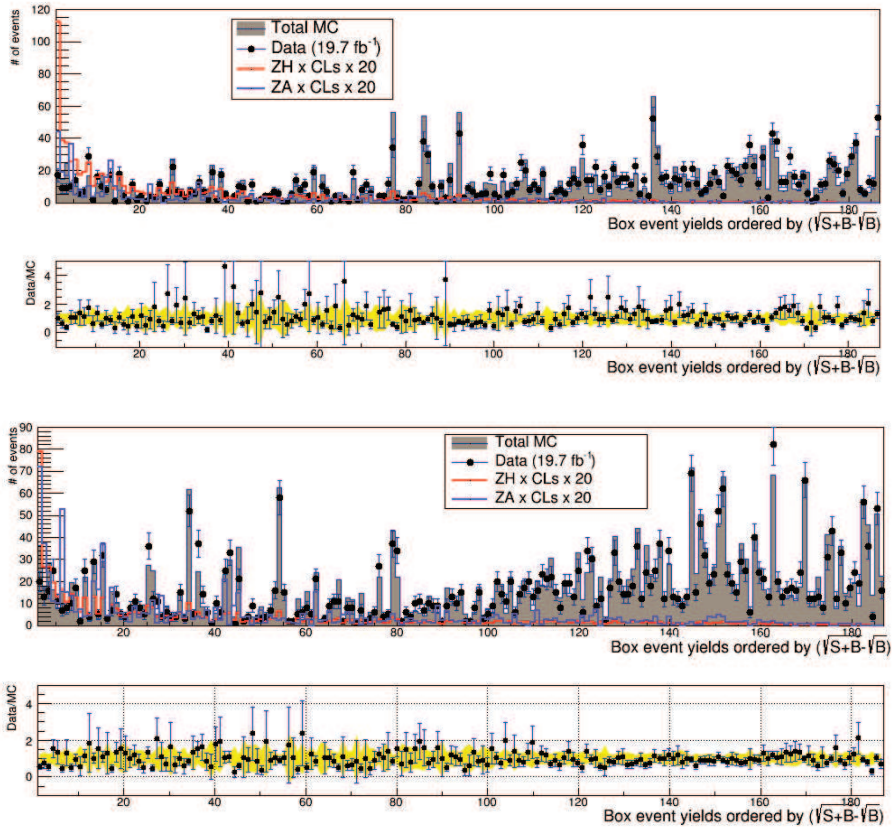

Figure 5.19: Distribution of the final discriminant used for the limit oh the ZH strength determination for the sum of the MC processes and the data, ordered by $(\sqrt{S+B}-\sqrt{B}$ for the b-like 2-jets boxes (top) and the b-like 3-jets boxes (bottom). The total MC distribution does not include the ZH and ZA processes. The yellow bands represent the statistical uncertainty on the MC distributions.

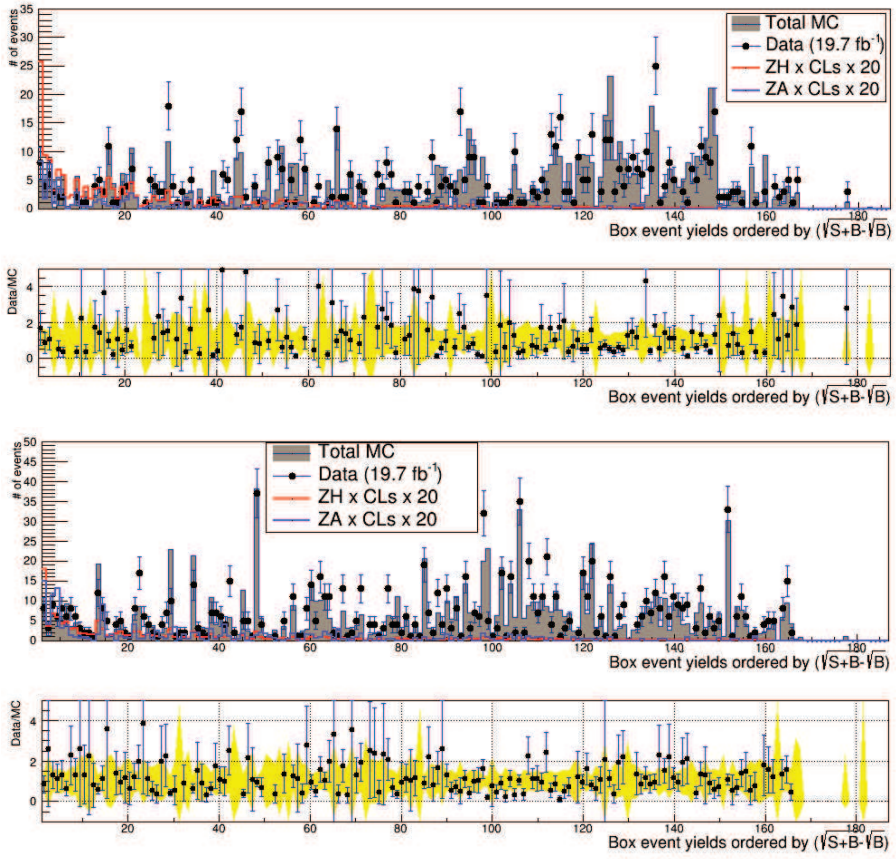

Figure 5.20: Distribution of the final discriminant used for the limit oh the ZH strength determination for the sum of the MC processes and the data, ordered by $(\sqrt{S+B}-\sqrt{B}$ for the light-like 2 jets boxes (top) and the light-like 3 jets boxes (bottom). The total MC distribution does not include the ZH and ZA processes. The yellow bands represent the statistical uncertainty on the MC distributions.

5.4 Conclusion

The goal of the method is to create boxes in which each SM process will be as background free as possible. It leads to the creation of several boxes, with very different background contributions: some boxes are highly populated by one specific background, making good control region boxes, and other boxes could be used to probe any discrepancy observed with respect to the SM predictions. As no input process is defined as a signal, it allows to probe the whole phase space without introducing any bias toward any model but the SM. Only a few free parameters define the tree returned by the method, used to build the final discriminant. These parameters can be tuned in order to find the best configuration to build the boxes.

So far, promising results have been produced, showing that the data/MC discrepancies observed are compatible with the SM predictions. Besides, when a signal is chosen, the obtained 95% CL exclusion limit on this signal is similar to the one computed using a dedicated search.

A lot of improvements can be achieved: for example, the N_{MC} threshold could be thought in a more flexible way to take into account the statistic limitation in the building of the tree, and more values of the free parameters could be probed.

An other important aspect is the phase space of study: here, a limitation is done to only study events with a Zbb final state. But by only removing the MET cut, a whole new phase space would be available for study, where the WW and WZ contributions would not be negligible anymore.

Conclusion

This thesis presents a detailed study of the Zbb final state with the Z boson decaying into two leptons, produced in the CMS detector at the LHC, after having collected an integrated luminosity of 19.7 fb^{-1} at a center-of-mass energy of 8 TeV. In order to tag this topology, specific di-lepton triggers have been used together with sophisticated b jet tagging techniques. Within this framework, this thesis presents the calibration and study at high energy of the so called Jet Probability tagger. This investigation is followed by the search for the associated production of the Higgs with a Z boson, using the Matrix Element Method and the Jet Probability tagger as well as the Combined Secondary Vertex tagger. Finally, the development of an analysis method aiming to develop a model-independent search of physics beyond the standard model is described.

Concerning the study of the Jet probability tagger, which has the advantage to be calibrated with the data, we demonstrated that this algorithm performs well up to 200 GeV. After this limit, a degradation of b-tagging efficiency is observed, caused by a loss of B tracks inside the b jet, in association with a high non B track contamination. In order to fix this issue, several methods have been tested. First, new high energy based categories have been created, leading to a slight improvement of the tagger performance. This study also gave the opportunity to perform an improvement of the framework used for the calibration. Another analysis was carried out in order to improve the B track purity inside the b jets entering the JP algorithm, using a Boosted Decision Tree. The first results are promising with a gain 4-7% of efficiency, paving the way for potential further investigations. In parallel to these developments, control plots of the b-tagging related variable have been produced for CMS, together with an improvement of the dedicated framework.

The study of the Zbb final state involves a state-of-the art analysis technique named the Matrix Element Method. In order to validate it, we have studied the possibility to observe the production of a Higgs boson in association with a Z boson. Although this search was already done in CMS, the present work involves a completely different approach than in the official CMS analysis and exploits the expertise and the performance of the aforementioned Jet Probability b-tagger. The Matrix Element Method is sophisticated and we have demonstrated its power: from the 4-vectors of the particles in final state, it produces an event-by-event discriminating variable, called a weight, that contains the maximal amount of theoretical information available from the hard process. In order to be able to access the kinematics of the partons in a given event, a transfer function giving the probability density to observe a specific detector information when the true kinematics of the event parton is given must be determined. The weights returned by the Matrix Element Method have been combined using a multivariate technique to produce final discriminator, used to compute an upper exclusion limit on the ZH production.
The event selection has been optimized to improve the signal sensitivity. Besides, two algorithms of b jet identification have been tested: the Jet Probability tagger and the Combined Secondary Vertex. A significant data/MC discrepancy has been observed, most likely coming from an incorrect modeling of the mis-tagging rate, mixed with NLO effects. To correct for these effects, scale factors have been estimated from two-dimensional fits in a control region to renormalize the background contributions. The main systematic uncertainties affecting the estimation of the limit arose from the normalization of the backgrounds and the limited size of the Monte Carlo samples. The final results give a blinded 95% expected C.L. limit on the ZH signal strength of $3.52\times\sigma_{SM}$ when using the JP tagger, and for CSV this limit is $3.39\times\sigma_{SM}$. The observed limits are $5.46\times\sigma_{SM}$ when using the CSV tagger and $4.89\times\sigma_{SM}$ when using the JP algorithm: a small excess of data is observed only for CSV. The performance of the two algorithms have been found to be comparable, and both of them can be used in a search analysis.

Having demonstrated the power of the Matrix Element Method with the ZH search and mastering the b jet tagging techniques, we have finally developed an analysis method for a search of physics beyond the standard model in a model-independent way. The goal of this model-independent search was to design an analysis, using the Matrix Element weights already determined, to discriminate all the SM processes in order to categorize the Zbb phase space. The method is based on a recursive approach, creating decision trees from the prior knowledge of these weights. Free parameters needed to be tuned in order to find the best approach when building the tree, and this was done using, as a reference, the exclusion limits on the ZH process. For the best configuration, the blinded 95% expected C.L. limit on the cross section is

$3.58 \times \sigma_{SM}$, a limit comparable with the one found using the dedicated search of Chapter 4 ($3.36 \times \sigma_{SM}$). In order to confirm the power of the method, a new physics signal process, a specific final state ZA motivated by 2HDM models, was used as a signal. The blinded 95% expected C.L. limit on the cross section is found to be $0.42 \times \sigma_{BSM}$, which is again similar the one obtained using a specific search ($0.83 \times \sigma_{BSM}$).
This study is still a prospect since a lot of improvements can be achieved by, for instance, finding more elegant ways to take into account the statistics limitation in the building of the tree, or by probing more values of the free parameters inherent to the construction of the decision tree. However, promising results have already been produced. Another important preliminary conclusion of this study is that the agreement between standard model and the data is rather good in the phase space region where b quark jets are clearly identified.

Acronyms

H_T Hadronic Activity. 123

2HDM 2-Higgs-Doublet Model. 190

ALICE A Large Ion Collider Experiment. 38

APD Avalanche Photo Diodes. 47

ATLAS A Toroidal LHC ApparatuS. 38

BDT Boosted Decision Tree. 105, 155

BEH Brout Englert Higgs. 119

CERN Centre Européen pour la Recherche Nucléaire. 37

CKM Cabbibo-Kobayashi-Maskawa. 23

CL Confidence Levels. 161

CMS Compact Muon Solenoid. 38

CMSSW CMS SoftWare. 60, 123

CR Control Region. 129

CRAB Cms Remote Analysis Builder. 41, 82

CSC Cathode Strip Chamber. 56

CSV Combined Secondary Vertex. 73

CTEQ Coordinated Theoretical-Experimental Project on QCD. 28

DL Decay Length. 78

DT Drift Tube. 56

DY Drell-Yann. 120, 123

ECAL Electromagnetic CAlorimeter. 41, 47

FR Full Region. 128

FSR Final State Radiation. 31, 127

GS Gluon Splitting. 98

GSF Gaussian Sum Filter. 48

HCAL Hadronic CAlorimeter. 41, 51

HL-LHC High Luminosity Large Hadron Collider. 40

HLT High Level Trigger. 60

IP Impact Parameter. 73

ISR Initial State Radiation. 31

JBP Jet B-Probability. 74

JER Jet Energy Resolution. 54, 160

JES Jet Energy Scale. 160

JP Jet Probability. 73

LEP Large Electron Positron. 37

LHC Large Hadron Collider. 37

LHC*b* Large Hadron Collider beauty experiment. 38

LINAC LINear particle ACcelerator. 38

LO Leading Order. 30

LS Long Shutdown. 40

MC Monte Carlo. 41

MEM Matrix Element Method. 31, 150, 175

MET Missing Transverse Energy. 54

MVA Multi Variate Analysis. 31, 105

NLO Next to Leading Order. 30

NN Neural Network. 135

PaS Parton Shower. 19

PAT Physics Analysis Tools. 123

PDF Proton Density Function. 28, 88

PF Particle Flow. 55

PS Proton Synchrotron. 38

PU Pile-Up. 31

PV Primary Vertex. 73, 126

QCD Quantum ChromoDynamics. 18

QED Quantum ElectroDynamics. 17

RPC Resistive Plate Chamber. 56

SF Scale Factor. 87

SM Standard Model. 15, 37

SPS Super Proton Synchrotron. 38

SR Signal Region. 129

SSVHE Simple Secondary Vertex High Efficiency. 74

SSVHP Simple Secondary Vertex High Purity. 74

SV Secondary Vertex. 73

TCHE Track Counting High Efficiency. 74

TCHP Track Counting High Purity. 73

TF Transfer Function. 32

TMVA Toolkit for Multivariate Data Analysis. 271

VBF Vector Boson Fusion. 25

WP Working Point. 109

Appendix A

Transfer functions plots

A.1 Fitted transfer functions

In this section, plots related to the fitted electrons and jets transfer functions are shown. First, the cross check plots for the jets in central and backward/forward regions are shown on Fig. A.1. The fitted parameters are displayed in Table A.1 and Table A.2, respectively for the two η regions.

Jets

Table A.1: Parameters of the central jets ($0 < |\eta| < 1.6$) TF extracted by maximizing an unbinned likelihood fit for jets in the detector acceptance. A double Gaussian parametrization depending on E is chosen here, with $a_i = a_{i,0} + a_{i,1} \times E + a_{i,2} \times \sqrt{E}$.

	Independent term	E **term**	$\sqrt{E}$ **term**
a_1	$a_{10} = 1.05 \pm 0.01$	$a_{11} = -0.02 \pm 0.00$	$a_{12} = 0.00$
a_2	$a_{20} = 0.00$	$a_{21} = 0.04 \pm 0.00$	$a_{22} = 0.83 \pm 0.02$
a_3	$a_{30} = 0.00$	$a_{31} = 1.23 \times 10^{-3} \pm 3.00 \times 10^{-5}$	$a_{32} = 1.00 \times 10^{-3} \pm 2.00 \times 10^{-4}$
a_4	$a_{40} = 5.02 \pm 0.04$	$a_{41} = 0.00$	$a_{42} = 2.40 \pm 0.34$
a_5	$a_{50} = 0.00$	$a_{51} = 0.05 \pm 0.01$	$a_{52} = 0.93 \pm 0.03$

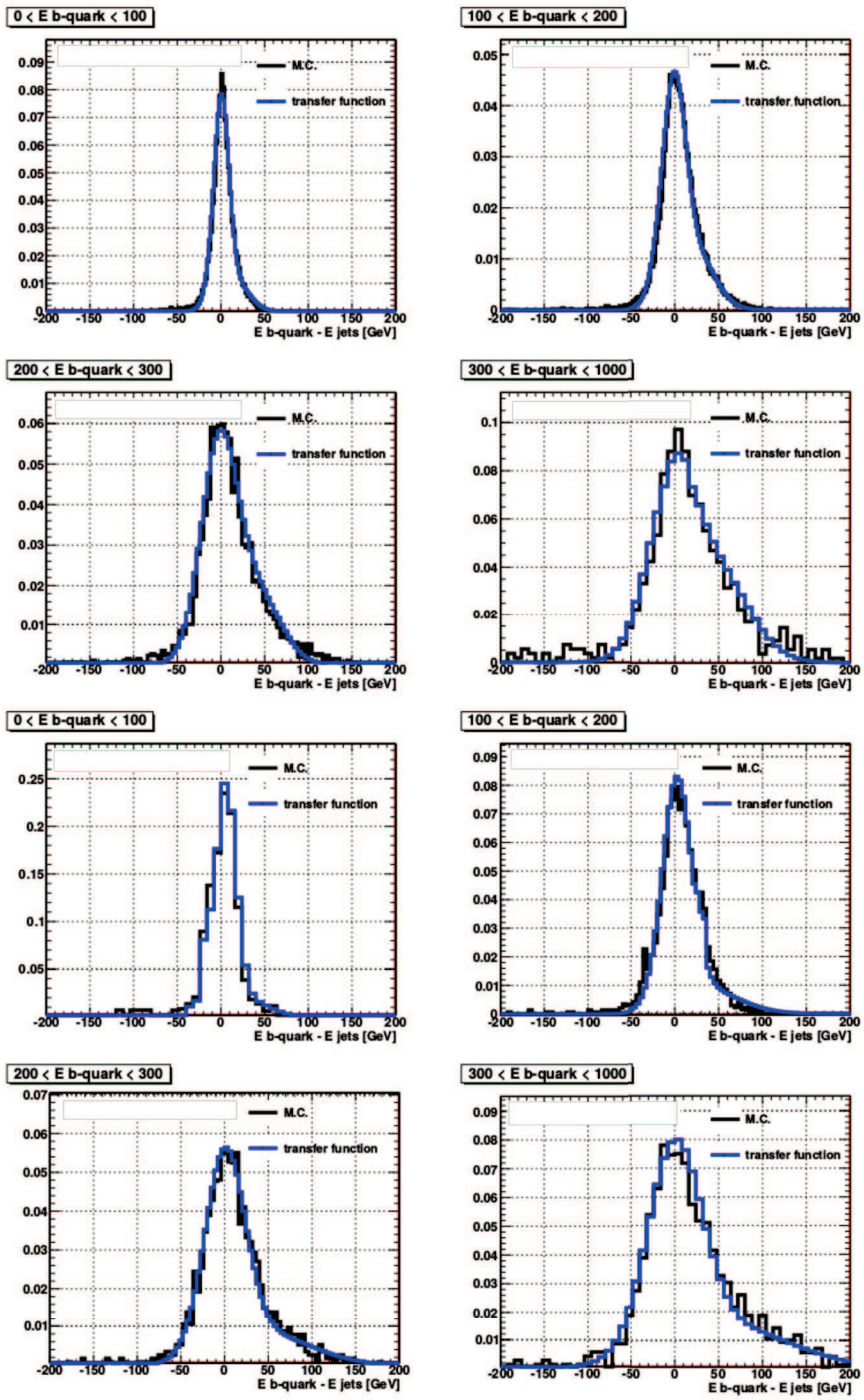

Figure A.1: Comparison of the TF obtained for jets in the central region (top) and in the forward-backward region (bottom), with the expected $\Delta E = E - E^{vis}$ distribution, for different b quark energy ranges.

Table A.2: Parameters of the forward/backward jets ($1.6 < |\eta| < 2.4$) TF extracted by maximizing an un-binned likelihood fit for jets in the detector acceptance. A double Gaussian parametrization depending on E is chosen here, with $a_i = a_{i,0} + a_{i,1} \times E + a_{i,2} \times \sqrt{E}$.

	Independent term	E **term**	$\sqrt{E}$ **term**
a_1	$a_{10} = 3.40 \pm 0.01$	$a_{11} = -0.02 \pm 0.00$	$a_{12} = 0.00$
a_2	$a_{20} = 0.00$	$a_{21} = 0.04 \pm 0.00$	$a_{22} = 0.88 \pm 0.03$
a_3	$a_{30} = 0.00$	$a_{31} = 3.90 \times 10^{-4} \pm 5.00 \times 10^{-5}$	$a_{32} = 3.00 \times 10^{-3} \pm 4.00 \times 10^{-4}$
a_4	$a_{40} = 5.01 \pm 0.05$	$a_{41} = 2.90 \pm 0.20$	$a_{42} = 0.00$
a_5	$a_{50} = 0.00$	$a_{51} = 0.15 \pm 0.03$	$a_{52} = 0.91 \pm 0.04$

Electrons

The cross check plots for the electrons in central and backward/forward regions are shown on Fig. A.2. The fitted parameters are displayed in Table A.3 and Table A.4, respectively for the two η regions.

Table A.3: Parameters of the central electrons ($0 < |\eta| < 1.5$) TF extracted by maximizing an un-binned likelihood fit for electrons in the detector acceptance. A double Gaussian parametrization depending on the energy is chosen here, with $a_i = a_{i,0} + a_{i,1} \times E + a_{i,2} \times \sqrt{E}$.

	Independent term	E **term**	$\sqrt{E}$ **term**
a_1	$a_{10} = -0.20 \pm 0.04$	$a_{11} = 1.05 \times 10^{-3} \pm 4.00 \times 10^{-5}$	$a_{12} = 0.00$
a_2	$a_{20} = 0.30 \pm 0.02$	$a_{21} = 5.10 \times 10^{-4} \pm 2.00 \times 10^{-5}$	$a_{22} = 0.12 \pm 0.02$
a_3	$a_{30} = 0.00$	$a_{31} = 1.20 \times 10^{-3} \pm 3.00 \times 10^{-4}$	$a_{32} = 0.02 \pm 0.00$
a_4	$a_{40} = 2.00 \pm 0.34$	$a_{41} = 0.01 \pm 0.00$	$a_{42} = 0.01 \pm 0.00$
a_5	$a_{50} = 0.00$	$a_{51} = 0.03 \pm 0.01$	$a_{52} = 0.00$

A.2 Binned transfer functions

In this section, cross check plots related to the fitted electrons and muons transfer functions are shown, respectively on Fig. A.3 and Fig. A.4.

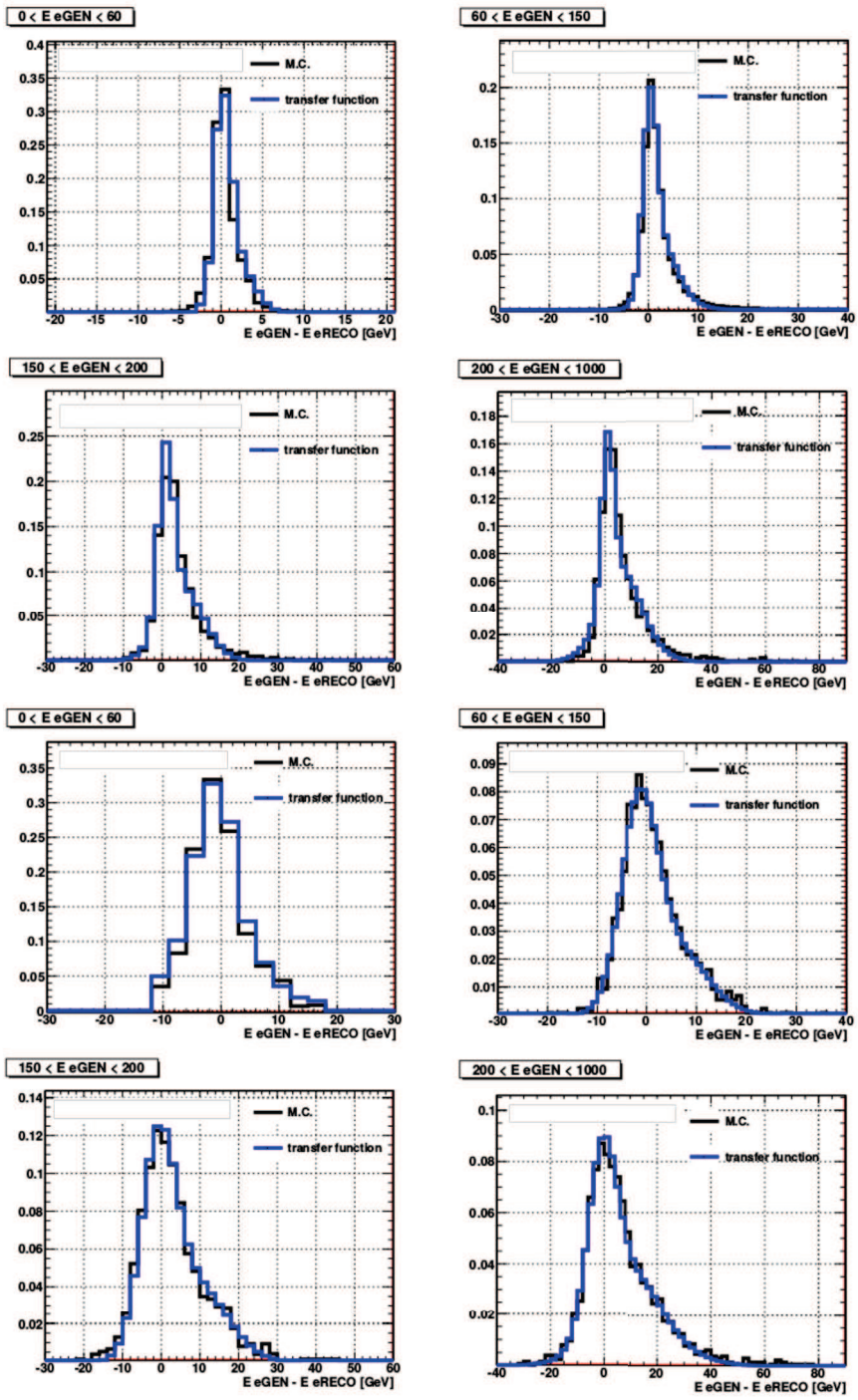

Figure A.2: Comparison of the TF obtained for electrons in the central region (top) and in the forward-backward regions (bottom), with the expected $\Delta E = E - E^{vis}$ distribution, for different generated electron energy ranges.

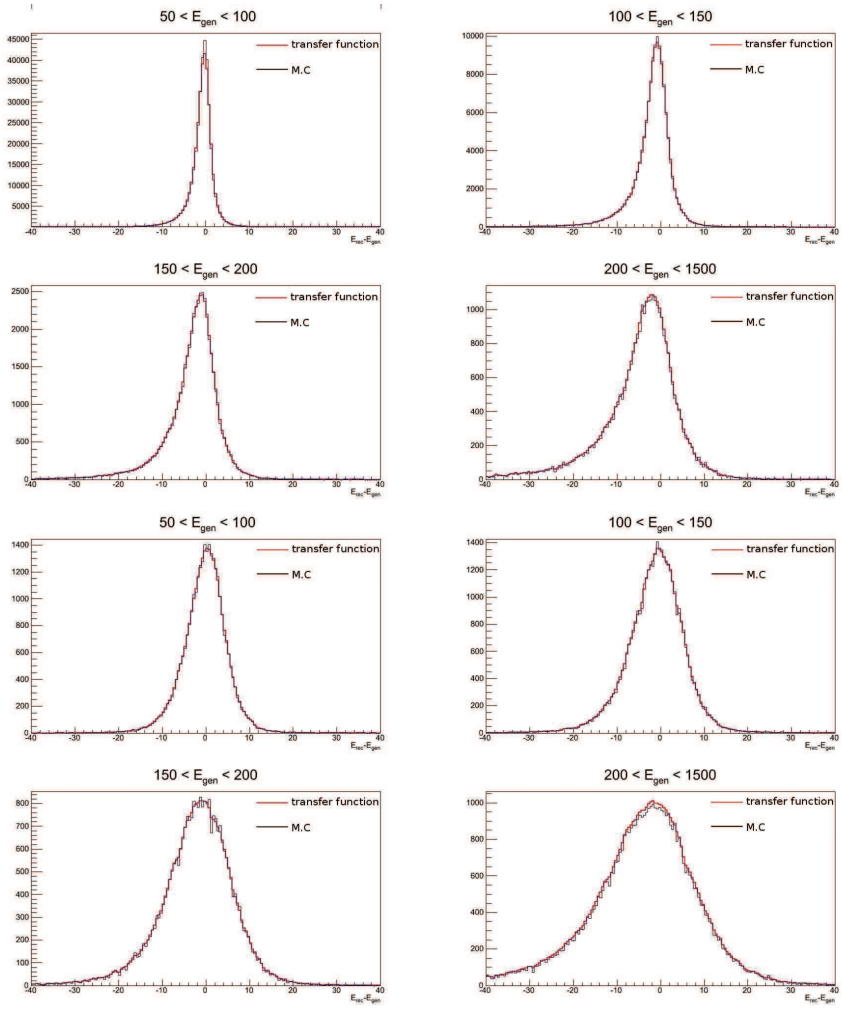

Figure A.3: Comparison of the TF obtained for electrons in the central region (top) and in the forward-backward region (bottom), with the expected $\Delta E = E - E^{vis}$ distribution, for different generated electron energy ranges.

Table A.4: Parameters of the forward/backward electrons ($1.5 < |\eta| < 2.4$) TF extracted by maximizing an un-binned likelihood fit for electrons in the detector acceptance. A double Gaussian parametrization depending on the energy is chosen here, with $a_i = a_{i,0} + a_{i,1} \times E + a_{i,2} \times \sqrt{E}$.

	Independent term	E **term**	$\sqrt{E}$ **term**
a_1	$a_{10} = -1.90 \pm 0.01$	$a_{11} = 1.10 \times 10^{-3} \pm 1.00 \times 10^{-4}$	$a_{12} = 0.00$
a_2	$a_{20} = 2.30 \pm 0.04$	$a_{21} = 0.01 \pm 0.00$	$a_{22} = 0.06 \pm 0.00$
a_3	$a_{30} = 0.00$	$a_{31} = 8.90 \times 10^{-4} \pm 5.00 \times 10^{-5}$	$a_{32} = 0.01 \pm 0.00$
a_4	$a_{40} = 5.01 \pm 0.50$	$a_{41} = 0.01 \pm 0.00$	$a_{42} = 0.11 \pm 0.02$
a_5	$a_{50} = 0.00$	$a_{51} = 0.03 \pm 0.00$	$a_{52} = 0.03 \pm 0.00$

A.3 Performance comparison

In this section, the plots showing the performance comparison between the parametrized fitted TF and the new binned TF are shown, for the jets (Fig. A.5) and for the electrons (Fig. A.6).

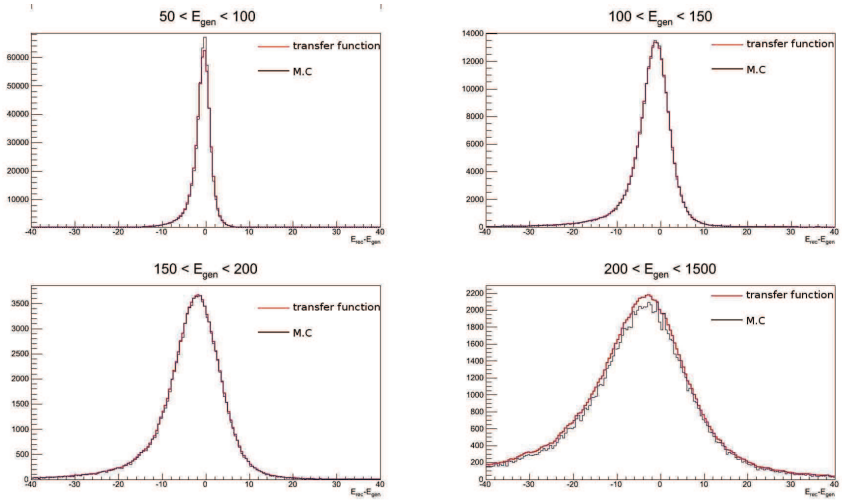

Figure A.4: Comparison of the TF obtained for muons, with the expected $\Delta E = E - E^{vis}$ distribution, for different generated muon energy ranges.

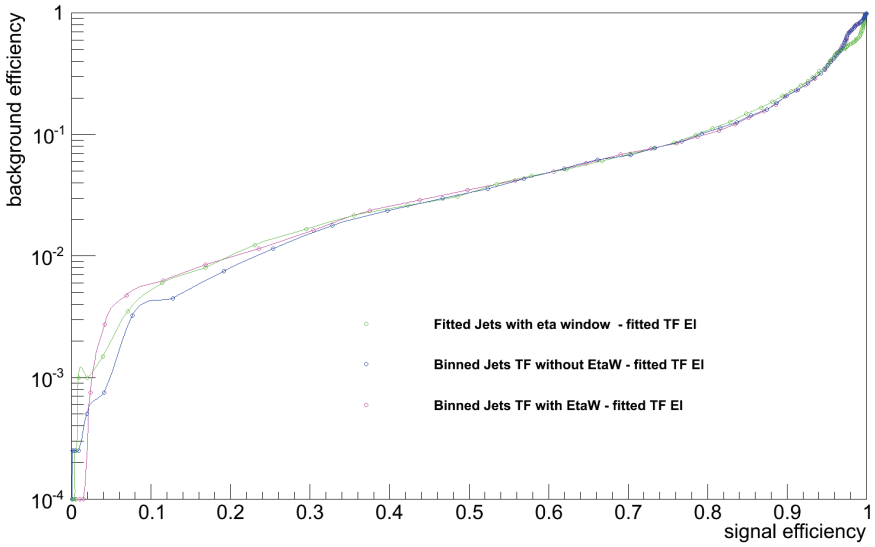

Figure A.5: Comparison of performance for ZH (signal) versus $t\bar{t}$ events, using parametrized fitted TF (green and pink curves) and binned TF (blue curve) for the jets.

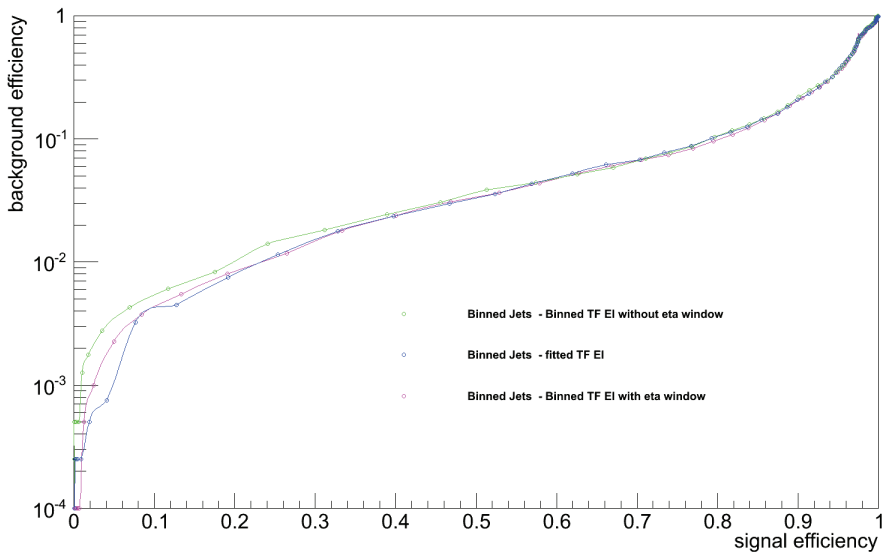

Figure A.6: Comparison of performance for ZH (signal) versus $t\bar{t}$ events, using parametrized fitted TF (green and pink curves) and binned TF (blue curve) for the electrons.

Appendix B

More plots about b-tagging in CMS

B.1 More discriminating variables

Several variables have been found to give a good discrimination between the real B tracks and the non real B tracks. In this Appendix, more are shown, such as the normalized χ^2 of the track (Fig. B.1), the number of hits in the SiStrip detector (Fig. B.2), the invariant mass between a track and its closest track (Fig. B.3), and the distance of the track to the z plane (Fig. B.4).

B.2 Use of a BDT for jets with $450 < p_T < 550$ GeV

This section presents the results of the study done on JP, using an additional BDT cut for the track selection (Section 3.4.2), for jets with $450 < p_T < 550$ GeV. The

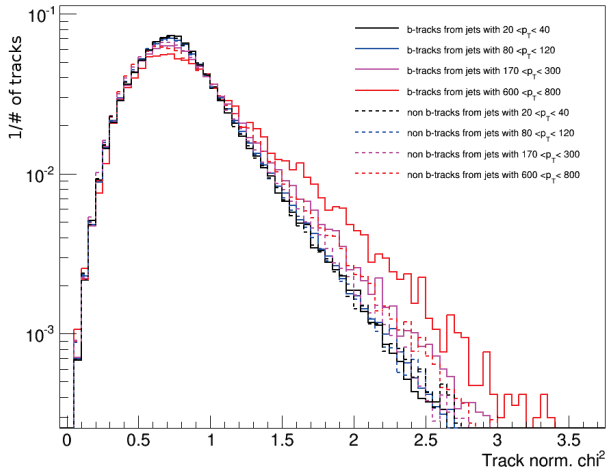

Figure B.1: Distribution the of the normalized χ^2, for real B tracks (plain lines) and for non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

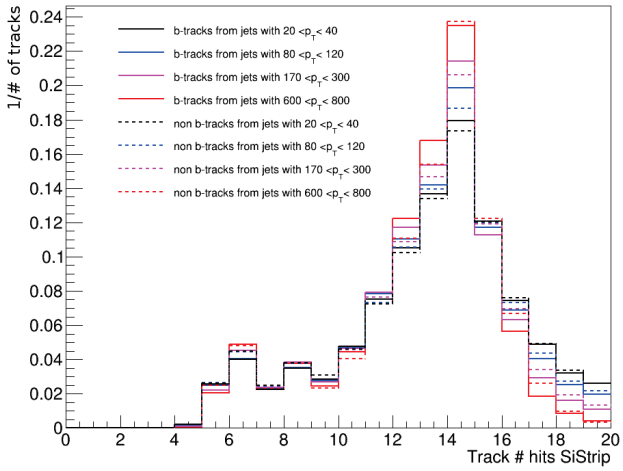

Figure B.2: Distribution of the number of hits in the SiStrip detector, for B tracks (full lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

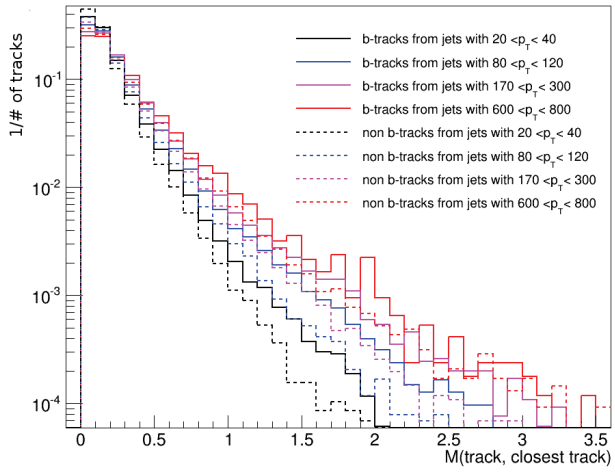

Figure B.3: Distribution of the invariant mass between a track and its closest track in the b jet, for B tracks (full lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

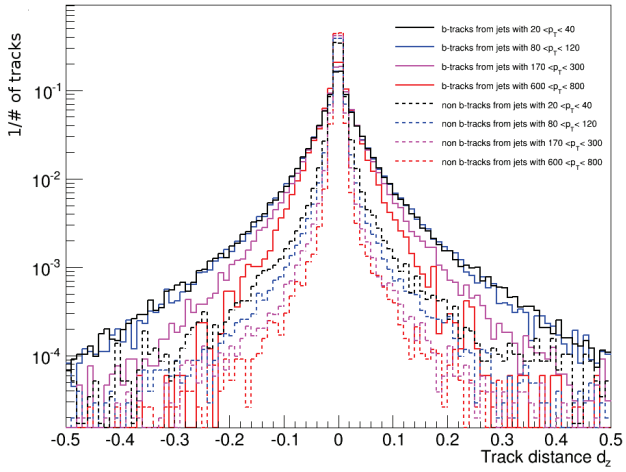

Figure B.4: Distribution of the d_z, for B tracks (full lines) and non B tracks (dashed lines), for different jet p_T ranges. The distributions have been renormalized to unity.

optimization of the BDT output cut can be found on Fig. B.5 for JP and on Fig. B.6 for JBP.
For JBP, a good gain of b-tagging efficiency can be achieved when applying a high BDT cut of 0.2 and the performance curves can be seen on Fig. B.7. All the results have been summarized in Tab. B.1.

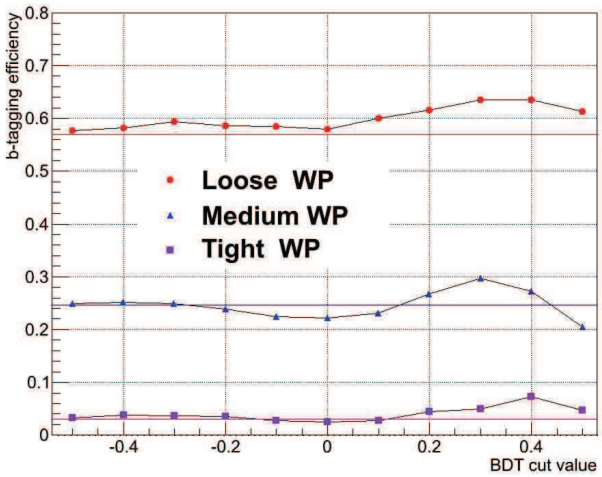

Figure B.5: b-tagging efficiency for JP as a function of the BDT cut value, for the three working points: Loose (red), Medium (blue) and Tight (purple), for jets with $450 < p_T < 550$ GeV.

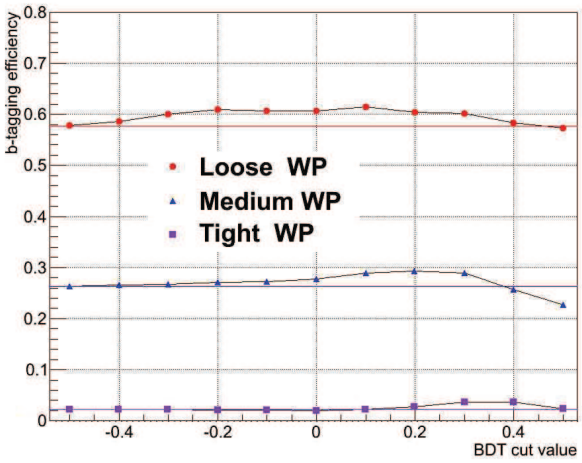

Figure B.6: b-tagging efficiency for JBP, as a function of BDT cut value, for the three working points: Loose (red), Medium (blue) and Tight (purple), for jets with 450 $< p_T <$ 550 GeV.

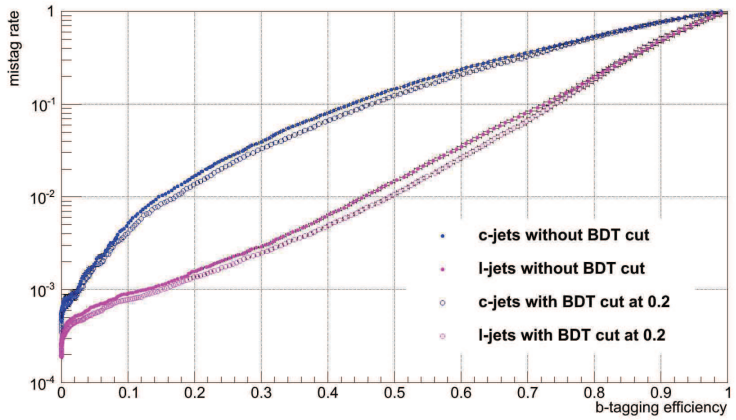

Figure B.7: JBP performance curves for c-jets (blue) and light-jets (pink), using selected tracks (plain markers) and selected tracks passing an additional BDT cut of 0.2 applied (empty markers).

Table B.1: Gain of b-tagging effciency (in %) for JP and JBP for the different WP, applying a specific BDT cut, using jets with $450 < p_T < 550$ GeV.

	JP			JBP		
BDT cut	**Loose WP**	**Medium WP**	**Tight WP**	**Loose WP**	**Medium WP**	**Tight WP**
-0.5	0.7	0.4	0.2	0.1	0	0
-0.4	1.3	0.5	0.8	0.9	0.3	0
-0.3	2.5	0.4	0.7	2.3	0.5	0
-0.2	1.7	-0.6	0.5	3.3	0.8	-0.2
-0.1	1.7	-2.2	-0.2	2.9	0.9	-0.2
0	1.1	-2.4	-0.4	2.9	1.4	-0.3
0.1	3.1	-1.6	-0.2	3.7	2.7	0
0.2	4.6	2.1	1.5	2.6	3	0.4
0.3	6.6	4.9	2	2.4	2.7	1.3
0.4	6.6	2.6	4.4	0.6	-0.5	1.4
0.5	4.3	-4.1	1.7	-0.4	-3.6	0.1

B.3 More plots for JP calibration using new categories

In this section, more plots concerning the new tested categories are shown. On Fig. B.8, the effect of the track's number of hits in the pixel detector is illustrated. New calibrations have been performed using additional categories in track decay length and in trach momentum, and their effect can be seen respectively on Fig. B.9 and Fig. B.10.

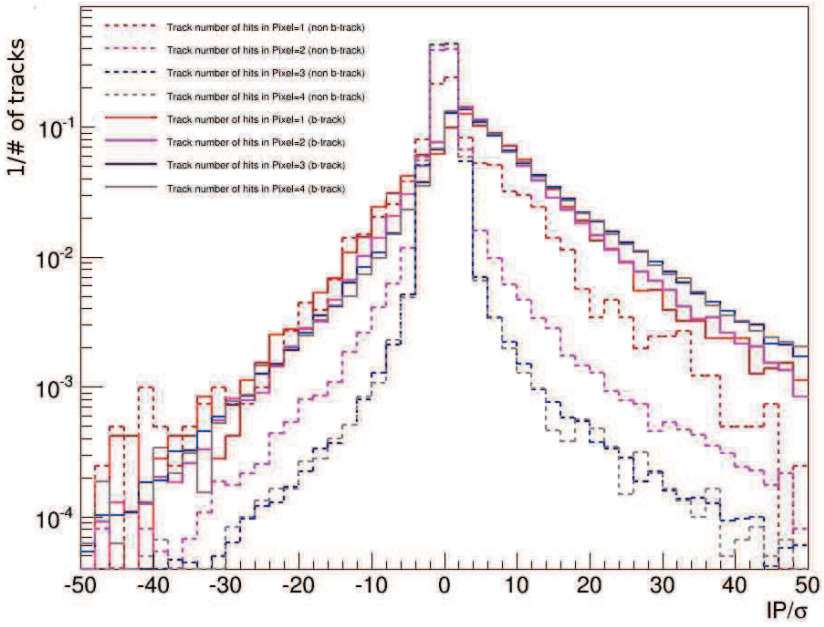

Figure B.8: Distribution of the IP/σ for tracks with different number of hits in the pixel detector, for B tracks (solid lines) and non B tracks (dashed lines).

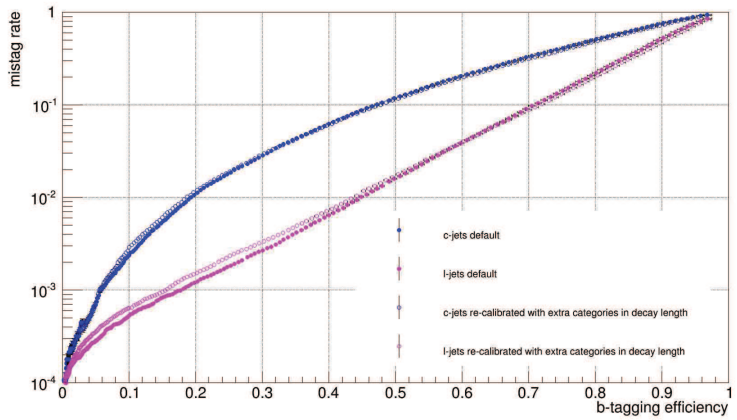

Figure B.9: JP performance curves for c-jets (blue) and light-jets (pink), using the nominal calibration (plain markers) and the new calibration with new ranges in track decay length (empty markers).

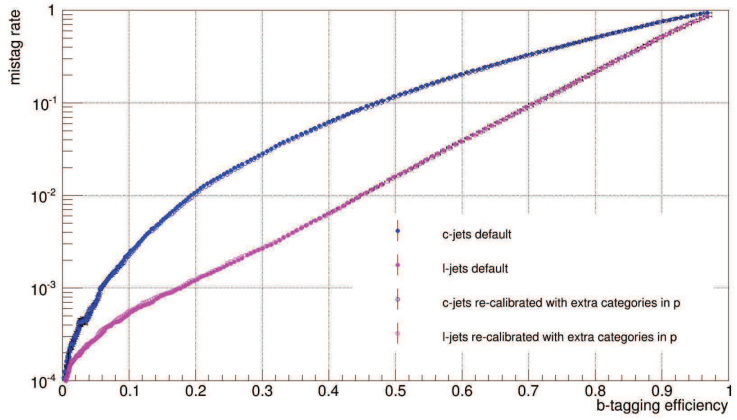

Figure B.10: JP performance curves for c-jets (blue) and light-jets (pink), using the nominal calibration (plain markers) and the new calibration with new ranges in track p (empty markers).

Appendix C

More plots for the comparison of the CSV and JP performance in the Z(ll)H(bb) final state using a Matrix Element Method

C.1 Merging of the DY samples

The different DY samples used in this analysis, generated with a given $p_{T(Z)}$ and H_T, are combined according to a reweighting procedure accounts for the different cross section and the effective number of events processed for each exclusive sample. First, the $p_{T(Z)}$ samples are combined by mean of a weight w_i^{pt}:

$$w_i^{pt} = \frac{\sigma_i}{\sigma_{incl}} \times \frac{N_{tot}^{pt}}{N_i^{pt}} \tag{C.1}$$

where σ_i is the absolute cross section for each processes, N_i^{pt} is the number of events genereated for the exclusive sample, σ_{incl} is the total cross section for the DY inclusive sample and N_{tot}^{pt} is the sum of events for the DY inclusive sample and all the $p_{T(Z)}$

samples. Then, the H_T sample are reweighted similarly but taking into account the previous $p_{T(Z)}$ reweighting:

$$w_j^{HT} = \frac{\sigma_j}{\sigma_{incl}} \times \frac{N_{tot}^{HT}}{N_j^{HT-pt}} \tag{C.2}$$

where $N_j^{HT-pt} = N_j^{HT} + \sum_i w_i^{pt} \times N_{ij}^{pt}$ accounts for the number of events from the i^{th} sample going into the j^{th} H_T bin and N_{tot}^{HT} is the total number of events from all the H_T samples. A final weight w_{ij} for each event extracted is extracted by computing the ratio of events falling simultaneously into the two dimensional i^{th} p_T-j^{th} H_T bin, and total number of generated events in that bin:

$$w_{ij} = \frac{(N_{ij}^{pt} \times w_i^{pt} + N_j^{HT}) \times w_j^{HT}}{N_{ij}^{gen}} \tag{C.3}$$

C.2 Control Region plots

In this section, the plots of the 3-jets category used for the background fit in the Control Region selection are shown: the product of the leading b jets JP discriminator values (Fig. C.1), the product of the leading b jets CSV discriminator values (Fig. C.2) and the NN output used to discriminate DY events from $t\bar{t}$ events (Fig. C.3).

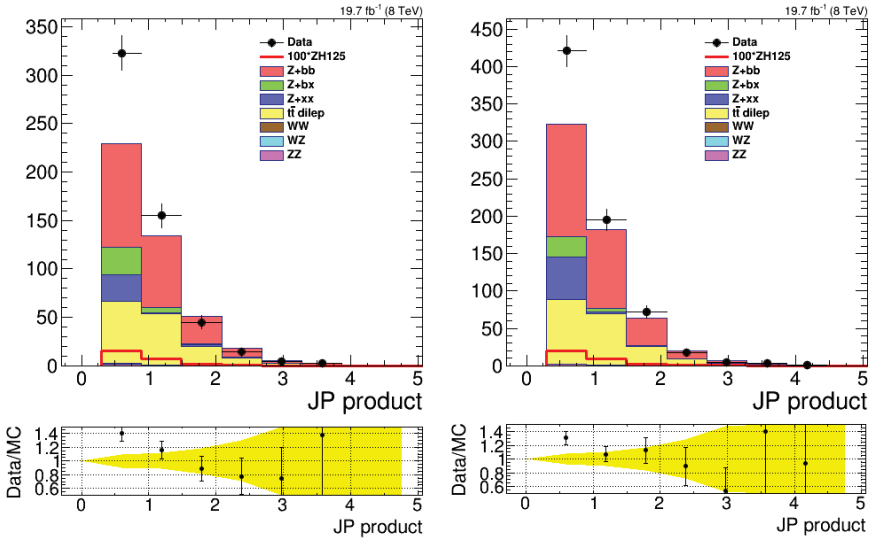

Figure C.1: Product of the two leading b jets JP discriminator values, used to discriminate b jets from light jets contributions, in the CR. The left plot is in the di-electron channel and the right plot in the di-muon channel, for the 3-jets category.

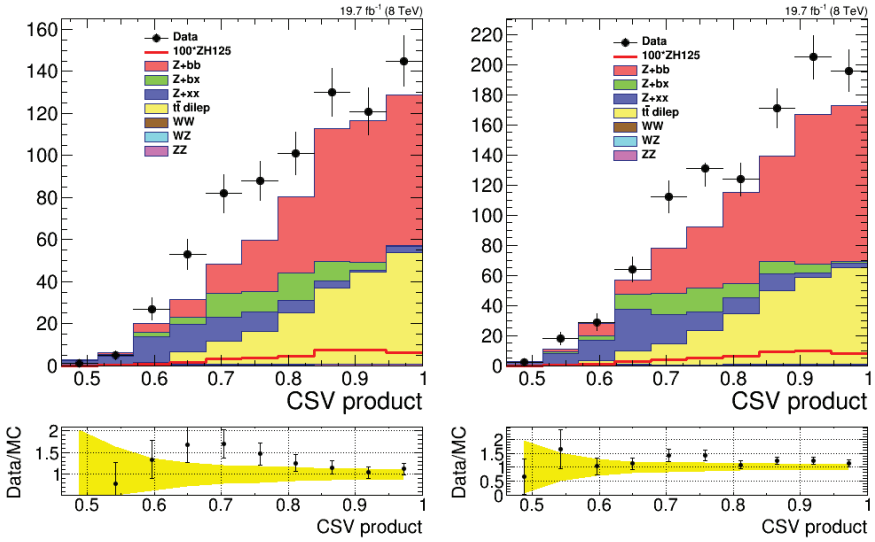

Figure C.2: Product of the two leading b jets CSV discriminator values, used to discriminate b jets from light jets contributions, in the CR. The left plot is in the di-electron channel and the right plot in the di-muon channel, for the 3-jets category.

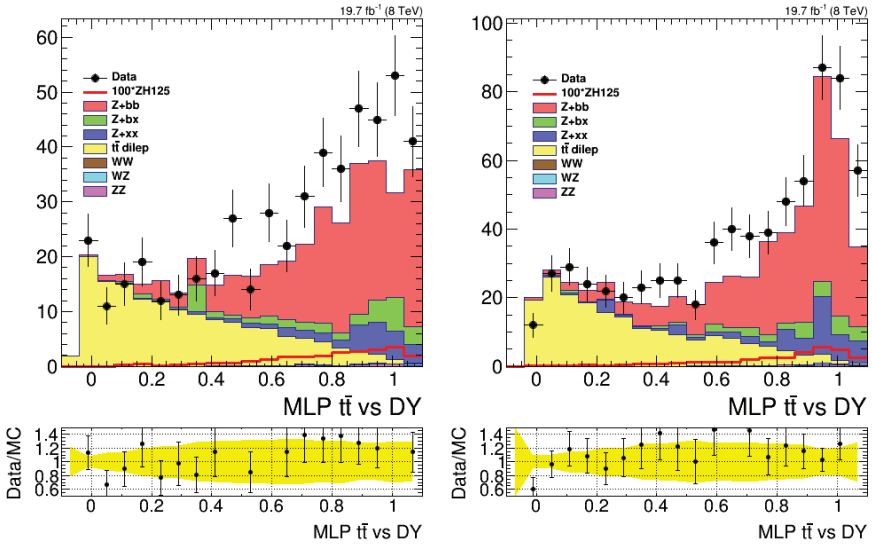

Figure C.3: Neural-Net output used to discriminate DY events from $t\bar{t}$ events, in the CR. Left plots represent the di-electron channel, right plots the di-muon channel, for the 3-jets category.

C.3 Full Region plots

In this section, plots of the main kinematic variables are displayed, with the Full Region selection for the CSV tagger: the p_T of the leading and sub-leading b jet (lepton) on Fig. C.4 (Fig. C.5), the p_T of the two leading b jets and leptons on Fig. C.6, the ΔR between the two leptons and jets (Fig. C.7), the b-jets pair invariant mass and the MET significance (Fig. C.8) and the jet/lepton multiplicity (Fig. C.9). The SF obtained by the background fit procedure are not applied here.

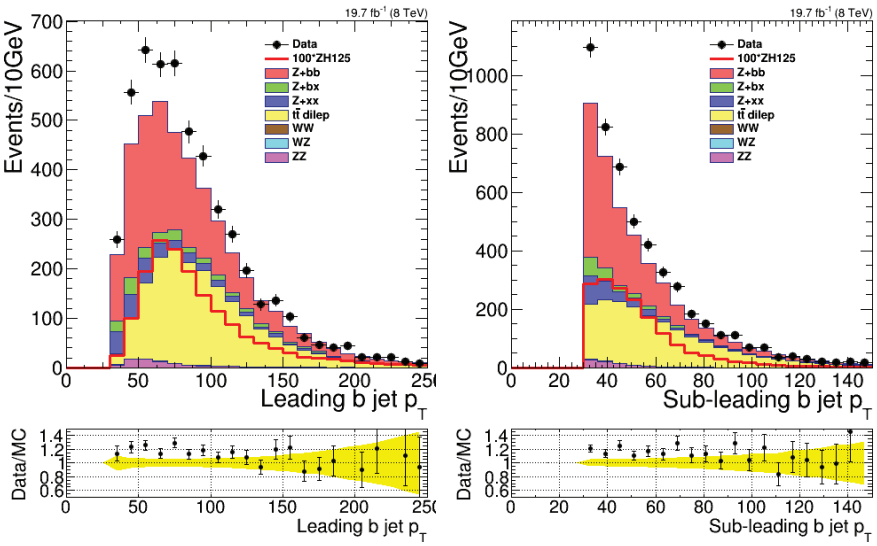

Figure C.4: Distributions of the p_T of the leading (left) and sub-leading (right) b-jet, in the FR. The SF from the background reweighting have not been applied.

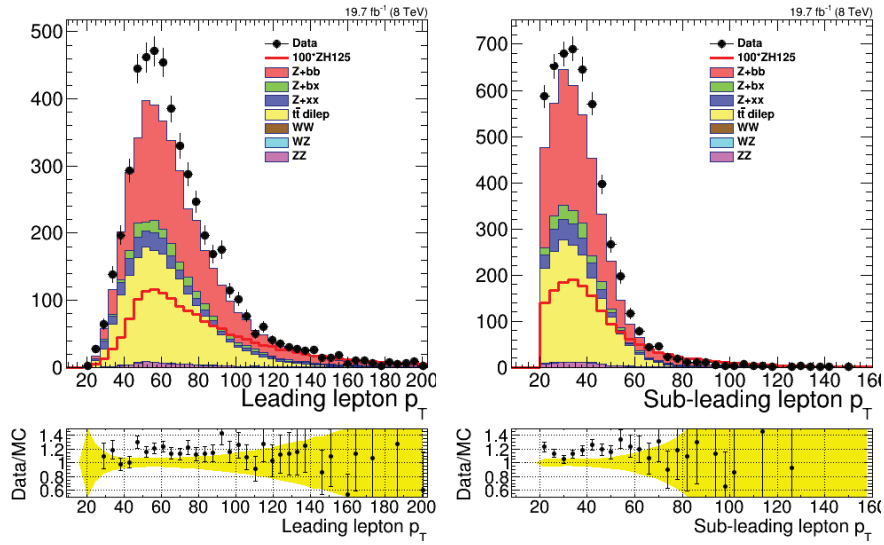

Figure C.5: Distributions of the p_T of the leading (left) and sub-leading (right) lepton, in the FR. The SF from the background reweighting have not been applied.

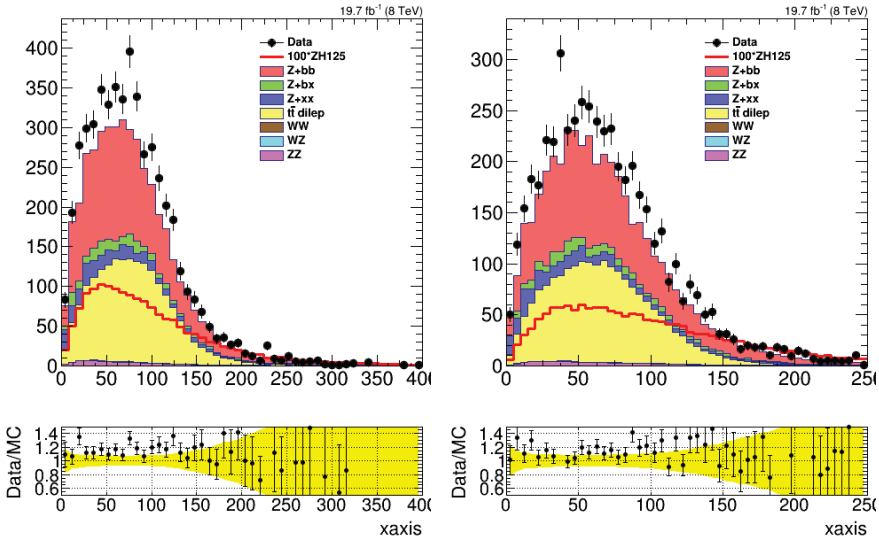

Figure C.6: Distributions of the p_T of the two leading jets (left) and leptons (right), in the FR. The SF from the background reweighting have not been applied.

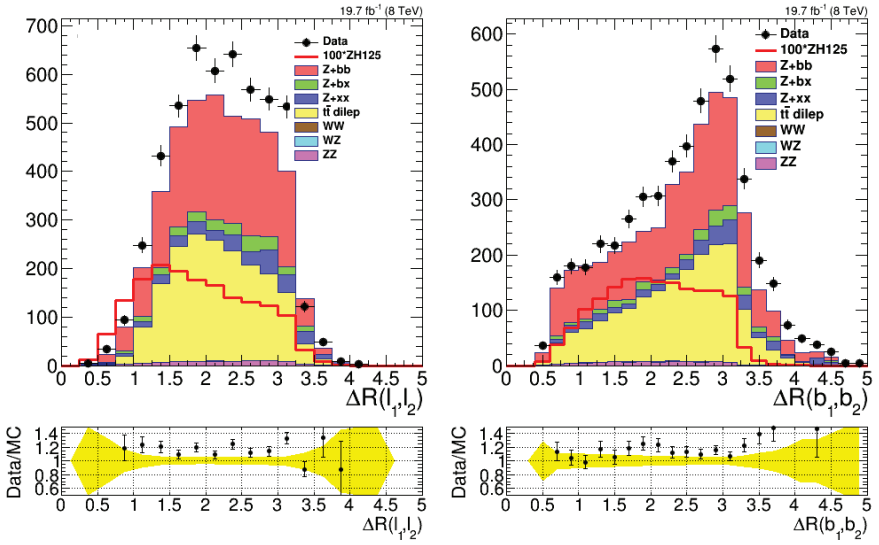

Figure C.7: Distributions of the ΔR between the two leptons (left) and between the two b-jets (right) lepton, in the FR. The SF from the background reweighting have not been applied.

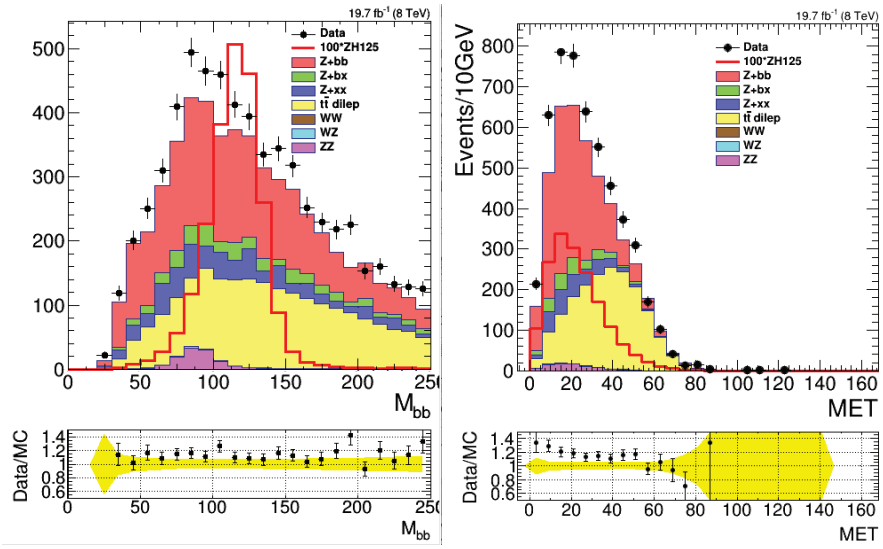

Figure C.8: Distributions of the b-jets invariant mass (left) and of the MET significance (right), in the FR. The SF from the background reweighting have not been applied.

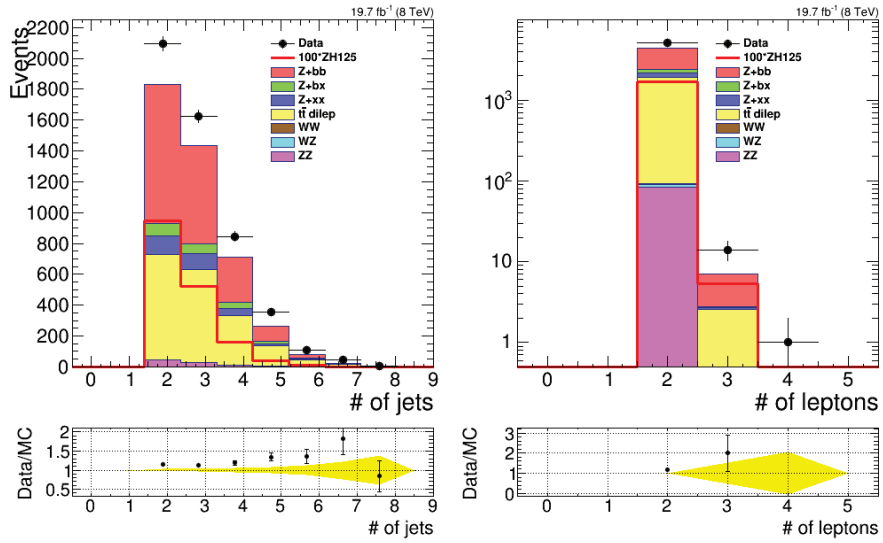

Figure C.9: Distributions of the multiplicity of jets (left) and of leptons (right), in the FR. The SF from the background reweighting have not been applied.

C.4 ll+jj+X plots

In this section, plots of the main kinematic variables are displayed, for the “ll+jj+X” selection: two leptons and two jets, plus “X” meaning that no cut is applied on the MET or for extra jets. The following plots are displayed: the p_T of the leading and sub-leading jet (Fig. C.10), the η of the leading and sub-leading jet (Fig. C.11), the p_T of the leading and sub-leading lepton (Fig. C.12), the η of the leading (left) and sub-leading lepton (Fig. C.13), the $\Delta R(ll)$ and $\Delta R(jj)$ (Fig. C.14), the leading jets invariant mass and of the multiplicity of jets (Fig. C.15) and the multiplicity of leptons (Fig. C.16).

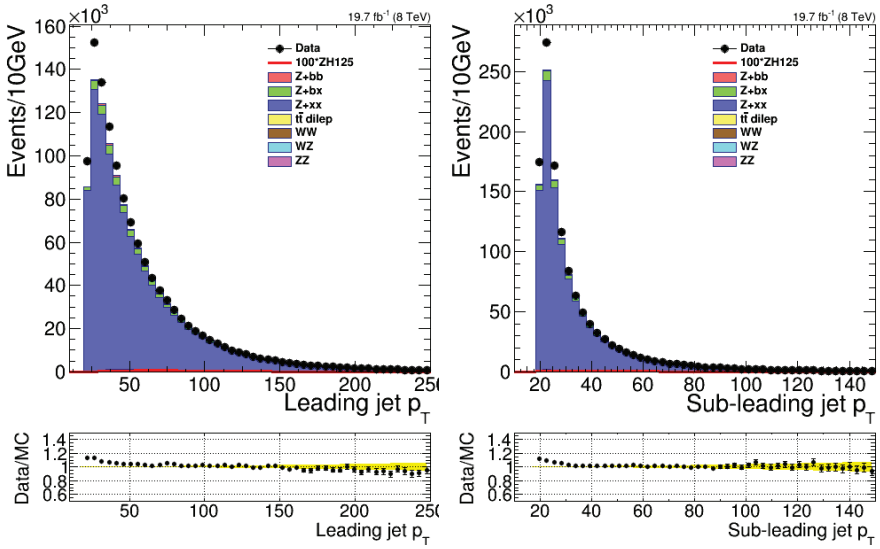

Figure C.10: Distributions of the p_T of the leading (left) and sub-leading jet (right), at "ll+jj+X" stage, with a FR like selection.

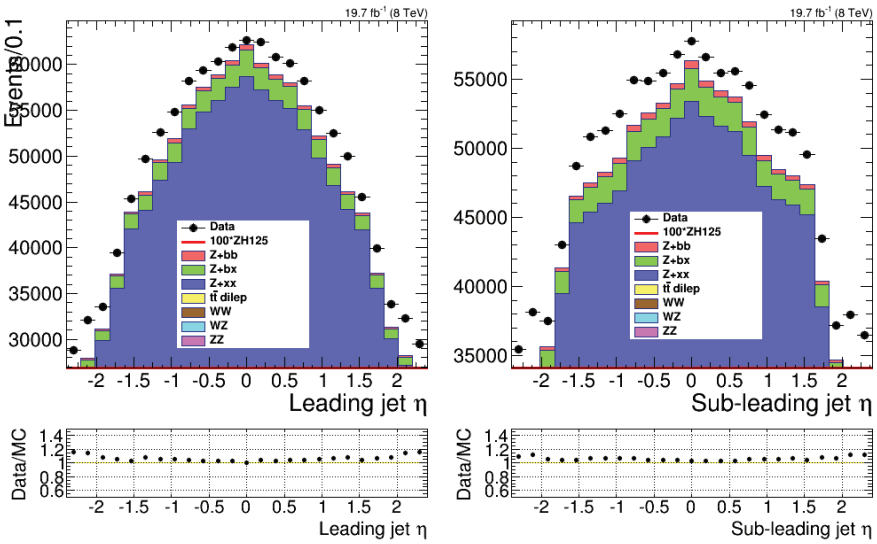

Figure C.11: Distributions of the η of the leading (left) and sub-leading jet (right), at "ll+jj+X" stage, with a FR like selection.

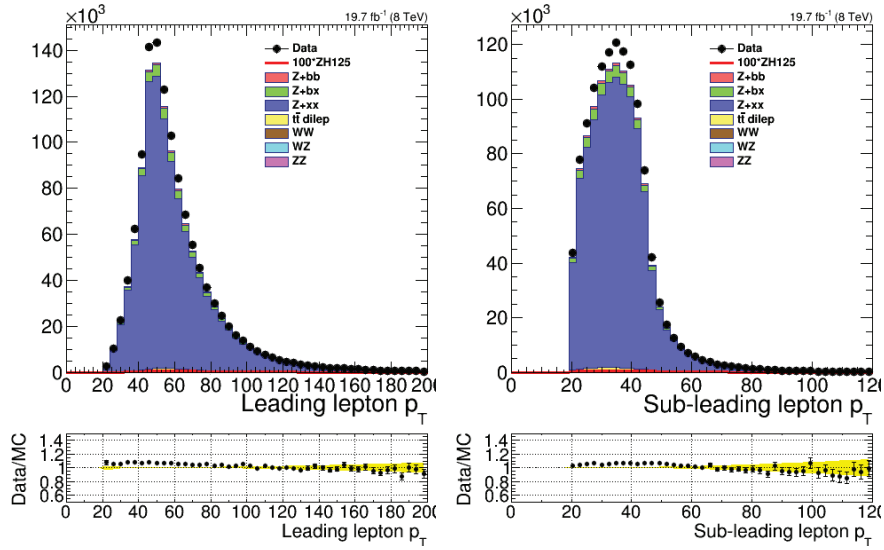

Figure C.12: Distributions of the p_T of the leading (left) and sub-leading lepton (right), at "ll+jj+X" stage, with a FR like selection.

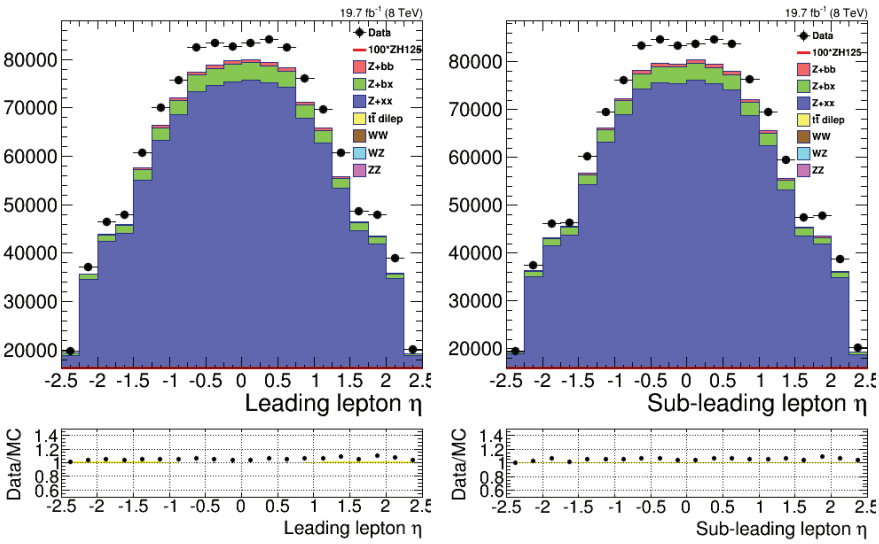

Figure C.13: Distributions of the η of the leading (left) and sub-leading lepton (right), at "ll+jj+X" stage, with a FR like selection.

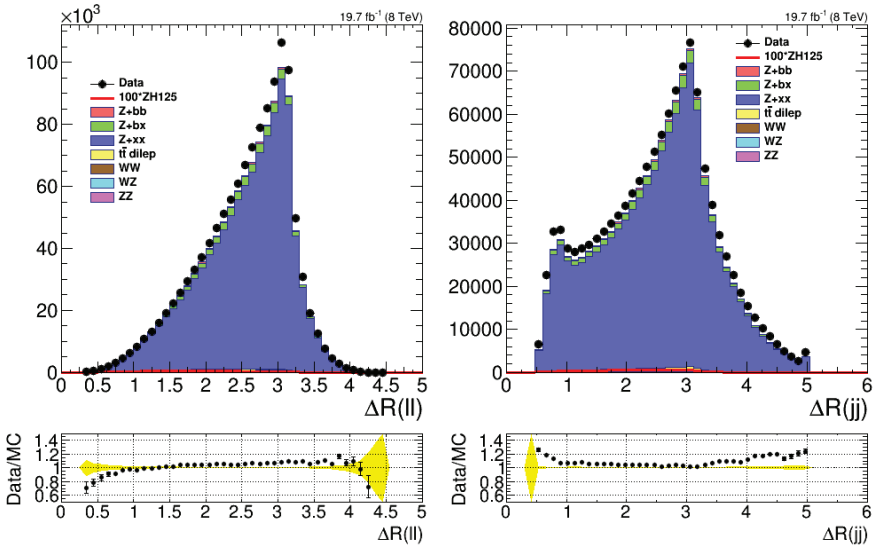

Figure C.14: Distributions of the $\Delta R(ll)$ and $\Delta R(jj)$ (right), at "ll+jj+X" stage, with a FR like selection.

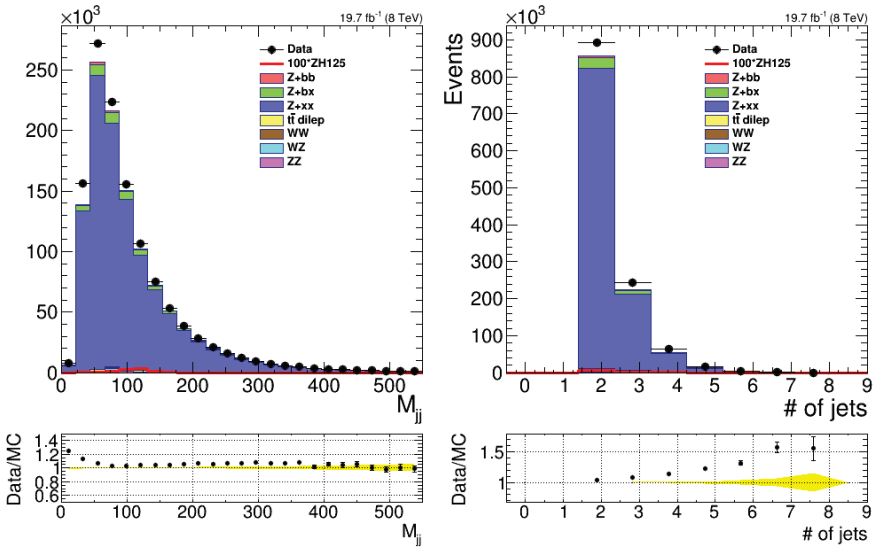

Figure C.15: Distributions of the leading jets invariant mass (left) and of the multiplicity of jets (right), at "ll+jj+X" stage, with a FR like selection.

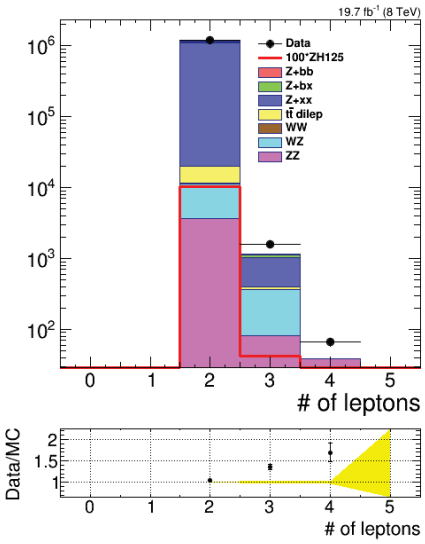

Figure C.16: Distributions of the multiplicity of leptons, at "ll+jj+X" stage, with a FR like selection.

C.5 Signal Region plots

In this section, plots of the main kinematic variables are displayed for the Signal region, first when using the JP tagger, then when using the CSV tagger.
For the JP tagger, the folling plots are shown for both categories of jets: the ΔR between the two leptons (Fig. C.17), the ΔR between the two b jets (Fig. C.18), the leading lepton p_T (Fig. C.19), the leading lepton η (Fig. C.20), the two b jets p_T (Fig. C.21), the two leptons p_T (Fig. C.22) and finally the jets mulitplicity in the 3-jets category (Fig. C.23).

C.5.1 JP tagger

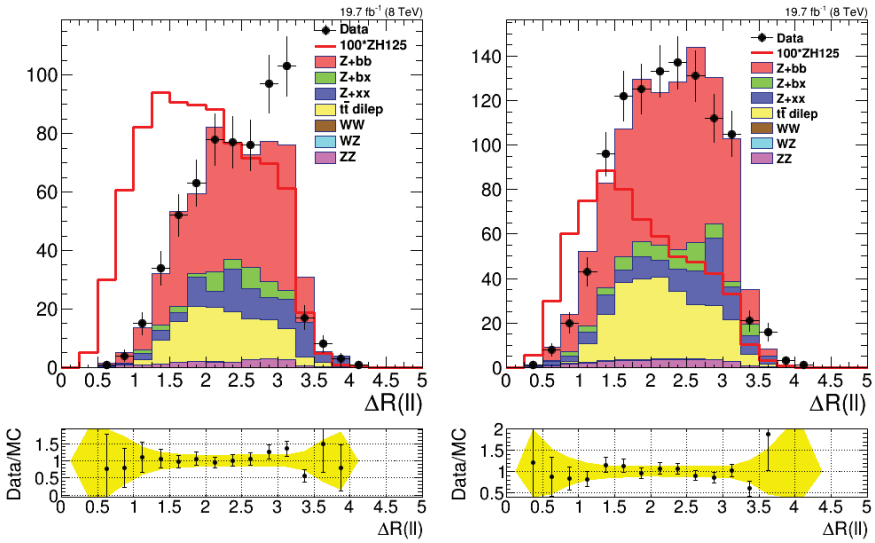

Figure C.17: Distribution of the ΔR between the two leptons, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

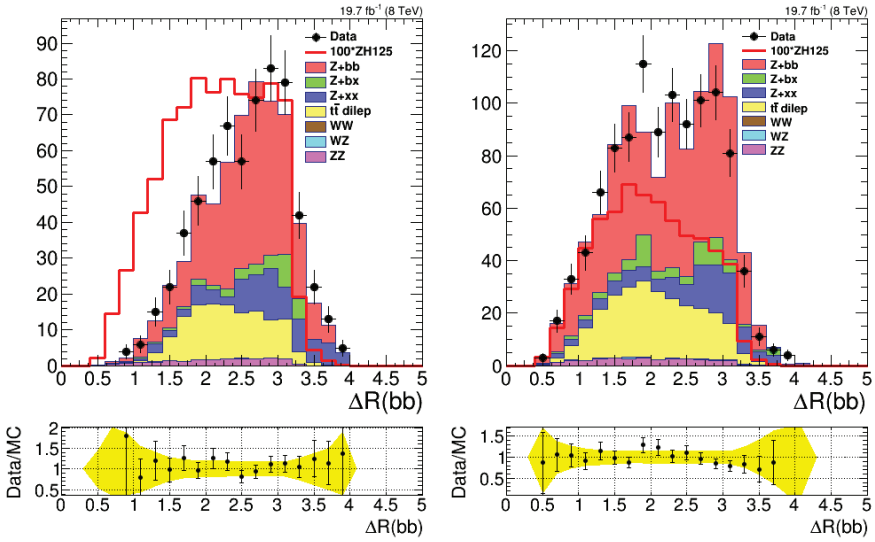

Figure C.18: Distribution of the ΔR between the two b jets, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

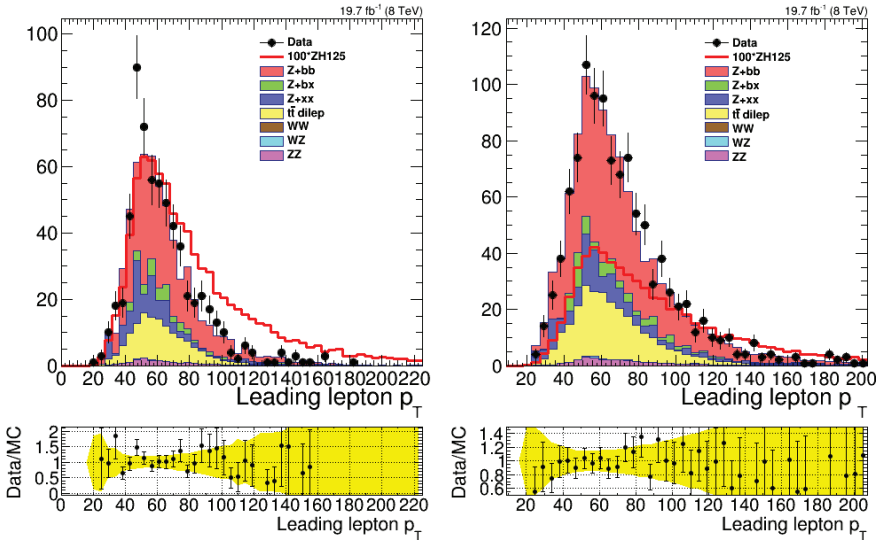

Figure C.19: Distribution of the leading lepton's p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

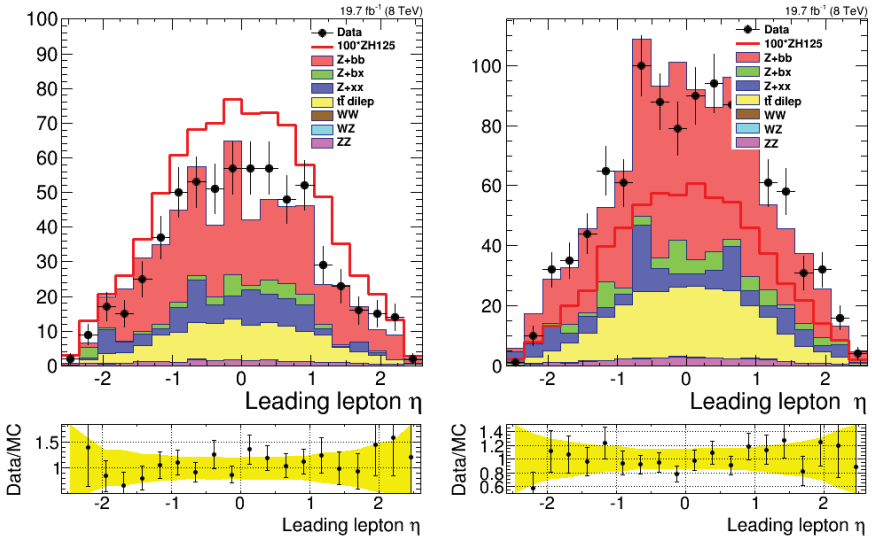

Figure C.20: Distribution of the leading lepton's η, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

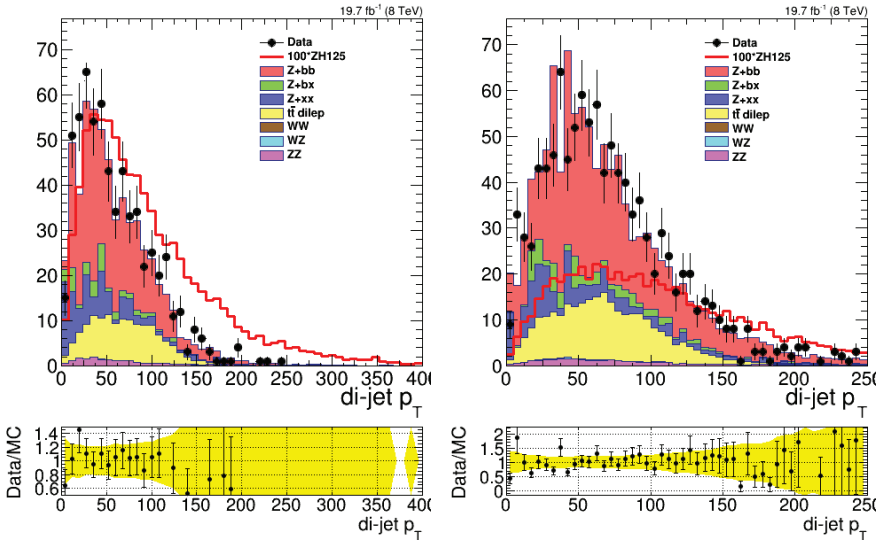

Figure C.21: Distribution of the two b jets p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

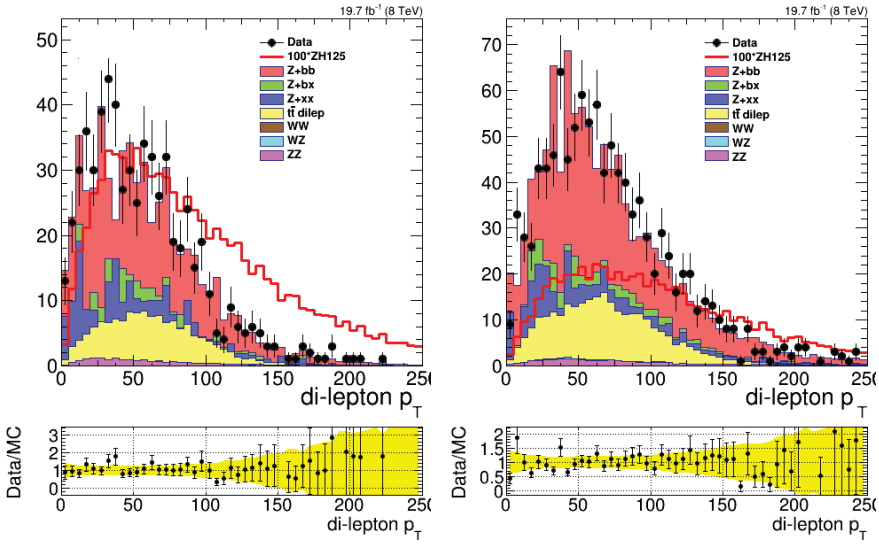

Figure C.22: Distribution of the two leptons p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

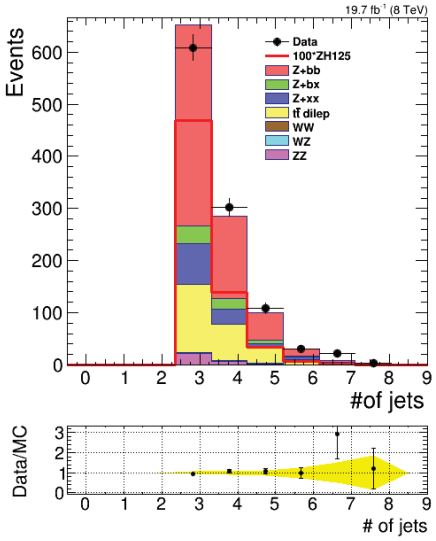

Figure C.23: Distribution of the jet multiplicity the 3-jet category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

C.5.2 CSV tagger

For the CSV tagger, the following plots are displayed for both jets categories: the di-lepton (jet) invariant mass on Fig. C.24 (Fig. C.25), the MET significance (Fig. C.26), the (sub-)leading b jet p_T on Fig. C.27 (Fig. C.28), the distribution of the ΔR between the two leptons (b jets) on Fig. C.29 (Fig. C.30), the distribution of the leading lepton's p_T (η) on Fig. C.31 (Fig. C.32), of the two b jets (leptons) p_T on Fig. C.33 (Fig. C.34) and finally the jet multiplicity in the 3-jets category on Fig. C.35.

The control plots for the MEM weights can be found on Fig. C.36 for the $t\bar{t}$ hypothesis, on Fig. C.37 for the Zbb_{gg} hypothesis, on Fig. C.38 for the Zbb_{qq} hypothesis, on Fig. C.39 (Fig. C.40) for the ZZ_{cor0} (ZZ_{cor3}) hypothesis and on Fig. C.41 (Fig. C.42) for the ZH_{cor0} (ZH_{cor3}) hypothesis, for both jets categories.

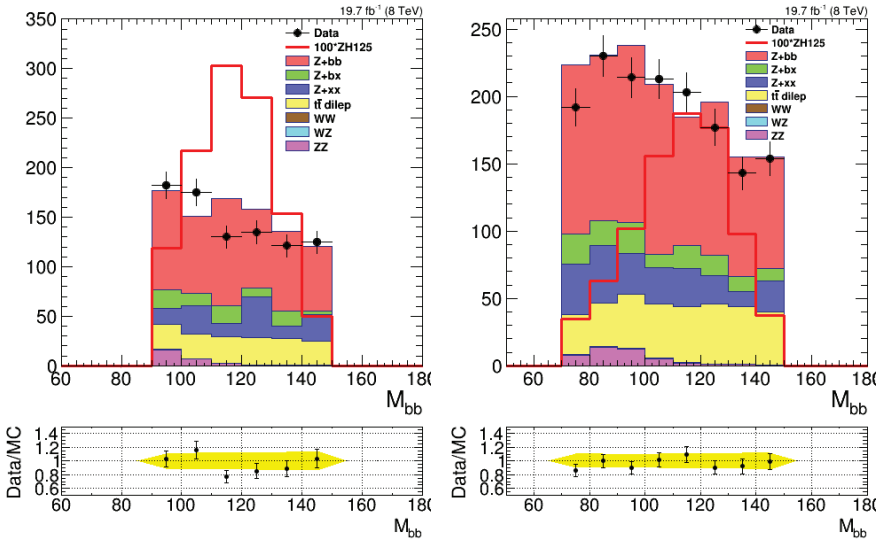

Figure C.25: Di-jet invariant mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

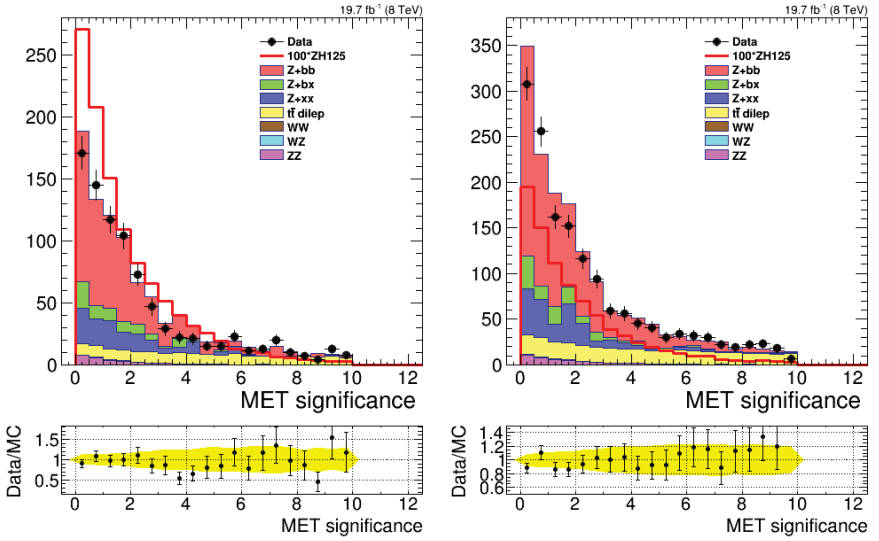

Figure C.26: MET significance mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

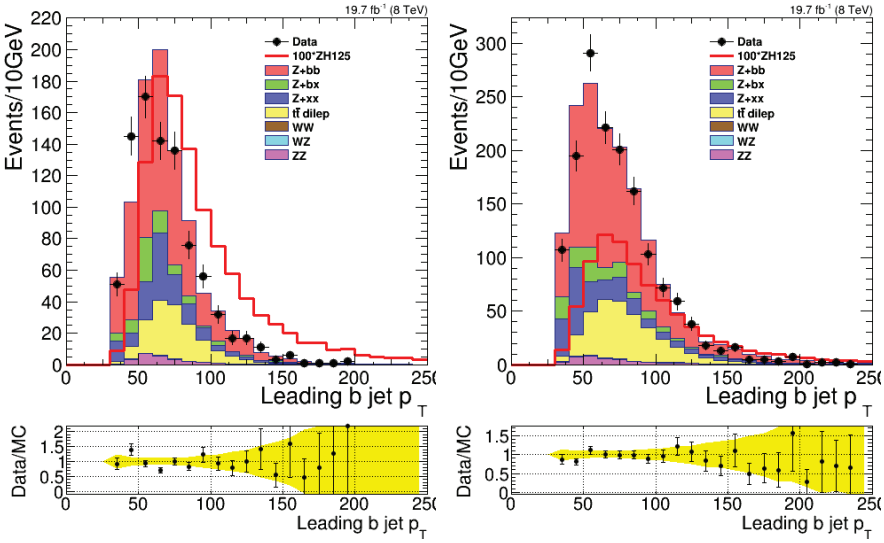

Figure C.27: Leading b jet p_T in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

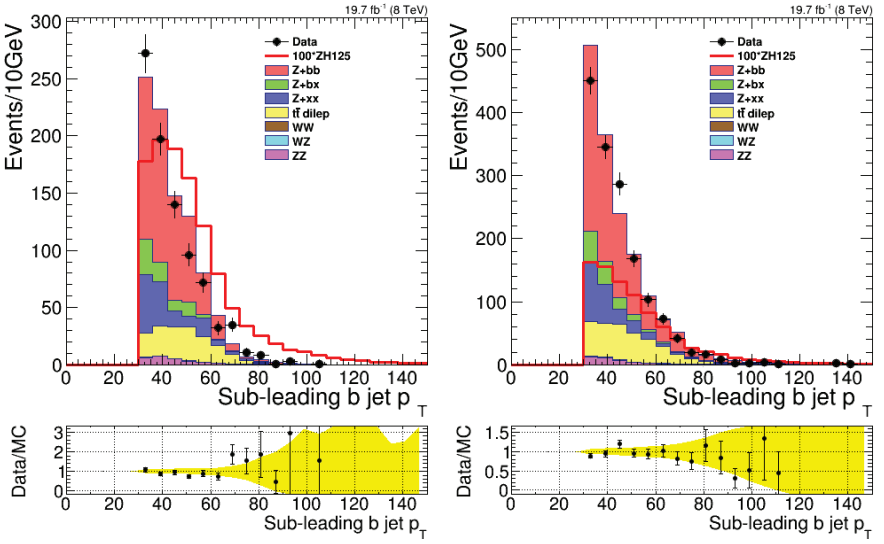

Figure C.28: Sub-leading b jet p_T in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

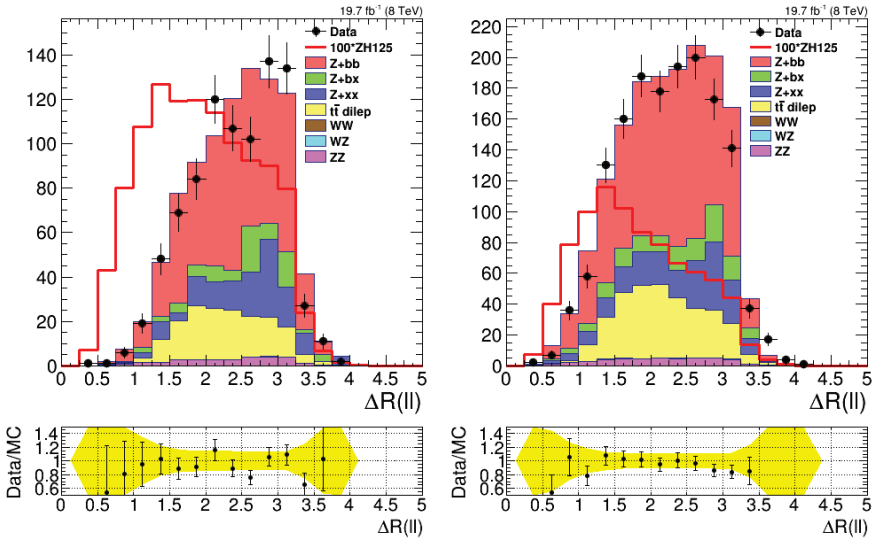

Figure C.29: Distribution of the ΔR between the two leptons, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

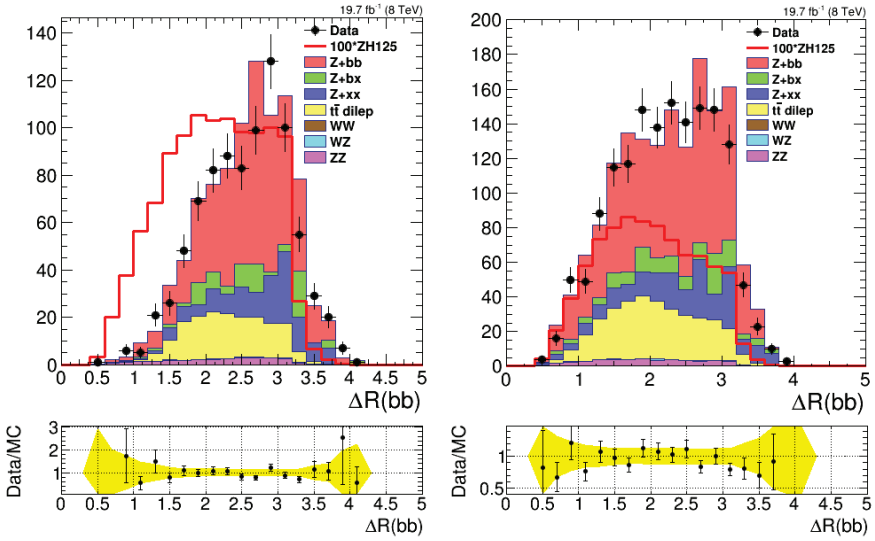

Figure C.30: Distribution of the ΔR between the two b jets, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

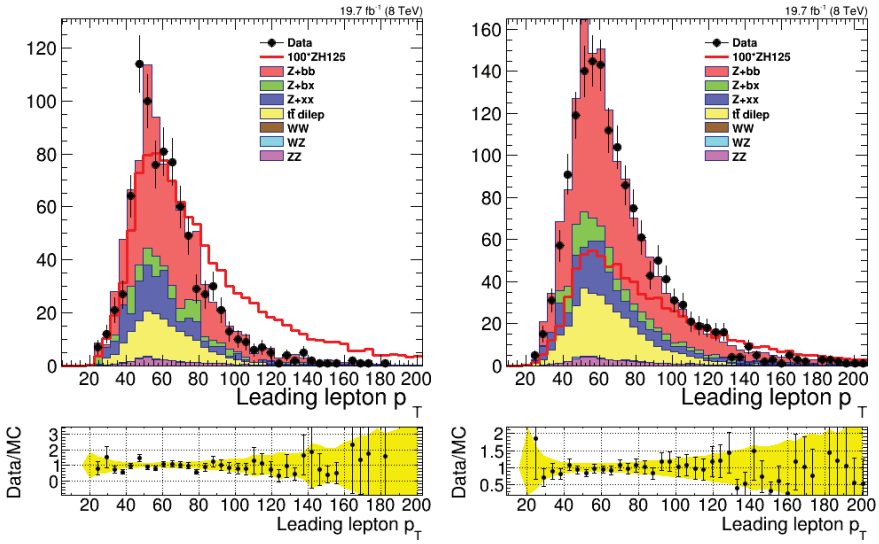

Figure C.31: Distribution of the leading lepton's p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

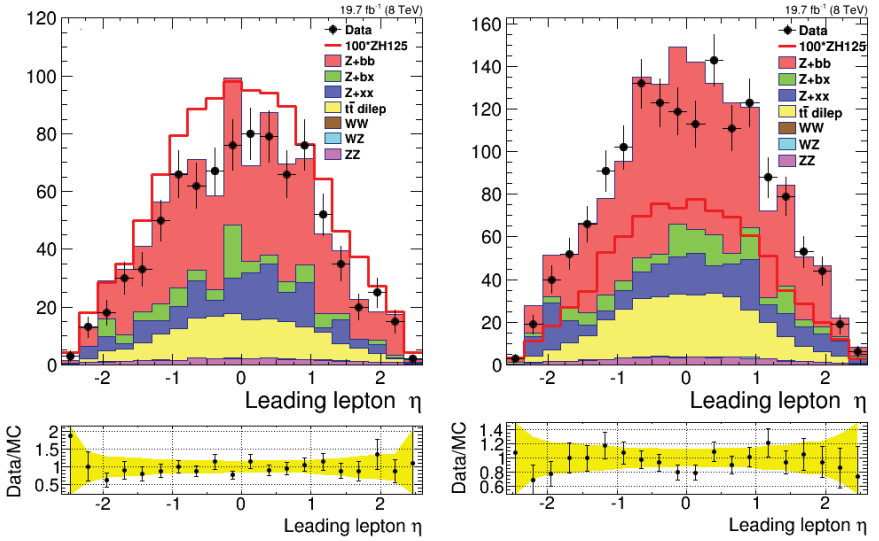

Figure C.32: Distribution of the leading lepton's η, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

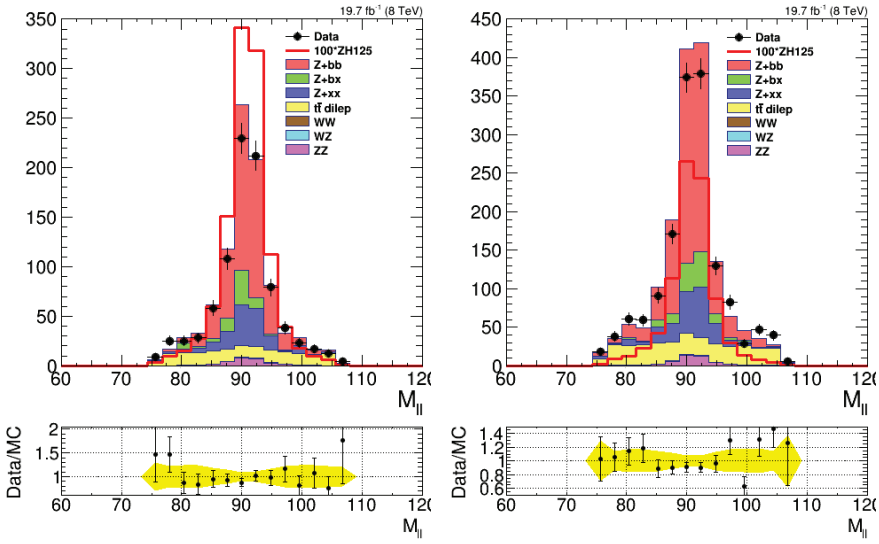

Figure C.24: Di-lepton invariant mass in the signal region, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factors are applied.

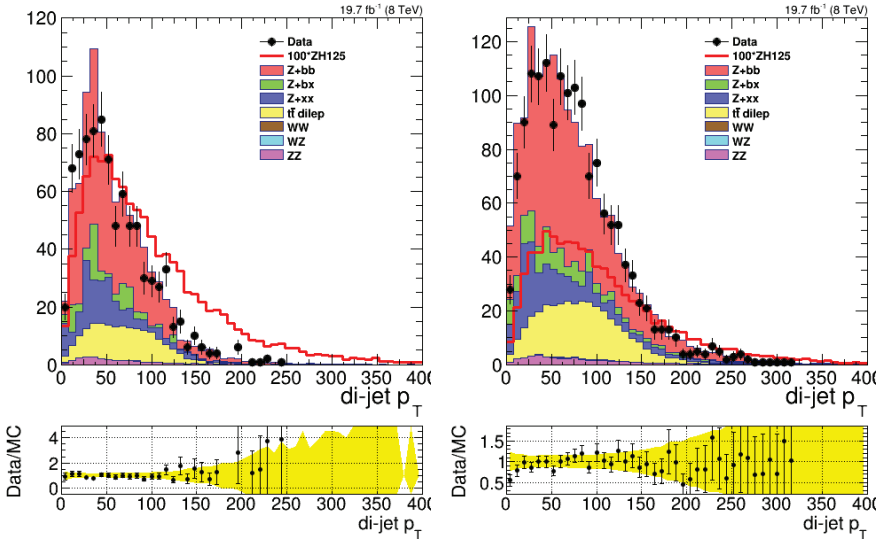

Figure C.33: Distribution of the two b jets p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

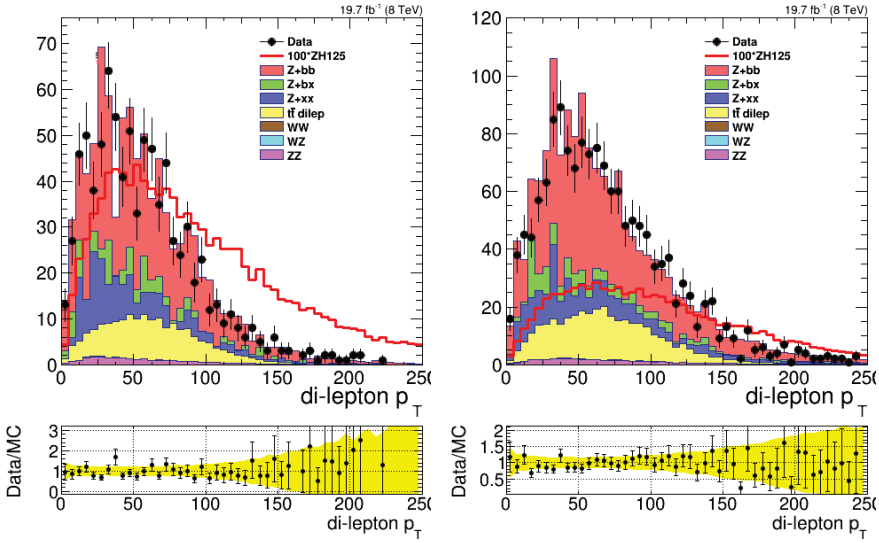

Figure C.34: Distribution of the two leptons p_T, for the 2-jets (left) and the 3-jets (right) category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

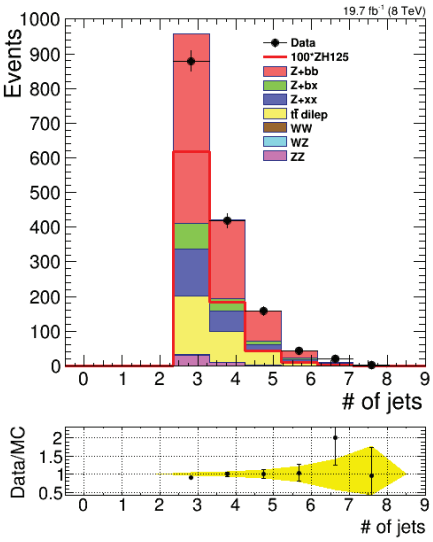

Figure C.35: Distribution of the jet multiplicity in the 3-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied.

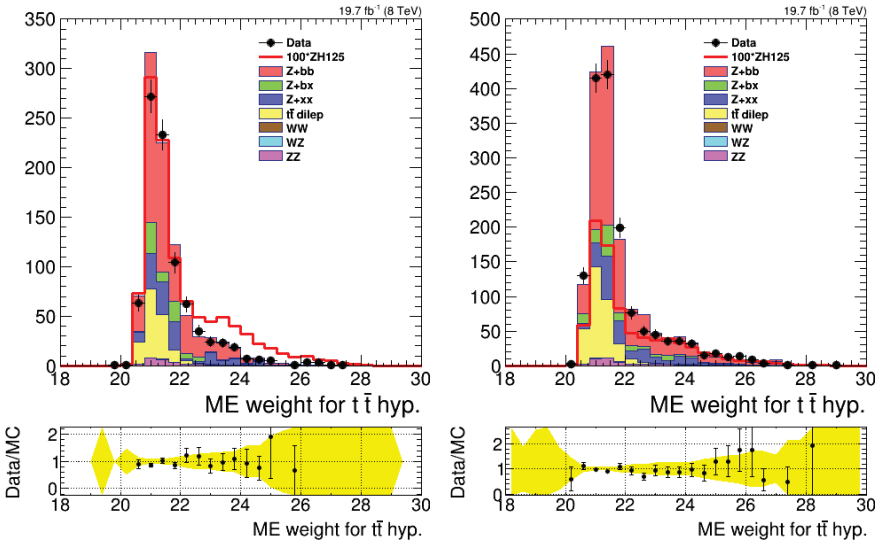

Figure C.36: ME weights for the $t\bar{t}$ hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

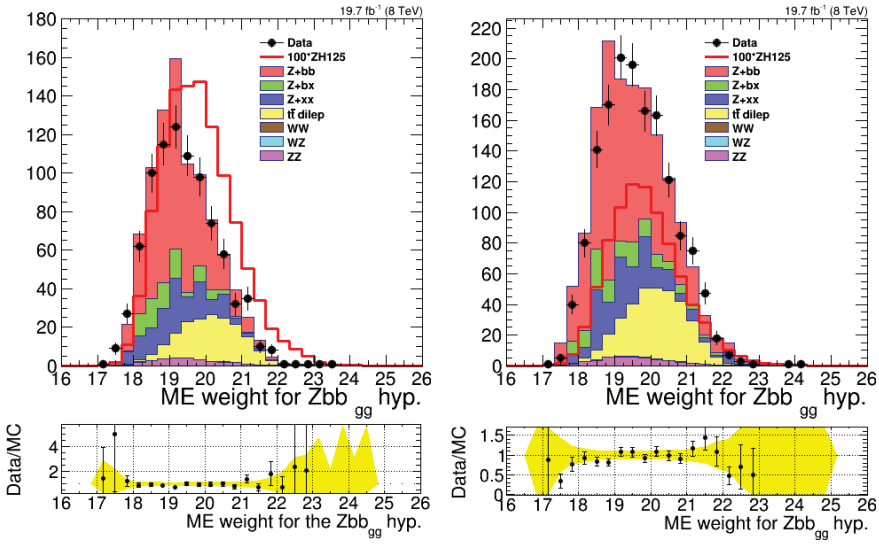

Figure C.37: ME weights for the Zbb induced by gluon-gluon hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

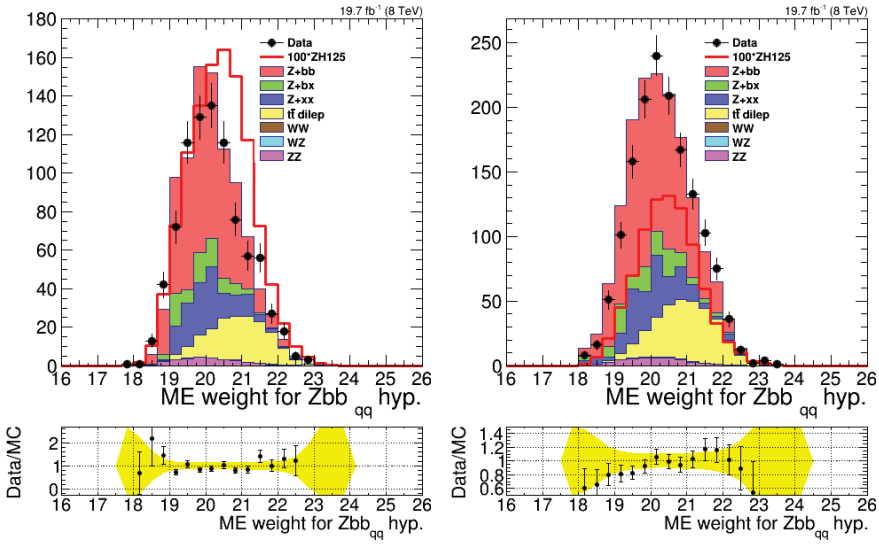

Figure C.38: ME weights for the Zbb induced by $q\bar{q}$ hypothesis, for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

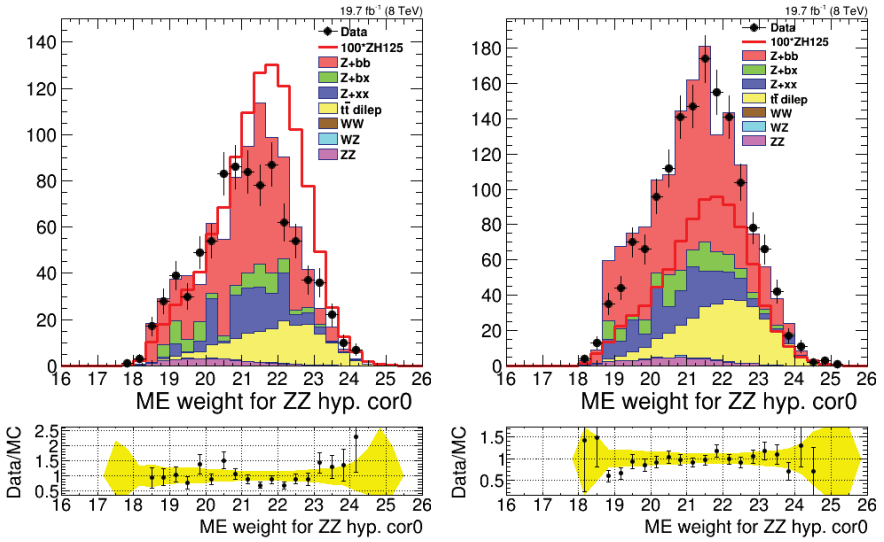

Figure C.39: ME weights for the ZZ hypothesis with E-p conservation (cor_0), for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

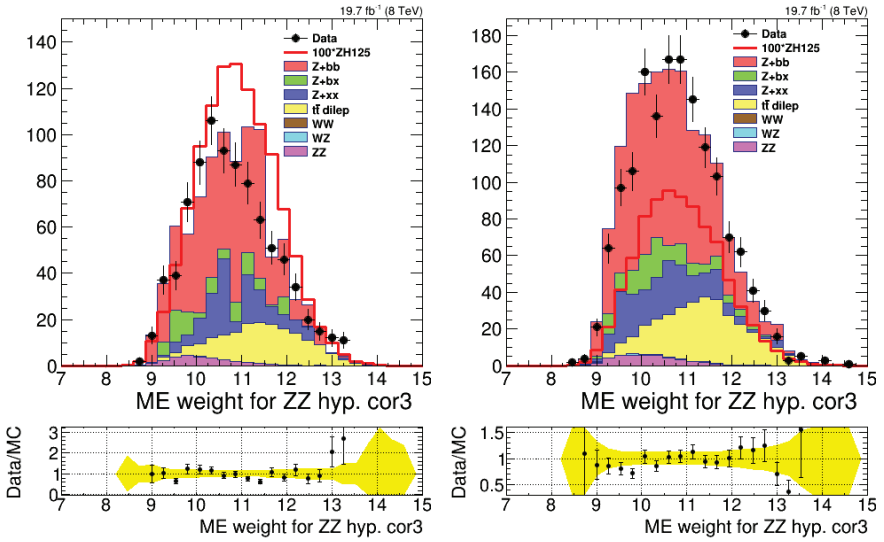

Figure C.40: ME weights for the ZZ hypothesis without E-p conservation (cor_3), for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

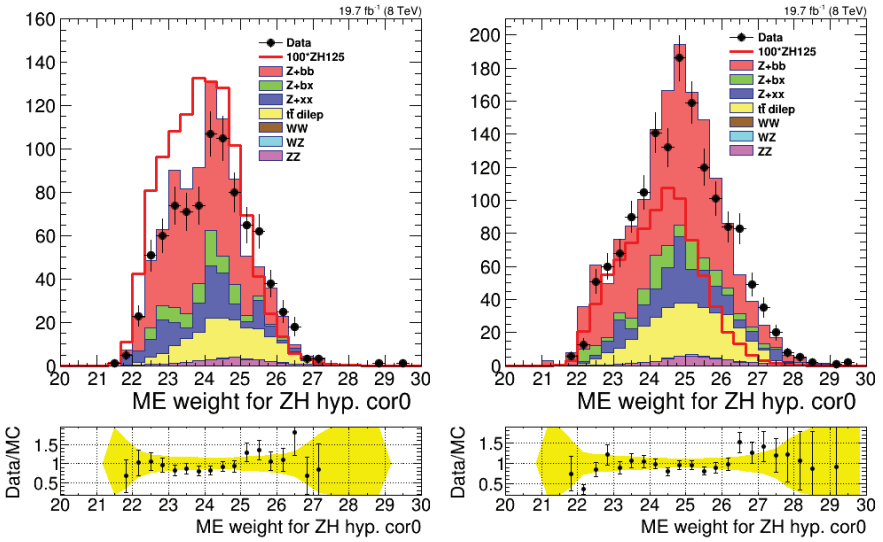

Figure C.41: ME weights for the ZH hypothesis with E-p conservation (cor_0), for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

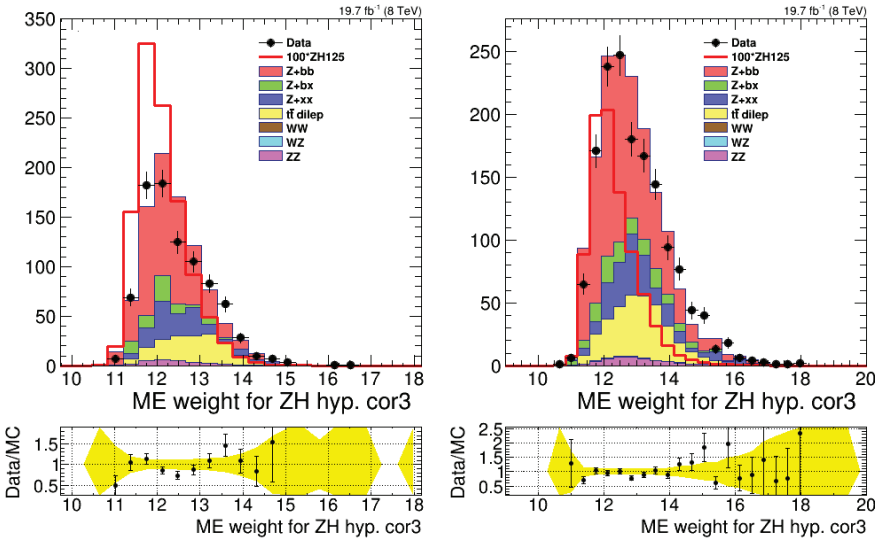

Figure C.42: ME weights for the ZH hypothesis without E-p conservation (cor_3), for the 2-jets (left) and the 3-jets (right) categories, for the CSV tagger. Events have been renormalized according to their cross section. All the correction scale factor have been applied.

C.6 Use of a MLP

The use of a Multi Layer Perceptron (MLP) was also performed in this analysis. The procedure to build the discriminators for the ZH signal is divided in two steps, because of the limited available statistics (MLPs seem to be more sensitive to this criteria than BDTs).

First, intermediate MLPs are trained in order to separate the ZH signal from a given background process. The inputs of those MLPs are the available MEM weights corresponding to the signal and background event hypotheses involved. The signal ZH weights with and without imposing energy-momentum conservation are used in all the MLPs trainings. The background weights used for the three considered intermediate MLPs are listed below:

- ZH versus DY exploits the MEM weights for the Zbb_{gg}, Zbb_{qq}, ZH_{cor0} and ZH_{cor3} hypotheses. All the DY samples are used in the training, with all the Zbb, Zbx and Zxx contributions. This is motivated by the fact that the Zbx and Zxx events tend to behave like the Zbb events, it allows to keep most of the statistics;
- ZH versus $t\bar{t}$ uses the MEM weights for the $t\bar{t}$, ZH_{cor0} and ZH_{cor3} hypotheses;
- ZH versus ZZ exploits the MEM weights for the ZZ_{cor0}, ZZ_{cor3}, ZH_{cor0} and ZH_{cor3} hypotheses.

Separate trainings are performed for the 2-jets and 3-jets categories and for muons and electrons together. The SR selection is applied.

In a second step, the three intermediate MLPs are used as the only inputs of the final MLP. Similarly to intermediate MLPs, separate trainings are performed for the 2 categories of jets, and the same event selection is applied. In this manner, the training background sample is formed by the addition of the DY, $t\bar{t}$, and ZZ samples used for the intermediate MLPs. For these samples, the same weights are applied as in the BDT case: 0.9, 0.085 and 0.015, respectively.

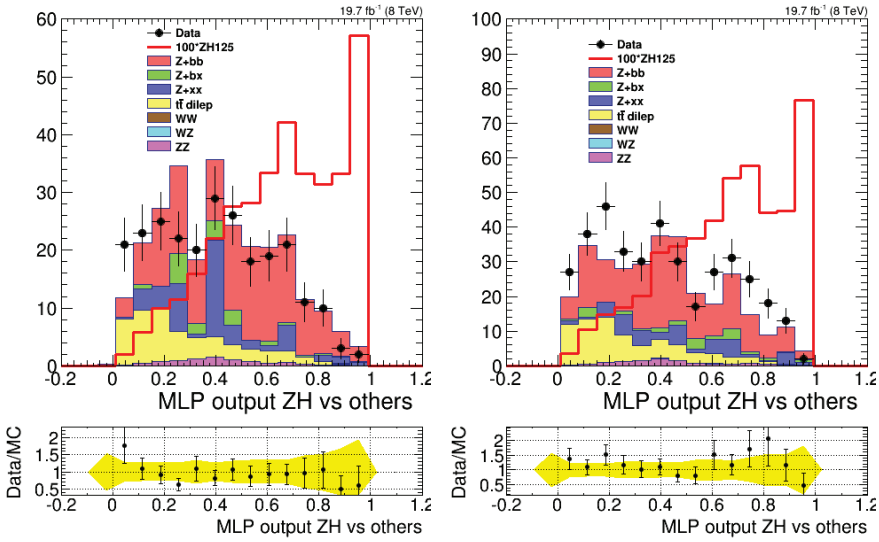

Figure C.43: MLP output in the SR for the electrons (left) and muons (right) in the 2-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is JP.

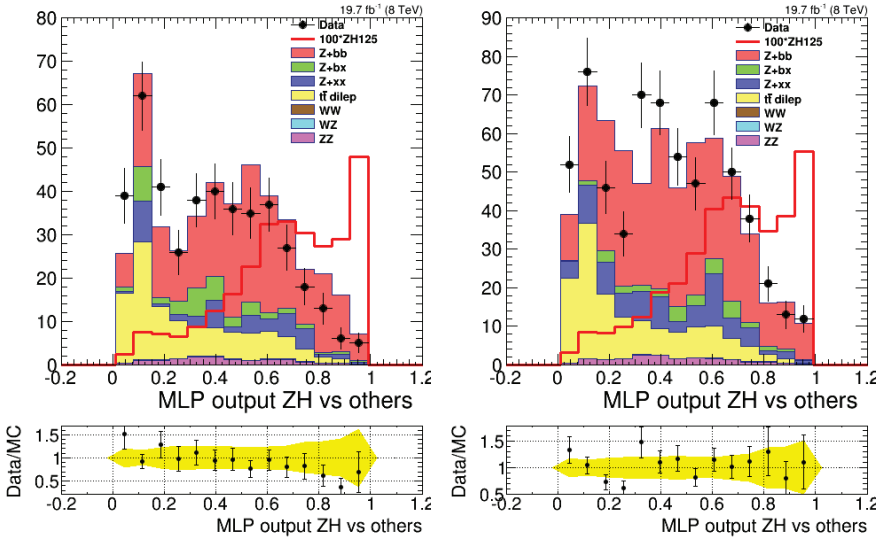

Figure C.44: MLP output in the SR for the electrons (left) and muons (right) in the 3-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is JP.

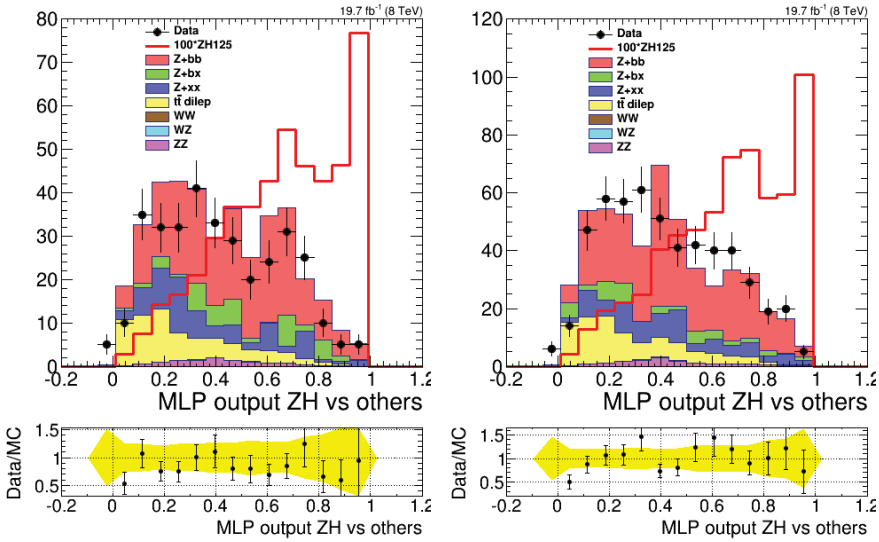

Figure C.45: MLP output in the SR for the electrons (left) and muons (right) in the 2-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is CSV.

The plot of the MLP output can be seen for JP (CSV) on Fig.C.43 (Fig. C.44) and Fig.C.45 (Fig. C.46), for both jets categories and both muon and electron channels, in the signal region.

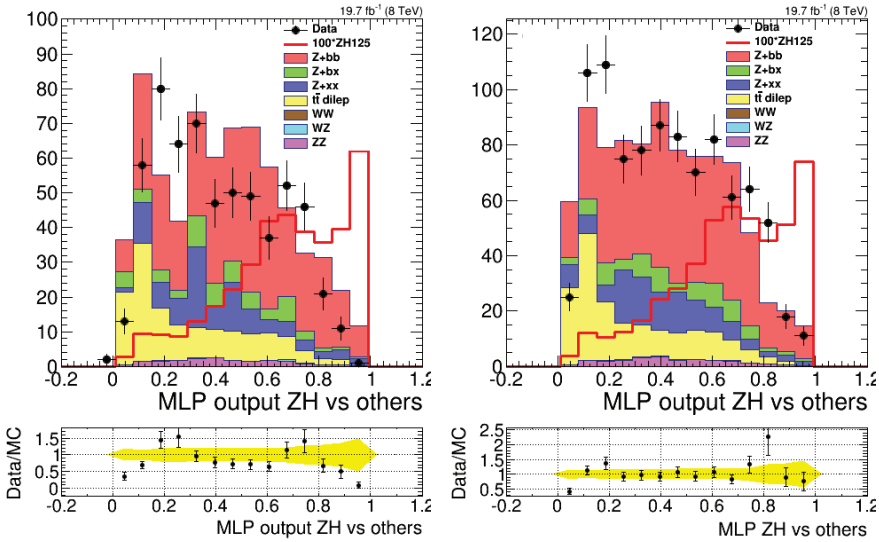

Figure C.46: MLP output in the SR for the electrons (left) and muons (right) in the 3-jets category. Events have been renormalized according to their cross section and all the correction scale factor have been applied. The tagger used is CSV.

C.7 Training plots

The TMVA tool produces several control plots: the output of the BDT/MLP, showing the final discrimination between the signal and the background(s), when the events have not been yet renormalized to their corresponding cross section. For the MLP, it is also possible to display the training convergence as well as the neuronal architecture used.

C.7.1 DY versus $t\bar{t}$

MLP output, convergence and architecture for both muon and electron channels are respectively shown on Fig. C.47, Fig. C.48 and Fig. C.49.

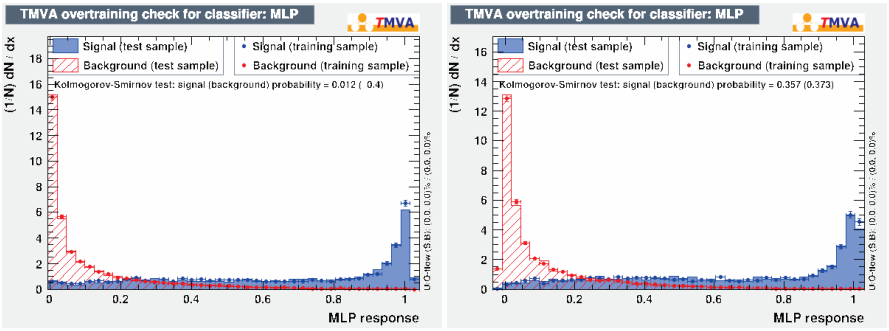

Figure C.47: TMVA MLP output for DY versus $t\bar{t}$, for the muons (left) and the electrons (right).

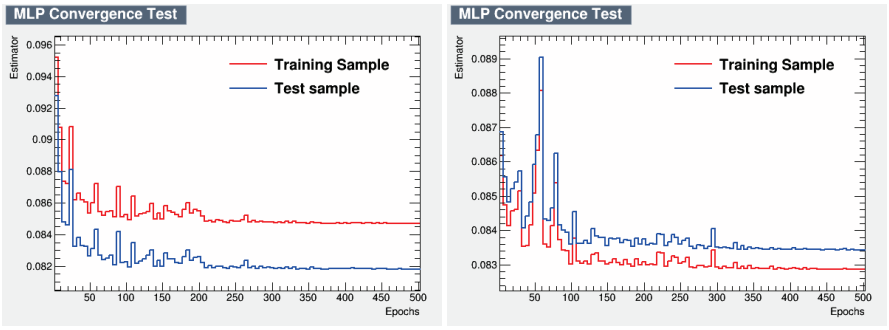

Figure C.48: TMVA MLP training convergence for DY versus $t\bar{t}$, for the muons (left) and the electrons (right).

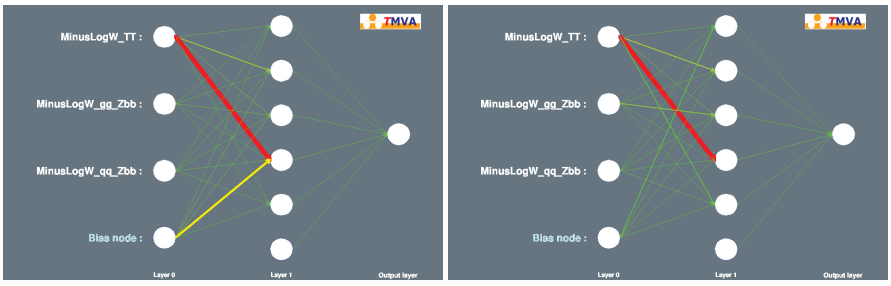

Figure C.49: TMVA MLP training architecture for DY versus $t\bar{t}$, for the muon (left) and the electron (right).

C.7.2 ZH versus other backgrounds

On Fig. C.50, the BDT output ZH versus all backgrounds is displayed for both jets categories.

The DY versus ZH MLP output, convergence and architecture are respectively shown on Fig. C.51, Fig. C.52 and Fig. C.53, the ZZ versus ZH MLP output, convergence and architecture are respectively shown on Fig. C.54, Fig. C.55 and Fig. C.56 and finally the $t\bar{t}$ versus ZH MLP output, convergence and architecture are respectively shown on Fig. C.57, Fig. C.58 and Fig. C.59, for both jets categories.

In the end, ZH versusall backgrounds MLP output and convergence are respectively shown on Fig. C.60 and Fig. C.61 for both jets categories.

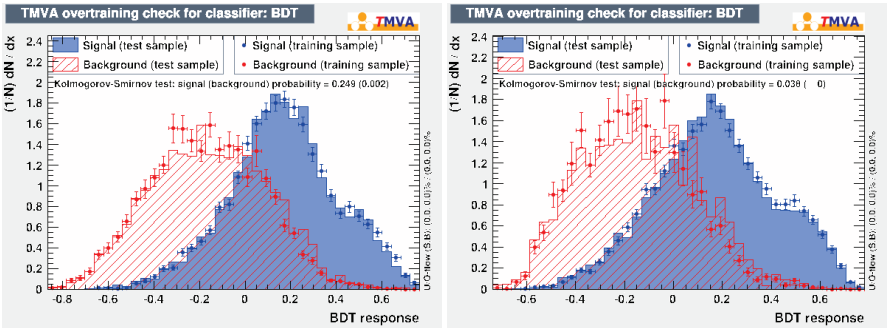

Figure C.50: TMVA BDT output ZH versus all backgrounds, for the 2-jets (left) and 3-jets (right) categories.

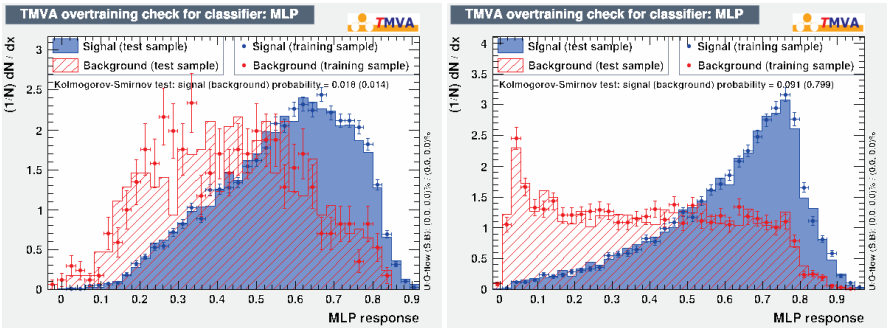

Figure C.51: TMVA MLP output DY versus ZH, for the 2-jets (left) and 3-jets (right) categories.

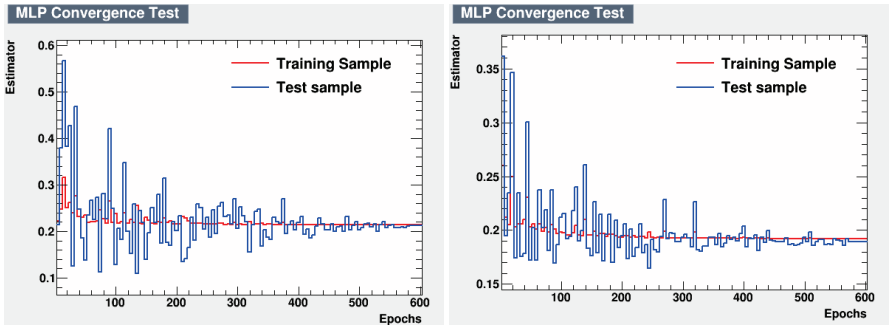

Figure C.52: TMVA MLP training convergence, for DY versus ZH, for the 2-jets (left) and 3-jets (right) categories.

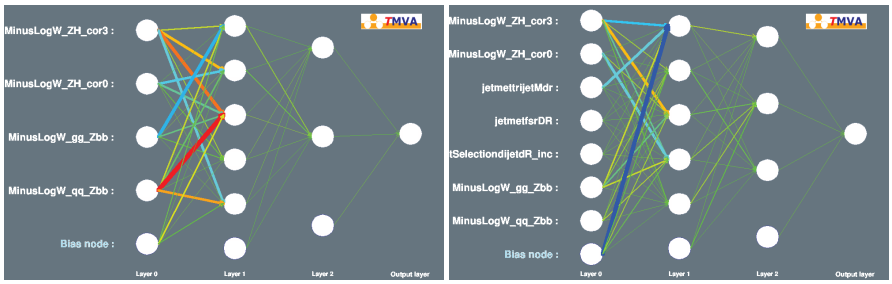

Figure C.53: TMVA MLP architecture DY versus ZH, for the 2-jets (left) and 3-jets (right) categories.

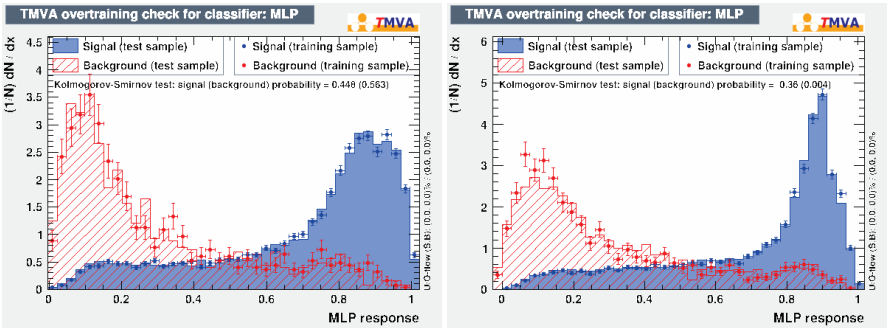

Figure C.54: TMVA MLP output ZZ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

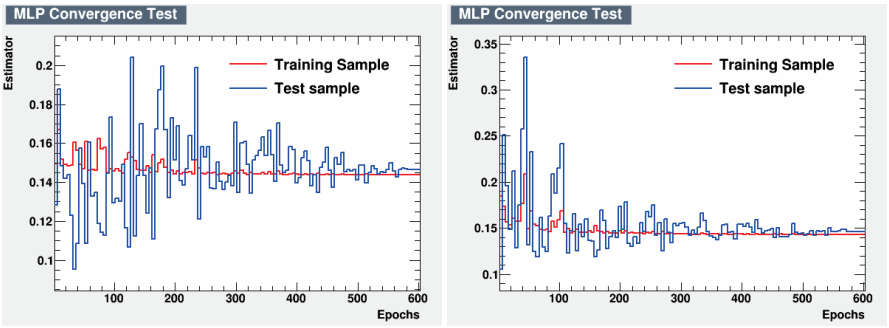

Figure C.55: TMVA MLP training convergence, for ZZ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

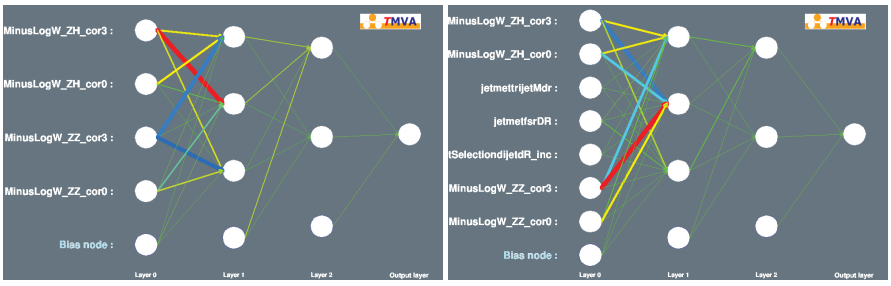

Figure C.56: TMVA MLP architecture ZZ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

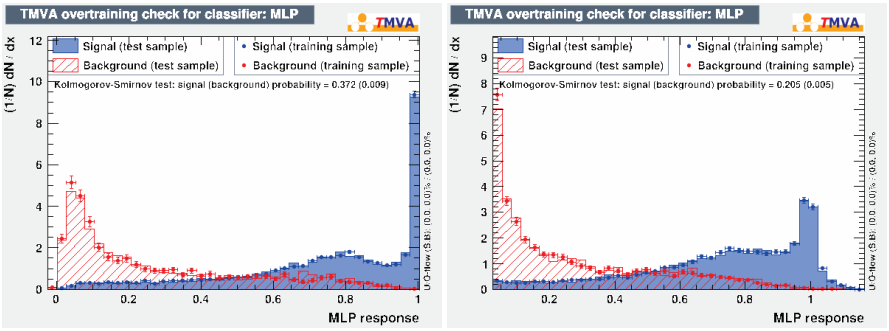

Figure C.57: TMVA MLP output $t\bar{t}$ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

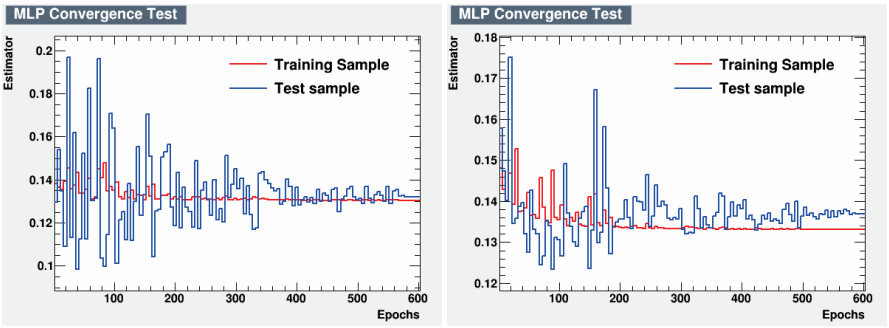

Figure C.58: TMVA MLP training convergence, for $t\bar{t}$ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

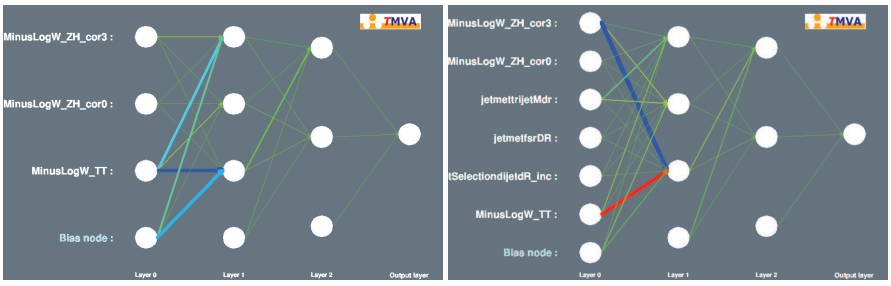

Figure C.59: TMVA MLP architecture $t\bar{t}$ versus ZH, for the 2-jets (left) and 3-jets (right) categories.

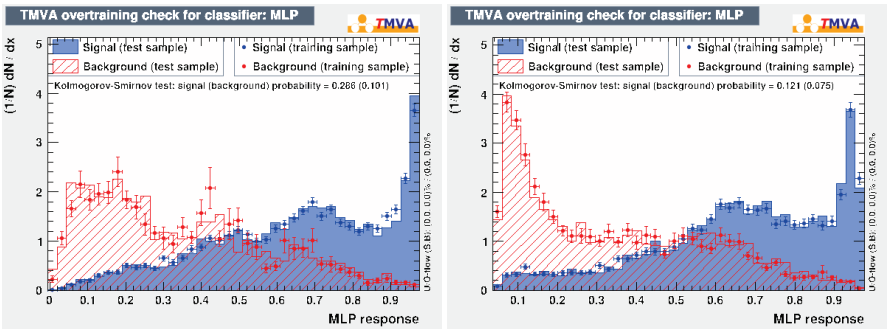

Figure C.60: TMVA MLP output ZH versus all backgrounds, for the 2-jets (left) and 3-jets (right) categories.

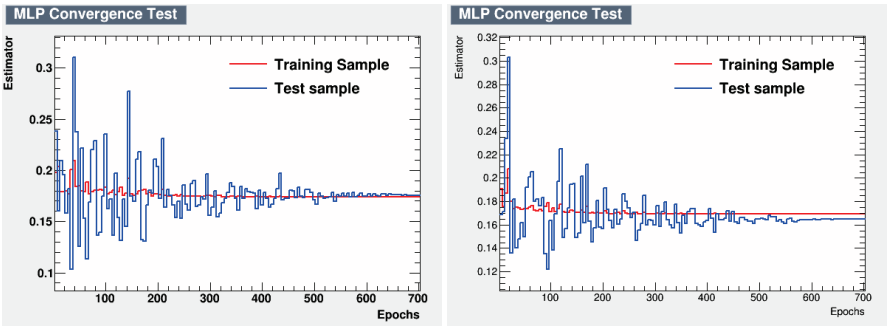

Figure C.61: TMVA MLP training convergence, for ZH versus all backgrounds, for the 2-jets (left) and 3-jets (right) categories.

C.8 Pool plots for CSV

Distribution of the fit performed to obtain the limit on the ZH signal strength, before (left plots) and after(right plots) including the systematics, for the CSV selection are shown on Fig. C.62 and Fig. C.63, respectively for electrons and muons.

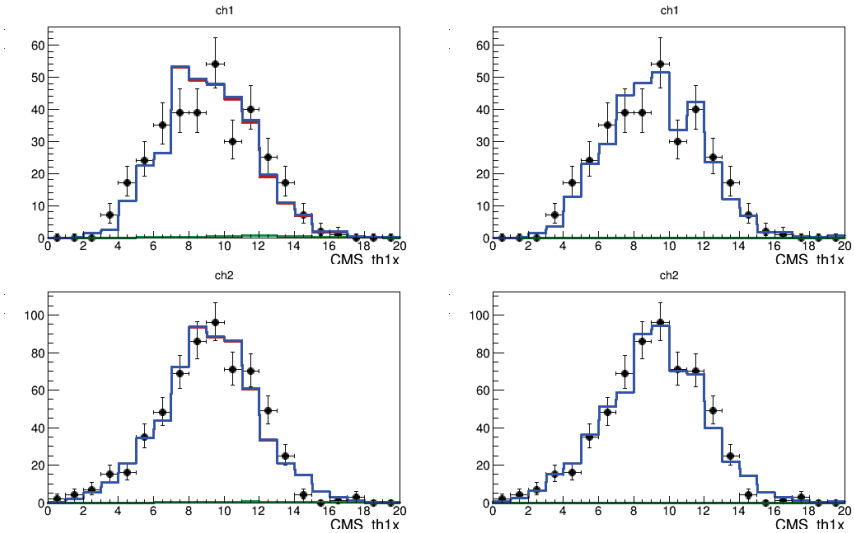

Figure C.62: Distribution of the final discriminant before (left) and after (right) fit, for the electron channel in the 2-jets (top) and 3-jets category, using the CSV tagger.

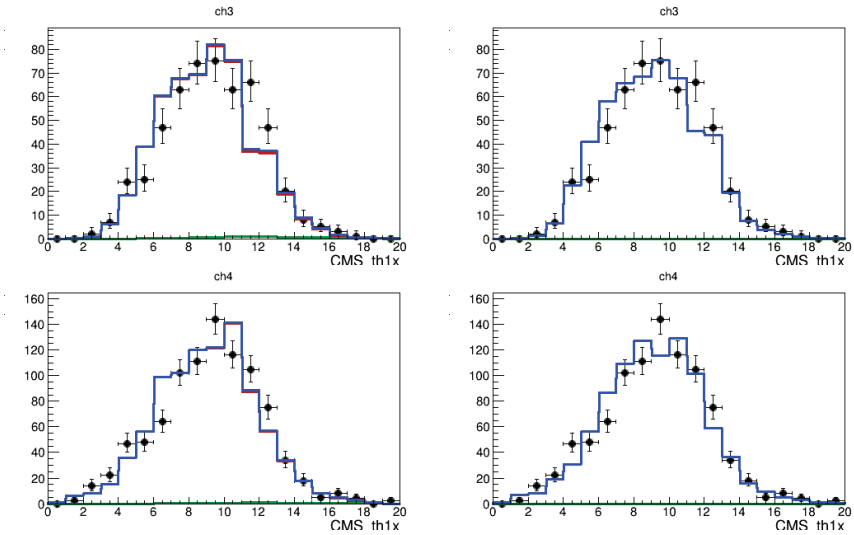

Figure C.63: Distribution of the final discriminant before (left) and after (right) fit, for the muon channel in the 2-jets (top) and 3-jets category, using the CSV tagger.

C.9 Systematics extra information

C.9.1 Yields for systematics

In this section , the yields obtained for the different systematics are shown: for the variation up/down of the Jet Energy Scale (JES), of the Jet energy Resolution (JER), of the scale factor applied to correct for the b-tagging efficiency ($Btag\_bc$) and for mis-tagging efficiency ($Btag\_light$), for JP (Table C.1) and CSV (Table C.2).

Table C.1: Data yields for the FR, normalized to the theoretical cross section. The variation from the nominal yields in the FR is indicated. The tagger used is **JP**.

	Z+bb	Z+bx	Z+xx	$t\bar{t}$	ZZ
JES^{+}	1969.0 $\pm$ 44.4	210.8 $\pm$ 14.5	300.8 $\pm$ 17.3	928.2 $\pm$ 30.5	83.8 $\pm$ 9.2
Variation (%)	+5.2	+7.4	+7.8	+0.3	+4.5
JES^{-}	1858.2 $\pm$ 43.1	199.3 $\pm$ 14.1	279.0 $\pm$ 16.7	902.8 $\pm$ 30.0	76.6 $\pm$ 8.8
Variation (%)	-0.3	+1.5	-	-2.5	-4.7
JER^{+}	1963.4 $\pm$ 44.3	215.5 $\pm$ 14.7	304.9 $\pm$ 17.5	913.9 $\pm$ 30.2	80.1 $\pm$ 8.9
Variation (%)	+5.3	+9.8	+9.3	-1.3	-
JER^{-}	1978.5 $\pm$ 44.5	209.7 $\pm$ 14.5	295.6 $\pm$ 17.2	917.2 $\pm$ 30.3	80.6 $\pm$ 9.0
Variation (%)	+6.1	+6.9	+6.0	-0.9	+0.5
$Btag\_bc^{+}$	1993.3 $\pm$ 44.6	212.9 $\pm$ 14.6	299.0 $\pm$ 17.3	978.3 $\pm$ 31.3	86.3 $\pm$ 9.3
Variation (%)	+6.9	+8.5	+7.2	+5.7	+7.6
$Btag\_bc^{-}$	1733.1 $\pm$ 41.6	186.1 $\pm$ 13.6	260.1 $\pm$ 16.1	855.2 $\pm$ 29.2	74.4 $\pm$ 8.6
Variation (%)	-7.6	-5.4	-7.2	-8.2	-8.0
$Btag\_light^{+}$	1858.4 $\pm$ 43.1	90.4 $\pm$ 9.5	118.1 $\pm$ 10.9	909.0 $\pm$ 30.1	78.8 $\pm$ 8.9
Variation (%)	-0.3	-54.0	-57.7	-1.8	-1.8
$Btag\_light^{-}$	1858.4 $\pm$ 43.1	90.4 $\pm$ 9.5	118.1 $\pm$ 10.9	909.0 $\pm$ 30.1	78.8 $\pm$ 8.9
Variation (%)	-0.3	-54.0	-57.7	-1.8	-1.8

C.9.2 Background fit for systematics

In this section , the SF obtained for the different systematics are shown on Table C.3: for the variation up/down of the Jet Energy Scale (JES), of the Jet energy Resolution (JER), of the scale factor applied to correct for the b-tagging efficiency ($Btag\_bc$) and for mis-tagging efficiency ($Btag\_light$). For the MEM related systematics, the SF are shown only for the JP selection on Table C.4.

Table C.2: Data yields for the FR, normalized to the theoretical cross section. The variation from the nominal yields in the FR is indicated. The tagger used is **CSV**.

	Z+bb	Z+bx	Z+xx	$t\bar{t}$	ZZ
JES^{+}	2717.1 $\pm$ 52.1	392.6 $\pm$ 19.8	664.0 $\pm$ 25.7	1219.9 $\pm$ 34.9	118.7 $\pm$ 10.9
Variation (%)	+6.5	+7.7	+9.9	+0.5	+5.0
JES^{-}	2400.8 $\pm$ 49.0	341.6 $\pm$ 18.5	555.9 $\pm$ 23.6	1176.4 $\pm$ 34.3	107.4 $\pm$ 10.4
Variation (%)	-6.3	-6.8	-8.7	-2.6	-5.3
JER^{+}	2549.2 $\pm$ 50.5	377.1 $\pm$ 19.4	614.7 $\pm$ 24.8	1193.1 $\pm$ 34.5	112.8 $\pm$ 10.6
Variation (%)	+-0.1	+3.4	+1.8	-1.2	-0.3
JER^{-}	2540.8 $\pm$ 50.4	361.2 $\pm$ 19.0	604.3 $\pm$ 24.6	1196.8 $\pm$ 34.6	113.6 $\pm$10.7
Variation (%)	-0.4	-1.0	-	-0.9	-
$Btag\_bc^{+}$	2656.1 $\pm$ 51.5	381.6 $\pm$ 19.5	629.5 $\pm$ 25.1	1227.1 $\pm$ 35.0	118.3 $\pm$ 10.9
Variation (%)	+4.0	+4.7	+4.2	+1.6	+4.6
$Btag\_bc^{-}$	2443.5 $\pm$ 49.4	354.7 $\pm$ 18.8	576.4 $\pm$ 24.0	1147.7 $\pm$ 33.9	108.0 $\pm$ 10.4
Variation (%)	-4.4	-2.8	-4.8	-5.2	-4.7
$Btag\_light^{+}$	2552.3 $\pm$ 50.5	392.1 $\pm$ 19.8	668.6 $\pm$ 25.9	1182.7 $\pm$ 34.4	113.7 $\pm$ 10.7
Variation (%)	-	+7.5	+10.7	-2.1	+0.5
$Btag\_light^{-}$	2545.0 $\pm$ 50.4	344.0 $\pm$ 18.5	544.6 $\pm$ 23.3	1191.9 $\pm$ 34.5	112.5 $\pm$ 10.6
Variation (%)	-0.3	-6.0	-10.9	-1.3	-0.5

C.10 Background normalization uncertainty

A data-driven technique has been used in order to obtain the four background-normalization scale factors SF_{Zbb}, SF_{Zbx}, SF_{Zxx} and $SF_{t\bar{t}}$, as described in Section 4.3.

The SF can be considered as a set of four vectorial variables that define a non-orthogonal basis. Their uncertainties are given by the following matrix:

$$\epsilon = \begin{pmatrix} \sigma_{t\bar{t}} & 0 & 0 & 0 \\ 0 & \sigma_{Zbb} & 0 & 0 \\ 0 & 0 & \sigma_{Zbx} & 0 \\ 0 & 0 & 0 & \sigma_{Zxx} \end{pmatrix} \tag{C.4}$$

where $\sigma_{t\bar{t}}$, σ_{Zbb}, σ_{Zbx} and σ_{Zxx} represent the relative statistical uncertainties on the scale factors, correlated between them. The goal is transformed into a second set of uncorrelated uncertainties, each of them affecting several of the considered contributions.

The correlation factors between the SF are given as an output of the fit and represented by the diagonal matrix:

$$Cor = \begin{pmatrix} 1 & c_{12} & c_{13} & c_{14} \\ - & 1 & c_{23} & c_{24} \\ - & - & 1 & c_{34} \\ - & - & - & 1 \end{pmatrix} \tag{C.5}$$

For JP,

$$Cor_{JP} = \begin{pmatrix} 1 & -0.135 & -0.191 & -0.018 \\ - & 1 & 0.093 & -0.324 \\ - & - & 1 & -0.438 \\ - & - & - & 1 \end{pmatrix} \tag{C.6}$$

For CSV,

$$Cor_{CSV} = \begin{pmatrix} 1 & -0.316 & -0.424 & 0.055 \\ - & 1 & 0.194 & -0.320 \\ - & - & 1 & -0.461 \\ - & - & - & 1 \end{pmatrix} \tag{C.7}$$

The diagonal covariant matrix can then be expressed:

$$Cov = \begin{pmatrix} \sigma^2_{t\bar{t}} & c_{12} \times \sigma_{t\bar{t}} \times \sigma_{Zbb} & c_{13} \times \sigma_{t\bar{t}} \times \sigma_{Zbx} & c_{14} \times \sigma_{t\bar{t}} \times \sigma_{Zxx} \\ - & \sigma^2_{Zbb} & c_{23} \times \sigma_{Zbb} \times \sigma_{Zbx} & c_{24} \times \sigma_{Zbb} \times \sigma_{Zxx} \\ - & - & \sigma^2_{Zbx} & c_{34} \times \sigma_{Zbx} \times \sigma_{Zxx} \\ - & - & - & \sigma^2_{Zxx} \end{pmatrix} \tag{C.8}$$

A new basis defined by orthogonal vectors can be obtained introducing the transformation matrix T that diagonalizes the matrix Cov such as $T{-}1CovT$ =D, where D is the diagonal matrix with eigenvalues of Cov in the diagonal. The transformation matrix allows then to express the scale factors in a new basis, where they are uncorrelated, such as:

$$\epsilon^{'} = T^{-1}\epsilon T \tag{C.9}$$

The final $\epsilon^{'}$ matrix is then used for the systematic determination. For the JP tagger,

$$\epsilon^{'}_{JP} = \begin{pmatrix} 0.999986 & 0.993127 & 0.992372 & 1.14628 \\ 1.00537 & 0.947894 & 1.00842 & 0.985156 \\ 1.01491 & 0.995875 & 0.955903 & 0.977526 \\ 1.03197 & 1.01067 & 1.01915 & 1.01304 \end{pmatrix} \tag{C.10}$$

and for the CSV tagger

$$\epsilon^{'}_{CSV} = \begin{pmatrix} 1.00249 & 0.99046 & 0.98851 & 1.09533 \\ 1.02739 & 0.970966 & 0.981492 & 0.973434 \\ 1.01102 & 1.03613 & 0.970716 & 0.99687 \\ 1.02966 & 1.0142 & 1.02894 & 1.01769 \end{pmatrix} \tag{C.11}$$

Table C.3: Scale factors obtain by the 2D simultaneous fit, for the events selected with the JP/CSV tagger, for a given systematic.

JES^{+}	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.092 ± 0.06	1.292 ± 0.06	1.63 ± 0.24	1.056 ± 0.04
	CSV			
	1.173 ± 0.06	1.216 ± 0.06	1.245 ± 0.14	0.0982 ± 0.04
JES^{-}	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.025 ± 0.06	1.217 ± 0.06	1.30 ± 0.24	0.984 ± 0.04
	CSV			
	1.119 ± 0.06	1.482 ± 0.06	1.489 ± 0.14	1.019 ± 0.04
JER^{+}	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.141 ± 0.06	1.158 ± 0.06	1.429 ± 0.24	1.061 ± 0.04
	CSV			
	1.187 ± 0.06	1.326 ± 0.06	1.295 ± 0.14	1.004 ± 0.04
JER^{-}	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.069 ± 0.06	1.189 ± 0.06	1.586 ± 0.24	1.055 ± 0.04
	CSV			
	1.151 ± 0.06	1.356 ± 0.06	1.333 ± 0.14	0.996 ± 0.04
$Btag\_bc^{+}$	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.257 ± 0.06	1.185 ± 0.06	1.54 ± 0.24	0.976 ± 0.04
	CSV			
	1.095 ± 0.06	1.300 ± 0.06	1.297 ± 0.14	0.992 ± 0.04
$Btag\_bc^{-}$	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.191 ± 0.06	1.372 ± 0.06	1.73 ± 0.24	1.116 ± 0.04
	CSV			
	1.194 ± 0.06	1.409 ± 0.06	1.43 ± 0.14	1.057 ± 0.04
$Btag\_light^{+}$	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.190 ± 0.06	1.230 ± 0.06	4.33 ± 0.24	1.030 ± 0.04
	CSV			
	1.137 ± 0.06	1.345 ± 0.06	1.204 ± 0.14	1.026 ± 0.04
$Btag\_light^{-}$	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
	JP			
	1.190 ± 0.06	1.280 ± 0.06	4.33 ± 0.24	1.030 ± 0.04
	CSV			
	1.149 ± 0.06	1.358 ± 0.06	1.54 ± 0.14	1.021 ± 0.04

Table C.4: Scale factors obtain by the 2D simultaneous fit, for the events selected with the JP tagger, for MEM related systematic.

	SF_{Zbb}	SF_{Zbx}	SF_{Zxx}	$SF_{t\bar{t}}$
MW JES^{+}	1.111 $\pm$ 0.06	1.218 $\pm$ 0.06	1.50 $\pm$ 0.24	1.065 $\pm$ 0.04
MW JES^{-}	1.086 $\pm$ 0.06	1.218 $\pm$ 0.06	1.50 $\pm$ 0.24	1.039 $\pm$ 0.04
MW JER^{+}	1.143 $\pm$ 0.06	1.212 $\pm$ 0.06	1.50 $\pm$ 0.24	1.061 $\pm$ 0.04
MW JER^{-}	1.064 $\pm$ 0.06	1.238 $\pm$ 0.06	1.52 $\pm$ 0.24	1.057 $\pm$ 0.04

Appendix D

Multi-Variate Tools

In this section, the two MVA tools used in this thesis are introduced. They have been exploited via the Toolkit for Multivariate Data Analysis (TMVA) tool of ROOT. More details about the TMVA tool can be found in [88]

D.1 Boosted Decision Tree

A decision tree is a sequence of binary splits of the data, that has been trained with a set of known events. The results are measured using a different set of testing events. From one "node" composed by all the data events, a best cut is found to separate signal from background events, creating two new nodes. The process is repeated on these two nodes and is continued until a given number of final ending nodes, called "leaves", is obtained, or until all leaves are pure or one node has too few events.
If events are considered having a weight W_i, the purity P of the node is defined as the weight of signal events on the leaf divided by the total weight of events on that node. Then, for each node, a criterion C is defined such as $C = P(1-P)\sum_i W_i$. C is 0 when P=1 or P=0, and the best split is chosen such as $C_{daughter-node1} + C_{daughter-node2}$ is minimized, and the following leaf split is chosen by finding the one whose splitting maximize the change in C. In this way, a decision tree is built, and leaves with P $\geq$0.5 are signal leaves while the rest are background leaves.
However, decision tree are unstable and a small change in the training tree can lead

to a sustainable change in the final tree. This is remedied by the use of boosting: the training events that are misidentified have their weights boosted and a new tree is formed. This procedure is then repeated for the new tree. In this way, several trees are built; the score of the nth individual tree T_n is taken as +1 if the event falls on a signal leaf and -1 if the event falls on a background leaf. The final score is taken as a weighted sum of the scores of the individual leaves, and used as a weight.

D.2 Neural Network

A MLP is a simple feed-forward network made of neurons characterized by a bias and weighted links between them, as it can be seen on Fig. D.1. The input variables are associated to the input neurons, which normalize them and forward them to the first hidden layer. Neurons in any other layer than the input compute a linear combination of the outputs of the previous layer. The output of a neuron is given by a function of this linear combination. This function is a sigmoid for the hidden layers, while it is a linear function for the output neurons.

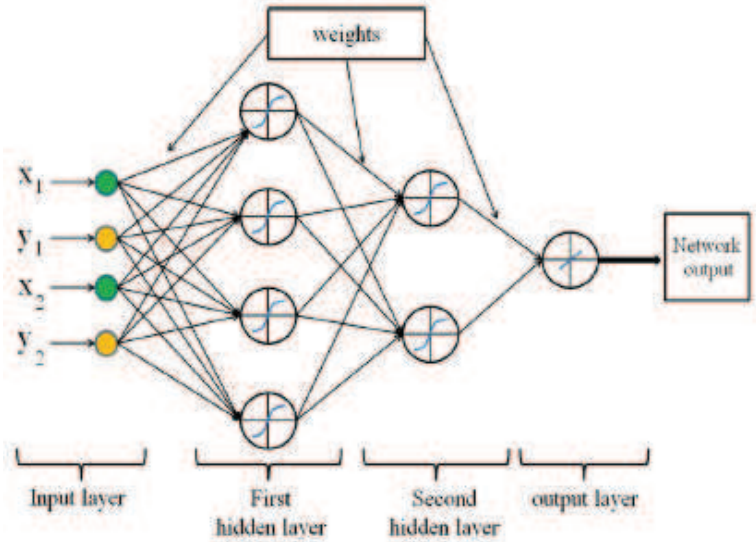

Figure D.1: Structure of a MLP-type Neural Network.

ring the learning process, a total error, defined as the sum in quadrature of the error on each individual output neuron, is minimized.

Bibliography

[1] Steven Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* ***19*** *(1967) 1264*.

[2] Sheldon L. Glashow, "Partial-symmetries of weak interactions", *Nucl. Phys.* ***22*** *(1961) 579*.

[3] J. Goldstone A. Salam and S. Weinberg, "Broken Symmetries", *Phys. Rev.* ***127*** *(1962) 965*.

[4] K. Nakamura et al (Particle Data Group), "Review of Particle Physics", *Journal of Physics G* ***37*** *075021 (2010)*.

[5] UA1 Collaboration, "Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s}$=540 GeV", *Phys. Lett. B 122 (1983) 103-116*.

[6] UA2 Collaboration, "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN $p\bar{p}$ collider", *Phys.Lett. B122 (1983)*.

[7] F. Abe et al. (CDF Collaboration), "Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab", *Phys.Rev.Lett.74:2626-2631,1995*.

[8] The 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*.

[9] The 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*.

[10] The 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*.

[11] Sven-Olaf Moch, "Quantum Chromodynamics", *BND school 2012, QCD lecture I*.

[12] Jörg Resag, "Private web-page", `http://www.joergresag.privat.t-online.de/`.

[13] T. D. Lee and C. N. Yang, "Question of Parity Conservation in Weak Interactions", *Phys. Rev.B 104 254 1956*.

[14] P.W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* ***12*** *(1964) 132*.

[15] P.W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons", *Phys. Rev.* ***145*** *(1966) 1156*.

[16] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* ***13*** *(1964) 321*.

[17] Luis Alvarez-Gaume and John Ellis, "Eyes on a prize particle", *Nature Physics (2011)*.

[18] The CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV ", *Eur. Phys. J. C* ***75*** *(2015) 212*.

[19] The CMS Collaboration, "On the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs.", *Phys.Rev.Lett. 110:081803*.

[20] LHC Higgs Cross Section Working Group, "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables", *CERN-2011-002*.

[21] J. Pumplin D.R. Stump J. Huston H.L. Lai Pavel M. Nadolsky and W.K. Tung, "New generation of parton distributions with uncertainties from global QCD analysis ", *JHEP 0207 (2002) 012*.

[22] George Sterman) John C. Collins, Davison E. Soper, "Factorization of Hard Processes in QCD", *Adv.Ser.Direct.High Energy Phys.5:1-91,1988*.

[23] J. Alwall R. Frederix S. Frixione V. Hirschi F. Maltoni O. Mattelaer H.-S. Shao T. Stelzer P. Torrielli M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP07(2014)079*.

[24] T. Gleisberg S. Hoeche F. Krauss M. Schoenherr S. Schumann F. Siegert J. Winter, "Event generation with SHERPA 1.1.", *JHEP02 (2009) 007*.

[25] Torbjörn Sjöstrand Stephen Mrenna and Peter Skands, "PYTHIA 6.4 physics and manual", *JHEP05(2006)*.

[26] M.L. Mangano M. Moretti F. Piccinini R. Pittau FA. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions", *JHEP 0307:001,2003*.

[27] D0 Collaboration, "A precision measurement of the mass of the top quark", *Nature 429, 638-642*.

[28] S. Chatrchyan et al. (CMS Collaboration), "Study of the Mass and Spin-Parity of the Higgs Boson Candidate via its Decays to Z Boson Pairs", *Phys. Rev. Lett.110, 081803*.

[29] The CMS Collaboration, "Search for $t\bar{t}$H events in the H$\rightarrow b\bar{b}$ final state using the Matrix Element Method", *CMS-PAS-HIG-14-010*.

[30] Pierre Artoisenet Vincent Lemaître Fabio Maltoni Olivier Mattelaer, "Automation of the matrix element reweighting method", *Journal of High Energy Physics 2010:68*.

[31] T. Ohl, "Vegas revisited: Adaptive Monte Carlo integration beyond factorization", *Computer Physics Communications 120 (1): 13â19*.

[32] C. Beluffi, "The Matrix Element Method within CMS", 2014.

[33] "ATLAS collaboration web-site", `http://atlas.ch/`.

[34] "CMS collaboration web-site", `http://cms.web.cern.ch/`.

[35] "ALICE collaboration web-site", `http://aliceinfo.cern.ch/`.

[36] "LHC-b collaboration web-site", `http://lhcb-public.web.cern.ch/lhcb-public/`.

[37] "CERN web-site", `http://home.web.cern.ch/`.

[38] "HL-LHC project web-site", `http://hilumilhc.web.cern.ch/`.

[39] Rene Brun and Fons Rademakers, "ROOT - An Object Oriented Data Analysis Framework", *Sep. 1996, Nucl. Inst. Meth. in Phys. Res. A 389 (1997) 81-86*.

[40] The CMS Collaboration, "CMS Physics : Technical Design Report Volume 1: Detector Performance and Software", *CERN-LHCC-2006-001,*, 2 February 2006.

[41] The CMS Collaboration, "CMS Physics : Technical Design Report Volume 2: Physics Performance", 25 June 2006.

[42] The CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST **9** (2014) P10009*.

[43] The CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST **9** (2014) P10009*.

[44] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. and Methods A 262, 444 (1987*.

[45] V. Veszpremi on behalf of the CMS collaboration, "Operation and performance of the cms tracker", *JINST 9 C03005*.

[46] The CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$=8 TeV", *CMS-EGM-13-001*.

[47] The CMS Collaboration, "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC", *CMS-NOTE-2005-001*.

[48] The CMS Collaboration, "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV", *CMS-EGM-14-001*.

[49] The CMS Collaboration, "Electron performance with 19.6 fb^{-1} of data collected at $sqrts$=8 TeV with the CMS detector", *CMS DP -2013/003*.

[50] Matteo Cacciari Gavin P. Salam Gregory Soyez, "The anti-k_t jet clustering algorithm", *JHEP 0804:063,2008*.

[51] The CMS Collaboration, "Determination of the Relative Jet Energy Scale at CMS from Dijet Balance", *CMS-PAS-JME-08-003*.

[52] The CMS Collaboration, "Jet Energy Correction Using Z ($\rightarrow e^+e^-$) + Jet p_T Balance and the Method for Combining Data Driven Corrections", *CMS-AN-2009/004*.

[53] The CMS Collaboration, "Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s}$ = 8 TeV", *J. Instrum. 10 (2015) P02006*.

[54] The CMS Collaboration, "Description and Performance of $\vec{E}_t^{miss}$ Significance in 2012 Data ", *CMS-AN-13-173*.

[55] N. Adam J. Berryhill V. Halyo A. Hunt and K. Mishra, "Generic Tag and Probe Tool for Measuring Efficiency at CMS with Early Data", *CMS-AN-2009/111*.

[56] The CMS Collaboration, "The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s}$ = 7 TeV at the LHC ", *JINST **8** (2013) P11002*.

[57] The CMS Collaboration, "First two months of data taking at 7 TeV: $J/\psi \to \mu\mu$, W $\to \mu\nu$, Z $\to \mu\mu$ mass plots and displays of Z $\to \mu\mu$ candidates".

[58] J. Brooke on behalf of the CMS Collaboration, "Performance of the CMS Level-1 Trigger", *CMS-CR-2012-322*.

[59] Daniele Trocino on behalf of the CMS Collaboration, "The CMS High Level Trigger", *Journal of Physics: Conference Series 513 (2014) 012036*.

[60] S. Agostinelliae J. Allisonas K. Amakoe and al., "Geant4 - a simulation toolkit", *Nuclear Instruments and Methods in Physics Research A 506 (2003) 250-303*.

[61] The CMS Collaboration, "The CMS Particle Flow Algorithm", *Proceedings of the CHEF2013 Conference*.

[62] The CMS Collaboration, "Particle-flow Event Reconstruction in CMS and Performance for Jets, Taus, and E_T^{miss}", *CMS PAS PFT-09-001*.

[63] Boris Mangano, "The CMS particle flow algorithm in CMS".

[64] C. Beluffi R. Castello A. Caudron C. Delaere T.A. du Pree V. Lemaitre A. Pin and J. Vizan, "A Search for the SM Scalar Boson in the Z(ll)H(bb) Final State using a Matrix Element Method", *CMS-AN-12-476*.

[65] The DELPHES collaboration J. de Favereau C. Delaere P. Demin A. Giammanco V. Lemaître A. Mertens and M. Selvaggi), "DELPHES 3: a modular framework for fast simulation of a generic collider experiment", *JHEP **02** (2014) 057*.

[66] Jérôme H. Friedman, "Data analysis techniques for high energy particle physics", *Lectures presented at the CERN School of Computing, Godoysund, Norway, August 11-24, 1974*.

[67] The CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST **8** (2013) P04013*.

[68] The CMS Collaboration, "Performance of b tagging at $\sqrt{s}$= 8 TeV in multijet, $t\bar{t}$ and boosted topology events", *CMS-BTV-13-001*.

[69] C. Beluffi, "b-jet identification in CMS", *CMS CR-2014/192*, 2014.

[70] D. Spiga S. Lacaprara W. Bacchi and al., "The CMS Remote Analysis Builder (CRAB)", *High Performance Computing - HiPC 2007, pp 580-586*.

[71] "Jet Probability Calibration Twiki Page", https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagJetProbabilityCalibration.

[72] C. Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", *CMS NOTE 2006/014*.

[73] "Commissioning Twiki Page", https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagCommissioning2012.

[74] "Pileup reweighting utilities twiki page", https://twiki.cern.ch/twiki/bin/view/CMS/PileupMCReweightingUtilities.

[75] "Trackhistory detail information", https://twiki.cern.ch/twiki/bin/viewauth/CMS/TrackHistoryDetailInfo.

[76] C. Beluffi and J. Andrea, "Calibration of the jet probability and prospects at high p_T", *CMS AN-2014/237*.

[77] Arnaud Pin, "The Matrix Element Method at the LHC: a search for the associated production of Higgs and Z bosons".

[78] The 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*.

[79] Alexander Mitov Michal Czakon, Paul Fiedler, "The total top quark pair production cross-section at hadron colliders through O(α_S^4)", *Phys. Rev. Lett. **110** (2013) 252004*.

[80] The CMS Collaboration, "Measurement of ZZ production cross section in ZZ to 2l2l' decay channel in pp collisions at $\sqrt{s}$ = 8 TeV", *Technical Report CMS-PAS-SMP-12-014, CERN, Geneva, (2012)*.

[81] The LHC Higgs Cross Section Working Group, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties".

[82] "Methods to apply b-tagging efficiency scale factors", https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagSFMethods.

[83] The CMS Collaboration, "Search for H/A decaying into Z and A/H, with Z to ll and A/H to fermion pair", *CMS PAS HIG-15-001*.

[84] Glen Cowan Kyle Cranmer Eilam Gross and Ofer Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J **C71** (2011) 1554*.

[85] The 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*** *012003 (2014)*.

[86] The CMS Collaboration, "Measurement of the W^+W^- cross section in pp collisions at $\sqrt{s}$ = 8 TeV and limits on anomalous gauge couplings", *CERN-PH-EP-2015-122*.

[87] The CMS Collaboration, "Measurement of WZ production rate", *CMS-PAS-SMP-12-006*.

[88] A. Hoecker P. Speckmayer J. Stelzer J. Therhaag E. von Toerne and H. Voss, "TMVA Toolkit for Multivariate Data Analysis with ROOT - User guide", *CERN-OPEN-2007-007*.

Recherche de processus rares avec la signature Z+bb au LHC, à l'aide de la Méthode et Éléments de Matrice: **Résumé en français**

Camille Beluffi

Thèse soutenue le 14 octobre 2015

1 Contexte théorique et expérimental

1.1 Contexte théorique et axe de recherche

Le Modèle Standard (MS) [1] [2] [3] est une théorie développée dans les années 60 et basée sur la mécanique quantique, permettant de décrire les interactions entre les particules fondamentales. Ces particules sont divisées en deux groupes : les fermions, composés de trois générations de particules de spin $\frac{1}{2}$, et les bosons, vecteurs des interactions, de spin entier. Ces interactions sont régies par les trois forces présentes : force forte, véhiculée par les gluons, force électromagnétique échangée par l'intermédiaire de photons, et force électrofaible, véhiculée par les bosons Z et W. Ces derniers acquièrent une masse grâce au mécanisme de Brout-Englert-Higgs [4] [5] [6], qui donne également naissance à une particule appelée le boson de Higgs.

Cette particule a été découverte par les expériences CMS et ATLAS le 4 juillet 2013 [7] [8] [9]. Depuis, ses propriétés ont été étudiées et mesurées, et les résultats sont jusqu'à présent compatibles avec ceux attendus pour le boson de Higgs prédit par le Modèle Standard [10] [11]. Cependant, seul son couplage avec les bosons a été mis en évidence, le couplage avec les fermions étant plus difficile à établir. Ce dernier demeure toutefois fondamental à prouver et à mesurer pour confirmer la nature Modèle Standard du boson de Higgs. En particulier, le couplage aux quarks b est un sujet d'étude pertinent afin de confirmer le couplage du boson de Higgs aux quarks de type "bas". Il est à noter que le boson de Higgs se désintègre préférentiellement en deux quarks b lorsqu'il est produit à une masse de 125 GeV.

Malgré le succès du Modèle Standard, des questions subsistent : pourquoi y a t-il trois générations de fermions ? Pourquoi y a t-il plus de matière que d'antimatière dans l'Univers ? Autant de questions auxquelles on ne peut répondre pour le moment et qui nécessitent de rechercher des phénomènes de nouvelle physique.

1.2 Le LHC et CMS

Le plus grand accélérateur de particules actuel est le Grand Collisionneur de Hadron (Larhe Hadron Collider, LHC en anglais) situé à Genève au CERN (Centre Européen pour la Recherche Nucléaire), où des faisceaux de protons entrent en collisions à très hautes énergies : 7 TeV au démarrage du LHC, 8 TeV en 2012, et depuis le printemps 2015, 13 TeV. Pour cette thèse, les données produites à une énergie de 8 TeV uniquement ont été exploitées. Autour de l'anneau du LHC sont placés quatre détecteurs qui enregistrent les produits de ces collisions.

CMS (Compact Muon Solenoïd) [12] est l'un de ces détecteurs, dédié entre autres à la recherche du boson de Higgs. Ce détecteur est composé de plusieurs sous-détecteurs, disposés en couches afin d'identifier les particules les traversant et de mesurer leurs propriétés [13] [14]. Au centre se trouve le trajectographe, qui détecte le passage des particules chargées. Autour, le calorimètre électromagnétique (ECAL) permet de mesurer l'énergie déposée par les électrons et photons. Le calorimètre hadronique (HCAL) est disposé juste après

afin de déterminer l'énergie des jets, générés lors de l'hadronisation des quarks. La partie la plus externe du détecteur est dédiée à l'identification des muons par l'intermédiaire de différentes chambres à muons. Entre ces dispositifs et HCAL se trouve un aimant supra-conducteur produisant un champ magnétique de 3.8 Tesla, permettant ainsi de courber la trajectoire des particules chargées.

Les données sont enregistrées à l'aide d'un système de déclenchement, puis des algorithmes complexes (tels que l'algorithme de Particle Flow [15] ou anti-k_T [16] pour la reconstruction des jets) permettent de reconstruire les différentes particules produites durant les collisions et de déterminer leurs propriétés. Ces données sont enfin stockées dans différents centres à travers le monde et rendues accessibles aux scientifiques via le système de la grille.

1.3 La Méthode des Éléments de Matrice

Afin de distinguer le signal des événements de bruit de fond, un outil appelé la Méthode des Éléments de Matrice (MEM) est utilisé. Cette méthode permet de calculer la probabilité qu'un événement reconstruit corresponde à une hypothèse théorique donnée. Cette probabilité contient le maximum d'informations concernant le processus dur, via l'élément de matrice associé à l'hypothèse testée et les fonctions de probabilité traduisant la composition des protons entrants. Des informations relatives aux effets de reconstruction sont introduites par le biais des fonctions de transfert qui représentent la transition de l'état final généré à l'état final reconstruit (ces fonctions prennent donc en compte, entre autres, les effets de reconstruction, de résolution et l'hadronisation des jets).
La probabilité P renvoyée par la MEM est calculée à l'aide du logiciel MadWeight [17]. Par la suite, la variable utilisée est le "poids", proportionnel à $-\log(P)$.

Deux lots de fonctions de transfert ont été créés lors de cette thèse. Pour le premier lot, les fonctions sont déterminées en ajustant une double Guaussienne aux distribuons attendues de différence d'énergie entre l'état généré et l'état reconstruit, pour plusieurs énergies générées. Ces fonctions de transfert ont l'avantage de très bien reproduire les queues de distributions mais leur détermination est complexe, de part le grand nombre de paramètres à ajuster. C'est pourquoi un deuxième lot a été élaboré et préféré pour cette thèse : le lot de fonctions de transfert binées. Celle ci sont directement déterminées à partir d'un histogramme 2D représentant la différence d'énergie entre état généré et reconstruit en fonction de l'énergie générée. Une procédure de lissage est appliquée à l'histogramme pour éliminer les pics résiduels, puis il est renormalisé à l'unité pour chaque tranche en énergie générée.

2 Étude de l'étiquetage des jets b dans l'expérience CMS

L'identification des jets provenant de l'hadronisation des quarks b est crucial dans CMS et pour cette thèse, car ces jets sont attendus dans de nombreuses signatures de nouvelle physique. Cette identification est possible grâce aux caractéristiques très particulières que ces jets présentent. En effet, le hadron B

produit lors de l'hadronisation du quark b va parcourir une distance importante avant de se désintégrer, et cette distance est entachée d'une erreur plus petite que sa valeur. Un vertex secondaire est donc visible, et les traces provenant de celui ci présentent un important paramètre d'impact, correspondant à la distance entre la trace et le vertex de première interaction.

Ces propriétés sont utilisées afin de construire des algorithmes permettant l'identification des jets b. Dans CMS, Combined Secondary Vertex (CSV) [18] est l'algorithme qui donne actuellement les meilleures performances. Il exploite les propriétés liées au vertex secondaire ainsi que le paramètre d'impact des traces, et les combine à l'aide d'une méthode multivariée pour renvoyer une valeur appelée discriminant. Si le candidat jet possède une valeur de discriminant au dessus d'une valeur seuil, il est étiqueté comme jet b.

2.1 Algorithme de probabilité de jets

Un autre algorithme montrant de très bonnes performances est Jet Probability (JP). Cet algorithme est basé sur la partie négative de la distribution du paramètre d'impact des traces pour construire des fonctions de résolution, grâce auxquelles la probabilité que le jet vienne du vertex primaire est évaluée. Puis, pour toutes les traces du jet, ces probabilités sont combinées afin de renvoyer la probabilité que le jet vienne du vertex primaire. Si cette probabilité est faible, le jet est considéré comme étant un b jet.
Un aspect important de JP est sa calibration : en effet, la distribution du paramètre d'impact des traces, utilisée pour définir les fonctions de résolution, varie sensiblement en fonction de la qualité des traces. Afin de prendre en compte cet effet, différentes catégories de traces ont été créées, basées sur des critères relatifs à la qualité de la trace. Un avantage non négligeable découle du fait que cette calibration peut être effectuée sur des événements de données. Un nouveau code a été créé pendant cette thèse et permet de facilement ajouter de nouvelles catégories pour la calibration de JP [19].

2.2 Étude à hautes énergies

Tous les algorithmes d'étiquetage des jets b présentent une décroissance d'efficacité importante lorsque les jets possèdent une impulsion transverse (p_T) au delà de 200 GeV [20]. Étant donné que la nouvelle physique est attendue à hautes énergies, il est important de comprendre la cause de cette perte d'efficacité pour chercher à la réduire.

Dans le cas de JP, la perte d'efficacité observée semble provenir de la combinaison de deux effets : d'une part des traces rattachées au hadron B sont perdues lors du processus de reconstruction et sélection des traces, et d'autre part une augmentation du nombre de traces qui viennent du vertex primaire est observée. Ce sont ces dernières traces qui vont détériorer les performances de l'algorithme à hautes énergies, lorsque la contamination se fait importante.
Un des axes de recherche qui a été suivi consiste à tenter d'éliminer ces traces de l'algorithme JP. Pour cela, une méthode multivariée d'arbre de décision (Boosted

Decision Tree - BDT) est exploitée. Cette méthode combine différentes variables pour lesquelles un bon pouvoir de discrimination entre les traces venant du hadron B et les traces de contamination a été trouvé. Elle renvoie une variable de sortie qui est utilisée comme critère de sélection supplémentaire : si la trace possède une valeur au delà d'un certain seuil, elle sera utilisée dans l'algorithme de JP. Ce faisant, une amélioration des performances à hauteur de 4 % a pu être observée pour des jets avec un p_T moyen de 300 GeV.
Une autre étude a été menée en modifiant l'algorithme Jet B Probability : dans celui ci, les quatre traces avec le plus haut paramètre d'impact ont une plus forte pondération lors du calcul du discriminant. Ces quatre traces sont alors remplacées par les quatre traces ayant la plus haute valeur de sortie de BDT, menant à une amélioration de l'efficacité de l'étiquetage des jets de quelques pour-cent.
Finalement, tirant avantage du nouveau code de calibration élaboré, de nouvelles calibrations de JP ont été testées en utilisant de nouvelles catégories de trace. Ces catégories ont été définies en affinant les coupures appliquées et en introduisant de nouvelles variables, affectant la distributions du paramètre d'impact des traces. Cependant, aucune amélioration notable n'a pu être notée.

3 Recherche du boson de Higgs Modèle Standard produit en association avec un boson Z

3.1 Problématique

Cette thèse propose une étude complémentaire et comparative pour la recherche du boson de Higgs se désintégrant en deux quarks b, basée sur la mesure de section efficace du processus Zbb [21]. Afin de s'affranchir un maximum du bruit de fond généré par les événements de Quantum Chromo Dynamics (QCD), le canal de production du boson de Higgs en association avec un boson Z est préféré (canal ZH) ; en effet ces événements disposent d'une signature propre grâce aux deux leptons provenant de la désintégration du boson Z. Cependant, ce canal présente une section efficace très faible : pour le lot de données enregistré à 8 TeV, environ 500 évènements de signal ont été produits, mais une fois l'acceptance du détecteur, les effets de reconstruction et d'efficacité de sélection pris en compte, seulement moins de 4 % des ces événements seront gardés. Il sera donc difficile de clamer une découverte dans ce canal. Toutefois, la MEM ayant produit des résultats intéressants, elle est utilisée ici pour discriminer le signal de ses principaux bruits de fond. Cette étude est également l'occasion de comparer les performances des algorithmes CSV et JP, afin de déterminer lequel est le plus pertinent à utiliser pour une recherche de nouvelle physique.

3.2 Sélection des événements

Plusieurs processus du MS présentent une signature similaire à celle du signal ZH. Le principal bruit de fond est composé d'événements Drell-Yann (DY) : la production d'un boson Z en association avec deux jets b (événements Zbb). Ces événements peuvent être catégorisés en fonction du nombre de jets b correctement étiquetés. Le deuxième processus de bruit de fond est la production d'une paire de quarks top suivie DE la désintégration leptonique des deux bosons W,

le processus $t\bar{t}$ di-leptonique. Le dernier bruit de fond pris en compte est la production de deux bosons Z, l'un se désintégrant en deux leptons, l'autre en deux quarks b (processus ZZ). D'autres événements de bruits de fond peuplent l'espace de phase d'intérêt mais leur contribution étant très faible, ils ont été négligés dans cette étude.

Afin de sélectionner un maximum de signal en gardant le moins de bruit de fond possible, la sélection suivante est appliquée : les évènements contenant deux leptons dans l'acceptance du détecteur et du système de déclenchement, avec une masse invariante compatible avec celle d'un boson Z, sont gardés. De plus, ces événements doivent contenir deux jets ayant été étiquetés par l'algorithme CSV ou JP avec un point de fonctionnement correspondant à un taux de mauvaise identification de 1 %. Une catégorisation des événements est faite en fonction du nombre de jets supplémentaires : si l'événement ne contient pas de jets additionnels, il entre dans la catégorie "2 jets". Si il en contient au moins un, il peuple la catégorie "3 jets". Finalement, une coupure sur l'énergie transverse manquante est appliquée afin d'éliminer le bruit de fond $t\bar{t}$.
Plusieurs régions d'étude sont définies : une région large, une région de contrôle dans laquelle sera effectuée une renormalisation des bruits de fond, et une région signal dans laquelle le signal sera mesuré. Dans cette dernière région, une coupure sur la masse invariante des deux jets b est appliquée afin de sélectionner 90 % du signal dans les deux catégories.

Une fois la sélection appliquée, une comparaison entre le nombre d'évènements attendus pour les principaux bruits de fond et les données est effectuée, pour les deux algorithmes. Il apparaît que JP est un algorithme plus pur : il sélectionne en moyenne 35 % d'évènements en moins que CSV, lorsque ceux ci contiennent deux vrais jets b. Cet écart se creuse encore lorsque les événements contiennent au moins un jet ayant été mal étiqueté (jusqu'à trois fois moins d'évènements sélectionnés). Cependant, pour les deux algorithmes, un excès de données de 16 % est observé. Cet excès ne semble pas être dû à la présence de nouvelle physique, puisqu'il est situé au niveau du pic du boson Z dans la distribution de la masse invariante di-leptonique, et à basse énergie transverse manquante. L'excès est également observé sur les distributions du discriminant renvoyé par les algorithmes d'étiquetage, à basses valeurs. Cela tente à incriminer une mauvaise modélisation des événements DY, en particulier ceux qui contiennent au moins un jet mal identifié.

3.3 Renormalisation des bruits de fond

Ainsi, avant de continuer cette étude, il est nécessaire d'effectuer une renormalisation des bruits de fond aux données. Pour se faire, un ajustement est réalisé simultanément sur deux distributions dans les canaux di-électrons et di-muons. Les distributions utilisées sont la sortie d'un réseau de neurones discriminant les processus DY et $t\bar{t}$, et le produit des discriminants des deux b jets ; la première permet de contraindre les processus DY et $t\bar{t}$ entre eux, et la deuxième permet de séparer les différentes composantes des événements DY (les événements contenant des jets mal étiquetés peuplent les basses valeurs).
Pour les deux sélections avec CSV et JP, les facteurs de renormalisation obtenus

pour les processus $t\bar{t}$ et DY sont compatibles avec 1, tandis que ceux correspondant aux contributions DY sont plus élevés.

3.4 Résultats

Finalement, le signal est extrait. Afin de rejeter les évènements de bruit de fond, les poids calculés avec la MEM pour les différentes hypothèses signal et bruit de fond sont utilisés. Ils sont combinés en utilisant un arbre de décision qui renvoie une variable de sortie ; la distribution de cette variable est utilisé pour calculer une limite d'exclusion sur la production du signal (σ_{MS}). Celle ci est déterminée en utilisant la méthode CL_s [22], avec un niveau de confiance de 95 %. Les systématiques liées à cette étude sont inclues dans le calcul de cette limite, la systématique ayant le plus d'impact sur le résultat final étant celle induite par la statistique Monte Carlo (MC) disponible (son impact est de près de 10 %). Lorsque l'algorithme CSV est utilisé, la limite finale observée obtenue est de $5.46 \times \sigma_{MS}$, tandis que lorsque l'algorithme JP est utilisé, la limite calculée est de $4.89 \times \sigma_{MS}$. Cela correspond à un léger excès de données dans le cas de CSV, tandis que les observations sont en accord avec les prédictions dans le cas de JP. Ces résultats sont compatibles avec les autres résultats produits par la collaboration CMS [23] [24], et montrent que les deux algorithmes présentent des performances très similaires.

4 Recherche de nouvelle physique modèle-indépendante

4.1 Enjeux et méthode

L'analyse précédente a permis de mettre en place de nombreux outils et un grand nombre de poids ont été calculés. Sur base du travail déjà effectué, une nouvelle analyse a été établie, ayant pour but d'utiliser les poids calculés par la MEM pour construire une discrimination des processus MS appartenant à l'espace de phase Z+bb (état final composé de deux leptons venant de la désintégration d'un boson Z, de deux jets b et sans énergie transverse manquante). En utilisant une approche récursive, des "boîtes" sont construites afin de catégoriser l'espace de phase d'intérêt. L'algorithme d'étiquetage des jets b utilisé pour cette analyse est CSV, étant donné que cet algorithme a montré des performances légèrement meilleures lors de la précédente analyse.

A partir des distributions de poids W calculés pour les différentes hypothèses théoriques MS (le ZH est inclus, ainsi que les processus di-bosoniques WW et WZ), une variable discriminante D est construite afin de séparer tout processus a d'un processus b :

$$D = \frac{ArcTan(W_a - W_b) + \frac{\pi}{2}}{\pi} \tag{1}$$

Cette définition permet d'avoir une distribution comprise entre 0 et 1. Cette variable constitue la base de l'approche récursive.

Prenons un exemple simple dans lequel l'espace de phase MS Z+bb n'est peuplé que par les trois processus a, b et c, par ordre croissant d'importance.

La méthode commence par discriminer les processus les moins importants entre eux, donc ici a et b. Le processus a étant le plus rare, il sera appelé "signal". La variable D est construite afin de séparer ces deux processus et une tentative de création de deux nouvelles boîtes va avoir lieu. En effet, si en découpant la distribution de D de manière à créer deux boîtes, une dans laquelle une fraction f du signal est gardée, l'autre contenant le reste, on parvient à améliorer la pureté du signal au delà d'un certain seuil cut_{Pur}, la méthode créée ces deux boîtes filles. Sinon, les boîtes ne sont pas générées et la méthode va tenter de discriminer a et c, puis b et c, et si aucun de ces tests n'est fructueux, le processus s'arrête. De façon récursive, une arborescence est ainsi obtenue.

La méthode contient donc trois paramètres libres qui seront ajustés en prenant comme figure de mérite la limite d'exclusion sur la production d'événements ZH : la configuration donnant la meilleure limite sera gardée. Avant cela, des paramètres nominaux sont fixés afin de créer un nombre de boîtes filles raisonnable.

4.2 Catégorisation additionnelle

Une fois les boîtes filles obtenues, une catégorisation est effectuée en fonction de deux facteurs. Tout d'abord, la distribution du produit des discriminants CSV des deux jets b est utilisée de manière à créer deux nouvelles boîtes : une composée d'événements peuplant les basses valeurs de la distribution (boite de saveur "légère"), l'autre boite avec la seconde moitié des événements (boite type "b"). Puis, une catégorisation additionnelle suit : si l'événement contient uniquement les deux jets b, il est placé dans une boite "2 jets", tandis que s'il contient au moins un jet supplémentaire, il va peupler la boite "3 jets".
Finalement, le nombre final de boites est multiplié par quatre. Un histogramme est ensuite rempli, contenant le nombre d'événements de chaque boite finale, les boites ayant été regroupées suivant leur catégorie ("légère-2 jets", "type b-3 jets", etc...). Cet histogramme sera utilisée comme discriminant final.

4.3 Sensibilité de la méthode

Une méthode Monte Carlo est utilisée afin de calculer la précision statistique que l'on peut attendre lorsque l'on veut mesurer les différents processus MS. A partir de la distribution des contributions MC additionnées, des points de pseudo-données sont générés afin d'être compatibles avec l'erreur statistique de chaque bin. Puis, un ajustement est effectué pour mesurer l'accord entre les données et la simulation, et des facteurs de renormalisation sont extraits pour chaque processus MS. Cette expérience est renouvelée 1000 fois et les distributions finales révèlent qu'en moyenne un bon accord est observé entre les données et la prédiction (ce qui est attendu) et qu'il est possible d'extraire les principaux processus du MS avec une très bonne précision. Pour le processus ZZ, cette précision est 53.7 % et pour le processus ZH, elle est de 256.5 %, compatible avec la limite calculée lors de l'analyse précédente.

Ensuite, cette expérience est réalisée mais cette fois ci la distribution des données à 8 TeV est utilisée. L'accord entre données et simulation trouvé n'est pas satisfaisant, mais à ce stade aucune erreur systématique n'a été introduite.

Par ailleurs, tous les facteurs de renormalisation extraits pour les différents processus MS sont compatibles avec 1, ce qui permet de conclure qu'aucun désaccord entre le MS et les données n'est observé grâce à cette méthode.

Finalement, différentes valeurs pour les paramètres libres de la méthode, exposés plus haut, sont testés afin d'obtenir la meilleure limite sur la section efficace du processus ZH. Ce choix donne lieu à une arborescence finale qui discrimine au mieux le processus Z+bb le plus rare, et permet de comparer le résultat final avec celui de l'analyse précédente. Cette limite, incluant toutes les erreurs systématiques du chapitre antérieur recalculées, est de $3.58 \times \sigma_{MS}$, proche de celle trouvée par l'analyse exclusive. L'impact de la systématique principale, liée à la statistique MC, est ici réduit de moitié.

Le dernier test effectué consiste à injecter un signal de nouvelle physique dans l'arborescence qui renvoie la meilleure limite pour le processus ZH. Ce signal est motivé par les modèles 2HDM ; dans celui ci, un boson de Higgs lourd se désintègre en un boson Z et en un boson de Higgs pseudo-scalaire léger. Ce dernier se désintègre à son tour en deux quarks b, tandis que le boson Z donne naissance à deux leptons. Ce processus est appelé processus ZA. La limite obtenue pour ce signal est alors de $0.42 \times \sigma_{ZA}$ (ici on suppose que l'impact des systématiques est le même que pour le cas du ZH). Cette limite est comparable à celle trouvée par une analyse dédiée à la recherche de ce signal [25], qui est de $0.83 \times \sigma_{ZA}$.

Pour conclure, cette méthode dite modèle-indépendante est une méthode d'analyse inédite qui peut encore être améliorée mais présente d'ors et déjà des résultats intéressants : aucun désaccord entre les données et le MS n'est observé pour l'heure, et une sensibilité à des processus rares tels que les processus ZH et ZA a pu être notée.

Références

[1] Steven Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* ***19*** *(1967) 1264.*

[2] Sheldon L. Glashow, "Partial-symmetries of weak interactions", *Nucl. Phys.* ***22*** *(1961) 579.*

[3] J. Goldstone A. Salam and S. Weinberg, "Broken Symmetries", *Phys. Rev.* ***127*** *(1962) 965.*

[4] P.W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* ***12*** *(1964) 132.*

[5] P.W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons", *Phys. Rev.* ***145*** *(1966) 1156.*

[6] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* ***13*** *(1964) 321.*

[7] The 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.*

[8] The 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.*

[9] The 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.*

[10] The CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV ", *Eur. Phys. J. C* ***75*** *(2015) 212.*

[11] The CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", *CERN-PH-EP/2014-288.*

[12] "CMS collaboration web-site", `http://cms.web.cern.ch/`.

[13] The CMS Collaboration, "CMS Physics : Technical Design Report Volume 1 : Detector Performance and Software", *CERN-LHCC-2006-001,*, 2 February 2006.

[14] The CMS Collaboration, "CMS Physics : Technical Design Report Volume 2 : Physics Performance", 25 June 2006.

[15] The CMS Collaboration, "Particle-flow Event Reconstruction in CMS and Performance for Jets, Taus, and E_T^{miss}", *CMS PAS PFT-09-001.*

[16] Matteo Cacciari Gavin P. Salam Gregory Soyez, "The anti-k_t jet clustering algorithm", *JHEP 0804 :063,2008.*

[17] Pierre Artoisenet Vincent Lemaître Fabio Maltoni Olivier Mattelaer, "Automation of the matrix element reweighting method", *Journal of High Energy Physics 2010 :68.*

[18] C. Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", *CMS NOTE 2006/014.*

[19] C. Beluffi and J. Andrea, "Calibration of the jet probability and prospects at high p_T", *CMS AN-2014/237.*

[20] The CMS Collaboration, "Performance of b tagging at $\sqrt{s}$= 8 TeV in multijet, $t\bar{t}$ and boosted topology events", *CMS-BTV-13-001.*

[21] The 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.*

[22] Glen Cowan Kyle Cranmer Eilam Gross and Ofer Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J* ***C71*** *(2011) 1554.*

[23] The 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*** *012003 (2014).*

[24] C. Beluffi R. Castello A. Caudron C. Delaere T.A. du Pree V. Lemaitre A. Pin and J. Vizan, "A Search for the SM Scalar Boson in the Z(ll)H(bb) Final State using a Matrix Element Method", *CMS-AN-12-476.*

[25] The CMS Collaboration, "Search for H/A decaying into Z and A/H, with Z to ll and A/H to fermion pair", *CMS PAS HIG-15-001.*