DISS: ETH NO. 26621

Search for the standard model Higgs boson decaying into bottom quarks, produced in association with a W and Z boson decaying leptonically.

A thesis submitted to attain the degree of

Doctor of Sciences of ETH Zurich

(Dr. sc. ETH Zurich)

Presented by

Gaël Ludovic Perrin

MSc. ETH Physik

Born on 19.09.1990

Citizen of Payerne (VD)

Accepted on the recommendation of

Prof. Dr. Christophorus Grab
Prof. Dr. Rainer Wallny

2020

Abstract

A search for the Standard Model Higgs boson decaying into a bottom quark pair when produced in association with an electroweak vector boson decaying leptonically is presented for the Z($\mu\mu$)H(bb) and Z(ee)H(bb) processes. The search is performed in data samples corresponding to an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$ at a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ recorded by the CMS experiment at the LHC during Run 2 in 2016. An excess of events is observed in data compared to the expectation in the absence of a $H \to b\bar{b}$ signal. The significance of this excess is 3.1σ, where the expectation from a Standard Model Higgs boson production is 1.8σ. The signal strength corresponding to this excess, relative to that of the Standard Model Higgs boson production, is 1.8 ± 0.6.

A similar search for the W($\mu\nu$)H(bb) and W(eν)H(bb) processes on the same 2016 data samples is performed in a boosted regime, where the W and Higgs bosons are required to have a transverse momentum above $250\,\mathrm{GeV}$. An excess of events is observed in data compared to the expectation in the absence of a $H \to b\bar{b}$ signal. The significance of this excess is 0.21σ, where the expectation from Standard Model Higgs boson production is 0.42σ. The signal strength corresponding to this excess, relative to that of the Standard Model Higgs boson production, is $0.5^{+2.4}_{-0.5}$.

Résumé

Une étude d'un boson de Higgs du modèle standard se désintégrant en une paire de quarks b, associé à un boson Z ou W se désintégrant en lepton, est présentée pour les processus Z($\mu\mu$)H(bb) et Z(ee)H(bb). Les données collectées pendant l'année 2016 par le détecteur CMS situé au LHC, à une luminosité de $35.9\,\mathrm{fb}^{-1}$ et une energie au centre de masse de $\sqrt{s} = 13\,\mathrm{TeV}$, sont considérées pour cette étude. Un excès du nombre d'événements par rapport l'hypothèse excluant la présence d'un signal de type $H \to b\bar{b}$ a été observé. La signification statistique de cet excès est de 3.1σ, alors que la signification statistique attendue est de 1.8σ. L'amplitude du signal de cet excès correspond à la section efficace prédite par le modèle standard multipliée par un facteur 1.8 ± 0.6.

Une étude similaire est conduite pour les processus W($\mu\nu$)H(bb) et W(eν)H(bb) sur les mêmes données. Dans le cadre de cette seconde étude, le boson de Higgs et le boson W sont requis d'avoir une quantité de mouvement transverse supérieur à $250\,\mathrm{GeV}$. Un excès du nombre d'événements par rapport l'hypothèse excluant la présence d'un signal de type $H \to b\bar{b}$ a été observé. La signification statistique de cet excès est de 0.21σ, alors que la signification statistique attendue est de 0.42σ. L'amplitude du signal de cet excès correspond à la section efficace prédite par le modèle standard multipliée par un facteur $0.5^{+2.4}_{-0.5}$.

Contents

Part I

Introduction

1 Standard Model

1.1 Introduction

The Standard Model of particles physics is the theory describing the fundamental interaction between particles. It is built on a set of elementary particles, *bosons* and *fermions*, that interact through the *electromagnetic*, *strong nuclear* and *nuclear weak* forces. Massive elementary particles acquire their masses through another interaction, which involves a particular particle of the standard model: the Higgs Boson. The mathematical formalism of the standard model is based on *quantum field theory*. This chapter gives an overview of the Standard Model, starting with properties of the fundamental particles and then followed by an outline of quantum field theory. Many textbooks have been written about this subject and the material from this chapter has been taken from [1][2].

1.2 The Fundamental Particles

Fundamental particles can be separated into two groups: fermions, that have an half integer spin and follow *Fermi-Dirac statistics*, and bosons, that have an integer spin and follow *Bose-Einstein statistics* [3].

1.2.1 Fermions

Not all fermions can interact through strong nuclear forces. This motivate to separate *leptons*, which are not sensitive to strong nuclear interaction from *quarks*.

- **Leptons**: There are six fermions in total. Three charged fermions: the *electron* e^-, *muon* μ^- and *tau* τ^- and three corresponding *neutrinos*: the *electron-neutrino*, *muon-neutrino* and *tau-neutrino*. Charged fermions interact through

electromagnetic and weak interaction (which are both unified in the electroweak interaction, described later in this chapter), while neutrinos interact exclusively through the electroweak interaction. The masses of each charged fermion has been measured. Neutrino are also massive particles, as they can oscillate between different families, but the value of their masses hasn't been established.

- **Quarks**: There are six quarks in total. Three *up-type* quarks, which are the *up*, *charm* and *top* quarks, have an electric charge of $+2/3$, and three *down-type* quarks, which are the *down*, *strange* and *bottom* have an electric charge of $-1/3$. Through strong interaction, quarks can combine to form composite particles referred to as *hadrons* such as the proton and neutron, each one composed of three quarks. Quarks can only be found in such bound state, single quarks are not observed in nature (this property will be explained in section 1.3.2). The masses of the quarks have a broad spectrum, the mass of the top quark ($173.0 \pm 0.4\,\mathrm{GeV}$) being approximately five orders of magnitude larger than the mass of the up quark $2.2^{+0.5}_{-0.4}\,\mathrm{MeV}$.

Because of the way they interact through the weak force, fermions can be classified in three *generations* or *families*, as shown in the Table 1.1. Fermions are also separated by helicity in this Table, as *right-handed* neutrinos cannot interact weakly. The weak equivalent of the electric charge for the electromagnetic interaction is the *weak isospin*, T_3. The right-handed fermions don't have any weak isospin charge.

	Left-Handed					Right-Handed				
	Generation			Charges		Generation			Charges	
	1'st	2'nd	3'rd	Q	T_3	1'st	2'nd	3'rd	Q	T_3
Leptons	$\begin{pmatrix} e^- \\ \nu_e \end{pmatrix}_L$	$\begin{pmatrix} \mu^- \\ \nu_\mu \end{pmatrix}_L$	$\begin{pmatrix} \tau^- \\ \nu_\tau \end{pmatrix}_L$	-1 0	-1/2 +1/2	e^-_R	μ^-_R	τ^-_R	-1/2 +1/2	0 0
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}_L$	$\begin{pmatrix} c \\ s \end{pmatrix}_L$	$\begin{pmatrix} t \\ b \end{pmatrix}_L$	+2/3 -1/3	+1/2 -1/2	u_R d_R	c_L s_R	t_R b_R	+2/3 -1/3	0 0

Table 1.1: List of all the fermions of the Standard Model and the corresponding electric Q and weak isospin T_3 charges. The fermions have been separated in three generations (or families) for the left- and right-handed helicity case. Right-handed neutrinos are omitted in this table, as they are not subjected to any of the Standard Model interactions.

1.2.2 Boson

Bosons have an integer spin (0, 1, 2, etc). In the Standard Model, the particles having the role of force carrier are vector bosons with a spin of 1. Each vector boson is the mediator of one of the three fundamental interactions, as listed below.

- The photon γ is the mediator of the electromagnetic interaction between two electrically charged particles

- The Z, W^+ and W^- boson are the mediators of the weak interaction. They are massive, the mass of the Z boson beeing $91.1876 \pm 0.0021\,\mathrm{GeV}$ and the mass of the W being $80.379 \pm 0.012\,\mathrm{GeV}$. The $W^{+(-)}$ boson has an electric charge of $+1$ (-1), repectivly.

- Eight massless $gluons$ (g) are the mediators of the strong interaction. They are electrically neutral.

The *Higgs boson* is the only scalar boson (spin 0) in the standard model. It is not a force carrier, like the vector bosons listed above. It has a unique role in the standard model, as it's interaction with other elementary particles gives them their masses. It has been discovered by the Atlas and CMS collaboration in 2012 [4–6].

1.3 Quantum Field Theory

The description of the fundamental interactions between the particles listed in the previous section relies on the formalism of quantum field theory. All the fundamental particles are treated as excited states of their corresponding field, whose interactions are described by a corresponding Lagrangian. Group theory has an important role in building the standard model, as requiring the theory to remain invariant under a specific set of symmetries translates into constraints on its Lagrangian. In particular, all interactions must be Lorentz invariant and locally gauge invariant. Each of the fundamental forces corresponds to a group representation of the local gauge symmetry. The full standard model Lagrangian is invariant under the $SU(3)_C \times SU(2)_L \times U(1)_Y$ representation of the gauge group. The letter C and Y in this expression refer to the charges corresponding the symmetry group, the color and weak hypercharge, respectively. Those charges will be introduced in the following sections. The label L indicates that the $SU(2)$ group only acts on left-handed particles.

1.3.1 Quantum Electrodynamics

The quantum field theory description of electrodynamics, refeered to as *quantum electrodynamics* (QED), is described in this section.

A Lorentz transformation on a field $\Phi(x)$ can be written as $M(\lambda)\,\phi\left(\lambda^{-1}x\right)$, where $M(\lambda)$ is a representation of the Lorentz group. Choosing a particular representation

$M(\lambda)$ is equivalent to choose the Lie generators of the group. For QED, the 4×4 *Weyl* representation is commonly used. The corresponding generators are

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma^i \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \quad , \quad i = 1, 2, 3 \tag{1.1}$$

where σ^i are the Pauli matrices. The *Dirac Lagrangian* is invariant under this representation of the Lorentz group.

$$\mathcal{L}_{\text{Dirac}} = \overline{\psi}\left(i\gamma^\mu\partial_\mu - m\right)\psi,$$

where ψ is the fermionic field, and $\overline{\psi} = \phi^\dagger\gamma^0$.

The choice of Lagrangian above is motivated by observation, as it can be shown that the Dirac fermions corresponding to the field ψ have a spin of $\frac{\hbar}{2}$ like the standard model fermions (section1.2.1).

The Dirac Lagrangian describes a free field, in the sense that there are no interactions between the fermions. A local gauge transformation on the fermionic field can be written as $e^{i\alpha(x)}\psi(x)$. A locally gauge invariant Lagrangian is obtained by introducing a new vector field A_μ and replacing the derivative ∂_μ by $D_\mu = \partial_\mu + ieA_\mu$. The field A_μ transforms as $A_\mu \to A_\mu - \frac{1}{e}\partial_\mu\alpha(x)$. This leads to the Lagrangian of QED

$$\mathcal{L}_{\text{QED}} = \overline{\psi}\left(i\gamma^\mu\partial_\mu - m\right)\psi + e\overline{\psi}\gamma^\mu A_\mu\psi - F_{\mu\nu}F^{\mu\nu},$$

where the term $F_{\mu\nu}F^{\mu\nu}$ is the inner product of the electromagnetic field tensor $F^{\mu\nu}$. The vector field A_μ is in fact the four vector potential of the electromagnetic field. The term $e\overline{\psi}\gamma^\mu A_\mu\psi$ is the interaction term between the fermionic and vector fields, where term e can be identified as the electric charge of the fermion. The fermions are not longer free as in the Dirac Lagrangian; they can now interact through electromagnetic forces.

1.3.2 Quantum Chromodynamics

The strong interaction between the quarks is descriebed by *Quantum Chromodynamics* (QCD). The main difference with respect to QED is that its Lagrangian is locally invariant under the $SU(3)$ gauge group. The corresponding gauge transformation on the field ψ is

$$\psi \to e^{i\alpha_a(x)T_a}\psi(x), \quad i = 1, ..., 8,$$

where T_a are the 3×3 generators of the $SU(3)$ representation. Unlike the $U(1)$ gauge group used in QED, $SU(3)$ is a non-Abelian group. The generators satisfy the relation $[T_a, T_b] = if_{abc}$, where f_{abc} are the structure constant of the group. The QCD Lagrangian is

$$\mathcal{L}_{QCD} = \overline{\psi}\left(i\gamma^{\mu}\partial_{\mu} - m\right)\psi + g_S\overline{\psi}\gamma^{\mu}T_a G^a_{\mu}\psi - G^a_{\mu\nu}G^{\mu\nu}_a,$$

where $\psi(x)$ is the quark field, G^a_{μ} are gauge boson fields corresponding to the gluons, g_S is the *strong coupling constant* (analogous to e in the QED Lagrangian), and $G^a_{\mu\nu}$ is the strong field tensor (analogue to $F^{\mu\nu}$ in the QED Lagrangian). The expression of the strong field is not the same as in QED. The non-Abelien property of $SU(3)$ adds a third term: $G^a_{\mu\nu} = \partial_{\mu}G^a_{\nu} + \partial_{\nu}G^a_{\mu} - g_S f_{abc}G^b_{\mu}G^c_{\nu}$.

The equivalent of the electric charge in QCD are *colors* charges. As the QCD gauge group has dimension three, there are three color charges in total: *red*, *blue* or *green*. Quarks carry a single color charge, and gluons two. Since gluons carry color charges, they can also directly interact with other gluon fields. This comes from the structure constant term in the Lagrangian. The the Lagrangian can be expended only considering the terms containing the vector field G^a_{μ}, which describe the interactions of quarks and gluons:

$$\mathcal{L}_{\text{QCD}} = \cdots + \text{“}g_S\overline{\psi}G\psi\text{”} + \text{“}g_S G^3\text{”} + \text{“}g_S^2 G^4\text{”},$$

The term $g\overline{\psi}G\psi$ corresponds to the interaction between a quark and a gluon. It is similar to the $e\overline{\psi}\gamma^{\mu}A_{\mu}\psi$ part of the QED Lagrangian. The two terms $g_S G^3$ and $g_S^2 G^4$ are coming from the structure constant f_{abc} and have no analogue in QED. They correspond to interactions between three and four gluons, respectively. The three terms can be visualized in terms of *Feynman diagrams* that are depicted in Figure 1.1. The first diagram represent a quark radiating a gluon. The second and third diagram represent a three and four gluon interaction vertex, respectively.

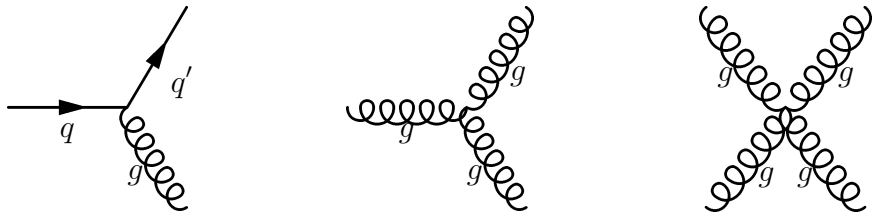


Figure 1.1: Feynman diagrams corresponding to the terms $e\overline{\psi}\gamma^{\mu}A_{\mu}\psi$, $g_S G^3$ and $g_S^2 G^4$.

In quantum field theory, theoretical predictions for a physical observable rely on a perturbative approach. For QCD, the perturbative serie is expanded in terms of α_S, where $\alpha_S = \frac{g_S^2}{4\pi}$ is the *strong coupling*. Some of the terms with large momenta integrals diverge. Those divergence can be treated with a *renormalisation* procedure, were the Lagrangian parameters are not consider as constant and absorb the divergence. This procedure introduce a dependence on the *energy scale* μ of the process. For the QCD

coupling, this dependence is:

$$\alpha_S(\mu) = \frac{4\pi}{b_0 \frac{\ln \mu^2}{\Lambda^2_{QCD}}}, \tag{1.2}$$

where $b_0 > 0$ and Λ_{QCD} was measured to be around $200\,\mathrm{MeV}$. The positive value of b_0 is coming from the additional three and four gluon vertex interactions[1] and leads to two important properties of QCD:

- **Asymptotic freedom:** At large energy scales or short distance: $\alpha_S << 1$. The strong interaction is reduced for quarks within a bound state such as a proton.
- **Color confinement:** At small energy scales or short distance: $\alpha_S >> 1$. The strong interaction confines quarks and gluons into colorless hadronic bound states.

The prediction from the theory depends on the choice of an unphysical *renormalization scale* μ_R to evaluate the value of $\alpha_S(\mu_R)$ in the perturbative serie. This may seem to introduce an arbitrariness in the prediction. However, it can be shown that, if one computes the n'th first terms of the serie, the scale dependence is of order of α_s^{n+1}, so it is greatly reduced with an increasing of the perturbation order [7]. At the scale of the Z boson mass, the QCD coupling is measured to be $\alpha_s(M_Z) = 0.1181 \pm 0.0011$. Pertubative QCD calculation are therefore valid for high energy colliders.

1.3.3 Electroweak Theory

The QED and QCD theory remain invariant under three discrete symmetries: charge conjugation C (a charge q becomes $-q$), parity transformation P ($\psi(t,x) \to \psi(t,-x)$) and time reversal T ($\psi(t,x) \to \psi(-t,x)$). Experimental observations such as the β decay have shown that the P symmetry was not conserved in this processes [8]. The weak interaction violates the P and C conservation, as well as the CP and T symmetry (but much weakly than the C or P separately). The electromagnetic and weak interactions are both unified in the *electroweak theory*, described in this section. The content of this section is taken from [9].

The gauge group of the electroweak theory is $SU(2)_L \otimes U(1)_Y$. For the representation of $SU(2)_L$, the fermionic fields ψ are organized in the left-handed doublets

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L \quad \text{or} \quad \begin{pmatrix} \nu_l \\ l \end{pmatrix}_L,$$

[1] In QED, where there are no self interaction between the vector boson field A_μ, the value of b_0 is below 0.

where the u and d fermionic fields correspond to up and down-type quarks, ν_l and l correspond to a neutrino and a lepton within the same generation (for example ν_e and e). The representation of $U(1)_Y$ acts on the right-handed fields

$$\psi_2 = u_R \quad \text{or} \quad \nu_{lR}$$

$$\psi_3 = d_R \quad \text{or} \quad l_R,$$

which are analogous to the $SU(2)$ representation except the R subscript is added, as the fields are right-handed.

The corresponding local gauge transformations on the fields $\psi(x)$ are

$$\begin{aligned}\psi_1(x) &= e^{iy_1\beta(x)}e^{i\alpha(x)\frac{\sigma_i}{2}}\psi_1(x)\\ \psi_2(2) &= e^{iy_2\beta(x)}\psi_2(x)\\ \psi_3(2) &= e^{iy_3\beta(x)}\psi_3(x),\end{aligned}$$

where y_i, $i = 1, 2, 3$, are is the *hypercharge* and σ_i, $i = 1, 2, 3$, are the Pauli matrices. The parameters $\beta(x)$ and $\alpha(x)$ parameterize the local $SU(2)_L \otimes U(1)_Y$ gauge transformation. The $U(1)_Y$ gauge transformation is analogous to QED, and the $SU(2)_L$ transformation is analogous to the non-Abelien gauge group of QCD. The electroweak Lagrangian has therefore as structure similar to QED and QCD. Introducing the gauge fields W^i_μ for $SU(2)_L$ and B_μ for $U(1)_Y$, the electroweak Lagrangian can be expressed as

$$\mathcal{L}_{\text{EWK}} = \overline{\psi}_i i\gamma^\mu \partial_\mu \psi_i + g y_i \overline{\psi}_i \gamma^\mu \frac{\sigma_i}{2} B_\mu \psi_i + g' \overline{\psi}_i \gamma^\mu \frac{\sigma_i}{2} W_{i\mu}\psi_i - \frac{1}{4} B_{\mu\nu}B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu},$$

where $i = 1, 2, 3$ and $W_{i\mu\nu}$ and $B_{\mu\nu}$ are the field strength tensor of the gauge fields, whose definition is analogous to the QED and QCD field strength tensor. The terms g and g' are the $U(1)_Y$ and $SU(2)_L$ couplings, respectively. The electroweak Lagrangian doesn't contain any mass terms such as $\overline{\psi} m \psi$ as they would mix the right- and left-handed components and hence break the local gauge invariance. The presence of mass terms for the gauge fields would also break the local gauge symmetry and are not present in the Lagrangian.

The vector boson fields B_μ and $W_{i\mu}$ are not physical objects, in the sense that they don't correspond to the vector bosons γ, Z and $W^\pm$. This Lagrangian can be separated in a kinematic component (Kin), a charged current (CC) and a neutral current (NC) component as

$\mathcal{L}_{\mathrm{EWK}} = \mathcal{L}_{\mathrm{NC}} + \mathcal{L}_{\mathrm{CC}} + \mathcal{L}_{\mathrm{Kin}}$,
where

$$\mathcal{L}_{\mathrm{NC}} = \sum_j \overline{\psi}_j \gamma^\mu \left\{ A_\mu \left[g\frac{\sigma_3}{2} \sin\theta_W + g' y_j \cos\theta_W \right] + Z_\mu \left[g\frac{\sigma_3}{2} \cos\theta_W - g' y_j \sin\theta_W \right] \right\} \psi_j$$

$$\mathcal{L}_{\mathrm{CC}} = \frac{g}{2\sqrt{2}} \left\{ W_\mu^\dagger \left[\overline{u}\gamma^\mu (1-\gamma_5) d + \nu_l \gamma^\mu (1-\gamma_5) l \right] + h.c \right\},$$

$$\mathcal{L}_{\mathrm{Kin}} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu},$$

where $h.c$ stands for hermitian conjugate and the gauge fields $W_\mu^\dagger = (W_\mu^1 - i\mathrm{W}_\mu^2)$. The gauge fields Z_μ a A_μ are defined by the rotation

$$\begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} Z_\mu^3 \\ A_\mu \end{pmatrix}, \tag{1.3}$$

where θ_W it the *Weinberg angle*. The vector boson fields W, $W^\dagger$, Z and A_μ correspond to the W^-, W^+, Z boson and the photon γ, respectively. To impose that the term with A_μ in $\mathcal{L}_{\mathrm{NC}}$ corresponds to the QED gauge field, it follows that

$$\begin{aligned} g \sin\theta_W &= g' \cos\theta_W = e \\ Y &= Q - T_3, \end{aligned}$$

where Y, Q and T_3 are the hypercharge, electric charge and weak isospin operators, respectively. The last equation implies that the hypercharge of a right-handed neutrino is 0. They therefore cannot interact through the weak or electromagnetic interaction, as expected. Using the last equation and the rotation relation 1.3, the correspondence between the physical fields $W^\pm$, Z and the gauge fields B_μ and W_μ^i can be summarized as

$$\begin{aligned} W_\mu^\pm &= (W_\mu^1 \mp i\mathrm{W}_\mu^2) \\ A_\mu &= \frac{g' W_\mu^3 + g B_\mu}{\sqrt{g^2 + g'^2}} \\ Z_\mu &= \frac{g W_\mu^3 - g' B_\mu}{\sqrt{g^2 + g'^2}}, \end{aligned}$$

The electroweak Lagrangian presented in this section unifies the electromagnetic and weak interaction. However, it cannot include mass terms for the fermions and

massive gauge boson without breaking the local gauge symmetry. The solution to this problem involves the inclusion of a new scalar boson, the Higgs boson, and is described in the next section.

1.3.4 The Higgs Mechanism

Mass terms are introduced to the electroweak Lagrangian through the mechanism of *spontaneous symmetry breaking* (SSB) that preserves local gauge invariance. To generate the masses of the $W^{\pm}$ and Z^0, the following Lagrangian is added to (1.3.3)

$$\mathcal{L}_{\mathrm{S}} = |(i\partial_\mu - ig\frac{\sigma_i}{2}W^{i\mu} - ig'\frac{y_\phi}{2}B_\mu)\phi|^2 - V(\phi),$$

where the field ϕ is an isospin doublet with weak hypercharge $y_\phi = 1$

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

and the potential $V(\phi)$ is defined as

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2, \quad (\lambda > 0, \mu^2 < 0)$$

which has it's minimum for fields satisfying $\phi^\dagger\phi = \frac{\mu^2}{2\lambda}$. A particular choice of minimum is of $V(\phi)$ is

$$\phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ \nu\end{pmatrix}, \tag{1.4}$$

where the term $\nu \sim 246\,\mathrm{GeV}$ is referred to as the *scalar vaccum expectation value.*

The Lagrangian $\mathcal{L}_{\mathrm{S}}$ is invariant under a local $SU(2)_L \otimes U(1)_Y$. However, this symmetry is no longer apparent if the Lagrangian is rewritten by expanding ϕ_0 about a minimum ϕ_0 as [2]

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ \nu + H(x)\end{pmatrix}, \tag{1.5}$$

where $H(x)$ is a scalar field. Both the $SU(2)_L$ are $U(1)_Y$ "spontaneously broken" by this procedure, introducing mass terms for the gauge field corresponding to each broken generator, here $W^{\pm}$ and Z. This is the Higgs (or Brout-Englert-Higgs) mechanism. The value $y_\phi = 1$ is chosen such that the photon remains masseless. Since the gauge group of the electromagnetic interaction $U(1)_{EM}$ is not broken, as ϕ_0 has no electric charges and is invariant under a local $U(1)_{EM}$ symmetry, no mass terms are

[2] One could also choose a general expansion $\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}\theta_1(x) + i\theta_2(x)\\ \nu + h(x) + i\theta_3(x)\end{pmatrix}$ but the fields $\theta_1(x)$, $\theta_2(x)$, $\theta_3(x)$ can be gauge out with a local $SU(2)_L$ gauge transformation.

generated for the photon. The presence of mass terms becomes clear when substituting the Lagrangian $\mathcal{L}_{\mathrm{S}}$ in terms of the $\phi(x)$ expansion around the vacuum value. The Lagrangian $\mathcal{L}_{\mathrm{S}}$ can be re-written in terms of the fields H, $W_{\mu}^{\pm}$, Z and separated in two terms, $\mathcal{L}_{H}$ and $\mathcal{L}_{HG^2}$.

$$\mathcal{L}_{S} = \frac{1}{4}h\nu^4 + \mathcal{L}_{H} + \mathcal{L}_{HG^2},$$

where

$$\mathcal{L}_{H} = \frac{1}{2}\partial_{\mu}H\partial^{\mu}H - \frac{1}{2}M_{H}^{2}H^{2} - \frac{M_{H}^{2}}{2\nu}H^{3} - \frac{M_{H}^{2}}{8\nu^{2}}H^{4},$$

$$\mathcal{L}_{HG^2} = M_{W}^{2}W_{\mu}^{\dagger}W^{\mu}\left\{1 + \frac{2}{\nu}H + \frac{H^{2}}{\nu^{2}}\right\} + \frac{1}{2}M_{Z}^{2}Z_{\mu}Z^{\mu}\left\{1 + \frac{2}{\nu}H + \frac{H^{2}}{\nu^{2}}\right\}.$$

The first term includes the kinematic component and self-interaction of the Higgs field. The second term includes the coupling between the Higgs and gauge boson, sometime refereed to as the *Yukawa coupling*. $\mathcal{L}_{H}$ is the kinematic term of the field H and $\mathcal{L}_{HG^2}$ the interaction between H and the gauge bosons. The mass terms of the gauge boson are $M_{A} = 0$, $M_{W} = \frac{1}{2}\nu g$, $M_{Z} = \frac{1}{2\cos\theta_{W}}\nu$. By adding the term $\mathcal{L}_{S}$ to the $SU(2)_{L} \otimes U(1)_{Y}$ Lagrangian, the vacuum expectation value ν has generated the mass of terms for the M_{Z} and M_{W} boson while keeping the local gauge symmetry. A new scalar boson is also introduced in the Lagrangian, the Higgs boson, whose corresponding field is H. The Higgs boson mass is given by $M_{H} = \sqrt{-2\mu^{2}} = \sqrt{h\nu}$. As neither the value of μ nor h from the potential $V(\phi)$ are know, the Higgs boson mass is not predicted by the theory. As it can be seen from the interaction terms in $\mathcal{L}_{S}$, the Higgs boson interaction with the coupled gauge boson are proportional to the boson mass squared divided by the vacuum expectation value ν.

The lepton mass terms can be generated in the electroweak Lagrangian in a similar way. Using the same isospin doublet ϕ and adding the following Lagrangian to (1.3.3)

$$\mathcal{L}_{\mathrm{Lep}} = -c_{l}\left[(\overline{\nu}_{l}, \bar{l})_{L}\begin{pmatrix}\phi^{+}\\ \phi^{0}\end{pmatrix} l_{R} + \bar{l}_{R}(\phi^{-}, \overline{\phi}^{0})\right], \tag{1.6}$$

where the value of the input parameters c_{l} are not predicted by the theory.

Similarly, the mass terms for the quarks are included with the following Lagrangian

$$\mathcal{L}_{\mathrm{quarks}} = c_{d}(\overline{u}, \overline{d})_{L}\begin{pmatrix}\phi^{+}\\ \phi^{0}\end{pmatrix} d_{R} + c_{u}\begin{pmatrix}\phi^{(0)*}\\ -\phi^{-}\end{pmatrix} d_{R}u_{R}, \tag{1.7}$$

where the parameters c_{d} and c_{u} are not predicted by the theory and the quark fields u and d correspond to the mass eigenstate, which are not the same as the weak inter-

action eigenstate. The relation between the two is given by the *Cabibbo-Kobayashi-Maskawa* (CMK) matrix

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{tb} & V_{ts} & V_{tb} \end{pmatrix} \times \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1.8}$$

where d', s', b' are weak eigenstate and d, s, b the mass eigenstates of the three generation of down-type quarks.

After the spontaneous symmetry breaking, the corresponding mass term and the Yukawa coupling for the fermions are

$$m_d = -c_d \frac{\nu}{\sqrt{2}}, \quad m_u = -c_u \frac{\nu}{\sqrt{2}}, \quad m_l = -c_l \frac{\nu}{\sqrt{2}},$$
$$\mathcal{L}_Y = -\left(1 + \frac{H}{v}\right)\left\{m_d \bar{d}d + m_u \bar{u}u + m_e \bar{l}l\right\},$$

where $\mathcal{L}_Y$ is the Yukawa coupling between the fermions and the Higgs boson, m_d, m_u, m_l are the masses of the down-type, up-type quarks and the lepton, respectively. The masses of the fermions are not predicted by the theory, as the constant c_d, c_u and c_l are not known and their value come from experimental observations. A remarkable property that can be seen from the term $\mathcal{L}_Y$ is that the coupling between the Higgs field and the fermions is proportional to the fermion's mass.

As each of the Lagrangian $\mathcal{L}_S$, $\mathcal{L}_{\text{quarks}}$ and $\mathcal{L}_{\text{lep}}$ are invariant under a local $SU(2)_L \otimes U(1)_Y$ gauge symmetry, the Higgs mechanism introduce mass terms for gauge bosons and the fermions and conserves gauge invariance. This electroweak theory that incorporates the SSB in the $SU(2)_L \otimes U(1)_Y$ is the Glashow-Weinberg-Salam model.

The measurements of the Higgs boson have been performed using proton-proton collision data from the *Large Hadron collider*, based at CERN. There the productions of the Higgs boson can occur through a variety of processes, the most important ones are described in a later chapter, see section 5.2.

1.3.5 Matrix Element

An essential quantity to calculate the cross-section for a given interaction process is the amplitude of the *matrix element*, $|M_{fi}|^2$, which gives the probability of a set of initial state particles i to interact and create the final state particles f.

The expression of M_{fi} is calculated using a perturbative approach. The perturbative serie is expanded in terms of the coupling terms g as

$$M_{fi} = \sum_{n=0}^{N} c_n g^n = c_0 + c_1 g + c_2 g^2 + \ldots + c_N g^N.$$

The integer N fixes the order of the prediction. The first order is referred to as the *leading order* (LO), the second order is referred to as the *next-to-leading order* (NLO), the third order as *next-to-next-to-leading order* (NNLO), and so on. If the coupling g is small, the serie converges and the precision on the cross-section increases with the number of additional terms.

The calculation of the terms c_i is organized in terms of Feynman diagram. Each order of the perturbation serie correspond to a list of Feynman diagram, which give a pictural representation of the interactions taking place. The terms c_i can be derived by following a set of rule, the *Feynman rules*, coming from the quantization of the Lagrangian of the theory.

Example of Feynman diagrams for the $q\overline{q} \to ZH$ process are depicted in Figure 1.2. This process corresponds to two incoming quarks, producing a Z and Higgs boson from the scattering. The features of this process are discussed in a later chapter for the case where the Higgs boson decays to a bottom and anti-bottom quark pair, see section 5.3.0.1.

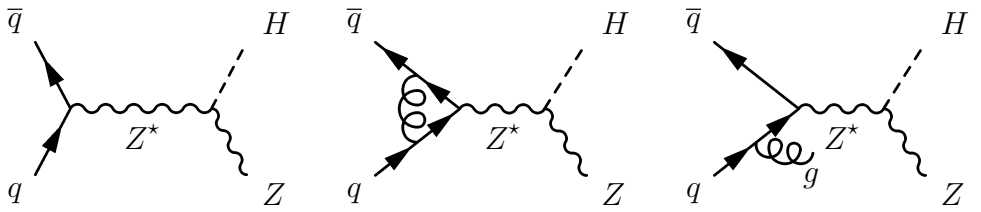


Figure 1.2: Example of one leading order (left) and two NLO (middle and right) Feynman diagrams for the $q\overline{q} \to ZH$ process.

2
Monte Carlo Simulation

2.1 Introduction

Simulations are essential to study the data produced at the LHC. In the analysis presented in this thesis, they are a key ingredient to study the various processes from the proton-proton collision and to extract the contribution from a standard model Higgs boson. This chapters gives a short overview of the *event generator* tools used to simulate proton-proton collisions at the LHC.

One of the key assumption of event generation is *factorization*, which breaks down the simulation of a collision event in different steps. The *hard process* is the heart of the collision and describes the high momentum transfer of the proton constituent. The internal structure of the incoming protons is modeled by the *parton distribution functions* (PDF). Evolution of the outgoing hadrons is described by *parton showers* and *hadronisation*. An additional interactions to the hard process can also take place, referred to as *underlying event*. A variety of tools are available to perform those different steps, that are then combined in the simulation of the full scattering process. All those aspect of an event simulation are described in the next sections. The Figure 2.1 give a representation of an event produced by an event generator.

2.2 Parton Distribution Function

The substructure of the protons is described by the *parton model*, which refers to the constituents (quarks and gluons) of the proton as *partons*. The *parton distribution functions* (PDF), $f_i(x, \mu_f)$, gives the probability that the parton i carries a fraction x of the proton momentum. The μ_f is the *factorization scale* of the process. The dependence on the factorization scale is introduced to treat divergent term occurring in the perturbative calculation due to collinear (low-angle) emissions.

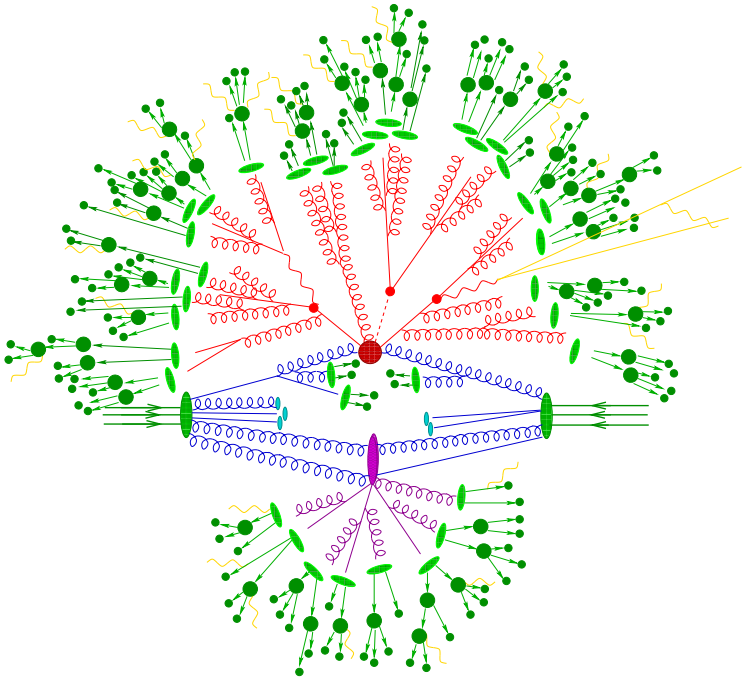


Figure 2.1: Pictorial representation of a an event as produced by an event generator. Two incoming protons are colliding (green arrows on right and left side). The hard interaction is depicted by the big dark red blob at the center. It is followed by additional hard QCD radiations described by the parton shower (blobs and lines in lighter red). The final state partons then hadronise (light green blobs) and decay decay (dark green blobs). The figure is taken from [10].

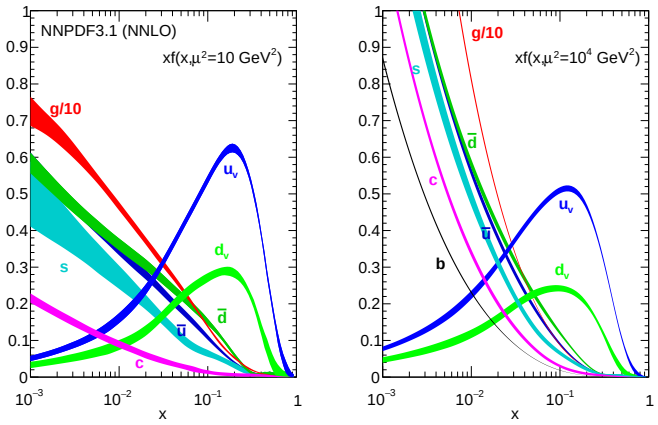


Figure 2.2: The parton distribution functions according to the NNPDF3.1 NNLO, evaluated at $\mu_F^2 = 10\,\text{GeV}^2$ (left) and $\mu^2 = 104\,\text{GeV}^2$ (right). The fraction of the proton energy x multiplied by the parton distribution function is shown for various partons as a function of x. Valence quarks as u_V and d_V for the up and down quark, respectively, do not include virtual quarks from gluon emission. The quarks without the V subscript include virtual contributions from the so-called gluon sea. The figure is taken from [11].

Data collected at deep inelatic scattering experiments, as well as the Tevatron and LHC are used as input the PDF extraction. This analysis presented in this thesis uses PDFs from NNPDF3.0 and NNPDF3.1 [11]. An example of PDFs distribution for NNPDF3.1 can be seen in Figure 2.2.

2.3 Hard Process

A proton-proton collision consists of initial states partons, producing a variety of outgoing final state particles. As events of interests at the LHC often involve large momentum, the simulation focuses on the description of a hard process, where the interaction between two initial state partons involves a large momentum transfer. The matrix element of the hard process is computed using perturbation theory, as mentioned in section 1.3.5.

The timescale in which the hard process occur is much sorter than the typical parton-parton interaction within the proton. This has for consequences that the interaction between the two partons involved in the hard process and the rest of the protons can be neglected. Using this property, proton substructure and the hard scattering can be factorised in the cross-section formula for a scattering $ab \to n$

$$\sigma = \sum_{a,b} \int_0^1 \mathrm{d}x_a \mathrm{d}x_b \int \mathrm{d}\Phi_n f_a^{h_1}(x_a, \mu_F)\, f_b^{h_2}(x_b, \mu_F)\, \frac{1}{F} \left|\mathcal{M}_{ab\to n}\right|^2 (\Phi_n; \mu_F, \mu_R)\,, \tag{2.1}$$

where the functions $f_a^h(x,\mu)$ are the PDFs, the final state phase space is denoted by Φ_n [1] and the sum runs over all the possible partons. In the denominator, the *flux factor* F corresponds to the number of protons that crosses each other per unit area and time. The matrix element squared of the hard scattering for the production of the final state n through initial state parton a and b is $|\mathcal{M}_{ab\to n}|^2(\Phi_n;\mu_F,\mu_R)$, and depends on the factorisation scale μ_F and the renormalisation scale μ_R. The equation 2.1 holds to all order in perturbation theory. To be an event generator, the program must not only integrate this function over the phase space, but provide events according to the fully differential distributions including the partons flavor, helicity, momentum, and color.

The cross-section from equation 2.1 depends on the choice of the renormalisation and factorisation scale, which cannot be derived by first principle. One often used prescription is to set $Q^2 = \mu_R = \mu_F$, where Q^2 is the scale of the hard scattering. The prescription to estimate the corresponding systematic uncertainty for μ_R is to compute the cross-sections corresponding to a factor 2 up and down variation in the choice of scale. The inclusion of those uncertainties in the analysis presented in the thesis are described in section 8.3.2.

As mentioned in section 1.3.2, the perturbative expansion in terms of Feynman diagrams are in powers of $\alpha_S(\mu_R)$. This motivates calculating a high order matrix element, as it reduces the dependence on the scale μ_R. The scale dependence has a large impact on the overall normalisation: the dependance on μ_S is linear at LO. On the other hand, a LO matrix element describes the shapes of the distribution rather accurately but loop calculations are technically much more difficult as they involve cancellation of divergences between different terms. One therefore often corrects the overall normalisation using a so-called *k-factor*, which is the ratio of a higher and lower order cross-section. This correction can be inclusive (a single k-factor) or differential, in which case multiple k-factors are provided in bins of a certain kinematic property of the event (for example, the p_T of the simulated Z boson). Most common event generator employ LO and NLO matrix elements.

2.4 Parton Shower

The final state partons that are produced in the hard process emit gluon radiation, just like accelerated charges emit photon radiation. As the gluon themselves carry color charges, they can further emit additional gluons, leading to a *parton shower* emitted by the final state hadron. A full description of this process would require much higher order than the ones employed in the matrix element calculation. The effect of all higher orders for cases of *soft* (when the gluon momentum is small) and *collinear* (when the

[1] The phase-space element is $\mathrm{d}\Phi_n = \prod_{i=1}^{n} \frac{\mathrm{d}^3 p_i}{(2\pi)^3 2E_i} \cdot (2\pi)^4 \delta^{(4)}\left(p_a + p_b - \sum_{i=1}^{n} p_i\right)$, where $p_{a(b)}$ is the momentum of the $a(b)$ parton and the index i runs over all the hard scattering final state.

angle θ between the emitting parton and the gluon is small) emissions can be simulated trough a *parton shower* algorithm. This algorithm is based on an approximation scheme, described below. The content of this section is taken from [12].

Starting with the collinear emission the cross section of two final state partons $\sigma_{q\bar{q}}$ and two final state partons plus one soft radiation $\sigma_{q\bar{q}g}$ are related through

$$\mathrm{d}\sigma_{q\bar{q}g} \approx \sigma_{q\bar{q}} C_F \frac{\alpha_S}{2\pi} \frac{\mathrm{d}\theta^2}{\theta^2} \mathrm{d}z \frac{1+(1-z)^2}{z}, \tag{2.2}$$

where $C_F = \frac{N_C^2-1}{2N_C}$ is a color factor that can be thought as the color-charge square of a quark, z is the momentum fraction of the emitting parton carried by the emitted gluon. The equation 2.2 is only valid in the collinear $\theta \to 0$ approximation. It points out that a gluon emission can be taken into account as a multiplicative factor in the cross section. This motivates the idea that multiple emission can be treated independently using an iterative approach. Generalizing this idea to any parton emission (not just gluon), the cross section for a hard cross section σ_0 to be accompanied by a parton j with a momentum fraction z is given by

$$\mathrm{d}\sigma \approx \sum_{partons,i} \sigma_{q\bar{q}} C_F \frac{\alpha_S}{2\pi} \frac{\mathrm{d}\theta^2}{\theta^2} \mathrm{d}z P_{i,j}(z,\phi) \mathrm{d}\phi, \tag{2.3}$$

where ϕ is the azimuth of parton j around parton i, $P_{ji}(z,\phi)$ are the splitting functions which describe the distribution of the fraction z of the energy of i carried by j [12]. There are four splitting functions depending on the nature of the emitted and emitting parton: $P_{q\to qg}$, $P_{q\to gq}$, $P_{g\to gg}$ and $P_{g\to q\bar{q}}$.

The equation 2.3 diverges for $z \to 0$, $z \to 1$ (the former divergence is apparent in equation 2.2). This can be treated by imposing a cutoff Q_0 such as $Q_0^2/q^2 < z < 1 - Q_0^2/q^2$, where q^2 is the virtuality (virtual mass squared) of the emitting parton. Two partons are then consider as resolved if their relative transverse momentum is under the scale Q_0. This would correspond for example when the separation of both partons is below the detector resolution.

The sequential implementation of the parton shower algorithm requires the definition of an *evolution variable* that is used to order and separate the emission, such as the virtuality q^2. The virtuality is the highest at the hard process and decreases at each emission until it is below the cutoff $Q_0 \approx 1\,\mathrm{GeV}$, at which point the shower is terminated and the hadronization begins.

One thing to emphasize is that equation 2.3 does not take into account virtual contributions (loop diagram). It can however be used to calculate the probability of non-emission given by the *Sudakov form factor*

$$\Delta_i(Q^2, Q_0^2) = \exp\Big\{ -\int_{Q_0^2}^{Q^2} \frac{\mathrm{d}k^2}{k^2} \frac{\alpha_S}{2\pi} \int_{Q_0^2/k^2}^{1-Q_0^2/k^2} \mathrm{d}z P_{ji}(z) \Big\}. \tag{2.4}$$

The Monte-Carlo implementation of 2.4 can be summarized as follows:

1. A random number ρ is chosen between 0 and 1 and the equation $\Delta_i(Q^2, q^2) = \rho$ is solved for q^2.

2. If the solution is below Q_0^2, then the splitting is unresolved and the showering of that parton is terminated.

3. If the solution is above Q_0^2, and the splitting is $j \to k + l$, start again from step 1. on the parton k and l. The variable z and ϕ for the splitting are using use the Monte Carlo method on $P_{ji}(z)$.

4. Once all splitting from the successive parton have fallen below Q_0^2, stop.

The procedure described above is only valid for collinear emission. In the case of soft emissions, that is when the momentum of the emitted parton is small, the factorization is no longer valid at cross-section level. Another factorization is taking place for the amplitude (at the Feynman diagram level). But as the cross section is calculated by summing all Feynman diagrams and squaring, interference are unavoidable and the factorization is not apparent at the cross-section level as in the collinear case. It can be shown that those interferences are treated correctly if the angle θ is used as an evolution variable, which is referred to as *coherent showering* (HERWIG [13] is a Monte-Carlo generator using coherent showering). In that case, the first emission of the shower are often soft wide-angle gluons.

The treatment of parton shower differs if the emission is developed from an outcoming parton or from the parton incoming to the hard process. Both cases are referred to as *final* and *intial state shower*, respectively. For initial state shower, the iteration is "reversed". The momentum of the parton is initialized at the hard process and then evolved backward, gaining energy at each evolution and reducing the virtuality until the scale of the incoming hardon constituent is reached.

2.5 Hadronisation Model

At an energy scale of Λ_{QCD}, the final state shower can no longer describe the outgoing parton, as the perturbative approach breaks down because of the running of coupling (see section 1.3.2). An non-perturbative approach relying on the main feature of QCD is then employed, where the outgoing parton lead to the observed hadrons through a hadronization process. The two main hadronization models are the *string* and *cluster* models.

The starting point of the string model is color confinement. Taking as an example the production of a quark-antiquark $q\overline{q}$ pair, the physical picture is a color flux tube being stretched while the two quarks move appart. The transverse dimension of the

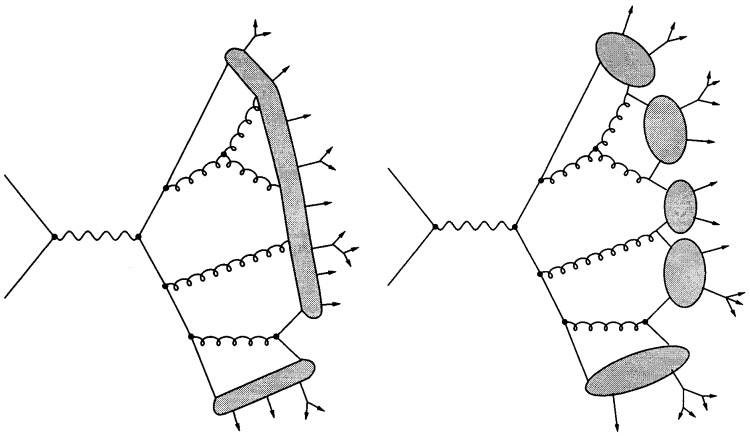


Figure 2.3: Schematic representation of a string (left) and cluster (right) hadronization model. Figures taken from [14].

tube are of a typical hadron size, roughly 1 fm. Assuming a tube with uniform density, the potential increase linearly as the $q\overline{q}$ quarks move apart until the separation is large enough to create a new $q'\overline{q}'$ pair by breaking the string. The color tube between the new pairs $q\overline{q}'$ and $q'\overline{q}$ increase as they move apart, and so on. At the end of the process, the string has broken to n $q\overline{q}$ pairs, and a hadron is formed for each of the adjacent $q\overline{q}$. The breaking of the string is modeled by the *fragmentation functions*, whose parameter can be tuned in the Monte-Carlo generator. The string is broken iteratively until no further breaking of the string can be associated to a final-state hadron.

The cluster hadronization model is based on the *color preconfinement* property of QCD: for evolution scales much smaller that the hard process $|q| < |Q|$, the partons in a shower are clustered in colorless groups whose invariant mass depend only on the scale q and Λ_{QCD}. The idea is to build colorless cluster at hadronization scale that then decay into the final-state hadron. Near the hadronization scale, gluons are forced to split into quark-antiquark pair and form a cluster with the corresponding color partner. The decay of the cluster into the final-state partons uses a simple isotropic quasi-two-body phase space model. A representation of both hadronization models is depicted in Figure 2.3.

The final state hadrons whose trajectory are close to each-other are collected into a same physical object, called *jet*, with it's own energy, momentum and trajectory. A method to merge the final state hadrons into the jet is described in a later chapter, see section 6.4.4.

2.6 Underlying Event

The hard process, parton shower and hadronization steps are sufficient to fully describe the final state of the hard interaction. However, this interaction involves the extraction of a colored parton from each of the hadrons, which are colorless bound states of many colored partons. It is therefore necessary to consider additional interactions between the remaining partons due to color connection [15].

This extra interaction from the hadron remnants is known as the *underlying event*. It it is described by *perturbative models*: the interaction of the hadron remnant is treated like multiple parton-parton interactions with their own hard process and parton shower.

One of the difficulty of the underlying events is the modeling of the color connection. Also, the jet cross-section is sensitive to rare underlying event fluctuation, and it is therefore important to have reliable underlying event models [15].

2.7 Matrix Element Matching

The parton shower method is based on a QCD approximation that is valid in the collinear and soft limits. It is therefore not reliable in a phase space with hard and well-separated jets, where a fixed order matrix element should provide a better description.

The best approach is to combine both methods: start with a fixed order matrix element and then evolve the jets through a parton shower program. This could however lead to double counting: in a event of $n + 1$ jets, one could have $n + 1$ jets from the matrix element (and no additional jets from the parton shower) or n from the matrix element and 1 additional jet from the parton shower. The treatment of those overlapping events is performed by a *matrix element matching*.

Several matching methods are used in the present analysis. The MLM [16] matching combines tree-level matrix elements for several jet multiplicities simultaneously, with parton showers describing the internal structure of the jets without double counting. The MC@NLO [17] and POWHEG [18–20] combine lowest-multiplicity NLO matrix elements with parton showers without double counting.

3
The Large Hadron Collider

3.1 Reference

The review of the Large Hadron Collider given in this chapter is directly quoted from [21]. The borrowed text is placed between inverted commas.

3.2 Introduction

"The Large Hadron Collider (LHC) is a proton-proton collider machine working at the nominal energy of $\sqrt{s} = 8\,\mathrm{TeV}$. It is installed at CERN in Geneva, about $100\,\mathrm{m}$ underground and is the most powerful particle accelerator ever built by human kind. It provides collisions between protons since November 2009 with increasing instantaneous luminosity and energy. The design of the machine is extremely complex and complicated: it exploits the leading edge technological findings from different fields of physics, engineering and computer science. Its success in running as designed could not be taken for granted. Nevertheless the machine exceeded its own designed possibilities in many sectors and achieved several world records. The time schedule for the machine is an alternate period of running (generally few years) providing collision data to the experiments, and a period of technical stop, where the machine and the detectors can be upgraded or, in case, repaired. During its first run LHC delivered an integrated luminosity of about $6\,\mathrm{fb}$ at $7\,\mathrm{TeV}$ and about $23\,\mathrm{fb}^{-1}$ at $8\,\mathrm{TeV}$ to the two main experiments ATLAS and CMS." During the second run period, the LHC delivered and integrated luminosity of $40.99\,\mathrm{fb}^{-1}$ at $13\,\mathrm{TeV}$ to both experiments during the year 2016."

3.2.1 The machine and its design

"The technical design report [22] describes in details the design of LHC. The following section summarises its content. The machine is composed by 1232 dipole magnets working at a temperature of 2.3 K. The very low temperature reached by the cryogenic system based on liquid helium allows the superconductive magnets to reach a magnetic field of $8.3\mathrm{T}$. Radio frequency cavities installed along the beam line accelerate the protons, with a gradient of $0.5\,\mathrm{MeV}$ per turn.

The instantaneous luminosity depends on several parameters:

$$L = \frac{\gamma f k_B N_p^2}{4\pi\varepsilon_n \beta^*} F, \tag{3.1}$$

In Equation 3.1 γ is the boost Lorentz factor, f is the revolution frequency, k_B and N_p refer respectively to the number of bunches and of protons per bunch. ϵ_n is the normalised transverse emittance while $\beta^\star$ is the betatron function at the interaction point[1]. Finally F is the reduction factor due to the crossing angle of the bunches." During the 2016 run the time between the bunches was $25\,\mathrm{ns}$ and the instantaneous luminosity peaked at about $6.17 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-2}$.

"The LHC ring is composed of two parallel adjacent beam pipes at the radial distance of $2.8\,\mathrm{cm}$, see Figure 3.1. Along the LHC tunnel, four big caverns accommodate a corresponding number of particle detectors: two multipurpose detectors, *A toroidal LHC Apparatus* (ATLAS) and the *Compact Muon Solenoid* (CMS), seek direct discoveries of new particles, deviations from standard model predictions and test new possible extensions of the standard model. In the other two cavities the *LHC Beauty experiment* (LHCb) and *A Large Ion Collider Experiment* (ALICE) investigate respectively the violation of the CP symmetry and the quark-gluon plasma production through heavy ion collisions."

3.2.2 The accelerator chain

"In order to reach the final proton energy at LHC, the particles go through a long pre-accelaration chain. The process starts with the ionisation of the hydrogen atoms. This produces protons, the primary source of the beam. The protons accelerate up to $50\,\mathrm{MeV}$. Then an injector moves the protons into a *Proton Synchrotron* (PS) ring where their energy increases up to $25\,\mathrm{GeV}$. The accumulation happens here. When the beam density is enough the protons pass to the *Super Proton Synchrotron* (SPS) where they boost up to $900\,\mathrm{GeV}$. Eventually an injector kicks the protons into LHC. Here radio frequencies rise the beam energy up to its final value. Quadrupoles and dipoles stabilise and focus the beam, so the collisions can start. The full pre-acceleration chain can be found in Figure 3.2."

[1] $\beta^\star$ is the distance between the focus point and the point where the beam width doubles.

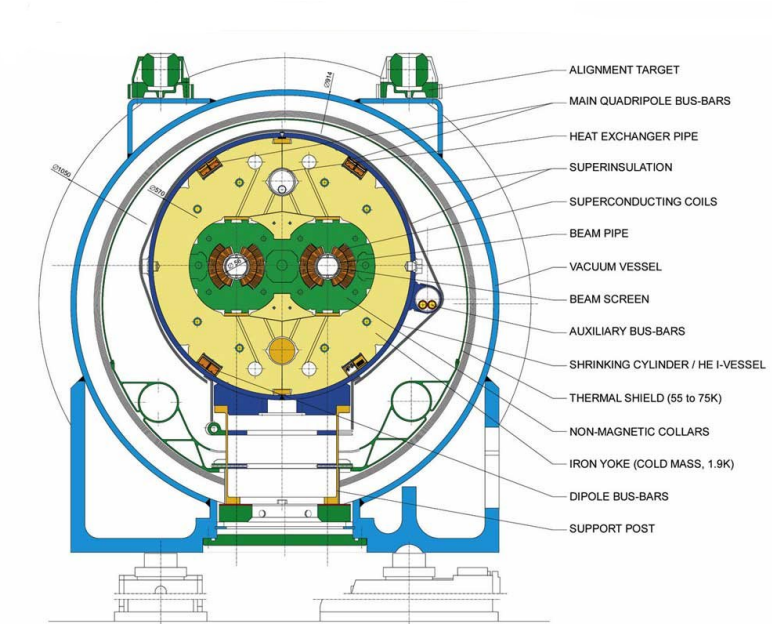


Figure 3.1: "Inside transverse view of the LHC. The two beam pipes are in the centre and special superconductive coils are attached to them. The iron yoke, kept at the temperature of $1.9\,\mathrm{K}$, surrounds beam pipes and coils."

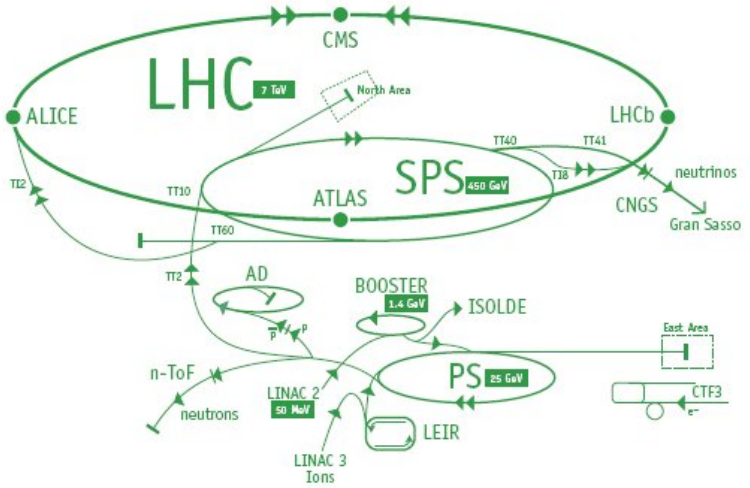


Figure 3.2: "Schematic view of the proton accelaration chain before the injection into LHC. Arrows indicates the proton moving direction."

4
The CMS detector

4.1 Reference

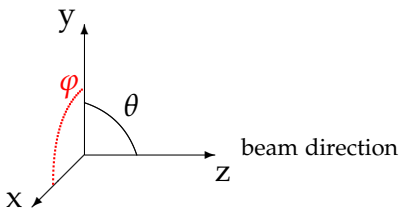
The review of the Large Hadron Collider given in this chapter is directly quoted from [23]. The borrowed text is placed between inverted commas.

4.2 Introduction

"The Compact Muon Solenoid, referred to as CMS, is one of the four particle detectors installed along the LHC. The purpose of the CMS experiment is to explore a new energy region exploiting hadronic collisions. The detector design optimises the detection and identification of the particles produced in the protons collisions. It has a cylindrical symmetry, the axis of which coincides with the beam pipe of the accelerator. It consists of a silicon tracking system, an electromagnetic and a hadronic calorimeter. A solenoid surrounds the trackers and the calorimeters and provides an intense magnetic field, essential for precise transverse momentum measurements of charged tracks. The outer part of the detector is composed by a redundant muon system. The intent of this chapter is to give a very brief overview of the detector and its performances. The detailed and complete description can be found in [24]."

4.3 General concept

"The CMS detector has a very compact design. Figure 4.1 shows the layout of the detector. It is 21.6 meters long and has a diameter of 14.6 meters. Figure 4.2 defines the CMS coordinate system. The triad (p_x, p_y, p_z) represents the momentum projections along the axes. An other largely used quantity is the p_T, the momentum projection on the $x - y$ plane, referred to as *transverse momentum*. An intense magnetic field



$$\eta = -ln\left[\tan\left(\frac{\theta}{2}\right)\right]$$

Figure 4.1: "The CMS coordinate system. CMS uses a right-handed coordinate system and places the origin at the nominal interaction point. The x axis points to the LHC centre, the y axis points up, and the z axis along the beam direction. The polar angle θ is measured from the positive z axis, and the azimuthal angle ϕ spans the $x - y$ plane."

used to bend charged tracks and muons is one of the points of strength of its design. The intent is to measure the momenta of the charged particles with the best accuracy. A superconducting solenoid allows to achieve a magnetic field of $3.8\,\mathrm{T}$ inside the detector. An iron yoke of $1.5\,\mathrm{m}$ closes the return field lines. Within the iron structure 4 layers of muon chambers increase the identification power and the momentum resolution of very energetic muons. The hadron and electromagnetic calorimeters sit between the solenoid and the tracking system. The calorimetry serves for the identification of photons, electrons and neutral hadrons and for the measurements of their energy. The inner part accommodates a full silicon tracking system, which measure the position of charged particles passing through."

4.3.1 Inner tracking system

"The CMS tracking system is composed, from inside to outside, by three cylindrical layers of pixels, followed by a silicon strip detector that incorporates ten layers which cover up to the end of the tracking system, at a radius of $1.1\,\mathrm{m}$. The detector extends up to $|\eta| = 2.5$ by endcaps disks: two disks of pixels and three plus nine disks in the silicon strips detector. The transverse view of the tracker is schematically shown in figure 4.3. A high granularity tracking system is essential to avoid ambiguities in the tracks reconstruction and in the primary vertex identification. The finest granularity layers are the ones closer to the beam pipe. The pixel read-out system components increase the material budget inducing multiple interactions and particle conversions. These effects reduce the identification efficiency and the overall energy resolution. A silicon micro-strips detector completes the tracker after the pixels. It has enough granularity to avoid efficiency losses and increases the material budget before the calorimeters. Thanks to the decreasing flux of particles, the outermost region of the silicon detector mounts larger-pitch silicon micro-strips."

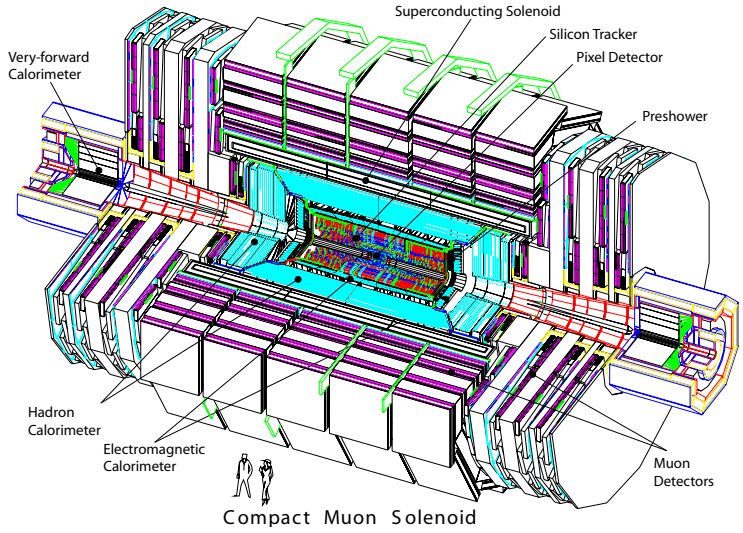


Figure 4.2: "An exploded view of the CMS detector. A pixel detector surrounds the beam pipe. From inside-out follow a silicon tracker system, an electromagnetic calorimeter, a hadron calorimeter, and eventually a muon system. The figure is taken from [25]."

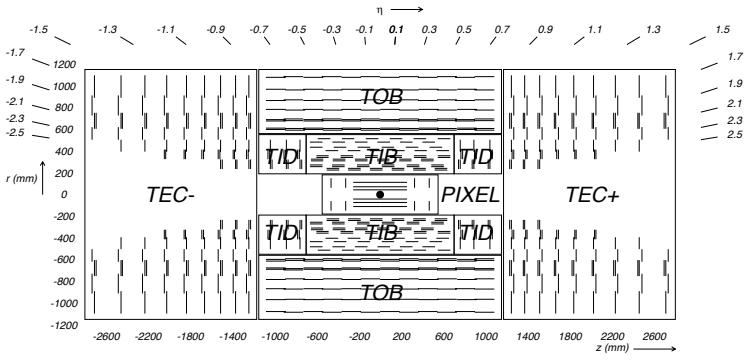


Figure 4.3: "The CMS tracker system. Figure taken from [25]."

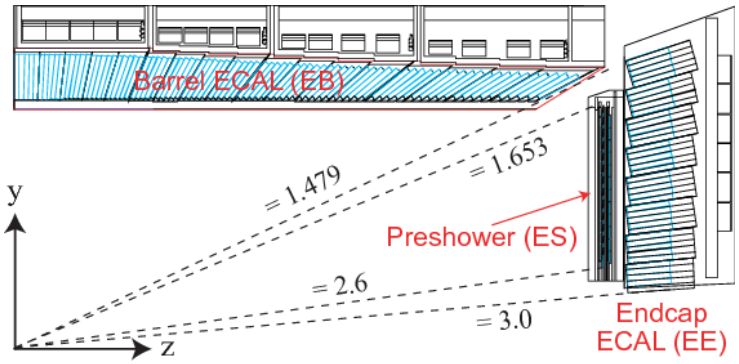


Figure 4.4: "A transverse view of the electromagnetic calorimeter of CMS. Figure taken from [25]."

4.3.2 Electromagnetic calorimeter

"The electromagnetic calorimeter is placed after the tracking system. It mounts 61200 crystals of lead tungstate ($PbWO_4$) in the barrel region and other 7324 in the endcaps. The geometry is cylindrical, the design is hermetic and homogeneous (see figure 4.4) in order to maximise the efficiency for the photon detection. The specific type of crystals assures good granularity important to reconstruct the direction of the photons and to suppress background coming form multiple interactions. Crystals have a fast response (in $25\,\mathrm{ns}$ they emit 80% of the light) and good radiation resistance necessary for the high luminosity of the LHC. In the barrel the part facing the beam pipe measures $22 \times 22\,\mathrm{mm}^2$ and is $23\,\mathrm{cm}$ long. In the endcaps the section is a bit bigger, $28.6 \times 28.6\,\mathrm{mm}^2$, but the length is shortened by $1\,\mathrm{cm}$. The light yield is on average 30 photons every MeV deposited in the crystal. Silicon Avalanche Photodiodes (APDs) collect the light from the crystals in the barrel, while Vacuum Photo Triodes (VPTs) are preferred in the endcaps."

4.3.3 Hadron calorimeter

"The hadron calorimeter is a substantial ingredient to successfully reconstruct proton collision events. Important objects to compare measurements with theory predictions are jets, well defined collection of particles. The HCAL identifies and measures the hadronic component of jets, including those initiated by b-flavoured quarks. Together with the ECAL, it gives indirect information about the production of neutrinos or hypothetical weakly interacting particles. Important characteristics for the hadron calorimeter are the minimisation of the non-Gaussian tails of the resolution and the maximisation of the solid angle coverage. These ensure a good energy resolution and accuracy for jets and missing transverse energy. The design of HCAL foresees the maximisation of the absorption of neutral and charged particles within the detector. Mounting the hadron calorimeter just before the solenoid avoids that hadronic shower

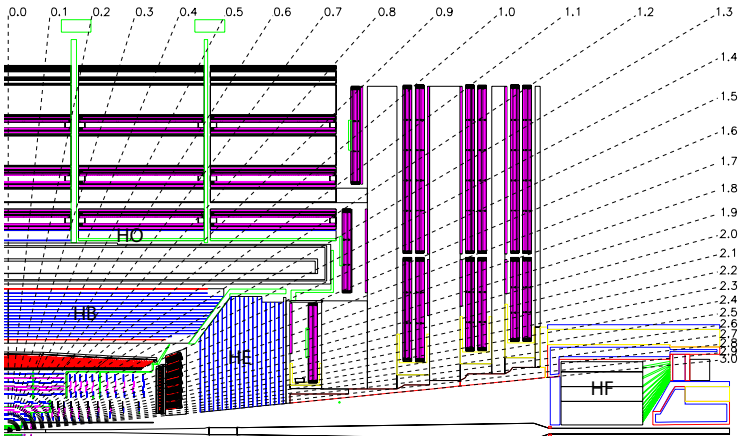


Figure 4.5: "Longitudinal view of CMS with a focus on the location of the hadron calorimeters. Figure taken from [25]."

remnants pollute the measurements of tracker and ECAL. An additional hadron calorimeter, or tail catcher, absorbes the remaining tails of the showers from behind the solenoid. The hadron calorimeter covers both regions: the barrel ($|\eta| < 1.3$) and the endcaps ($1.3 < |\eta| < 3$). In the central region the HB is a sampling calorimeter which consists in 36 identical azimuthal wedges made of plastic scintillator material alternate with brass used to induce the hadronic showers. In the endcaps the HE uses the same materials, but different thicknesses, to adapt to the expected higher tracks multiplicity. The layout of the hadron calorimeter is depicted in figure 4.5."

4.3.4 The muon system

"The muon system is a crucial part of the CMS design. Many interesting processes produced at LHC include muons in the final state. This signature cleans the high background expected by *multijet* production.

Muons are objects with extremely useful properties: their very small interaction with materials make them penetrate the detector without losses of information; they are faintly susceptible to radiative losses, characteristic which makes the measurement of their momentum more accurate. A reliable identification of muons has an additional, extremely important, advantage. It can act as a very efficient trigger, to reduce the very high event rate produced by LHC.

Three gaseous detectors identify and measure the properties of muons. The muon detectors follow the symmetry of the solenoidal field, and are split in a cylindrical barrel section and two planar endcap regions.

Standard *drift tubes* (DT) with rectangular cells are used in the barrel. These chambers intersperse the layers of the flux return. They form two groups installed perpendicularly: one set measures ϕ, the muon direction on the transverse plane, while the

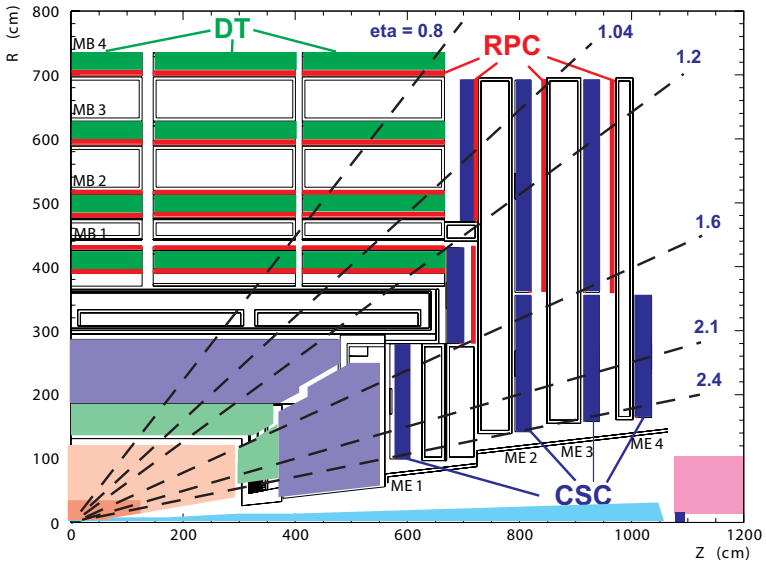


Figure 4.6: "The transvers view of the CMS muon system. Figure taken from [25]."

second set measures z, the component parallel to the beam line, giving a full three-dimensional information of the particle tracks.

In the endcap region, where the background is larger, *Cathod Strips Chambers* (CSC) are used instead. These detectors are faster in response and have a better segmentation with respect to the DT. These properties allow to reduce the background contamination and to have a better momentum resolution in a non constant magnetic field. As the DTs, the CSCs intersperse the endcap return plates, perpendicular to the beam line.

Both DT and CSC systems act as an independent trigger, a logic signal which activates the acquisition system and decides whether to store an event. Their p_T resolution at trigger level is about 15% in the barrel and 25% in the endcap. An additional detector technology reinforces the muon trigger system: the *Resistive Place Chambers* (RPC) are mounted in both, the barrel and the endcaps. These chambers provide a very fast response, but a courser position resolution than the DTs or CSCs. Figure 4.6 shows how the RPCs alternate the DT chambers in the barrel. In a very similar way the RPCs alternate the CSCs in the endcap regions. These extra information improve the time resolution, important for the bunch crossing identification, and the p_T resolution."

4.3.5 Trigger

"At the designed instantaneous luminosity LHC produces proton-proton collisions with a frequency of $40\,\mathrm{MHz}$. It is impractical, not cost effective, and not critical from

a physics point of view, to store all this information. A drastic reduction rate has to be achieved using a fast and flexible trigger system. The trigger system deployed in CMS consists in two parts: the *Level-1* (L1) trigger and the *High-Level Trigger* (HLT). A complete description of the CMS trigger system can be found in the dedicated technical design report [26]. The L1 consists of custom designed programmable electronics. The HLT is a software base filter which acts to reduce even further the rate after the first suppression at L1. The HLT reduces the rate down to about $100\,\mathrm{Hz}$, the maximum CMS storage rate capability. The designed output limit for the L1 is $100\,\mathrm{kHz}$, but in practise the output rate is around a factor three smaller. The L1 trigger accesses coarsely segmented data from the calorimeters and muons system. There is no time to reconstruct physics objects in a time span of $25\,\mathrm{ns}$. Contrary to L1, the HLT can access more sophisticated objects, similar to those used in off-line analyses. This allows to create more efficient trigger logics." The overall output rate of the L1 trigger and HLT can be adjusted by prescaling the number of events that pass the selection criteria of specific algorithms to keep the event rate below the threshold, in which case it is referred as a *prescaled trigger*. Trigger without an event prescaling is referred to as an *unprescaled trigger*.

4.4 Physics objects

"The CMS reconstruction software (CMSSW) has the goal of reconstructing and identifying physics objects such as electrons, muons, hadrons, photons and neutrinos. It takes as inputs the signals coming from the different subdetectors. From the signals in the muon chambers, it is possible to identify muons and to reconstruct their associated track. Using the electromagnetic calorimeter and the tracker information, the CMS algorithm distinguishes photons and electrons. The Hadron calorimeter deposits with a jet clustering algorithm give the possibility of defining jets and measuring their energy and position. More sophisticated algorithms allow to identify more complicated objects such as taus or b-flavoured jets. In the following section the reconstruction of these objects will be synthetically described. An extension to this description can be found in the technical design report of CMS [27]."

4.4.1 Vertices

"The innermost part of CMS is a silicon pixel detector. Its fine granularity allows to reconstruct the numerous primary vertices produced by the proton beams interactions. The identification of the primary vertex associated to the hardest interaction is essential to correctly reconstruct the objects of the event." The selection of the primary vertex in the analysis presented in this thesis is described in another chapter, see section 6.4.1.

4.4.2 Tracks

"The track reconstruction starts with the clusterisation in both pixel and strip detector. Here seeds with high signal-to-noise ratio generate clusters. The software then includes nearby strips and pixels deposits in the cluster using looser criteria. These clusters are the initial seeds for the trajectory. The seed can come internally from the tracker, but also externally from other subdetectors. Each seed in the tracker includes hits that are supposed to come from single charged particle deposits. The trajectory building starts with no less than five initial parameters: at least two hits and the beam constraint pattern are necessary. The main part of the trajectory building is the patter recognition. The method uses a combinatorial *Kalman Filter (KF)* [28]: the filter proceeds iteratively starting from the seed layer, which also provides a course estimate of the track parameters. Passing through successive tracker layers, from the inner to outer ones, more information is collected and the fit parameters improve. The fit is repeated every time a new hit is found compatible with the trajectory. The possibility that a track does not leave a signal in a specific layer is considered and implemented in the Kalman algorithm as an *invalid hit*. The propagation of the track between two layers takes in consideration the multiple scattering and the energy loss in the material as well as the local magnetic field. At this stage multiple track candidates can be created from a single seed. The algorithm ranks these candidates using their normalised χ^2 and on the number of valid and invalid hits. The track building then builds the track inwards searching for additional possible hits in the seeding layers and, where possible, in layers at smaller radii than the seeding layer. A dedicated algorithm solves the possible ambiguities left from the track building, and avoids double counting of tracks. A discriminating variable function of the shared hits is defined for any pair of tracks candidates with common hits. If they share more than the majority of their hits, the track with more hits wins. If the two tracks happen to have the same number of hits, the track with the largest χ^2 is discarded. This procedure is applied to all tracks and it eliminates all the possible ambiguities. A least squares algorithm estimates the final track parameters. Tuned cuts on final vertex compatibility and fit quality eventually cleans the track collection."

"Two special procedures reinforce the standard tracking algorithm: the *iterative tracking* and the *large impact parameters* track reconstruction. They increase the reconstruction efficiency for tracks with low transverse momentum and with a high impact parameter [1] respectively."

[1] The impact parameter is defined as the distance between the primary vertex and the point of closest approach of the track.

4.4.3 Electrons

"The electromagnetic calorimeter and tracker give the essential information to reconstruct electrons. The reconstruction software tries to match the tracker track to the ECAL deposit. In CMS the usual tracks are reconstructed using the Kalman Filter algorithm. This technique does not suit well the electron reconstruction because of the large kinks in the trajectories due the often energetic Bremsstrahlung photon emission. A dedicated electron track reconstruction is therefore implemented: it first relies on a KF-based pattern recognition, adopting relaxed criteria with respect to the usual one. Secondly a Gaussian-Sum Filter (GSF) [29] fit is employed. Here the photon emission is modelled by a Gaussian mixture and the fit can follow tracks with kinks in their trajectories. The GSF fit is computationally demanding and can therefore be implemented only on a limited number of seeds.

The seeds are provided by the ECAL energetic deposits. The barycenter of the ECAL cluster infers the position of the hits in the pixel detector and in the strips. Only track seeds that match an ECAL-inferred position trigger the GSF fit.

This method is not well suited for electrons inside jets though. For these, additional criteria to the ECAL deposits and a special track-driven seeding increase the reconstruction efficiency and reduce the fakes. Eventually the ECAL-driven and the tracker-driven seeds are merged together and used as inputs in the GSF tracking algorithm. More than one GSF track per electron happens to be reconstructed. Electrons emit photons which later convert in electron-positron pairs. Tracks from the conversion have generally a reconstructed η and ϕ close to the primary electron track. Therefore tracks very close to each other trigger the double counting cleaning algorithm. The primary track is eventually selected using information from both tracker and ECAL super-cluster and the final electron candidate is created."

Information to build observables for the electron identification come from all the CMS subdetectors. The most discriminating variables are: the ratio between the cluster energy and the track outer momentum, the ratio between the Bremsstrahlung photon energy as measured by the ECAL and by the tracker, and finally the ECAL energy matching with the inner track momentum. Other information from the tracker, the ECAL, and the cluster shapes discriminate further electrons from fake signals.

4.4.3.1 ECAL trigger

"The complicated algorithm used in the electron reconstruction and identification cannot run at trigger level. It is still possible to efficiently trigger on isolated electrons. The ECAL crystals are grouped in matrices of 5x5. At Level-1 these matrices are used as seed, if enough energetic.

Isolation of clusters is an important property to classify events: electrons produced directly in the primary interaction are not surrounded by other tracks or other calor-

imeter clusters. An isolation requirement may be applied by looking at the deposit around the crystal tower used as seed, and in the HCAL cells behind it. The seeds pass then to the HLT, which includes them in more sophisticated but slower logics. The ECAL energy is here clustered in super-clusters and a transverse energy threshold is applied. The survived seeds are matched to the expected pixel hits, otherwise rejected. The final step exploits the full tracker information: ECAL deposits have to match a track; the tracks enter is the isolation calculation. The event is finally rejected if the isolation in the energy requirement at HLT are not satisfied."

4.4.4 Muons

"Three different ways to reconstruct a muon are available in CMS. The muon system alone defines a *standalone muon*. A different approach uses the tracker information and extrapolates the tracker track, considering the magnetic field and the interaction with the material, to the muon system. If the extrapolated track matches a hit in one of the muon chambers, a *tracker muon* is created. The last and most complete muon reconstruction uses both information: the tracker track and the stand alone muon track. Starting from the muon system the stand alone muon is extrapolated backward to the tracker to check for possible matched tracks; in the positive cases the hits of the tracker track and the muon system are used for a global fit. The result is a *global muon*.

Figure 4.7 shows the energy resolution for the three different types of muons. The transverse momentum resolution for relatively soft muons is dominated by the elastic scattering with the detector volume before the muon chambers. For low momentum muons the tracker gives precious information which improves the energy resolution. For very energetic muons the intrinsic resolution of the muons chambers plays the major role on the uncertainty."

4.4.4.1 The Global muon trigger

"All the muon subsystems participate to the muon trigger system. The DT and CSC Level-1 electronics separately record the signals in their stations; a stand alone muon candidate is created in each sub-system The four highest p_T and best quality candidates are then sent to the *Global Muon Trigger* (GMT). Similarly for the RPC detectors: the recorded hits form muon candidates, ranked based on p_T and quality criteria, and then passed to the GMT. At this stage the GMP attempts to match all the candidates from the different subsystems. The deposits in the ECAL and HCAL are then used to determinethe isolation for each muon candidate. The HLT receives the four best candidates and applies additional selection criteria to reduce further the rate and to increase the signal purity."

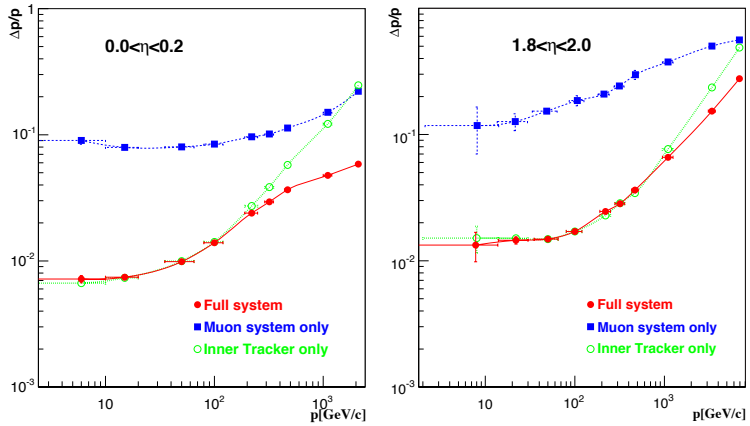


Figure 4.7: "Muon energy resolution for the different types of muon reconstructions. **Left:** central η bin. **Right:** forward eta bin. Figure is taken from [27]."

4.4.5 Jets

"Jet reconstruction aims to reconstruct and identify jets arising from the hadronisation of a scattered parton, with the final goal to accurately measure their direction and energy. Jets are composite objects; they need deterministic algorithmic clustering sequences, the jet algorithm, to be properly defined. They are a collection of particles of different flavours, color structures and charges. From an experimental point of view the constituents of a jet are not easily distinguishable. This reflects into an intrinsic uncertainty on the jet energy resolution and a bias in the energy scale. A specific calibration is therefore needed. An additional challenge is to distinguish the flavour of the parton which originated the jet. CMS adopts different, and very complex, algorithms in order to distinguish jets induced by b-quark (b-tagging), gluons (quark-gluon tagging) or light flavoured quarks." The b-tagging algorithm are discussed in a later chapter, see section 6.4.5.

4.4.5.1 Jet algorithms

"In order to be able to compare jet observables with theoretical predictions it is necessary to define a clustering procedure, which can be used in both cases. The inputs are different: in the experiments jets constituents can be either tracks or energy deposits in the calorimeter, and most likely a combination of the two. A theoretical prediction works instead with partons. Jet algorithms have to work in both cases. An additional requirement is that the jet definition should always avoid divergences in the theory predictions at all orders in the perturbative expansion. Any jet definition has to be *collinear* and *infrared safe*. The collinear safety is guaranteed if splitting an element of the jet of momentum p in two of momentum $p/2$ does not affect the clustering result;

being infrared safe requires that adding any additional soft element to the jet does not change the results. In CMS the most used jet clustering algorithm, and the one employed in this analysis, is the *anti-k_T* [30], which is infrared and collinear safe. The algorithm defines a distance between two elements i and j

$$d_{ij} = \min\left(\frac{1}{p^2_{\mathrm{T}_i}}, \frac{1}{p^2_{\mathrm{T}_j}}\right)\frac{\Delta R^2_{ij}}{R^2},$$

where R_{ij} is the distance in the $\eta-\phi$ plane and R defines the size of the jets. The size parameter used for this analysis is $R = 0.4$. The algorithm also defines the distance between the particle i and the beam

$$d_{i\mathrm{B}} = \frac{1}{p^2_{\mathrm{T}_i}}.$$

The iterative procedure is the following: for each couple of particles it calculates the distance d_{ij} and compares with d_{iB}. If $d_{iB} < d_{ij}$ then i is a jet and is removed from the list; otherwise it merges the two elements i and j and iterates further till only jets are left."

4.4.5.2 The particle flow

"The *particle flow* (PF) approach consists in the attempt of fully exploiting the redundancy of the CMS subdetectors. The PF algorithm first collects all the information from all the subdetectors, building its blocks. In a second time it tries to link them using appropriate linking algorithms. Each successful link forms a *PFCandidate*. These objects can be delivered to higher level algorithms that can reconstruct jets, hadronically-decaying taus, discriminating b-jets from light-jets, etc... CMS is especially indicated to successfully implement a particle flow approach: its compact design, the fine granularity of its calorimeter and its very performing tracking system make the track-to-cluster association effective."

4.4.5.3 Particle flow jets

"The jets constituents in CMS are tracker tracks and calorimeter deposits. The jet reconstruction exploits a *Particle Flow* (PF) approach which improves the jet energy resolution and the jet energy scale bias. The particle flow outperforms the calorimeter based approach and gives a substantial improvement to the jet reconstruction and missing energy measurement. Figure 4.8 shows the performance improvements compared to the calorimeter based jets. The performance gain is reduced when the energy increases because the tracker performances decrease. These results come from the first data collected by CMS at the centre of mass energy of $\sqrt{s} = 0.9$ and $2.36\,\mathrm{TeV}$. The complete documentation can be found in [31–33]."

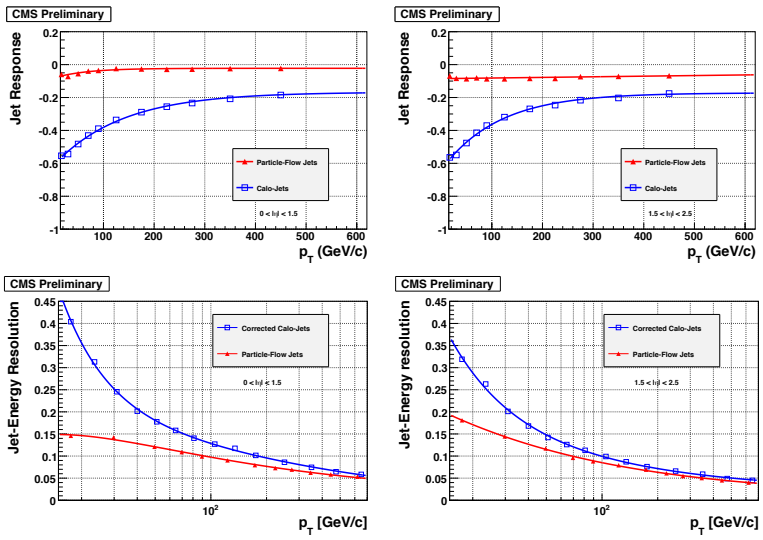


Figure 4.8: "Jet energy response (up) and resolution (bottom) for central (left) and forward (right) jets [31]."

4.4.6 Missing transverse energy

"The modern design of multipurpose detectors such as CMS or ATLAS foresees the almost complete coverage of the solid angle with calorime- try. The primary motivation is to be able to infer and measure the pres- ence of energetic neutrinos or weakly interacting particles traversing the detector. If such particles are produced, a substantial unbalance of the total momentum in the transverse plane appears in the event reconstruc- tion. This missing information to close the kinematic in the transverse plane is the *transverse missing energy* (MET). The missing energy is an important element for the analysis described in this thesis: it is a clear signature for top pair production and it plays a major role in the methodology used to improve the jet energy scale bias and resolution. The vector $p_{T,miss}$ is the opposite of the vectorial sum of all the particle-flow objects of the event, and is referred to by "MET" in the rest of this thesis The magnitude of this vector is referred to as pfMET. This value divided by the scalar sum of E_T of all the particle-flow objects is referred to as "pfMET significance"."

Part II

Analysis

5 Higgs boson at the LHC

The standard model Higgs boson can be produced through different *production mechanisms* (or *production modes*) at the LHC and can decay to a variety of final states. The branching ratios are summarized in section 5.1 and the production modes in section 5.2. The combination of the production and decay mode corresponding to the analysis presented in this thesis is described in section 5.3.

5.1 Branching ratios of the Higgs boson

The Higgs boson can decay into any massive particle. Decays to massless particles such as photons ($H \to \gamma\gamma$) and gluons ($H \to gg$) can also occur indirectly through a top or bottom quark loop [1].

The branching ratios for a Higgs mass around $125\,\mathrm{GeV}$ are given in Figure 5.1 and Table 5.1. The quoted uncertainties in the table are theoretical (THUs) and parametric from quark mass (PU(m_q)) and strong coupling (PU(α_q)). The $H \to b\bar{b}$ decay has the largest branching ratio, about 58%.

5.2 Production modes of the Higgs boson

There are four main Higgs production mechanisms at the LHC: *gluon fusion*, *vector boson fusion*, *associated production with top quarks* and *Higgs strahlung* (or *associated production with a vector boson*). The corresponding leading order diagrams are depicted in Figures 5.2-5.5.

[1] As the coupling between the Higgs boson and a massive particle is proportional to the mass squared of the particle, the top quark contribution is much larger than the bottom quark.

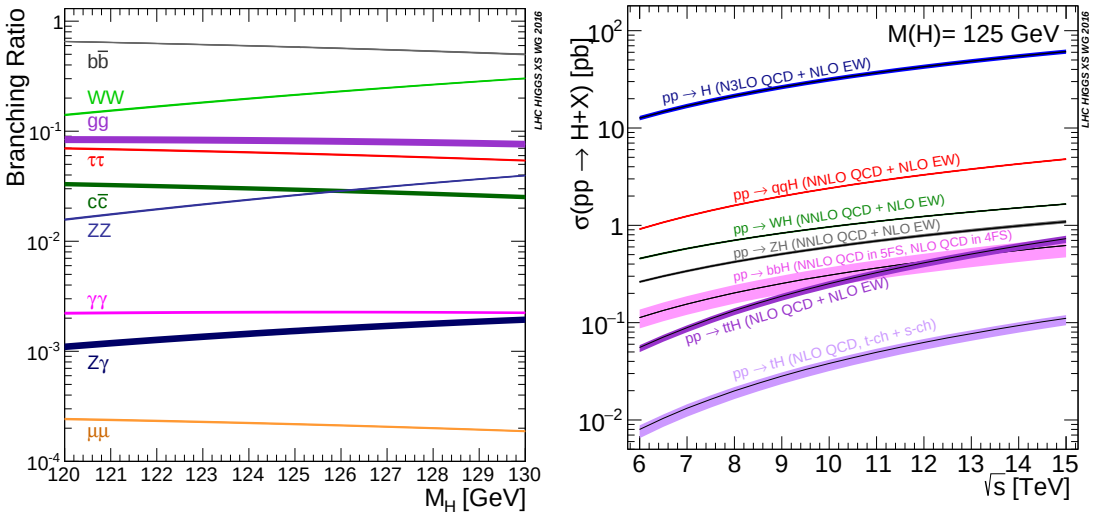


Figure 5.1: Higgs boson production and decay modes at the LHC. **Left:** Branching ratio for a Higgs boson with mass $m_H = 120 - 130\,\text{GeV}$. **Right:** Higgs production cross-section at center-of-mass energy $\sqrt{s} = 7 - 14\,\text{TeV}$ [34].

Decay	Branching Ratio [%]	THU +/− [%]	PU(m_q) +/− [%]	PU(α_s) +/− [%]
$H \to bb$	58.24	0.65/0.65	0.72/0.74	0.78/0.8
$H \to WW^\star$	21.37	0.99/0.99	0.99/0.98	0.66/0.63
$H \to gg$	8.187	3.4/3.41	1.12/1.13	3.69/3.61
$H \to \tau\tau$	6.272	1.17/1.16	0.98/0.99	0.62/0.62
$H \to c\bar{c}$	2.891	1.2/1.2	5.26/0.98	1.25/1.25
$H \to ZZ^*$	2.619	0.99/0.99	1/0.98	0.64/0.63
$H \to \gamma\gamma$	0.227	1.73/1.72	0.93/0.99	0.61/0.62
$H \to Z\gamma$	0.1533	5.71/5.71	0.98/1.01	0.58/0.65
$H \to \mu\mu$	0.02176	1.23/1.23	0.97/0.99	0.59/0.64

Table 5.1: Branching ratio of a 125 GeV standard model Higgs boson. The quoted uncertainties are theoretical uncertainty (THUs) and parametric uncertainties from quark mass (PU(m_q)) and strong coupling (PU(α_q)). The nominal values and uncertainties are given in %. The values are taken from [34]

Those are the Higgs production modes with the largest cross-section. Other production mechanisms are possible, such as the *Higgs production associated with bottom quarks* (bbH), a Higgs boson produced in association to a single top quark tH, or Higgs boson pair production. The cross-section for each production process is given in Figure 5.1, for a center-of-mass energy of $6-15\,\mathrm{TeV}$, and in Table 5.2, for a center-of-mass energy of $13\,\mathrm{TeV}$ and a Higgs boson mass of $125\,\mathrm{GeV}$.

For each production mode, there are processes with a real Higgs boson in the final state. Such processes are referred to as *signals*. *Background* processes have an event signature similar to the signal but without a real Higgs boson in the final state. The treatment and use of observable variables in order to distinguish signal from background processes are crucial when performing a Higgs boson measurement, as some of the backgrounds have a cross-section several order of magnitudes larger than the signal. The $H \to b\bar{b}$ decay in the four production modes, the corresponding signal, and backgrounds, are described in the next sections.

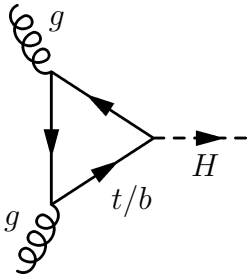


Figure 5.2: gluon fusion.

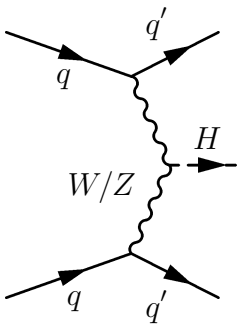


Figure 5.3: Vector boson fusion.

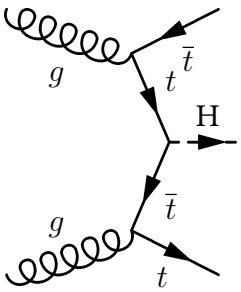


Figure 5.4: Associated production with top quarks.

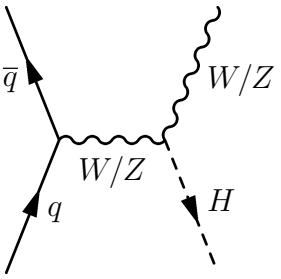


Figure 5.5: Higgs strahlung.

Figure 5.2-5.5: Leading order Feynman diagrams for main Higgs production processes at the LHC.

Higgs production mode	cross-section (pb)	QCD scale	PDF	α_s
Gluon fusion	44.14	$+7.6/-8.1$	± 1.8	± 2.5
Vector boson fusion	3.782	$+0.4/-0.3$	± 2.1	± 0.5
Higgs strahlung	0.8839	$+3.8/-3.1$	± 1.3	± 0.9
Associated production with top quarks	0.5071	$+5.8/-9.2$	± 3.0	± 2.0
Associated production with bottom quarks	0.488	$+20.2/-23.9$		
Associated production with a single top quark	0.07425	$+6.5/-14.9$	± 3.5	± 1.2
Higgs boson pair production	0.037	$+4.3/-6$	± 2.1	2.3

Table 5.2: Cross-sections and uncertainties for the main Higgs boson production modes at the LHC. The values are taken from [34]. The second column corresponds to the cross-section value in pb. The third, fourth and fifth column correspond to the uncertainties in % from the QCD scale, parton distribution functions and strong coupling constant α_S, respectively. For the associated production with bottom quarks mode, the three sources of uncertainties have been added in quadrature.

5.2.1 Gluon fusion production of the Higgs boson

The gluon fusion has the highest cross-section among all the Higgs production modes, about $48.58\,\mathrm{pb}$ at $13\,\mathrm{TeV}$. The leading order Feynman diagram is depicted in Figure 5.2. The dominant contribution involves a top quark loop, another contribution involving a bottom quark loop is also possible. The top quark loop allows to probe the Yukawa coupling between the Higgs boson and the top quark.

An $H \to b\bar{b}$ search in the gluon fusion mode used to be considered nearly impossible due to the overwhelming background of QCD multijet, seven orders of magnitude larger than the signal in terms of cross-section. In the case of QCD production of a $b\bar{b}$ pair, this background has the same final state particles as the signal, such that it becomes impossible to distinguish among the two processes. This type of background is referred to as an *irreducible* background (in opposite to *reducible* background, where final state particles differ from the signal).

However, progress has been made in the past years by exploring the boosted topology, where the Higgs boson transverse momentum is large (above $450\,\mathrm{GeV}$). Such a topology occurs when a hard (i.e. with a large momentum) gluon is radiated in addition to the Higgs boson. Variables describing the structure of the Higgs jets are used to separate the signal from the QCD background. The first results of this analysis can be found in [35]. For a Higgs boson mass of 125 GeV, an excess of events above the expected background is observed (expected) with a local significance [2] of 1.5σ (0.7σ) standard deviations.

[2] The concept of significance is described in a later chapter, see section 6.1.

5.2.2 Vector boson fusion production of the Higgs boson

The vector boson fusion has a cross-section of $1.975\,\mathrm{pb}$ at $13\,\mathrm{TeV}$. The leading order Feynman diagram is depicted in Figure 5.3.

An typical $H \to b\bar{b}$ search in this production mode requires four energetic jets in the final state: two b jets from the Higgs decay and two light jets (up or down) from the two vector bosons. Such a final state can be produced by QCD multijet, which is, in this case, the main source of irreducible background. The background reduction relies on certain characteristics of the signal, such as kinematics of the b jets, composition of the light jets (to remove jets originating from gluons) and the soft activity outside the jets (a low soft activity is expected in the rapidity η gap between the two b jets, see section 6.4.7). Nevertheless, such analysis is still very difficult due to the QCD background. The latest result of this analysis can be found in [36]. The observed (expected) limit on the cross section for Higgs boson production in association with top-quark pairs for a Higgs boson mass of 125 GeV is 5.8 (5.2) times the standard model expectation.

5.2.3 Associated production of the Higgs boson with top quarks

The associated production with top quarks has the smallest cross-section of all Higgs production modes, around $509\,\mathrm{fb}$ at $13\,\mathrm{TeV}$. The leading order Feynman diagram can be seen in Figure 5.5.

The final state can be separated into two categories depending on the W bosons decay: *hadronic*, where both W boson decay into light quarks, and *leptonic*, where at least one of the W bosons decay into a pair of lepton + neutrino. The presence of leptons in the final state removes most of the QCD multijet background, which is dominant in the hadronic case, about six orders of magnitude larger than the signal. The dominant background for the leptonic case is the QCD top-antitop multijet, four order of magnitude larger than the signal, and is irreducible when four b jets are produced in the final state. Also, as the signal produces four b jets, assigning two b jets that come from a Higgs boson decay is an ambiguous choice that leads to a background from combinatorics. The combinatorics of b jets, large backgrounds and a small production cross-section make this channel very challenging. The latest results of this analysis can be found in [37] [38]. The observed (expected) limit on the cross section for Higgs boson production in association with top-quark pairs for a Higgs boson mass of 125 GeV is 3.8 (3.1) times the standard model expectation for the hadronic analysis. For the leptonic case, an excess of events above the expected background is observed (expected) with a local significance of 1.6σ (2.2σ) standard deviations.

5.3 Higgs strahlung

The Higgs strahlung is the most suited production mode to perform an $H \to b\bar{b}$ measurement. The combination of this Higgs production and decay mode is referred to as the *VH(bb)* process, where the V stands for a vector boson (Z or W), produced in association with the Higgs boson. A measurement of the VH(bb) process is the goal of the analysis presented in this thesis.

When considering a leptonic decay of the vector boson, such as $Z \to l^+l^-$, $Z \to \nu_l\overline{\nu}_l$ or $W^- \to l^-\overline{\nu}_l$, , the presence of leptons or large MET removes most of the QCD multijet background that is present in the other Higgs production modes. Unlike the production associated with a top quark, there is no combinatoric background, as there are only two b jets in the final state.

5.3.0.1 Signal definition

The VH(bb) production mode has a cross-section around $0.5824\,\mathrm{pb}$ at $13\,\mathrm{TeV}$. The signal final state for this analysis consists of two b jets from the Higgs boson decay, and one vector boson decaying leptonically. This production is quark induced when the vector boson is a W. For the Z case, it can be either quark or gluon induced. The latter contribution has a cross-section of $0.1227\,\mathrm{pb}$ at $13\,\mathrm{TeV}$. Examples of Feynman diagrams for both the quark (*qqVH process*) and gluon induced (*ggZH process*) contribution are depicted in Figure 5.6.

A typical event signature for the signal is a *dijet* system, consisting of the $b\bar{b}$ pair from the Higgs boson decay, back-to-back in momentum space to the vector boson [3]. The nature and decay of the vector boson can be separated into three *channels*. The *Z(ll)H(bb) channel* correspond to a Z boson decaying into a e^+e^- or $\mu^+\mu^-$ pair. The *W(lv)H(bb) channel* is the case when the associated W boson decays leptonically to an electron-neutrino or muon-neutrino pair. The *Z(vv)H(bb) channel* correspond to a Z boson decaying into two neutrinos. The signature from the vector boson is the presence of MET (caused by the two neutrinos), back-to-back to the dijet system.

The case of a vector boson decaying to a tau lepton is not explicitly included in the three channels mentioned above. The branching ratio of the purely leptonic tau decay is 17.82% for the electron and 17.39% for the muon [3]. If at least one tau decays leptonically, it partially contributes to the electron and muon channel. The hadronic decay of the tau has a branching ratio of 64.79%. In case of a hadronic decay of both tau leptons, the event would have to two light jets in addition of the b jets. As no dedicated reconstruction for the tau leptons in considered, the term *lepton* will refer to either electron and muon in the rest of this dissertation.

[3] The two bosons don't have completely back-to-back transverse momentum due to reconstruction effects and because the presence of additional jets from an initial state radiation can occur.

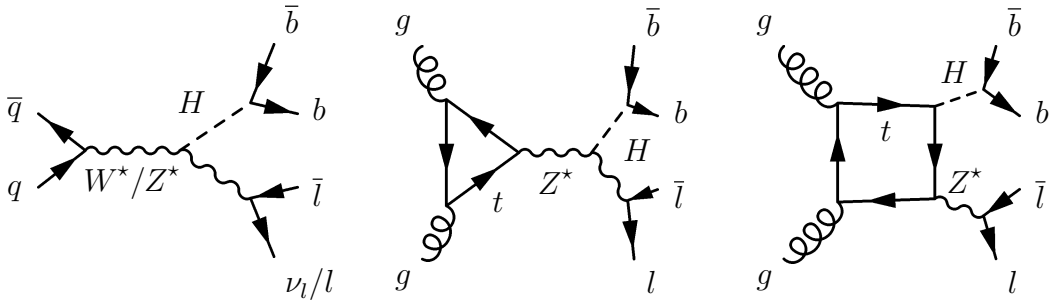


Figure 5.6: LO diagram for associated production with a vector boson. **Left:** qqVH production process. **Middle and right:** ggZH production process.

5.3.1 Backgrounds processes

Multiple backgrounds need to be considered when measuring the VH(bb) process. The last part of this chapter is dedicated to a description of the main background processes and the variables that can be used to separate them from the signal. Those variables are used in the analysis to define a region of phase-space with reduced background contribution and with a high signal efficiency, referred to as the *signal region*. The selection of the signal region is defined in a later chapter, see Table 8.1.

5.3.1.1 Diboson process

The diboson process, *VZ(bb)*, has the same final state particles are the signal. The bottom quark pair comes from a Z decay instead of the Higgs boson decay in the signal. An example of leading order Feynman diagram for the diboson process can be found in Figure 5.7. The main variable to separate the VH(bb) signal from the diboson is the invariant mass of the dijet system, which peaks around $125\,\mathrm{GeV}$ for the signal at $91\,\mathrm{GeV}$ for the background. The $13\,\mathrm{TeV}$ cross-section of the diboson production is $47.13\,\mathrm{pb}$ for a W boson and $16.52\,\mathrm{pb}$ for a Z boson.

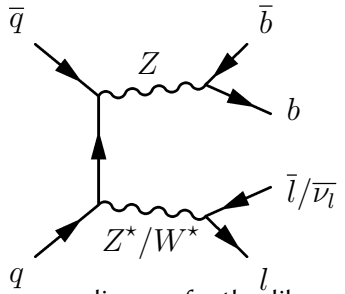


Figure 5.7: Example of LO Feynman diagram for the diboson process.

5.3.1.2 Vector boson + jets

The vector boson plus jets (*V + jets*) process corresponds to the production of a W or Z boson in association with jets. The case where the vector boson is a Z is referred to as the *Drell-Yan* process. Example of Feynman diagrams for V + 2 jets processes can be found in Figure 5.8.

If there are no or only one b jet, requiring at least two identified b jets significantly reduces the background contribution. In the case of two b jets, the background has a topology very close to the signal. This irreducible background has a softer boost in the vector boson momentum distribution than the signal and no invariant mass resonance of the dijet around the Higgs mass, which can be used to reduce its contribution. In the Z(ll)H(bb) channel, the reconstruction of the Z allows to further reduce the Z + jets contribution by excluding events in the side-bands (i.e. outside the resonance) of the Z mass distribution.

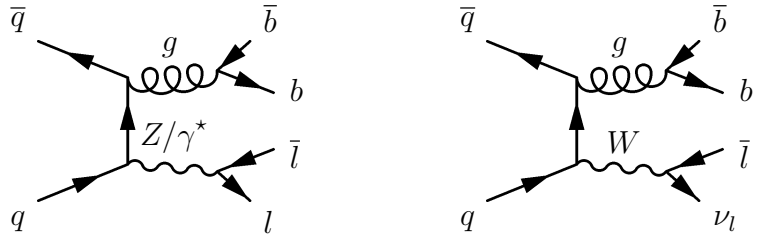


Figure 5.8: Example of LO Feynman diagram for V + 2 jet production at the LHC. **Left:** Drell-Yan + 2 jets. **Right:** W + 2 jets.

5.3.1.3 Top pair production

The LO Feynman diagram for a top and antitop quark production ($t\bar{t}$ process) is depicted in Figure 5.9. The event contains two W decays and at least two b jets. It has a cross-section of $831.76\,\mathrm{pb}$ at $13\,\mathrm{TeV}$.

The primary handle to reduce this background is topological. In a typical signal event topology, the dijet system and the vector boson are back-to-back. But for $t\bar{t}$, the azimuthal angular separation between the two systems is more broadly distributed. Another difference in topology is the presence of additional jets to the two b jets in $t\bar{t}$ events, originating from the hadronic decay of the W boson. In case of a leptonic decay of both W boson, the background separation in the Z(ll)H(bb) channel can benefit from the use of MET, which is larger for $t\bar{t}$. For the W(lv)H(bb) channel, reconstructing the mass of the top quark from the b quark and the MET can be used to eliminate background events. In the same channel, the MET and lepton have smaller angular separation for the signal than the $t\bar{t}$ background, which can also be used to reduce the $t\bar{t}$ background.

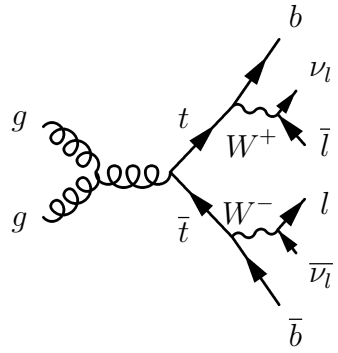


Figure 5.9: The LO Feynman diagram for $t\bar{t}$ process with the $b\bar{b}\,l\bar{l}$ final state.

5.3.1.4 Single top

Three channels contribute to the single top quark production (Figure 5.10). The corresponding cross-sections at $13\,\mathrm{TeV}$ are: $10.32\,\mathrm{fb}$ for the s-channel, $26.33\,\mathrm{fb}$ for the t-channel and $35.6\,\mathrm{fb}$ for the tW-channel.

This background can be reduced using the same variables as for $t\bar{t}$ described in the previous section. The main difference is that the Z(ll)H(bb) channel is very little affected by this background, as there is no pair of real lepton in the final state.

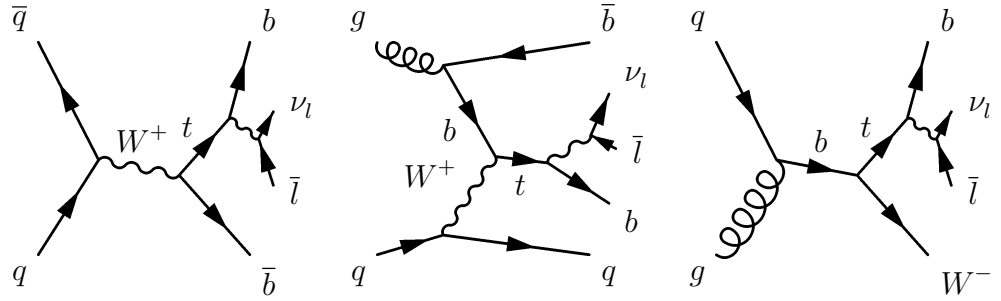


Figure 5.10: Example of LO Feynman diagrams for single-top. Left: s-channel. Middle: t-channel. Right: Wt-channel.

5.3.1.5 QCD multijet

An example LO Feynman diagram for the QCD multijet with a $b\bar{b}$ pair in the final state is depicted in Figure 5.11.

Due to the absence of leptons in the final state, this background is almost completely suppressed in the Z(ll)H(bb) and the W(lv)H(bb) channel. It is still relevant in the Z(vv)Hbb region, mainly due to MET from the mis-measurement of the jet energy.

MET due to the presence of high p_T neutrino in hadronic decays can also contribute to a lower extent. A common feature in both cases is the presence of one jet close to the direction of MET. A selection on the azimuthal angle between the MET and closest jet therefore reduces part of the QCD background. Also, as the QCD multijet is softer than the VH(bb) signal, requiring a large MET further reduces the QCD contribution.

The cross-section, topology and discriminating variables for the main backgrounds of the VH(bb) analysis are summarized in Table . The second column lists the cross-section of each process. The final state objects are listed in the third column, as well as the differences in the event topology with respect to the signal. In the fourth column, the main discriminating variables to reduce each background process are listed. For the vector boson + jets process, the discriminating variables are separated for the V + 0 or 1 b jet (# b jet < 2) and V + 2 or more b jets (# b jet ≥ 2) case. The topology and discriminating variables for the $t\bar{t}$ process are separated according to the decay of the W boson pair. The cross section for the QCD multijet corresponds to a bin of transverse hadron momentum between 100 and $200\,\mathrm{GeV}$.

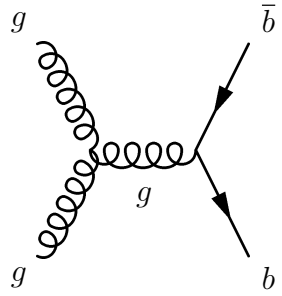


Figure 5.11: Example of LO feynman diagram for QCD $b\bar{b}$ production.

Process	Cross-section	Topology	Discriminating variables
VH(bb)	$0.7612\,\mathrm{pb}$ for qqZH(bb) $0.1227\,\mathrm{pb}$ for ggZH(bb) $1.373\,\mathrm{pb}$ for qqWH(bb)	• Vector boson and $b\bar{b}$ pair back-to-back in momentum space.	–
Background processes:			
Diboson	$47.13\,\mathrm{pb}$ for WZ(bb) $16.52\,\mathrm{pb}$ for ZZ(bb)	• Vector boson and $b\bar{b}$ pair back-to-back in momentum space. • Invariant mass of $b\bar{b}$ peaks at $91\,\mathrm{GeV}$.	Dijet invariant mass.
Vector boson + jets	$6100.8\,\mathrm{pb}$ for Drell-Yan $74447.307\,\mathrm{pb}$ for W + jets	• Vector boson with additional jets. • May not have $b\bar{b}$ jet pair. • Softer vector boson momentum than the signal. • More steeply falling dijet invariant mass than the signal.	• Vector boson with additional jets. • Number of identified b jets. • Dijet invariant mass. • Vector boson transverse momentum.
Top pair production	$831\,\mathrm{pb}$	• Presence of MET close to the leptons. • if $2 \times W \to l\overline{\nu}_l$: l^+l^- and $b\bar{b}$ pair. • if $W \to l\overline{\nu}_l$ and $W \to q\overline{q}$: one lepton, MET and $b\bar{b}$ pair.	• Angular separation between dijet and vector boson system. • if $2 \times W \to l\overline{\nu}_l$: MET. • if $W \to l\overline{\nu}_l$ and $W \to q\overline{q}$: MET. Reconstructed top mass. Angular separation between MET and lepton.
Single top	$35.6\,\mathrm{pb}$, $22.3\,\mathrm{pb}$ and $10.32\,\mathrm{pb}$ for tW-, t- and s-channel	• **s-channel:** one lepton, MET and $b\bar{b}$ pair. • **t-channel:** one lepton, MET and b jet + additional jet. • **Wt-channel:** similar to top pair production.	Same as top pair production
QCD multijet	$27990000\,\mathrm{pb}$	• $b\bar{b}$ pair and MET.	• Presence of leptons • Transverse momentum of vector and Higgs boson

Table 5.3: The cross-section, topology and discriminating variables for the main backgrounds of the VH(bb) analysis are summarized in Table . The second column lists the cross-section of each process. The final state objects are listed in the third column, as well as the differences in the event topology with respect to the signal. In the fourth column, the main discriminating variables to reduce each background process are listed. For the vector boson + jets process, the discriminating variables are separated for the V + 0 or 1 b jet (# b jet < 2) and V + 2 or more b jets (# b jet ≥ 2) case. The topology and discriminating variables for the $t\bar{t}$ process are separated according to the decay of the W boson pair. The cross section for the QCD multijet corresponds to a bin of transverse hadron momentum between 100 and $200\,\mathrm{GeV}$.

6
VH(bb) Analysis

This chapter is an overview of the various tools used in the VH(bb) analysis, focusing on the Z(ll)H(bb) channel. Section 6.1 gives a summary of the relevant statistical methods and multivariate analysis techniques. Section 6.2 is dedicated to the data and section 6.3 to the Monte-Carlo samples. Section 6.4 gives an overview of the main physics objects used in this analysis. It includes the reconstruction, selection and corresponding corrections of the final state objects and variables used for background rejection, for both data and Monte-Carlo samples.

The Z(ll)H(bb) channel can separated in two *sub-channels*, depending on the nature of the leptons. The Z(ee)H(bb) sub-channel correspond to the $Z \rightarrow e^{-}e^{+}$ and the Z($\mu\mu$)H(bb) to the $Z \rightarrow \mu^{-}\mu^{+}$ decay.

6.1 Statistical methods

A measurement of the Higgs boson cross-section requires a refined statistical methodology, taking into account all the systematic and statistical uncertainties. The statistical tools for the cross-section measurement and the signal extraction are reviewed below. More details can be found in [39, 40]. The signal-background discrimination is performed with *boosted decision trees*, a multivariate analysis technique, described below.

6.1.1 Statistical test

The standard procedure for the discovery of a new signal process is to test against an alternate *null hypothesis*, H_0. The hypothesis H_0 includes all standard model background processes without a Higgs boson contribution, and is therefor also referred to as the *background only hypothesis*. The *alternate hypothesis*, H_1, includes both the

background processes and the Higgs boson signal. A search for the presence of a Higgs boson signal doesn't give a statement about H_1 but about H_0, which could be rejected by the observation. For a given dataset x, one can estimate the probability p (the p-value) of finding a dataset of greater or equal incompatibly with the prediction H_0. One can consider the null hypothesis as excluded if its p-value is observed below a specified threshold.

The p-value can be converted into an equivalent significance, such that the probability to observe a random variable following a standard normal distribution above a value Z is equal to p. The significance Z is related to the p-value as

$$Z = \Phi^{-1}(1-p),$$

where Φ is a Gaussian cumulative distribution function.

In particle physics, one often refers to the p-value as "a significance of $Z \times \sigma$". By convention, the term *evidence* is employed when the p-value for the rejection of the null hypothesis is higher than 3σ and the term *discovery* when it is higher than 5σ.

The dataset x usually refers to a binned distribution of a physical observable in the data, such as the boosted decision trees score. The signal-background discrimination of this distribution has been optimized by the analysis. If the significance points to the presence of a Higgs boson signal, on of parameters of interest is the *signal strength*, a multiplicative factor applied to the signal Monte-Carlo simulation. It corresponds to the ratio

$$\mu = \frac{\sigma_{data}}{\sigma_{theory}},$$

where σ_{data} is the signal cross-section measured on the dataset and σ_{theory} the signal cross section predicted for a standard model Higgs boson. The measurement of μ quantifies how much the data deviates from the theoretical predictions. Additional parameters of the background model are the *nuisance parameters*, which parameterize the systematic uncertainties and whose values are not known a priori and have to be extracted from the data.

The test statistic to establish the discovery relies on a likelihood ratio. For a binned histogram, assuming the content of each bin follows a Poisson distribution, the likelihood function is

$$\mathcal{L}(x|\mu) = \prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-\mu s_i - b_i},$$

where s_i (b_i) is the signal (background) yield in the i'th bin, $n_i = s_i + b_i$ the total yield of the i'th bin and μ the signal strength modifier. The bin content, prior to the data entering the statistical analysis, is subject to multiple uncertainties that are handled by introducing nuisance parameters θ, so that signal and background expectations

become functions of the nuisance parameters: $s(\theta)$ and $b(\theta)$[1] [40]. The systematic uncertainty follows a probability distribution function (*pdf*) $\rho(\theta|\tilde{\theta})$, where $\tilde{\theta}$ is the best estimate from the nuisance parameter θ. The pdf follow a normal or log-normal distribution, see section 8.3.1. The values of $\tilde{\theta}$ are estimated from studies of the detector performance on the data. To include the nuisance parameter in the likelihood function, the pdf $\rho(\theta|\tilde{\theta})$ is interpreted at posteriory via the Bayes theorem

$$\rho(\theta|\tilde{\theta}) \sim p(\tilde{\theta}|\theta) \cdot \pi_\theta(\theta),$$

where $p(\tilde{\theta}|\theta)$ is the likelihood to measure the best estimate $\tilde{\theta}$ prior to the nuisance parameter θ and $\pi_\theta(\theta)$ is a flat prior. The likelihood $p(\tilde{\theta}|\theta)$ is now equal to $\rho(\theta|\tilde{\theta})$ and can be included in the likelihood function as

$$\mathcal{L}(x|\mu,\theta) = \prod_i \frac{(\mu s_i(\theta) + b_i(\theta))^{n_i}}{n_i!} e^{-\mu s_i(\theta) - b_i(\theta)} \cdot p(\widetilde{\theta}|\theta).$$

The test statistic t_μ to estimate the p-value is defined as

$$t_\mu = \begin{cases} -2\ln \frac{\mathcal{L}(x|\mu,\hat{\hat{\theta}}(\mu))}{\mathcal{L}(x|\hat{\mu},\hat{\theta})} & \text{if} \quad \hat{\mu} \geq 0 \\ -2\ln \frac{\mathcal{L}(x|\mu,\hat{\hat{\theta}}(\mu))}{\mathcal{L}(x|0,\hat{\hat{\theta}}(0))} & \text{if} \quad \hat{\mu} < 0 \end{cases}, \tag{6.1}$$

where $\hat{\mu}, \hat{\theta}$ are the unconditional maximum-likelihood estimator of the signal strength and nuisance parameters estimated from the maximum likelihood fit. The estimator $\hat{\hat{\theta}}(\mu)$ and $\hat{\hat{\theta}}(0)$ maximize the likelihood function for a fixed signal-strength value, μ and 0, respectively. The two cases in the definition of t_μ take into account the assumption that the presence of signal can only increase the event yield, i.e. $\mu \geq 0$.

This test statistic t_μ can be used to estimate the p-value of any null hypothesis that predicts a particular value of μ. The idea behind this definition is that higher values of t_μ correspond to increasing incompatibility between the data and μ.

6.1.2 The observed significance

For the background-only H_0 hypothesis mentioned at the beginning of this chapter, the corresponding test statistic is t_0, a particular case of Equation 6.1 for $\mu = 0$,

$$t_0 = \begin{cases} -2\ln \frac{\mathcal{L}(x|0,\hat{\hat{\theta}}(0))}{\mathcal{L}(x|\hat{\mu},\hat{\theta})} & \text{if} \quad \hat{\mu} \geq 0 \\ 0 & \text{if} \quad \hat{\mu} < 0 \end{cases}.$$

[1] The notation θ refers to all the nuisance parameters θ_1, θ_2, etc.

The p-value can be computed as

$$p_0 = \int_{t_{0,obs}}^{\infty} f(t_0|0)\mathrm{d}t_0, \tag{6.2}$$

where $t_{0,obs}$ is measured on the dataset. This calculation requires the knowledge of the pdf $f(t_0|0)$, which can be approximated as

$$f(t_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_0}}e^{-\frac{t_0}{2}} \tag{6.3}$$

in the asymptotic limit (for large sample size), where $\delta(q_0)$ is the Dirac delta function. Combining the equations 6.2 and 6.3, it can be shown that the observed significance $Z_{0,obs}$ in the asymptotic limit is

$$Z_{0,obs} = \sqrt{t_{0,obs}}.$$

In practice, it is known that such an asymptotic behavior works very well even for cases with very few expected events [40]. An alternative approach relies on Monte-Carlo toys generating pseudo-data, as described below.

6.1.3 The expected significance

The pdf of t_0 depends on the underlying hypothesis. In the previous case, $f(t_0|0)$ is defined prior to the background-only hypothesis. Other pdfs assuming a signal + background hypothesis with a particular value of the signal strength $\mu' > 0$, $f(t_0|\mu')$, can also be defined.

The case $\mu' = 1$, that is if the signal strength is equal to the standard model predictions, is particularly important. It is used to estimate the *expected* (or *median*) significance, which quantifies the expected sensitivity of the analysis before measuring the observed significance on the data.

Two versions of the expected significance are mainly used. The *prefit expected significance* is independent of the data. It is evaluated with the nuisance parameter values $\tilde{\theta}$ prior the maximum-likelihood fit and is the figure of merit to optimize the analysis sensitivity. The *postfit expected significance* is evaluated with the best fit nuisance parameters $\hat{\theta}$ prior to the maximum likelihood fit in the data. It is the expected significance quoted in the result of this analysis, together with the observed significance.

If $f(t_0|0)$ and the median t_{Med} of $f(t_0|1)$ are known, the expected sensitivity can be estimated. It corresponds the p-value of $f(t_0|0)$ at the median

$$p_{Exp} = \int_{t_{Med}}^{\infty} f(t_0|0)\mathrm{d}t_0,$$

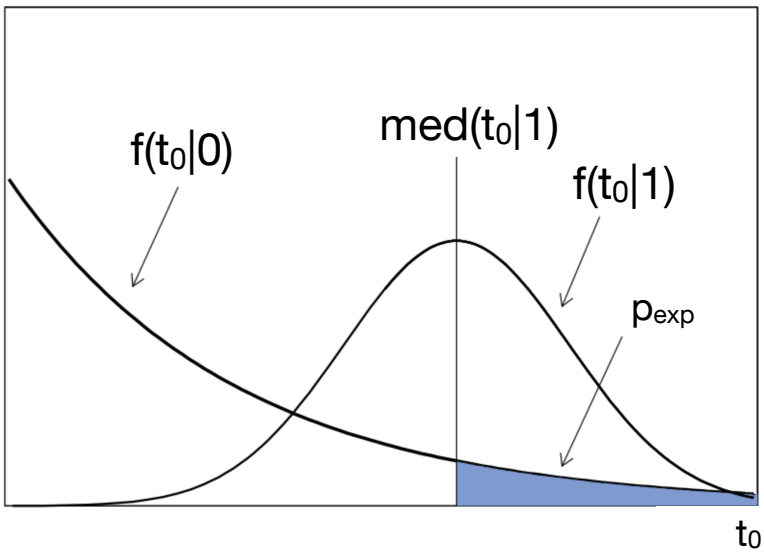


Figure 6.1: Illustration of the $f(t_0|0)$ and $f(t_0|1)$ distributions. The blue area corresponds to the expected significance p_{Exp}. It is evaluated at the median of $f(t_0|1)$. The figure is taken from [40].

where t_{Med} is the median of the $f(t_0|\mu')$ distribution and p_{Exp} is the p-value of the expected sensitivity.

The idea behind this approach is illustrated in Figure 6.1, which shows the $f(t_0|0)$ and $f(t_0|1)$ shapes. Large values of t_0 are less likely for $f(t_0|0)$, which therefore decreases with t_0. $f(t_0|1)$ is shifted to higher value of t_0. The expected significance corresponds to the p-value of $f(t_0|0)$ at the median of the $f(t_0|1)$ distribution.

6.1.4 Evaluation of the probability distribution function

6.1.4.1 Monte-Carlo toys

The pdf $f(t_0|\mu')$ can be estimated with Monte-Carlo generated pseudo-data. The nuisance parameters are fixed to the values $\tilde{\theta}$ ($\hat{\theta}$) in the pseudo-data generation for the prefit (postfit) expected significance, but are allowed to float in the likelihood fit evaluating the test statistic t_0. The event count in each bin follows a Poisson probability distribution assuming a signal strength μ'. This method gives a solid estimate of the pdf but may require a high computing capacity for low p_{Exp} values, as it requires to populate the tails of the distribution.

6.1.4.2 Asymptotic Limit

The alternative approach which is used for this analysis relies on the asymptotic limit approximation and the *Asimov dataset*. In the asymptotic limit, it can be shown that

$f(t_0|\mu')$ approaches the distribution

$$f\left(t_0|\mu'\right) = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta\left(q_0\right) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right] \quad (6.4)$$

where Φ the Gaussian cumulative distribution function and σ is the standard deviation of $\hat{\mu}$. The median of $f(t_0|\mu')$ can be estimated with the Asymov dataset, as described below.

The Asimov dataset is an artificial dataset, generated by Monte-Carlo assuming the signal strength μ' and nuisance parameters θ. It is defined as a binned dataset, where the number of events ν_i in a bin i is exactly the number of expected events in this bin, i.e.

$$\nu_i = \mu' s_i(\theta) + b_i(\theta),$$

where the values of ν_i are fixed by the underlying hypothesis μ' and the nuisance parameters θ are equal to $\tilde{\theta}$ or $\hat{\hat{\theta}}$, for the prefit or postfit significance, respectively. The test statistic $t_{\mu,A}$ can be evaluated on the Asimov dataset, following the definition 6.1. It can be shown that

$$t_{\mu',A} = \text{med}(t_0|\mu'). \quad (6.5)$$

i.e. the median of the pdf $f\left(t_0|\mu'\right)$ is equal to the test statistic evaluate on the Asimov dataset, $t_{\mu',A}$. The expected significance can therefore be estimated by evaluating the p-value of $f\left(t_0|0\right)$ at $t_{\mu'=1,A}$.

To summarize, in the Asymptotic limit,

- for $\mu' = 0$, Equation 6.4 becomes 6.3. After measuring the test statistic $t_{0,obs}$ in data, the observed significance is $Z_{0,obs} = \sqrt{t_{0,obs}}$.
- The median of the $f\left(q_0|1\right)$ distribution is equal to the test statistic evaluated on the Asimov dataset. Combining with the previous bullet, the expected significance can be estimated as $Z_{0,Exp} = \sqrt{t_{1,A}}$. When using this method, the expected significance is also referred to as the *Asymov significance*.

6.1.5 Signal strength extraction

If the dataset points to the presence of a signal, a measurement of the signal strength modifier μ can be performed to probe the theoretical predictions of a standard model Higgs boson. A maximum likelihood fit is performed on the profile likelihood λ_μ, defined as

$$\lambda_\mu = \frac{\mathcal{L}(x|\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(x|\hat{\mu}, \hat{\theta})},$$

where the parameters $\hat{\mu}$ and $\hat{\theta}$ are the best fit estimator, and $\hat{\hat{\theta}}(\mu)$ is best fit estimator with the fixed signal strength parameter μ.

The observed signal strength corresponds to best value of the signal strength, $\hat{\mu}$. The confidence intervals on μ, defined by the range μ_{down} and μ_{up}, can be determined in the asymptotic limit, where it can be be shown that the pdf of $-2\ln\lambda_\mu$ is distributed as a chisquare with 1 degree of freedom

$$f\left(-2\ln\lambda_\mu|\mu\right) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_\mu}}\cdot e^{-2\ln\lambda_\mu/2}.$$

For the 68% (95%) confidence level, $\mu_{up/down}$ must satisfy $-2\ln\lambda_\mu = 1$ ($-2\ln\lambda_\mu = 3.84$).

6.1.6 Boosted decision tree

Multiple variables are exploited for signal-background discrimination in the VH(bb) analysis, such as the dijet invariant mass or the MET, as mentioned in section 5.3. Rather than using each variable individually, one combines them using the classification feature of a multivariate analysis technique to build a single discriminant with a higher discrimination power. In the VH(bb) analysis, this is performed by boosted decision trees (BDT) [41], implemented in the TMVA package [42].

The BDT is a method for classification or regression problems, using supervised machine learning. The regression feature corresponds to a general classification problem that estimates the parameter values of a function, which predicts the value of a response variable in terms of the values of other input variables. An overview of the BDT classification and regression algorithm is given below.

6.1.6.1 Classification

Boosted decision trees are trained on part of the Monte-Carlo events to build a *weight file* for a set of pre-defined input variables during a *training step*. The weight files splits the phase-space of the input variables into cells with a *BDT score*, typically between -1 and 1 (although other ranges are sometime used). A low score means the event is likely to be background, a high score means the events is likely to be signal. Once the training is finished, this discriminant can be evaluated on a Monte-Carlo or data event by reading the weight file during the *evaluation step*.

The Monte-Carlo samples are split into two halves: one for the training of the BDT and one for the evaluation. This ensures statistical independence of the samples and avoids bias from *overtraining*. Overtraining can arise when the BDT learns features

of the training samples that are due to statistical fluctuations. A standard method to detect such cases is to compare the BDT score distribution of the training and the evaluation sample, which are to be compatible within their statistical uncertainties.

In the training procedure, information of the discriminating variables (DV) is combined to classify each event in a signal or background category. The events are split into two parts, depending on the value of the DV. The splitting value and DV variable are selected to give the best separation, such that one part contains mostly signal and the other mostly background events. This process is repeated sequentially, and is depicted by a tree-like diagram as in Figure 6.2: the initial sample (*node*) is separated in two parts (*branches*). The new sample (*node*) at the end of each branch is again separated into two branches, and so on, until a given number of final branches (*leaves*) are obtained, or until each leaf is pure signal or pure background, or has too few events to continue. The first node is referred to as the *root node*. All the nodes form a sequence of selections that is refereed to as a *decision tree* (or *tree*).

A small change in the training sample can give a large change in the tree and results. The *boosting* step accommodates for this instability. After a first tree as the one mentioned above is built, a second tree is evaluated after a re-weighting of the events. This reweighting increases the weights of misclassified event, where a signal (background) event ends in a background (signal) leaf. If it ends in a background (signal) leaf, a score of -1 (1) is assigned. After the second tree is built, events are reweighted again and the process is repeated. Those weights emphasize the misclassified events such that they are more likely to me classified correctly in the next tree. The weight $w_{i+1}(x_k)$ applied to the event x_k in the next tree is defined as

$$w_{i+1}(x_k) = w_i(x_k)\frac{1}{N}e^{\beta\cdot\ln((1-\epsilon)/\epsilon)I(x_k)},$$

where β is a parameter that controls the learning rate between the successive trees. A value of $\beta = 1$ is used in this analysis. Larger values increase the learning rate. ϵ is the *error rate* of the tree (number of misclassified events divided by the total numbers of event in the tree) and the function $I(x_k)$ is 0 (1) for a correcly classified (missclasified) event x_i. The normalization parameter N is defined such that the sum of events before and after the reweighting is constant.

A *forest* of N_{trees} trees is built this way. To built the final BDT classifier, a weight is applied to each tree of the forest, depending on the misclassification rate of the tree, to emphasize trees with a low error rate. The final BDT score of an event x_k is calculated by simultaneously applying all the trees on the event as

$$\text{BDT score}(x_k) = \frac{1}{N_{trees}}\sum_{l}^{N_{trees}}\ln\left((1-\epsilon)/\epsilon\right)\cdot h_l(x_k), \tag{6.6}$$

where N_{trees} is the number of decision trees in the forest and $h_l(x_k)$ is the BDT

score of the l'th tree on the event x_k. The weight definitions mentioned in this section correspond to the *AdaBoost* algorithm, which was used in the training of the VH(bb) analysis BDT classifier. More details about this classier are given in a later chapter, see 8.2.1.1.

6.1.6.2 Regression

The training of a BDT regression follows the same principle as for the classification. The main difference is that the target of the regression doesn't take only two values (signal or background). Instead, the regression process aims to approximate a target variable, not present in the data, for a given set of input variables. An example is the ratio between the reconstructed and generated transverse momentum of jets identified as originating from b quarks in the event. This ratio allows to correct for energy losses from a neutrino emission and inefficiencies in the tracker, ECAL and HCAL during the reconstruction process. Additional details about the regression procedure are given in section 6.4.9.1.

The node splitting during the training of the regression tree has to be redefined. It is performed on the variable that gives the maximum decrease in the average square error when attributing a constant value of the target variable in the output node. This average square error is defined as

$$\text{average squar error} = \frac{1}{N} \cdot \sum_{k}^{N} (y(x_k) - \hat{y})^2,$$

where $\hat{y}$ is the average value of the target variable in the output node, $y(x_k)$ is the value of the target variable of an event x_k in the output node and N is the number of events in the output node. The estimate of the target variable y^e in a leaf node is the average of the target variable of all the events in the node, $\hat{y}$.

The event weights for each successive tree in the boosting procedure are defined by estimating the individual *loss* of the event x_k, $L(x_k)$, defined as

$$L(x_k) = \left[\frac{|y^e(x_k) - y(x_k)|}{\max_{\text{events}} \left(\left| y^e\left(x'_k\right) - y\left(x'_k\right) \right| \right)} \right]^2 . \tag{6.7}$$

The denominator is the maximum deviation between the real value of the target variable and its estimated value, considering all the events in the tree. The numerator is the deviation between the real and the estimated value of the target variable for the event x_k. The variable $L(x_k)$ has larger values if y deviates for y^e for the event x_k. Such events are given a larger weight prior to the training of the next tree, defined as

$$w_{i+1}(x_k) = w_i(x_k) \frac{\overline{L}}{1 - \overline{L}}^{1 - L_i(x_k)}, \tag{6.8}$$

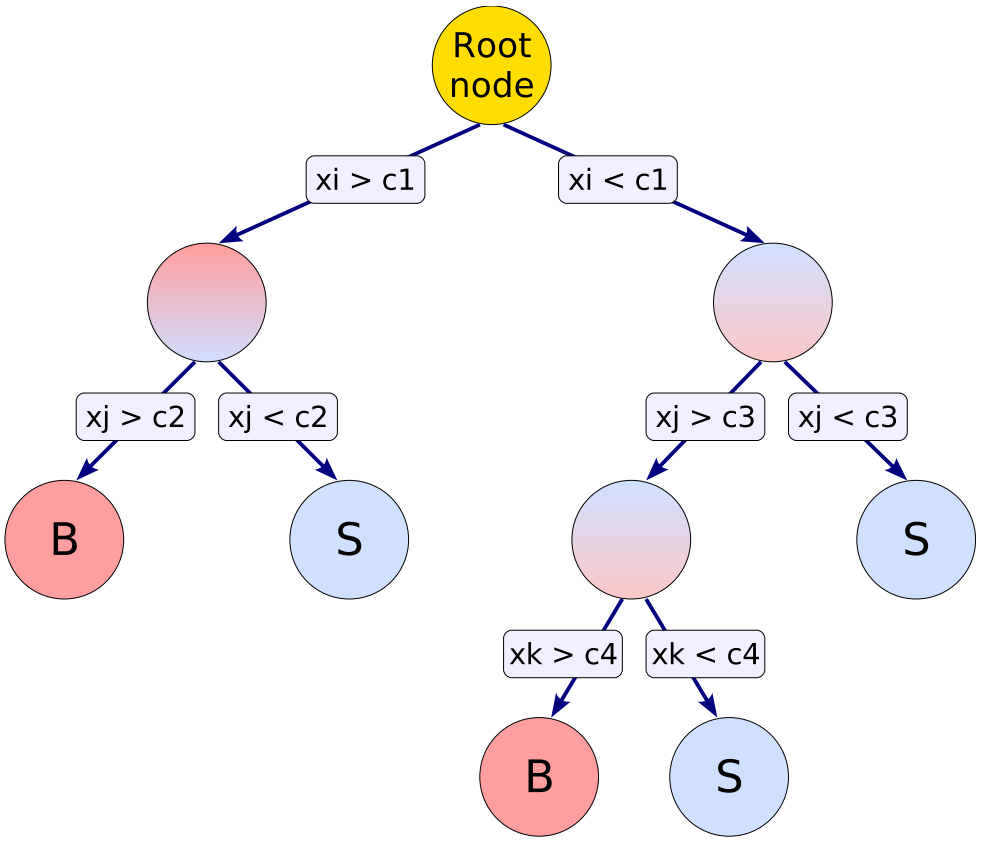


Figure 6.2: Schematic view of a decision or regression tree. Starting from the root node, a sequence of binary splits using the discriminating variable (DV) xi is applied to the data. For decision trees, each split uses the variable that at this node gives the best separation between signal and background when being cut on. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes. For regression trees, the node splitting is performed on the variable that gives the maximum decrease in the average squared error when attributing a constant value of the target variable as output of the node, given by the average of the training events in the corresponding (leaf) node. See section 6.1.6.2. The figure is taken from [42].

where $\overline{L}$ is the weighted averaged of the loss function over all the events in the tree. The larger the loss, the larger the weight, making this event more likely to be picked by the next tree.

For a particular event x_k, each tree h_l predicts a value $h_l(x_k)$. The final value of the target variable in the BDT regression is given by the discriminator h_f corresponding to weighted median of y_i. h_f is defined as

$$h_f = inf\{y \in Y : \sum_{t:h_t<y} \ln((1-\overline{L})/\overline{L}) \geq \frac{1}{2}\sum_t \ln((1-\overline{L})(\overline{L}))\}.$$

This is analogue to the BDT score for the classification case in equation 6.6. Instead of the mean, the median is used for the BDT regression. Intuitively, the weights $\ln((1-\overline{L})/\overline{L})$ can be seen as the equivalent of $\ln((1-\epsilon)/\epsilon)$. A classifier with an average large deviation between the true and the estimated value of the target value will have a lower weight. Additional information about the BDT regression algorithm can be found in [43].

6.2 Data taking

The dataset used in this analysis has been recorded during the year 2016 by the CMS detector at a center-of-mass energy of $13\,\mathrm{TeV}$, with a minimum bunch spacing of $25\,\mathrm{ns}$. It is recorded in different *run periods* (or *Run*) during the year, which are listed in Table 6.1. The integrated luminosity of all the run periods in 2016 correspond to $35.9\,\mathrm{fb}^{-1}$. Each channel uses a different *dataset* taken with a corresponding mix of triggers. The Z($\mu\mu$)H(bb) channel uses the *DoubleMuon* datasets and the Z(ee)H(bb) the *DoubleElectron* dataset.

The average number of pileup interactions is 23. The corresponding distribution of the pileup interactions per event of the 2016 dataset can be found in Figure 6.3. Dedicated pileup weights are applied on the Monte-Carlo simulations to correct the pileup distribution to data, see section 6.3.

6.2.1 High level trigger definitions

The analysis uses unprescaled HLT for the data acquisition, targeting to trigger on both leptons from the Z boson decay. Such HLT are referred to as *double-lepton triggers* (hence the corresponding DoubleElectron or DoubleMuon dataset name). They provide a larger signal efficiency with respect to *single-lepton triggers*, whose selection requires a single lepton in the final state, due to lower thresholds on the lepton p_T. The trigger definitions are listed in Table 6.2.

The double-muon HLT has a p_T threshold of $8\,\mathrm{GeV}$ and $17\,\mathrm{GeV}$ for the first and the second muon. A *tracker isolation* selection is applied on both muons, which improves

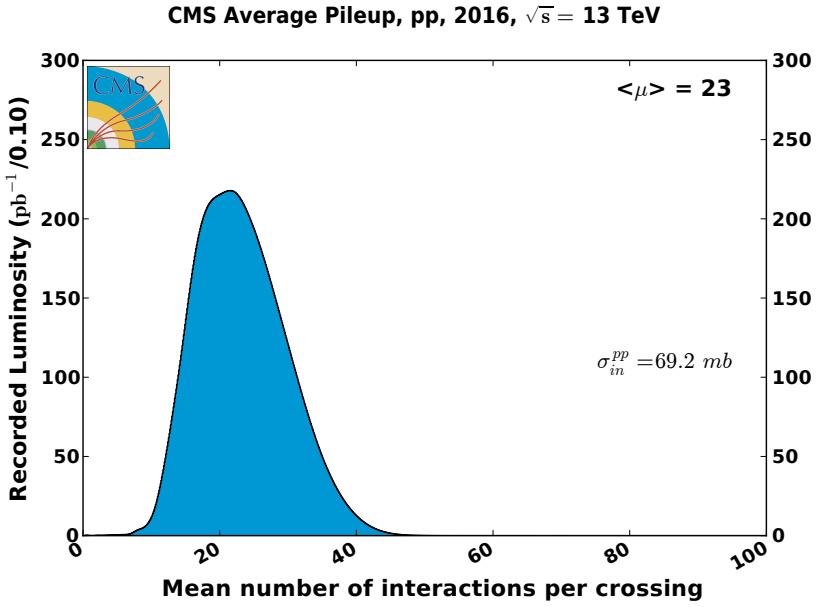


Figure 6.3: Pileup distribution of the 2016 dataset. The average pileup corresponds to 23 interactions per collision event. The figure is taken from [44].

Channel	**Run**	**Integrated Luminosity** (fb^{-1})
DoubleMuon DoubleElectron	Run2016B	~ 5.9
DoubleMuon DoubleElectron	Run2016C	~ 2.7
DoubleMuon DoubleElectron	Run2016D	~ 4.3
DoubleMuon DoubleElectron	Run2016E	~ 4.1
DoubleMuon DoubleElectron	Run2016F	~ 3.2
DoubleMuon DoubleElectron	Run2016G	~ 3.8
DoubleMuon DoubleElectron	Run2016H	~ 11.8
Total		35.9

Table 6.1: All runs of the 2016 dataset used in the Z(ll)H(bb) channel and the corresponding integrated luminosity. The Z($\mu\mu$)H(bb) sub-channel uses DoubleMuon datasets and the Z(ee)H(bb) sub-channel the DoubleElectron datasets.

Channel	L1 Seed	HLT Paths
Z($\mu\mu$)Hbb	SingleMu20	Mu17 TrkIsoVVL_Mu8 TrkIsoVVL Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL OR Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ OR Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ OR
Z(ee)Hbb	SingleEG30 SingleIsoEG22er OR SingleIsoEG24 OR DoubleEG_15_10	Ele23_Ele12 CaloIdL_TrackIdL_IsoVL_DZ

Table 6.2: List of L1 and HLT triggers used for the 2016 dataset, in the Z($\mu\mu$)H(bb) and the Z(ee)H(bb) sub-channels.

the signal purity while keeping a low p_T threshold. An overview of isolation-type selections is given in section 6.4.3.1. For Run 2016 H only, an additional filter is applied on the longitudinal impact parameter (d_Z) separation between the two muon candidate tracks which has to be lower than $2\,\mathrm{cm}$. This requirement is necessary to keep the trigger unprescaled, as the instantaneous luminosity of Run H is higher than for B, C, D, E and F in the 2016 dataset.

The double-electron HLT has a p_T threshold of $12\,\mathrm{GeV}$ and $23\,\mathrm{GeV}$ for the first and the second electron. Additional isolation and identifications selection improves the purity while keeping a low p_T threshold, like in the muon case. Additional information about the electron HLT can be found in section 10.2.2.1 of [24].

Due to mismodeling of the HLT input variables, the efficiency corresponding to the trigger selection can be different between data and Monte-Carlo simulations. Examples of such mismodeling can be found in [45] for the electron case. This effect is corrected by applying efficiency scale factors on the Monte-Carlo simulation after the HLT is applied. For muons, the values and method to derive those corrections are described in chapter 7 and [46]. For electrons, the efficiency corrections can be found in [45]. They are derived with the same method as for the muons.

6.3 Monte Carlo simulation samples

Samples of simulated background and signal events are produced using Monte-Carlo event generators, typically MG5AMC@NLO [47] and POWHEG [18–20]. All processes are interfaced with PYTHIA 8.212 [48] for the parton showering and hadronization. The PYTHIA parameters for the underlying event description correspond to the CUETP8M1 tune [49]. The PDFs used to produce NLO samples are the NLO NNPDF3.0, while the LO NNPDF3.0 set is used for the LO samples [11]. The detector response is modeled with GEANT4 [50]. The list of samples and the corresponding

Process	Generator	Order	Cross-section [pb]
qqZH	POWHEG (v2) [18–20] + MinLO [51, 51] + PYTHIA 8	NLO, rescaled to NLO+NLL QCD	$0.7612 \times 0.10974 \times 0.5824$
ggZH	POWHEG (v2) + PYTHIA 8	LO, rescaled to NNLO QCD	$0.1227 \times 0.10974 \times 0.5824$

Table 6.3: List of Monte-Carlo simulation samples for signal processes. The cross-sections are denoted as *ZH production* $\times$ $Z \to ll \times H \to b\bar{b}$. The cross-section values are taken from [34].

generator, cross-section for signal and background are given in section 6.3.1 and 6.3.2.

The Monte-Carlo samples contain distributions of the number of pileup interactions expected from the data-taking period. The number of primary vertices distribution is sensitive to reconstruction, underlying event modeling and offline selection criteria and is different between the data and Monte-Carlo simulations. To correct for those effects, the simulated numbers of pileup distribution is reweighted per bunch crossing and luminosity section. The pileup weight is extracted from data and depends on the minimum bias cross-section, which is chosen in such a way to reduce the mismodeling between data and Monte-Carlo samples.

6.3.1 Signal simulations

The signal samples are simulated for a $125\,\mathrm{GeV}$ Higgs boson. The qqZH sample is produced at NLO, and the cross-section is rescaled to NNLO QCD + NLO EW (electroweak) accuracy. The cross-section for the qqZH, $\sigma(\mathrm{qqZH})$, can be decomposed as

$$\sigma(\mathrm{qqZH}) = \sigma^{\mathrm{VH,DY}}_{\mathrm{QCD\,NNLO}} + (1 + \delta_{\mathrm{EW}}) + \sigma_{\text{t-loop}} + \sigma_{\gamma},$$

where $\sigma^{\mathrm{VH,DY}}_{\mathrm{QCD\,NNLO}}$ are the Drell-Yan like part of the QCD prediction for VH up to NNLO, δ_{EW} are the inclusive NLO EW corrections, $\sigma_{\text{t-loop}}$ are top loop corrections (not taking into account the contributions for qqZH) and σ_{γ} are contributions from photon-induced channel [34]. Instead of the inclusive NLO EW corrections $(1 + \delta_{\mathrm{EW}})$, NLO electroweak correction differential in the vector boson transverse momentum p_T are used to reweight the qqZH cross-section to NLO EW. The corresponding relative weight can be seen in Figure 6.4. The ggZH sample is produced at LO, and the cross-section is rescaled to NLO+NLL QCD accuracy. The corrections of the signal cross-sections are documented in [34].

The generators for the qqZH and ggZH processes are described in Table 6.3. The cross-sections are denoted as *ZH production* $\times$ $Z \to ll \times H \to b\bar{b}$.

6.3.2 Background simulation

The simulated background samples are listed in Table 6.4, with the corresponding generators and cross-sections. The production cross sections for the $t\bar{t}$ sample are rescaled to the NNLO with the NNLL prediction obtained with TOP++[52]. The Drell-

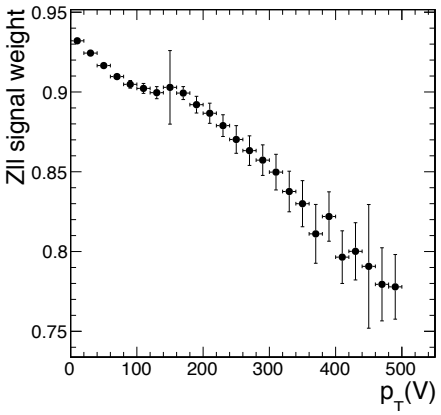


Figure 6.4: Multiplicative signal EW NLO correction applied in bins of Z boson p_T distribution on the qqZH process.

Yan + jets (DY + jets) are rescaled to the NLO cross sections using MG5AMC@NLO, which correspond to a 1.23 k-factor applied to the production cross-section.

To optimise the computation efficiency, the generation of DY + jets samples with an invariant Z boson mass above $50\,\mathrm{GeV}$ is split in different phase-spaces, depending on the number of jets, the jet transverse momentum H_T and the vector boson Z transverse momentum. Those samples are not mutually exclusive, which leads to double-counting events in overlapping regions of phase-space if employed together. To correct for this effect, a *stitching* procedure is performed. The Monte-Carlo samples in a overlapping region of the phases-space are reweighted such that the total yield is conserved. If both samples s_1 and s_2 have events in the overlaping region, they are reweighed by the weight w_1 and w_2, respectively. Those weights are calculated as

$$w_{1/2} = \frac{n_{1/2}}{n_1 + n_2},$$

where $n_{1/2}$ is the number of event in the overlap region from sample 1/2, respectively.

This procedure is validated by the two Z boson p_T distributions in Figure 6.5, performed in a region of phase-space enriched in DY + b jets. The left figure exclusively includes H_T-binned DY + jets samples and the right figure includes the stitched DY + jets samples. The lower part in each figure is the ratio between the data and the Monte-Carlo simulation. The black dots correspond to the values of the data/Monte-Carlo ratio. The hatching bands around 1 correspond to the statistical Poisson uncertainties of the Monte-Carlo samples propagated to this ratio. Those statistical uncertainties are reduced when moving from the left figure to the right figure. Hence including all the DY + jets samples in Table 6.4 increases the statistical power of the Monte-Carlo samples, as expected. In addition, the data and Monte-Carlo distributions agree

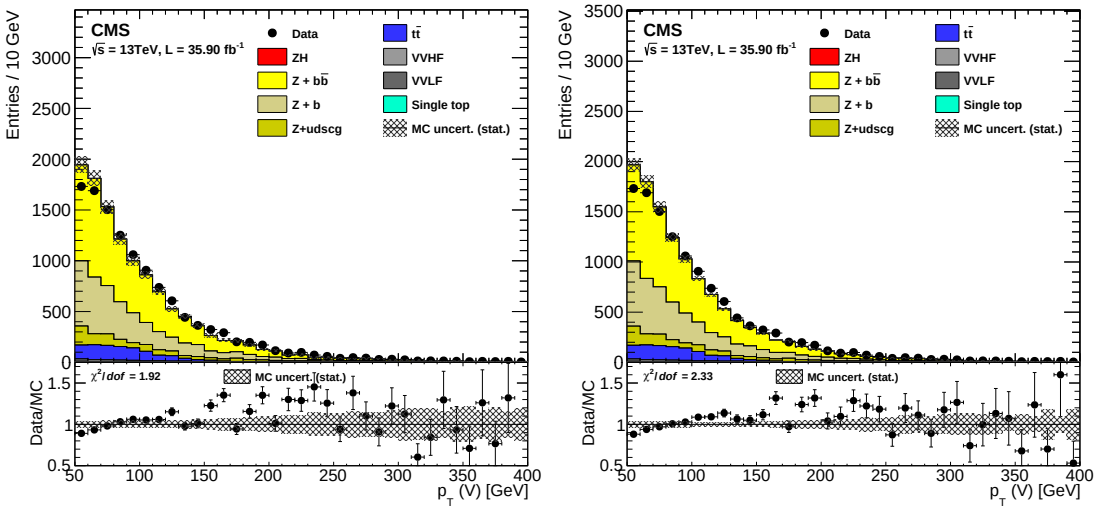


Figure 6.5: Z boson p_T distribution in a region of phase-space enriched in DY + b jets. **Left:** DY + jets include only H_T-binned samples. **Right:** DY + jets include all stitched samples.
The upper part of each figure shows the data and Monte-Carlo distributions. The Monte-Carlo distribution consists of multiple processes stacked together, which are listed in the legend. The lower part of each figure is the ratio between the data and the Monte-Carlo simulation. The black dots correspond to the values of the data/Monte-Carlo ratio. The hatching bands around 1 correspond to the statistical Poisson uncertainties of the Monte-Carlo samples propagated to this ratio. A χ^2 test for comparison between the Monte-Carlo and the data distribution is performed, and the corresponding $\chi^2/ndof$, where $ndof$ is the number of degree of freedom, is quoted in the ratio plot [53].
The statistical uncertainties in the ratio plot are reduced when moving from the left to the right figure, hence including all the DY + jets samples in Table 6.4 increases the statistical power of the Monte-Carlo samples, as expected. The data and Monte-Carlo distributions agree within uncertainties between the left and the right figure, which is a validation of the weights derived in the stitching procedure.

within uncertainties between the left and the right figure, which validates the weights derived from the stitching procedure.

6.4 Physics Objects

In this section, the reconstruction and selection of the physics objects, as well as derivation of Monte-Carlo sample corrections specific to the Z(ll)H(bb) channel, are reviewed. They are based on the particle flow algorithm, described in section 4.4.5.2.

Process	**Generator**	**Order**	**Cross-section** [pb]
Diboson (ZZ)	MG5AMC@NLO [47] + FxFx[54] + PYTHIA 8 [48]	NLO	16.523
DY + Jets, Z_{mass} = [10, 50]:	MADGRAPH 5 [55] + MLM[16]+ PYTHIA 8 [48]	LO, rescaled to NLO	-
DY + 1 Jet	"	"	725
DY + 2 Jet	"	"	394.5
DY + 3 Jet	"	"	96.47
DY + Jets, m(Z) > 50	"	"	
H_T inclusive	"	"	4960×1.23
H_T = [100, 200]	"	"	147.40×1.23
H_T = [200, 400]	"	"	40.99×1.23
H_T = [400, 600]	"	"	5.678×1.23
H_T = [600, 800]	"	"	1.367×1.23
H_T = [800, 1200]	"	"	0.6304×1.23
H_T = [1200, 2500]	"	"	0.1514×1.23
H_T = [2500, inf]	"	"	0.003565×1.23
DY + Jets, m(Z) > 50, b-renriched	"	"	
p_T inclusive	"	"	71.77×1.23
p_T(Z) = [100, 200]	"	"	3.027×1.23
p_T(Z) = [200, inf]	"	"	0.297×1.23
DY + Jets, m(Z) > 50, b gen. filter	"	"	
p_T inclusive	"	"	228.9*1.23
$t\bar{t}$	POWHEG [18–20] (v2) + PYTHIA 8 [48]	LO, rescaled to NNLO	831.76
single top	POWHEG [18–20]+ PYTHIA 8 [48]	LO	-
tW channel	POWHEG [18–20] (v1) + PYTHIA 8 [48]	"	35.6
tW channel, antitop	"	"	35.6
t-channel	POWHEG [18–20] (v2) + PYTHIA 8 [48]	"	44.2
t-channel, antitop	"	"	26.325
s-channel	"	"	10.32

Table 6.4: List of Monte-Carlo simulations for the background processes. The process is given in the first column. The generation of the DY + jets is split in different regions of phase-space, depending on the number of jets, the jet transverse momentum H_T and the vector boson Z transverse momentum. The second column describe the corresponding event generator. Differences between POWHEG v1 and v2 are summarized in [56]. The third column contains the order of the event generator, as well as an eventual rescaling of the cross-section. The fourth column lists the cross-section of each process. For the DY + jets background, whose cross-section has been rescaled from LO to NLO with a 1.23 k-factor, the cross-section is written as "cross-section $\times$ k-factor". A quote in one of the column (") indicates that this parameter is the same as for the sample above.

6.4.1 Primary vertex

The reconstructed variable corresponding to the hard process, the core of the proton-proton interaction, is the *primary vertex*. As it typically involves a large momentum transfer, the reconstructed vertex with the largest value of *summed physics-object* p_T^2 is defined to be the primary vertex. The physics objects used in the sum are the jets, MET and charged lepton reconstructed from the charged tracks associated to the primary vertex. This method has the advantage to take the MET into account, unlike a sum of the charged track p_T that was used in Run 1.

6.4.2 Pileup treatment

The number of additional vertices is related to the amount of pileup interactions. The presence of pileup affects the reconstruction of jets, in particular the energy resolution, MET and the lepton isolation (see section 6.4.3.1). Two techniques are used to mitigate those effects. The *Charged Hadron Subtraction* removes all charged hadrons not originating from the primary vertex during the PF jet reconstruction [57]. It is limited to outside of the tracker region ($2.5 < |\eta|$). The *FastJet* algorithm estimates the energy density per unit area ρ due to pileup in each event and uses it to subtract the pileup energy contribution for each jet [58].

6.4.3 Leptons

As mentioned in section 5.3.0.1, the term lepton refers to an electron or muon, as the vector boson decay to tau leptons is not directly covered by the VH(bb) analysis.

The leptons originating from the vector boson decay are produced at the interaction vertex due to the short lifetime of the vector boson. Such leptons are referred to as *prompt leptons*, or *signal leptons*, as they are the final state leptons of the VH(bb) process. The leptons present in background processes such as QCD multijet are referred to as *fake leptons*. Those can be real leptons, produced away from the primary vertex, or another objects that mimics a lepton signature in the detector, in which case no real lepton is present in the event.

Fake leptons can be produced within a *decay in flight*, where a non-prompt lepton candidate is produced within a c, b hadron or a tau lepton. For example, in the VH(bb) analysis, leptons from a b hadron decay within the dijet system are considered as fake. Additional sources of fake electrons include *jet misidentification*, where a significant shower in the ECAL is reconstructed as an electron, and *converted photons*, where a radiated photon decay into an electron-positron pair. Additional sources of fake muons include hadronic *punch-through*, *cosmic muons* and *duplicates*. In a hadronic punch-through, a hadronic remnant penetrate the muon system and can produce segment in the muon station, causing the reconstruction of fake muon tracks. In the cosmic muon case, muons from cosmic rays leaving a track in both the inner tracker and the

muon system can be reconstructed as a collision muon. Duplicates are caused by a single muon producing multiple track candidates that lead to multiple reconstructed muons.

To reduce fake lepton backgrounds in the final state, leptons must pass an *identification* and *isolation* selection. The identification selection depend on the nature of the lepton and is described in section 6.4.3.3 for electrons and in section 6.4.3.2 for muons. Isolation-type selections are described below.

6.4.3.1 Lepton isolation

The isolation selection mainly addresses fake leptons from decay in flight. Such leptons tend to have a significant flow of energy close to their reconstructed track. In contrast, the presence of other particles originating from the primary vertex near the trajectory of a signal lepton is much less likely. This motivates to use the particle flow *relative isolation* variable, which is defined as the ratio between the transverse momentum of all particles within a ΔR cone around the lepton trajectory and the lepton's transverse momentum

$$\text{Isolation} = \frac{\sum p_{T,CH} + \sum p_{T,NH} + \sum p_{T,\gamma} - PU_{corr}}{p_{T,\mu}},$$

where transverse momentum p_T correspond to charged particles assigned to the hard scattering for $p_{T,CH}$, neutral hadrons for $p_{T,NH}$, particle flow photons for $p_{T,\gamma}$ and the muons for $p_{T,\mu}$. The term PU_{corr} subtracts contributions from neutral pileup particles in the cone. In the muon case, it is performed by subtracting the charged hadron contribution multiplied by a 0.5 factor, corresponding to the measured ratio of neutral to charged hadron production in the hadronization process of pileup interactions [59]. For the electrons case, the pileup contribution is estimated by multiplying the effective area of the isolation cone by the energy density.

Other forms of isolation variables are available for the double-electron and double-muon HLT. In the *tracker-based isolation*, the numerator consists of the transverse momentum of tracks attributed to the primary vertex of a cone around the lepton. Other isolation variable specific to the double-electron HLT are the *ECAL* and *HCAL isolation*, measured within the calorimeters.

6.4.3.2 Muon selection

The muons used in the analysis are reconstructed as global muons. Both *muon candidates* (the muons attributed to a Z boson decay) must have a p_T of at least $20\,\mathrm{GeV}$ and be within the coverage of the muon system $|\eta| < 2.4$. In case of ambiguity due to more than two muons in the events, muons with the largest p_T satisfying the requirement above are selected.

Another requirement is to pass a *loose identification* quality selection. It requires that the muon passes a PF selection and is reconstructed as a global or tracker muon. This identification criteria is designed to have a high efficiency for the signal and also for muons from heavy and light quark decays. In addition, each muon must have an isolation score below 0.25, referred to as the *loose isolation*.

In the W(lv)H(bb) channel, the muon from the W boson decay must pass a tighter identification cut, referred to as the *tight identification* selection. It consists of a list of selection cuts on the muon objects: the χ^2 of the global muon track fit should be lower than 10, at least one muon chamber hit should be included in the global-muon track fit and muon segments should be present in at least two chambers. Those selections suppress background from hadronic punch-through. To suppress cosmic muons and decay in flight, the distance between the transverse impact parameter and the primary vertex has to be lower than $2\,\mathrm{mm}$, the longitudinal distance between the tracker track and the primary vertex has to be lower than $5\,\mathrm{mm}$, the inner track must have at least one pixel hit and at least six layers with hits. In addition, the muon has to be reconstructed as a global and particle flow muon. The isolation selection is tighter than in the Z(ll)H(bb) to remove the QCD multijet background. The muons must have an isolation score below 0.06.

The performances and efficiencies of the muon selections described in this section are documented in [46].

6.4.3.3 Electron selection

The electrons used in this analysis are GSF electrons, whose track is reconstructed with the GSF algorithm (see section 4.4.3). The electron candidates are selected in a similar way to the muon candidates, but the pseudorapidity range is extended to the full tracker coverage $|\eta| < 2.5$ and the ECAL gap between the barrel and the endcap, $1.444 < |\eta| < 1.566$, is excluded as the ECAL information is essential in the electron reconstruction. Both electrons must have p_T of a least $20\,\mathrm{GeV}$.

Additional electron variables related to the ECAL shower shape, electron track reconstruction, as well as geometry and energy matching between the track and the ECAL, are combined with boosted decision trees to further discriminate signal from fake electrons. This *BDT identification* is trained on a Monte-Carlo simulated DY + jets process. The corresponding working point for the Z(ll)H(bb) channel has a signal efficiency of 90%. For the W(lv)H(bb) channel, the electrons muss pass a tighter working point with a corresponding signal efficiency of 80%. An additional isolation selection is required to remove fake electrons from decay in flight, with an isolation value of 0.15 and a cone parameter of $\Delta R = 0.3$. An additional requirements rejects fake electrons from converted photons.

The performances and efficiencies of the electron selections described in this section are documented in [45] .

6.4.4 Jets

Jets are reconstructed from particle flow candidates with the anti-k_T algorithm with a cone radius of $\Delta R = 0.4$, referred to as *AK04 jets*. They are required to have a minimum p_T of $25\,\mathrm{GeV}$. For the two *b-jet candidates* (the two b-jets attributed to the dijet system), only jets within the tracker range are considered ($|\eta| < 2.5$).

In some cases, overlap of multiple pileup jets can give results in hard jets of tens of GeV in p_T. Those overlaps are referred to as pileup jets and can pass the $25\,\mathrm{GeV}$ selection of the analysis. A multivariate discriminant technique, (PUjetID), is used to identify such jets, relying on the jet shape variables and the fraction of charged particles contributed by pileup vertices.

The detector response to particles is not linear and therefore the reconstructed jet momentum and energy doesn't correspond to the one of the true particle or parton. A set of *jet energy corrections* (JEC) corrects the four-momentum of the reconstructed jet to obtain the correct jet energy scale [60]. The JEC consist of three levels of corrections applied sequentially. The first level removes the additional jet energy due to pileup. The second level corrects for non-linear detector response and is derived on Monte-Carlo simulation. The third level corrects the small (% level) differences between data and the Monte-Carlo simulation.

On top of the JEC, a smearing is applied on Monte-Carlo simulated jet p_T to correct for resolution differences between data and Monte-Carlo simulations. This is referred to as *jet energy resolution* (JER) correction. The smearing is performed on Monte-Carlo jets by rescaling the jet p_T by a factor C_{JER}. Two smearing methods are used to estimate the C_{JER} factor. In the *scaling method*, the reconstructed jet is first matched to a jet in generator-level particle. The factor C_{JER} is then computed as

$$c_{JER} = 1 + (S_{JER} - 1)\frac{p_T - p_T^{ptcl}}{p_T},$$

where p_T is the transverse momentum of the reconstructed jet, p_T^{ptcl} is the transverse momentum of the corresponding jet clustered from generator-level particles, and s_{JER} is the data-to-simulation resolution scale factor. In the *stochastic method*, no matching to the generator-level particle is performed. The factor C_{JER} is computed as

$$c_{JER} = 1 + \mathcal{N}(0, \sigma_{JER})\sqrt{\max(s_{JET}^2 - 1, 0)},$$

where σ_{JER} and s_{JER} are the relative p_T resolution in simulation and data-to-simulation scale factors, respectively, and $\mathcal{N}(0, \sigma_{JER})$ denotes a random number sampled from a normal distribution with a zero mean and variance σ_{JER}. In this analysis, the so-called *hybrid method* is used. When a particle-level jet is matched to the reconstructed jet, the scaling method is applied. If no particle-level jet is found, the stochastic method is applied instead.

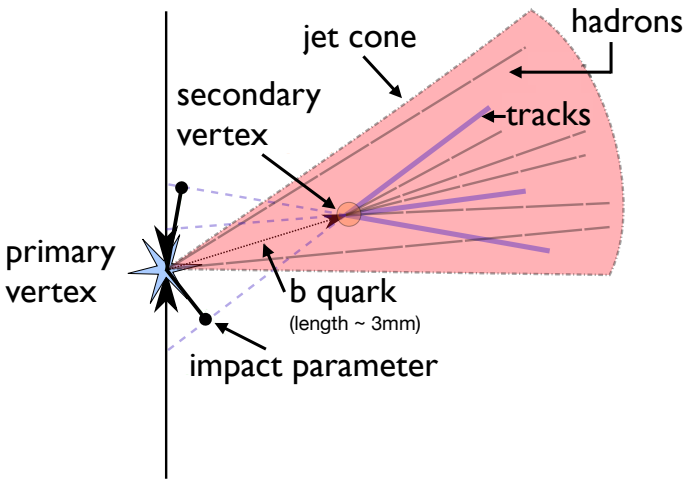


Figure 6.6: An illustration of the decay of a b quark, along with the definition of the secondary vertex and the impact parameter. The figure is taken from [63].

6.4.5 Identification of bottom jets

One of the crucial part of the analysis is to identify jets originating from b quarks. This is performed by a *b-tagger* that exploits particular properties from b hadrons to distinguish them from light (up, down, strange), charm, and gluon-initiated jets. The value of the b-tagger quantifies how likely the jet comes from a b hadron.

The discrimination algorithm exploits the long lifetime of the b hadrons ($\sim 10^{-12}$ s). This results in a displacement *secondary vertex* of a few millimeters between the b hadron decay and the primary vertex. The corresponding detector signature is the presence of displaced tracks within the b jets and a secondary vertex corresponding to the b quark decay. Also, b hadrons have a 20% probability to decay into a muon or an electron, which can be exploited to enhance the b jets purity of the sample and discriminate against light jet contributions, at the expenses of a smaller signal efficiency due to the branching ratio of soft lepton production in b/c-jets. The following description of the b-tagger algorithms has been taken from [61, 62].

The decay of a b quark is illustrated in Figure 6.6, as well as the secondary vertex and tracks' impact parameter, exploited for the track selection.

6.4.5.1 Track selection and vertex identification

The objects used by the b-tagger are reconstructed with the particle flow algorithm. Tracks used for the b-tagger need to pass quality selections. They must have a transverse momentum of at least $1\,\mathrm{GeV}$, a normalized χ^2 of the trajectory fit below 5, at

least 8 hits in the silicon tracker, of which at least two are in the pixel layer of the detector. Additional selections on the track impact parameters, the distance between tracks and the jet axis at the point of closest approach are also required. The decay length, which is the distance between the primary vertex and the point of closest approach between the jet axis and the track trajectory must be less than $5\,\mathrm{cm}$.

Two methods are available for reconstructing the secondary vertex. The *adaptive vertex reconstruction* (AVR) algorithm takes as input the tracks attributed to the jet, passing the selection described above. Additional criteria are then applied to remove secondary vertices not originating from a b hadron decay. The *inclusive vertex finder* (IVF) algorithm takes as input the collection of reconstructed tracks in the event. This algorithm is better suited for b hadron decays at small relative angle giving rise to overlapping, or completely merged, jets [64]. The tracks need to pass quality selections which are looser with respect to the AVR selection. The efficiency to reconstruct a secondary vertex for b jets using the IVF is about 10% higher than with the AVR algorithm, but also increases the fraction of vertices reconstructed for light jets by about 8%. About 60% of jets with an AVR vertex also have an IVF vertex, so both fitter provide independent information.

6.4.5.2 Bottom jets identification algorithm

Multiple methods are available for b jets identification. The b-tagger used in the VH(bb) analysis in Run I was CSVv2, based on the CSV algorithm. It combines information from the displaced tracks and secondary vertex associated to the jets using a multivariate technique. The training of the algorithm is performed on three independent vertex categories. The first category contains jets with at least one associated reconstructed secondary vertex, where the most discriminating variables are the vertex mass and the flight distance. The second category contains jets with a "pseudo vertex" (two good tracks but no vertex fit). The third category uses only the information of displaced tracks. A final discriminant combines the three categories with a likelihood ratio. The CSVv2 algorithm can make use of the AVR vertices CSVv2 (AVR) or the IVF vertices CSVv2 (IVF).

A new b-tagger, CMVAv2, was developed for Run 2. It uses a BDT to combine the CSVv2 (IVF) and the CSVv2 (AVR), as well as three other b-tagger method. The three other b-tagger involved in the combination are the *Jet Probability* (JP), the *Soft Electron* and the *Soft Muon* taggers. The JP tagger exploits the associated tracks to compute the likelihood of the jet to originate from the primary vertex. The track probabilities are then combined to obtain the jet probability as described in section 5.1.1 from [62]. The soft electron (muon) algorithm looks for a reconstructed electron (muon) within the jet, and exploits variables related to the impact parameter of the lepton, as well as the ratio of the transverse momentum of the lepton and the jet, and the transverse momentum of the lepton relative to the jet axis.

The performance of the CMVAv2 b-tagger is compared to the other discriminant in Figure 6.7, which shows the probability for non-b jets to be misidentified as b jets as a function of the efficiency to correctly identify b jets. The curves are obtained from a simulated semileptonic $t\bar{t}$ process. The CMVAv2 algorithm outperforms the other b jet identification algorithms for both c jets as well as for light parton and gluon jets (both treated as light jets).

Three standard working points are defined for each algorithm. They are labeled as *loose*, *medium*, *tight*, and correspond to a threshold on the discriminator after which the misidentification probability is around 10%, 1% and 0.1%, respectively, for light-flavour jets with a transverse momentum above $30\,\mathrm{GeV}$.

The b-tagger algorithm described in this chapter has been optimized for AK04 jets and can be used to b-tag multiple jets in an event, as the two b jets from the dijet invariant mass system in the VH(bb) analysis. It can also identify two overlapping AK04 b jets (i.e. with a ΔR separation below 0.8 between the two jet axis), assigning a separated b-tagging score for each jet. Such scenario are however more suited for a single AK08 jet instead of two AK04 jets and a *double-b tagger* to identify the presence of two b sub-jet within the AK08 jet, as an overlap reduces the reconstruction quality of the b-tagger. Additional details are given in the boosted analysis part, see section 10.2.3.7.

6.4.5.3 Bottom jets identification efficiency correction

Some differences of the b-tagger discriminant shape are observed between the data and Monte-Carlo simulations. This is partly due to the finite order calculations of the $t\bar{t}$ and single-top samples used in the b-tagger optimization, which impacts the modeling of the main b-tagging input variables, such as the track transverse momentum (see [61]). The b-tagger distribution has therefore to be corrected by a scale factor to take into account difference of efficiency between data and Monte-Carlo simulation. Fixed-cut operating cut scale factors to correct Monte-Carlo samples accordingly for the three working points mentioned in section 6.4.5.2 are available. They are obtained using a sample of jets enriched in b quark content by selecting dileptonic and semileptonic $t\bar{t}$ samples. The measurement of the scale factor relies on two methods, the *kinematic selection* for the dileptonic and a *tag and probe* technique on the semileptonic $t\bar{t}$ region, documented in [62]. The distributions of both b-tagging scale factors as a function of the jet p_T can be found in Figure 6.8 for the medium working point. The final scale factor is taken as the average between the two methods. The fixed-cut operating point scale factors are not sufficient to use the information from the full b-tagger distribution. For this purpose, efficiency corrections are extracted as a function of the b-tagger score. They are designed to correct only for shape differences and don't modify the normalization of the Monte-Carlo samples.

The method to extract those data-to-simulation scale factors use an iterative fit tech-

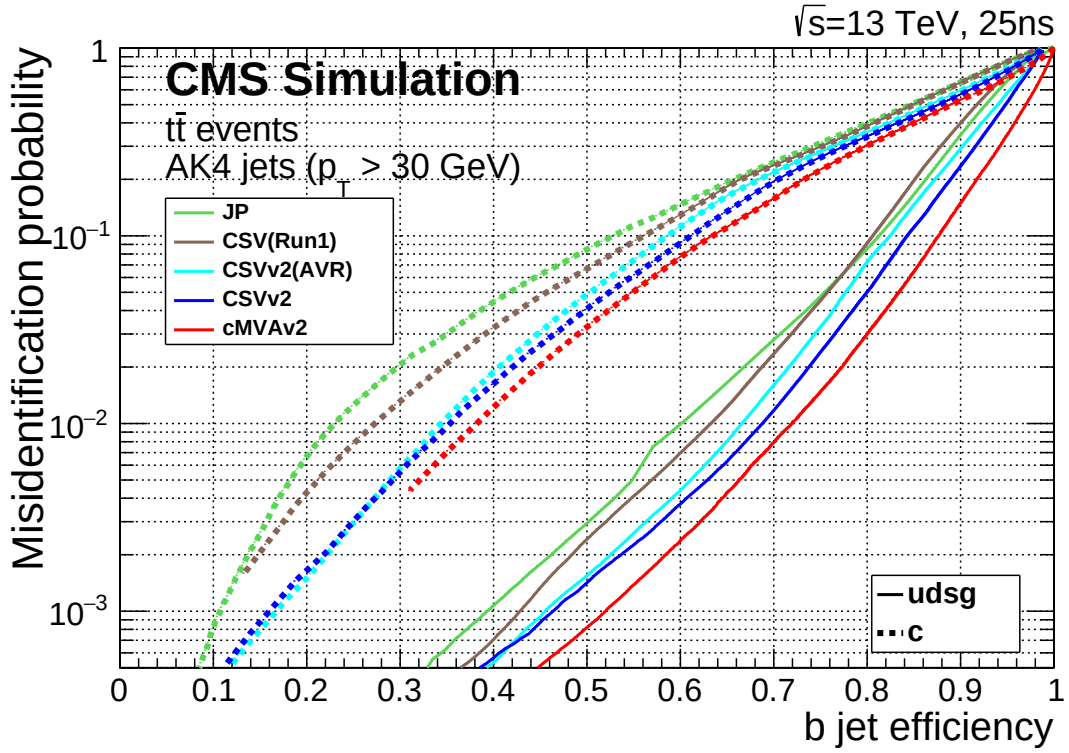


Figure 6.7: Performance of the b jet identification efficiency algorithms the probability for non-b jets to be misidentified as b jet as a function of the efficiency to correctly identify b jets. The curves are obtained on simulated $t\bar{t}$ events. The CMVAv2 algorithm outperforms the JP and CSVv2 algorithms for both c jets as well as light-parton and gluon jets. The improvement of the CSVv2 algorithm with respect to the Run 1 version of the algorithm is also shown. The figure is taken from [62].

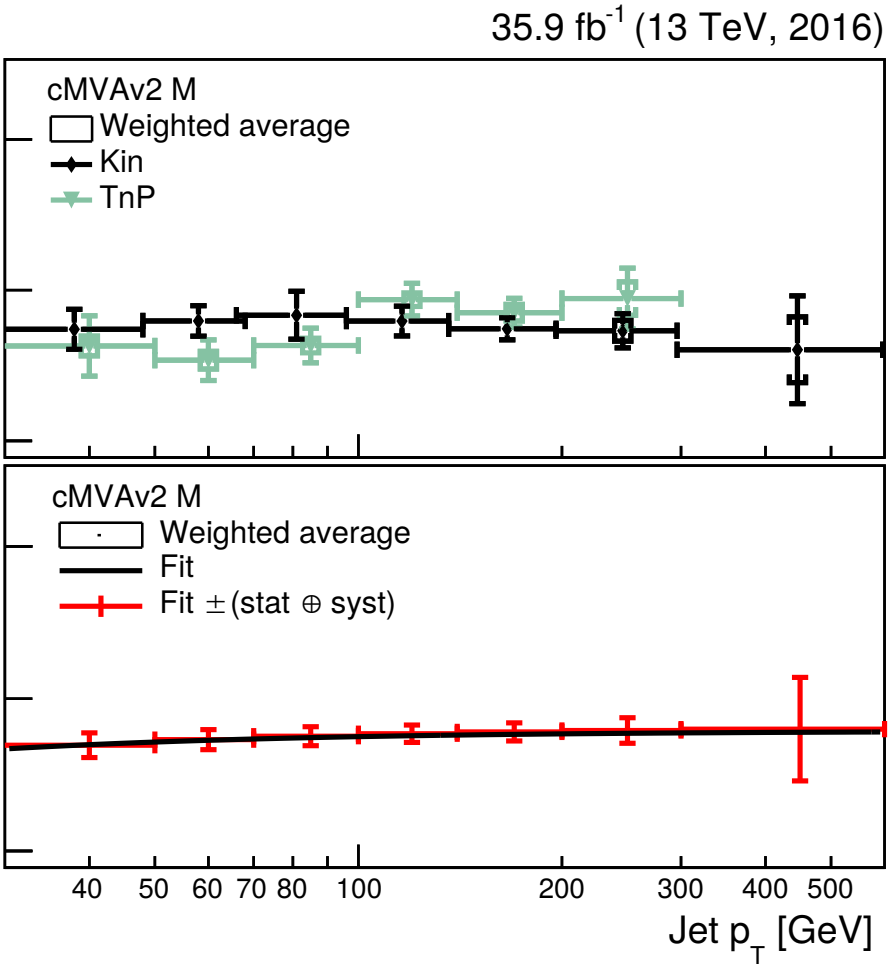


Figure 6.8: Data-to-simulation scale factors for b jets as a function of jet p_T for the medium CMVAv2 algorithms working points. The upper panels shows the scale factors for tagging b as function of jet p_T measured with the various methods. The inner error bars represent the statistical uncertainty, and the outer error bars the combined statistical and systematic uncertainty. The lower panels show the same combined scale factors with the result of a fit function (solid curve) superimposed.

nique [65]. The scale factors are separated in three categories depending on the jet flavor: light (up, down, strange or gluon), charm or bottom, and are extracted independently in bins of jet's p_T and η. Different selections are applied to obtain a phase-space enriched in light jets (from DY + jet process) and another phase space enriched in b quarks (from $t\bar{t}$ process). The b-tagger distributions are then compared between the data and the Monte-Carlo samples. To avoid any effect on the normalization, the Monte-Carlo distributions are normalized to data. When estimating the scale factors for b (light) jets, the contribution from non-b jets (b jets) are removed from the Monte-Carlo distributions and substracted from the data. A first iteration then determines the scale factors in bins of the b-tagger score on both jet categories (b and light jets). To take into account the impact of the scale factors from one category onto the derivation of the other category (coming from the subtraction mentioned above), the scale factors from the first iteration are applied on both categories and the process is repeated. This iterative procedure stops once the scale factors from the current iteration are stable with respect to the previous iteration. The systematic and statistical uncertainties are accounted for the whole procedure and included in the VH(bb) analysis. They are described in a later chapter dedicated to the analysis, see section 8.3.2.

Figure 6.9 shows the CMVAv2 b-tagger distribution for data and Monte-Carlo in a region of phase-space enriched in $t\bar{t}$ events before and after applying the efficiency scale factor corrections. The vector boson is required to have a transverse momentum above $50\,\mathrm{GeV}$. The rest of the $t\bar{t}$ selection is described in a later chapter, see section 8.2.2. Two b jets, one from each top quark decay, are selected this way. The CMVAv2 score in this distribution corresponds to the sub-leading b jet, the one having the lowest CMVAv2 score of the two. The Monte-Carlo modeling of the CMVAv2, without the efficiency corrections, presents some discrepancy with respect to the data, as it is shown in the right figure. This discrepancy is recovered after the efficiency scale factors are applied, as it is shown in the left figure. This becomes apparent when comparing the lower ratio plot in both figures, corresponding to the ratio between the data and Monte-Carlo simulations.

6.4.6 Missing Transverse Energy

The missing transverse energy is used in the Z(ll)H(bb) channel to discriminated the $t\bar{t}$ background. It is reconstructed from the list of particle flow objects as the negative vectorial sum from the list of all particle flow objects identified in the event [66]. The MET is corrected for the effect of the JEC that are applied in the vectorial sum. Another MET-related variable used in the W(lv)H(bb) channel is the *MET significance*, which is the MET value divided by square root of the the scalar sum of the transverse momentum of all the jets in the event with $p_T > 30\,\mathrm{GeV}$.

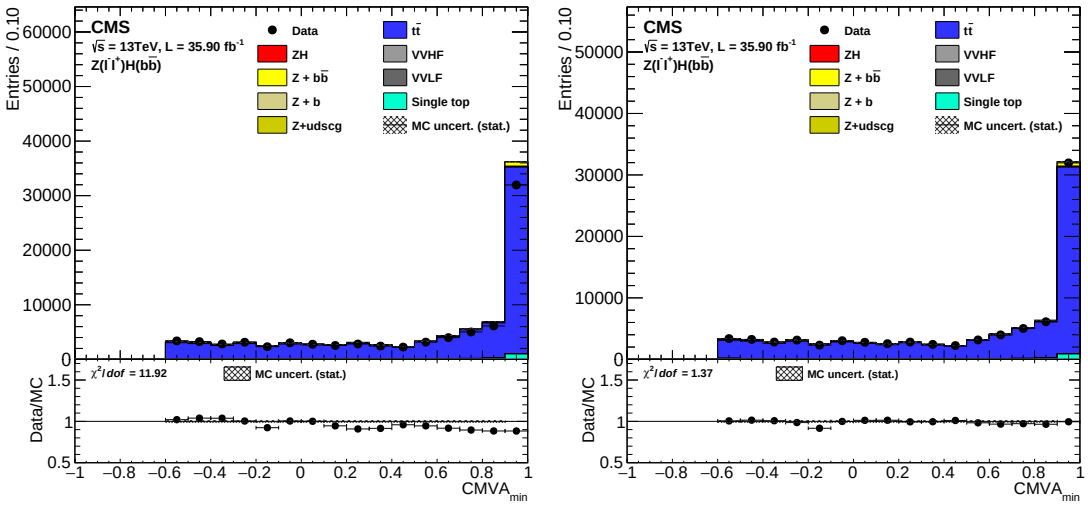


Figure 6.9: CMVAv2 score of the sub-leading b jet candidate in the Z(ll)H(bb) channel, in a phase-space enriched in $t\bar{t}$ events. **Left:** before applying the CMVAv2 efficiency corrections. **Right:** after applying the CMVAv2 efficiency corrections. The agreement between the Monte-Carlo simulation and the data in the CMVAv2 is improved after applying the CMVAv2 efficiency corrections, as it can be seen in the lower ratio plot. The composition of each figure is similar to what is shown in Figure 6.5 and is described in the legend.

6.4.7 Soft Activity

The Higgs boson is a color singlet that decays into two b quarks, each one carrying a color charge, in the VH(bb) signal final state. Not much additional hadronic activity is expected outside of the Z and Higgs boson system for a typical signal event. On the contrary, in background events (especially $t\bar{t}$) there is often the presence of additional, soft hadronic activity. A measurement of this *soft activity* provides additional discrimination for background rejection.

The soft activity reconstruction starts with a collection of tracks. The tracks must pass a high purity quality and $p_T > 300\,\mathrm{MeV}$ requirement [67], not be associated with the vector boson or the selected b jets in the event, have a minimum longitudinal impact parameter with respect to the main primary vertex rather than to other pileup interaction vertices, have the absolute value of the impact parameter below $2\,\mathrm{mm}$ and not be located in a region between the two b jets. The selected tracks are then clustered into "soft-track jets" using the anti-k_T clustering algorithm with a cone radius $\Delta R = 0.4$. The variable exploited for signal-background separation is the multiplicity of the soft-track jets with a transverse momentum larger than $5\,\mathrm{GeV}$.

6.4.8 Vector boson

The Z vector boson is reconstructed from the two same flavour, opposite sign lepton candidates passing the identification and isolation selection defined in 6.4.3. The invariant mass of the Z boson is required to be withing the range of 75 and 105 GeV. This additional requirement removes the contribution from $t\bar{t}$, as well as leptons coming from a $\gamma^{\star}$ in the the DY + jets background.

6.4.8.1 QCD and electroweak corrections

A difference in shape has been observed between data and the Monte-Carlo simulations, the latter showing a harder Z boson transverse momentum spectrum than the data.

This difference in shape is recovered after applying higher order QCD and electroweak corrections on the Monte-Carlo [68]. The QCD corrections are derived in different bins of H_T [68]. They correspond to a k-factor of 1.28, 1.17, 1.21, 0.93 for the H_T bins $100-200$, $200-400$, $400-600$ and $>600\,\mathrm{GeV}$, respectively. The electroweak correction is a function of the Z boson transverse momentum p_T:

$$f(p_T) = -0.18 + 6.04 * (p_T + 759)^{-0.24}, \quad \text{if } p_T > 100\,\mathrm{GeV},$$
$$f(p_T) = 1, \quad \text{if } p_T < 100\,\mathrm{GeV}.$$

6.4.9 Higgs boson

The Higgs boson candidate is reconstructed from the 4-vector of both b jet candidates. The relevant jet systematic uncertainties (like the JECs) are propagated to the Higgs candidate object.

The reconstruction of the Higgs boson starts from the selection of the two b jets candidates. A choice has to be made to address the combinatorics in the b jets selection, as there are often more then two jets in the events. Some options are to choose the two jets with: (i) the highest p_T, (ii) the highest CSVv2 b-tagger value or (iii) the highest value of the CMVAv2 discriminant. In the Run1 VHbb analysis, option (i) was found to have a lower signal efficiency and no improvement in background rejection with respect to jets selected by b-tag score, so option (i) was discarded [69]. The choice between (ii) and (iii) is evaluated by comparing the expected sensitivity, see section 9.1.1. This leads to the choice of the CMVAv2 discriminant.

6.4.9.1 Regression

The dijet invariant mass is one of the most important variable for signal-background separation. Its resolution is significantly improved by a regression technique, which increases the analysis sensitivity.

Regression variable	Description
p_T	transverse momentum of the jet after corrections
M_T	transverse mass of the jet after corrections
η	pseudorapidity of the jet
leading track p_T	transverse momentum of the leading track in the jet
vertex 3-d length	3-d flight length of the jet secondary vertex
vertex 3-d length error	error on the 3-d flight length of the jet secondary vertex
vertex p_T	transverse momentum of the jet secondary vertex
number of vertex tracks	number of tracks associated with the jet secondary vertex
neutral ECAL energy	fraction of jet constituents detected in the ECAL that have neutral charge
neutral HCAL energy	fraction of jet constituents detected in the HCAL that have neutral charge
number of PVs	number of primary vertices
soft lepton relative p_T	relative transverse momentum of soft lepton candidate in the jet
soft lepton p_T	transverse momentum of soft lepton candidate in the jet
soft lepton ΔR	distance in η-ϕ space of soft lepton candidate with respect to the jet axis

Table 6.5: List of input variables used for the training of the b jet energy regression.

The regression is performed using a BDT implemented in the TMVA package [42]. The regression training is summarized in section 6.1.6.2. It is trained on b jets from a simulated $t\bar{t}$ sample and the target variable is the ratio between the generated jet (including neutrinos) and the reconstructed jet transverse momentum. Once the regression is performed, this ratio is applied to the reconstructed p_T of each b jet in the event to recover for energy losses due to tracker, ECAL or HCAL inefficiencies or a neutrino emission within the jet.

The list of input variables for the training is given in Table 6.5. The most discriminating variables are the kinematic variables (momentum, angles). This is due to the fact that the largest corrections are coming from semileptonic decays of the b hadrons, where the missing energy due to the neutrino emission was not recovered by the reconstruction. The dijet mass distribution of the Z(ll)H(bb) signal before and after the regression are compared in Figure 6.10. Overall, the regression improves the dijet mass resolution by $\approx 15\%$. A combination of a Bernstein polynomial and a Crystal Ball function is used to fit the distribution and derive the full-width half-maximum and peak value.

Several studies have been performed to validate the regression technique. The modelling in data of the regression input variables has been verified using $t\bar{t}$ events in the VH(bb) phase-space. The dijet invariant mass distribution before and after the regression are compared in regions of phase-space enriched in background to demonstrate that the background shapes are not modified by the regression (see appendix B). Another validation uses the distribution of the $p_{T,balance}$ defined as

$$p_{T,balance} = \frac{p_T(jj)}{p_T(ll)}, \tag{6.9}$$

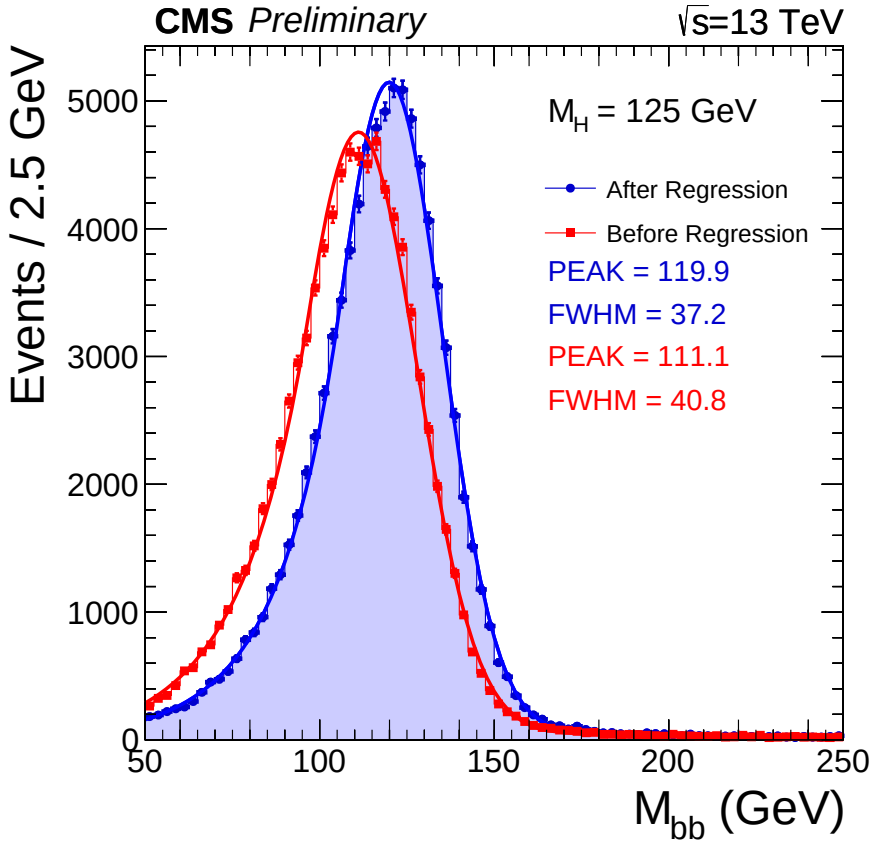


Figure 6.10: Distribution of the dijet invariant mass of from the $H \rightarrow b\bar{b}$ decay in the Z(ll)H(bb) channel before and after the regression is applied. A combination of a Bernstein polynomial and a Crystal Ball is used to fit the distribution and derive the full-width half-maximum and peak value.

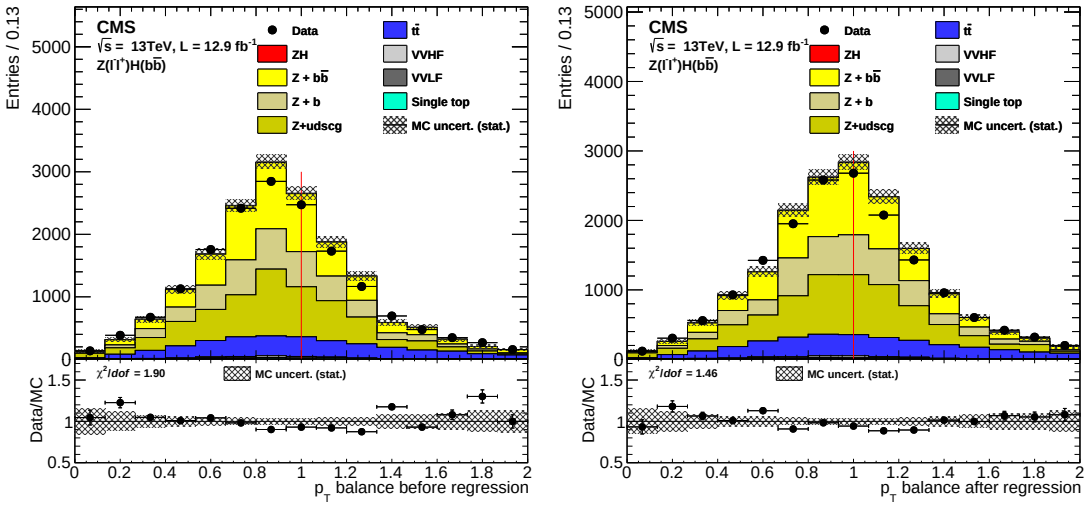


Figure 6.11: Distribution of the $p_{T,balance}$ for data and Monte-Carlo samples, in a subset of Z(ll)Hbb event enriched in Z + b jets background, rejecting the events with more than two b jets in the final state. **Left:** no regression applied to the jets. **Right:** regression applied to the jets. The red vertical line corresponds to a value of one. The regression correction shifts the peak of the distribution closer to unity. The composition of each figure is similar what is shown in Figure 6.5 and described in the legend.

where $p_T(jj)$ is the dijet transverse momentum and $p_T(ll)$ is the Z boson transverse momentum. It is performed on a subset of Z(ll)H(bb) events enriched in Z + b jets background, rejecting the events with more than two b jets in the final state. The mean of the distribution is sensitive to the dijet mass scale and the width to the resolution. The p_T balance is shown before and after the regression on Figure 6.11. An improvement of the resolution and agreement between data and Monte-Carlo simulation is observed after applying the regression. The regression correction shifts the mean closer to unity, from 0.939 ± 0.007 to 0.987 ± 0.009, and reduces the width from 0.327 ± 0.008 to 0.324 ± 0.007. The mean and width are extracted by fitting the Monte-Carlo distribution with a Gaussian function.

6.4.9.2 LO to NLO correction

A discrepancy has been observed in region enriched in DY + jets in dijet mass distribution between the data and the Monte-Carlo samples. Such a discrepancy is also present in the η separation between the two b-jet candidate, $\Delta\eta(jj)$, as the mass of the dijet system is related to the angular separation by $m_{jj} = \sqrt{2E_1 \cdot E_2(1 - \cos\theta_{12})}$, where $E_{1/2}$ is the energy of the first/second jet and θ_{12} the angular separation between both jets. Both distributions can be observed in Figure 6.12.

The description of the data is improved when using NLO DY + jets instead of the

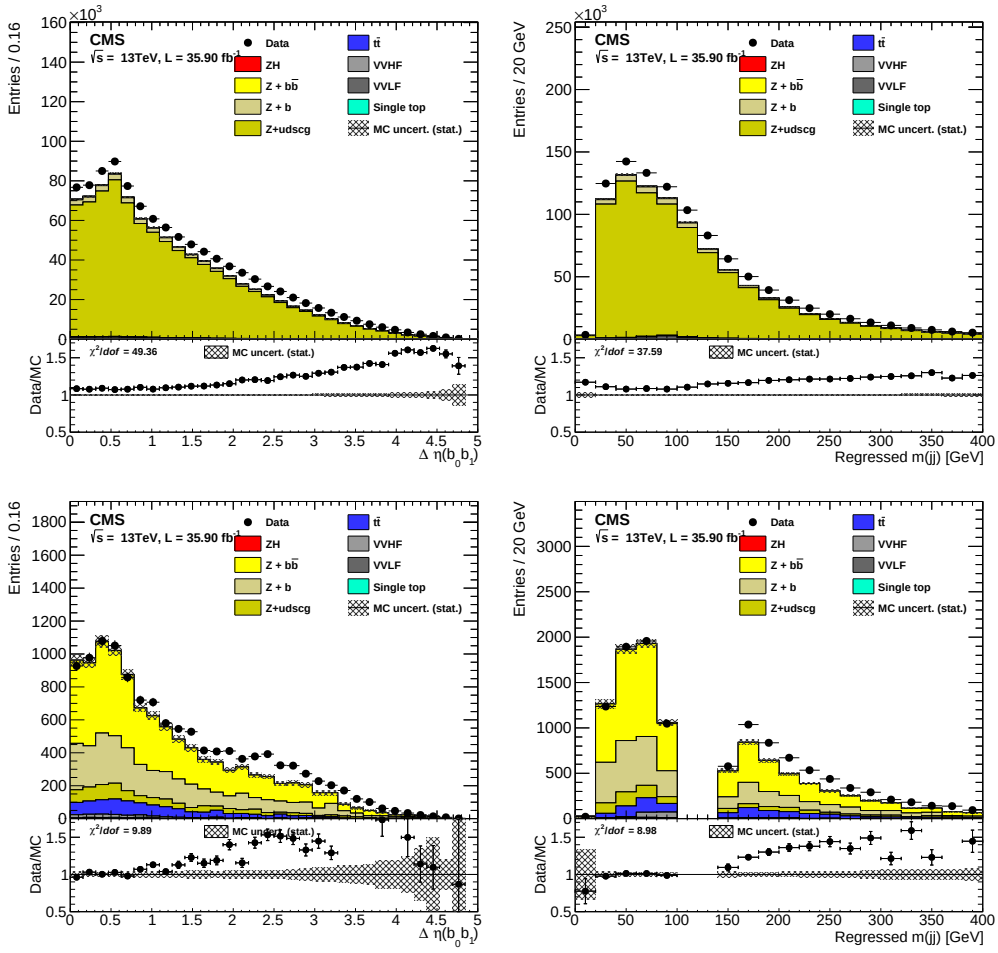


Figure 6.12: Distribution of $\Delta\eta(jj)$ (left column) and the regressed dijet mass (right column) using DY + jets Monte-Carlo samples. The first (second) row is a region of phase-space enriched in light jets (DY + 1 b jet and DY + 2 b jets). A discrepancy between the data and the Monte-Carlo samples can be observed in all the distributions. The composition of each figure is similar what is shown in Figure 6.5 and described in the legend.

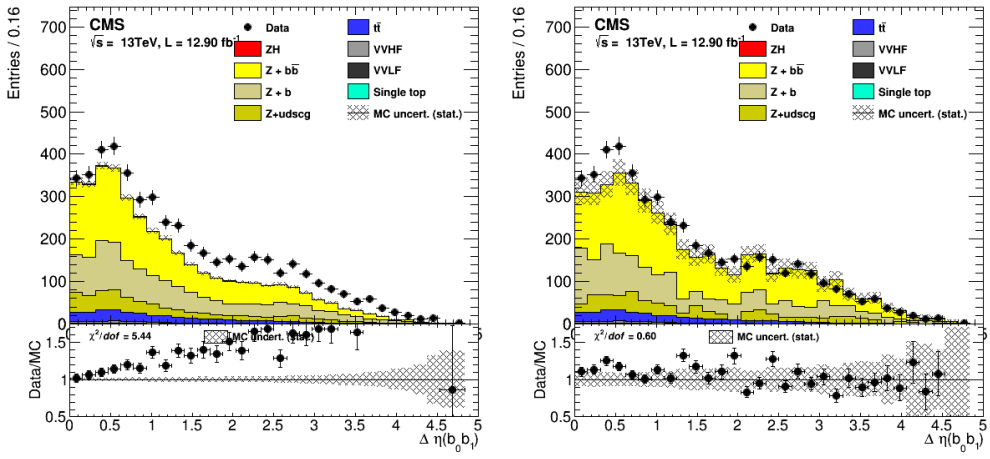


Figure 6.13: $\Delta\eta(jj)$ distributions for the LO (left) and NLO DY + jets (right) process in a region of phase-space enriched in DY + 1 or b jets. The NLO distributions provide a better modeling of the data than the LO. The composition of each figure is similar what is shown in Figure 6.5 and described in the legend.

LO samples both for the dijet mass and the $\Delta\eta(jj)$ distribution. The disadvantage of the NLO samples is the relatively low integrated luminosity with respect to the LO samples. Relying on NLO instead of the LO samples would provide a better modeling of the data but introduces statistical uncertainties that would impact the analysis. Figure 6.13 compares $\Delta\eta(jj)$ distributions for the LO and NLO DY + jets process in a region of phase-space enriched in DY + 1 or b jets. The NLO distributions provides a better modeling of the data than the LO.

To correct for the discrepancy between data and the LO DY + jets Monte-Carlo samples, a set of *LOtoNLO* weights are extracted by comparing the $\Delta\eta(jj)$ distribution between LO and NLO DY+jets samples. Both the LO and NLO samples are normalized. The ratio of the shapes (NLO DY + jets/LO DY + jets) is extracted in three jet flavour categories: DY without additional b jets (0b category), DY + one b jet (1b category), DY + at least two b jets (2b category). The ratios are then fitted using a fourth order polynomial for the 0b category and a third degree polynomial convoluted with a exponential function for the 1b and 2b category. The fitted distributions can be found in Figure 6.14.

The higher order QCD corrections (see 6.4.8.1) are not applied on the LO DY + jets during the evaluation of the *LOtoNLO* weights, as such correction are implicitly included in the higher order calculation of the the NLO DY + jets sample. Those QCD corrections have both an impact on the H_T shape and normalization. To take the normalization effect into account, a k-factor of 1.15 is then applied after the LOtoNLO weights.

The dijet mass and $\Delta\eta(jj)$ distribution using LO Monte-Carlo samples corrected

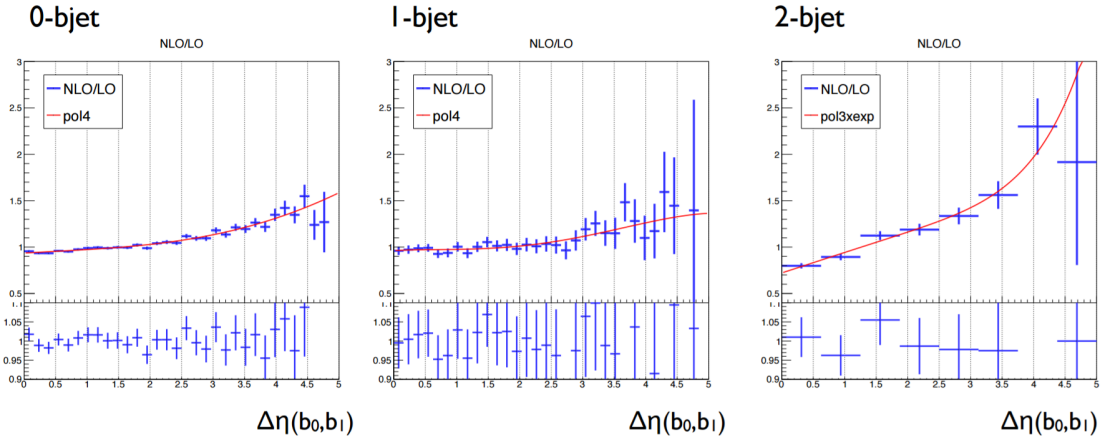


Figure 6.14: Fitted ratio of the LO and NLO DY + jets shapes, distributed in $\Delta\eta(jj)$ for the three jet flavour categories. **Left:** 0b category. **Middle:** 1b category. **Right:** 2b category.

with the LOtoNLO weights can be found in Figure 6.15. The first row corresponds to a region of phase-space enriched in DY + light (udscg) jets. The discrepancy between data and the Monte-Carlo sample is corrected for both the $\Delta\eta(jj)$ and m_{jj} distributions. In the second row, corresponding to a region of phase-space enriched in DY + 1 or 2 b jets, the data to Monte-Carlo discrepancies is partially recovered by the LOtoNLO weights, some discrepancy still remaining at large values of $\Delta\eta(jj)$ and m_{jj}.

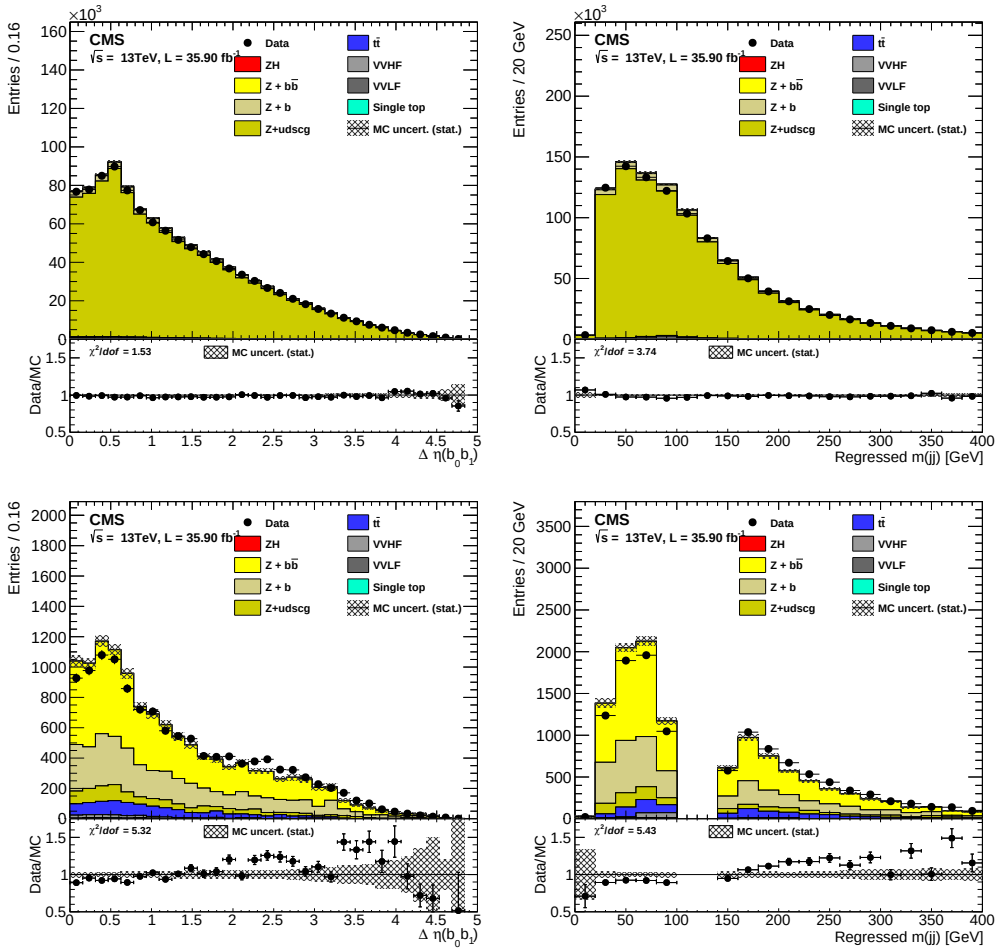


Figure 6.15: Same distributions as in Figure 6.12, but with the LOtoNLO weight applied to the Monte-Carlo samples. The first row correspond to a region of phase-space enriched in DY + light (udscg) jets. The discrepancy between data and the Monte-Carlo is corrected for the both the $\Delta\eta(jj)$ and m_{jj} distribution. In the second row, corresponding to a region of phase-space enriched in DY + 1 or 2 b jets, the data to Monte-Carlo discrepancy is partially recovered by the LOtoNLO weights, some discrepancy still remaining at large value of $\Delta\eta(jj)$ and m_{jj}. The composition of each figure is similar what is shown in Figure 6.5 and described in the legend.

7 Muon efficiency studies

As mentioned in section 6.4.3, muons in the VH(bb) analysis are required to pass identification and isolation cuts to reduce the background. The efficiency of those selections, as well as the trigger requirement, have a different impact on data and Monte-Carlo simulations. To address this discrepancy, the efficiency distribution of each of the aforementioned selection is evaluated on Monte-Carlo simulations and data to extract scale factors used to correct Monte-Carlo samples to equivalent efficiency in data. The scale factors have been studied on the 2016 data. Section 7.1 describes the method employed to the efficiency extraction on the 2016 data. The same approach is used for efficiency measurements on the 2017 and 2018 CMS data [70]. Section 7.2 is dedicated to the trigger efficiency studies. The results are summarized in section 7.3.

7.1 The Tag and Probe method

The efficiency of a particular muon selection corresponds to the ratio

$$\epsilon = \frac{N_p}{N_p + N_f}, \tag{7.1}$$

where N_p (N_f) are the number of signal-like muons passing (failing) the selection. The muons must pass a prior selection with respect to which the efficiency is calculated, referred to as the *denominator selection*. Assuming there is no correlation between the various levels of selections, efficiencies of multiple muon selections can then be factorized as

$$\epsilon_{total} = \epsilon_{track} \cdot \epsilon_{id|track} \cdot \epsilon_{iso|id\,+\,track} \cdot \epsilon_{trigg|id\,+\,track\,+\,iso}, \tag{7.2}$$

where the subscript corresponding to the muon track reconstruction efficiency is labeled as *track*, the identification efficiency *id*, the isolation efficiency *iso* and the trig-

ger efficiency *trigg*. This notation includes the numerator (*num. sel.*) and denominator selections (*den. sel.*) as $\epsilon_{\textit{den. sel.}|\textit{num. sel.}}$. For example, the isolation efficiency is calculated with respect to muons that pass a track and identification selection.

The signal-like muons for the efficiency calculation are selected with a *tag and probe* technique. The principle of this method is to select muons from a virtual photon or Z boson decay in DY + jets event both in data and Monte-Carlo simulations, relying on a fit of the invariant mass of the di-muon system to remove other processes. Events that have exactly one pair of opposite sign muons are selected through a background-subtraction technique. One of the two muons is referred to as the *tag* and the other the *probe*. Each event is used twice in the efficiency evaluation by switching role of the tag and the probe muon. The tag muon passes a single-muon HLT and a tight selection to reduce the fake muon contribution. The probe muon passes a looser selection corresponding to the denominator selection. Two invariant mass distributions are produced for the cases when the tag muon (1) passes the muon selection or (2) fails the muon selection. Both distributions are fitted by the sum of a signal model for the resonance at the Z peak and a background model for the background shape. The shapes used for the background and signal model are optimized for a given selection, denominator selection (as in Equation 7.2) and p_T-η bin of the probe muon to ensure a good quality of the fit. The background is modeled by an exponential shape or a *CMSshape* and the signal by a sum of two *Voigtians*. They are described in the sections 7.3.1, 7.3.2, 7.3.3 for the identification, isolation and trigger selection, respectively. The number of signal muons in each category is taken from the yield of the signal shape estimated from the fit and is used as input in the ratio in equation 7.2 to compute the efficiency. An example of such fits can be found in Figure 7.1. It corresponds to the tight identification selection with respect to muon tracks for a p_T bin of $40 - 50\,\mathrm{GeV}$. The signal is fitted using a sum of two Voigtians (a convolution of a Gaussian and Breit-Wigner distribution) and the background with a CMSshape (see section 7.3.1). No figure of merit such as the χ^2 score are used to evaluate the goodness of fit, which was done "by eye". Fits such as in Figure 7.1 are considered to have a good quality of fit.

Only statistical uncertainties have been considered in efficiency measurements on the 2016 dataset. Studies on the systematic uncertainties related to the fit procedure on the 2017 dataset, performed with the same methodology as the 2016 dataset, give an estimation of the fit bias. In this study, the scale factors have been varied by modifying the fit configuration. The various configuration include: a tightening and loosening of the tag muon isolation selection, using a single Voigtian instead of a double Voigtian for the signal shape, increasing and decreasing the bin size of the invariant mass distribution, reducing and increasing the mass window. The uncertainty is taken as the difference between the root mean square of all the variations and the nominal value of the SF. The average values of the systematic uncertainty is between 0.1% for the loose identification, $0.2 - 0.4\%$ for the tight identification and below 0.1% for the isolation.

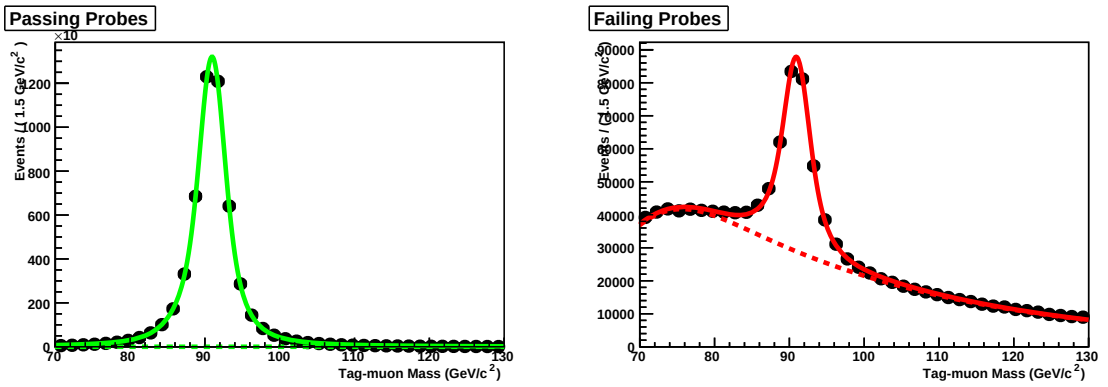


Figure 7.1: Example of invariant mass distribution fitted during the tag and probe efficiency estimation. The signal is fitted using a sum of two Voigtians (a convolution of a Gaussian and Breit-Wigner distribution) and the background with a CMSshape (see section 7.3.1). **Left:** probe muons passing the tight identification selection. **Right:** probe muons failing the tight identification selection.

For the efficiency measurement on Monte-Carlo samples, the invariant mass is built using a single sample of DY + jets (H_T inclusive, see Table 6.4). As the muon selection performance depends on the amount of pile-up, the number of primary vertices distribution in the Monte-Carlo sample is reweighted to match the one in data. The efficiency measurement is performed in different bins of p_T and η. A multiplicative correction factor, the muon efficiency scale factor, is calculated by dividing the data efficiency by the efficiency estimated on the Monte-Carlo sample in each p_T and η bin. This scale factor is applied as a multiplicative weight on all the Monte-Carlo samples to correct for the difference in muon selection efficiency between data and simulations. The values of those corrections are discussed in section 7.3.

7.2 Double muon trigger efficiency

The double-muon trigger on the 2016 dataset is described in section 6.2.1. It requires a different approach than the isolation and identification selections for the efficiency measurement.

The identification and isolation selections are applied to each muon individually and the corresponding efficiencies are measured on the probe muon, relying on the tag muon for the denominator selection. For the double-muon trigger, whose selection involves a muon pair, the efficiencies have to be studied simultaneously on the tag and the probe muon to take correlations between both muons into account. This procedure is relevant only for the Run H of the 2016 dataset, as it includes a d_Z filter for the double-muon HLT. If the tag muon passes the d_Z filter, meaning that the distance between the tag and the probe muon tracks is lower than $2\,\mathrm{cm}$, then the probe muon passes it as well, which would bias the efficiency study. For the other 2016 Runs

(BCDEFG), the efficiencies can be studied on the probe muon only, as it doesn't include a d_Z filter.

In the case of the Run BCDEFG (without the d_Z filter), the efficiency of the double muon path is the product of the efficiency of the two muons that composed the path. Those can be calculated using the tag and probe method as described in the previous sections (7.1). The efficiency of the d_Z filter introduces a correlation between the two muons. In that case, the efficiency is evaluated per-event: the tag and probe pair is required to pass the double-muon selection with the d_Z filter applied, and the denominator selection corresponds to the double muon selection without the d_Z filter [1]. The final per-event data efficiency $\epsilon_{Data,HLT}$ is then evaluated as

$$\epsilon_{Data,HLT} = \epsilon_{d_Z} \cdot \frac{\epsilon^2_{Data,8}(1) \cdot \epsilon_{Data,17}(2) + \epsilon^2_{Data,8}(2) \cdot \epsilon_{Data,17}(1)}{\epsilon_{Data,8}(1) + \epsilon_{Data,8}(2)},$$

where ϵ_{d_Z} is the efficiency of the d_Z filter (set to 1 for runs BCDEFG) and $\epsilon^2_{Data,8/17}(1/2)$ is the efficiency the $8/17\,\mathrm{GeV}$ path selection evaluated on the first/second muon in the event. The Monte-Carlo sample efficiency is estimated the same way, and the muon efficiency scale factors correction is taken as the ratio $\epsilon_{Data,HLT}/\epsilon_{MC,HLT}$.

7.3 Results

This section contains the result of the muon efficiency studies for the identification, isolation and trigger selection used in the VH(bb) analysis.

A decrease of muon hit reconstruction efficiency has been observed in late 2015 and part of the 2016 data-taking. This was due to the saturation effect in the pre-amplifier of the APV chip, the front-end amplifier of the CMS strip tracker [24]. Due to this APV behavior, the impact of the inefficiency increases with the luminosity and therefore with the pileup. It affects the 2016 runs B, C, D, E and F ($20.1\,\mathrm{fb}^{-1}$) and was fixed in the run G and H ($16.3\,\mathrm{fb}^{-1}$) by changing the APV settings [71]. The effect on the tight identification efficiency as a function of the number of vertices in the event can be seen in Figure 7.2 for the runs BCDEF and GH. The data efficiency in run BCDEF shows a dependence with the pile-up, which is mostly recovered in run GH.

Due to this effect, the efficiencies of the identification and isolation have been studies separately in Run BCDEF and GH. The overall data to Monte-Carlo simulation corrections for the full year 2016 have been taken as the luminosity-average between those two periods. For the electron case, the efficiency measurement has been performed simultaneously on the full 2016 dataset.

[1] In this scenario, each event is used only once to not introduce a bias. There is no switching between the role of the tag and the probe muon.

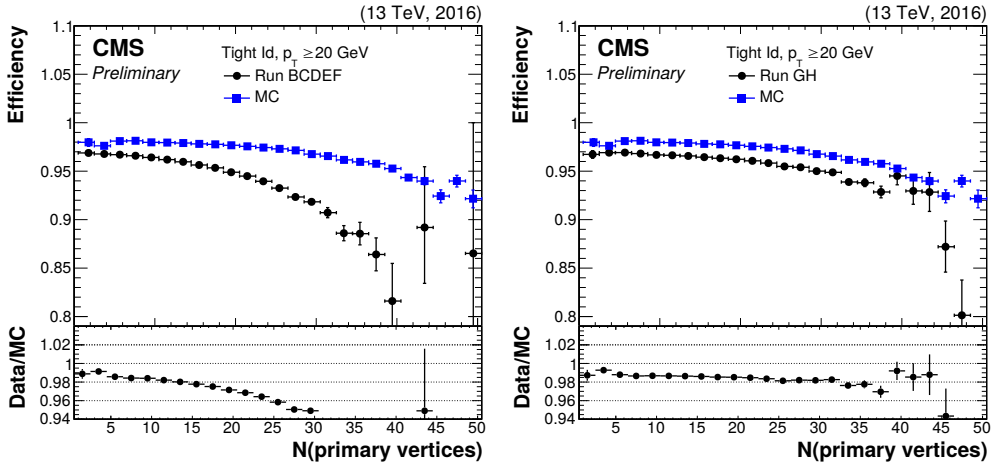


Figure 7.2: Efficiency distributions of the data and DY + jets Monte-Carlo sample in bins of the number of vertex for the tight identification. The uncertainties are statistical uncertainties from the tag and probe fit. **right:** efficiency distribution for run BCDEF. **left:** efficiency distribution for run GH.

7.3.1 Identification

In the tag and probe identification efficiency measurement, the di-lepton invariant mass is fitted within a range of $70 - 130\,\mathrm{GeV}$ in bins of $1\,\mathrm{GeV}$. The background distribution is fitted with a CMSshape [2] and the signal by a sum of two Voigtians (a convolution of a Gaussian and Breit-Wigner distribution). The width of the Voigtians is fixed to the Z boson width of $2.495\,\mathrm{GeV}$, while the peak of the Voigtians and variance of the Gaussians are left free floating in the fit. The usage of two Voigtians allows to properly fit the high p_T bins, above $60\mathrm{GeV}$. The tag muon is required to pass the tight identification selection, have an isolation value below 0.2 and have a p_T above $26\,\mathrm{GeV}$, and to pass the single-muon trigger selection. The probe corresponds to general muon tracks (tracker, standalone or global muon).

The data and MC p_T and η efficiency distributions for the loose identification can be found in Figure 7.3. The distributions are separated for the run BCDEF and GH. The ratio plot corresponds to the distribution of the muon scale factors. The uncertainties are statistical-only from the fit. For the loose identification, the efficiencies of the data and the Monte-Carlo sample are about 99%. The corresponding scale factors are ~ 1. No difference in performance can be observed before and after the APV fix [71]. Additional distributions in bins p_T-η can be found in appendix A.1, A.2.

The data and MC p_T and η efficiency distributions for the tight identification can be found in Figure 7.4. The average value of the muon scale factors is around $0.98 - 0.96$

[2] The CMSshape (f) is the product of a falling exponential and error function (erf). Beside normalization, $f(x, \mu, \alpha, \beta, \gamma) = (1 + \mathrm{erf}[(x - \alpha)\beta]) \cdot \exp[-\gamma(x - \mu)$, where $\mathrm{erf} = \int_0^y \exp(-\tau^2)\mathrm{d}\tau$.

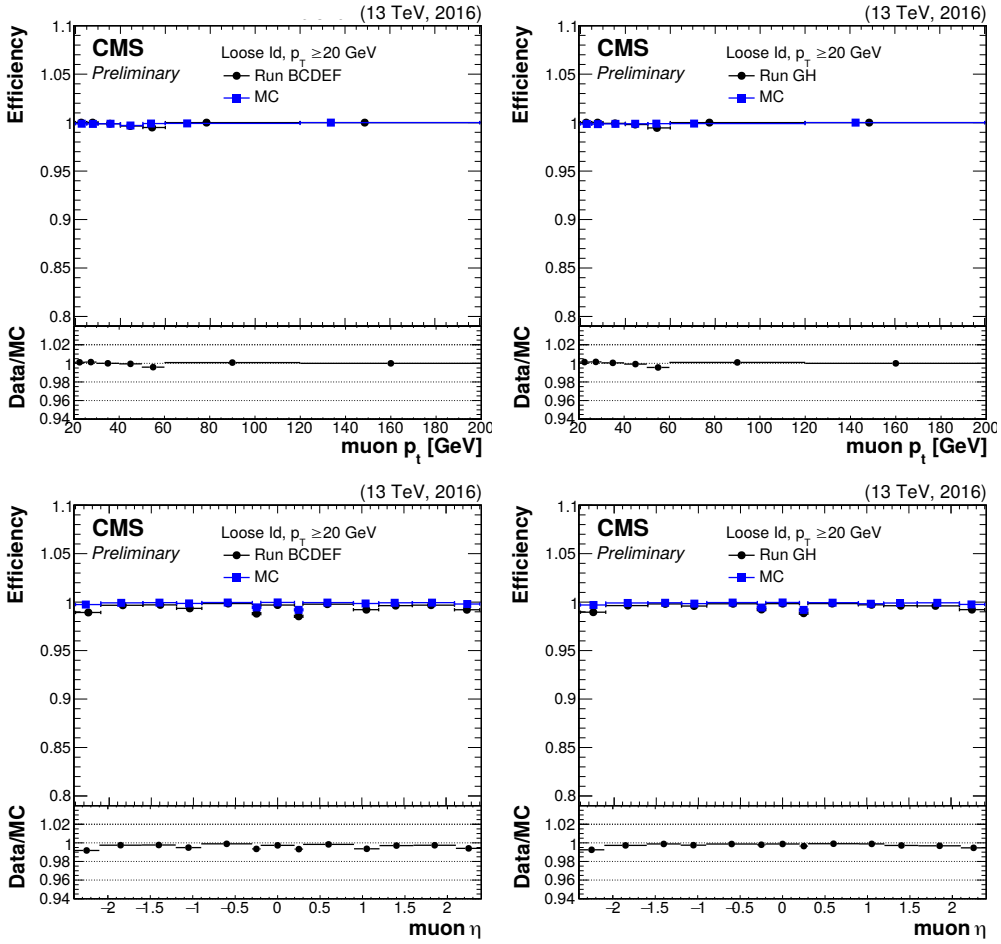


Figure 7.3: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the loose identification. The uncertainties are statistical uncertainties from the tag and probe fit. **upper row:** efficiency distribution in bins of p_T ($|\eta| < 2.4$). **lower row:** efficiency distribution in bins of η ($p_T > 20\,\text{GeV}$). **right column:** efficiency distributions for the runs B, C, D, E and F. **left column:** efficiency distributions for the run G and H.

for run BCDEF and $0.99 - 0.96$ for run GH. The efficiency drop in the two middle η bins, $[-0.3,-0.2]$ and $[0.2,0.3]$, corresponds to the gap between the two wheels of the barrel muon stations. The APV fix recovers between $1 - 2\%$ of the data efficiencies, depending on the η region. Additional distributions in bins p_T-η can be found in appendix A.3, A.4.

7.3.2 Isolation

In the tag and probe isolation efficiency measurements, the di-lepton invariant mass is fitted withing a range of $77 - 130\,\mathrm{GeV}$ in bins of $1\,\mathrm{GeV}$. The background distribution is fitted with an exponential shape and the signal by a sum of two Voigtians. The tag selection is the same as for the identification efficiency studies from section 7.3.1. There are two probe selections, depending on the denominator of the isolation efficiency.

For the *loose isolation* used in the Z(ll)H(bb) selection (below 0.25), the denominator is required to pass the loose identification. For the *tight isolation* used in the W(lv)H(bb) selection (below 0.06), the denominator is required to pass the tight identification.

The data and MC p_T and η efficiency distributions for the loose isolation can be found in Figure 7.5. A very good agreement is observed between data and MC; the muon scale factors are close to 1 for the full p_T, η range. Additional distributions in bins of p_T-η can be found in A.5, A.6.

The data and MC p_T and η efficiency distributions for the tight isolation can be found in Figure 7.6. The overall muon scale factors value is within $0.95 - 0.99$ for the BCDEF run and and within $0.97 - 0.99$ for the GH run. The largest recovery of the APV fix is observed in the $|\eta| > 2.1$ region, where the data efficiency is increased by $\sim 2\%$. Additional distributions in bins p_T-η can be found in A.7, A.8.

A turn-on of the p_T efficiency distributions is visible for both the loose and tight isolation, not present in the identification efficiencies. This behavior comes from the denominator in the isolation definition (muon p_T). For a same isolation cut and activity around the muon candidate, a lower p_T value increase the chances of failing the isolation selection.

7.3.3 Trigger

The studies of the double-muon trigger efficiency have been performed separately for runs BCDEFG (without d_Z filter) and run H (with d_Z filter). The muons in the denominator selections must pass the loose identification and the loose isolation selection.

For the run BCDEFG, the efficiency distributions for both the $17\,\mathrm{GeV}$ and $8\,\mathrm{GeV}$ muon can be found in the appendix A.9 and A.10, respectively. The efficiency scale factors are flat in p_T at a value within $0.98 - 0.99$ for the $8\,\mathrm{GeV}$ muon. For the $17\,\mathrm{GeV}$

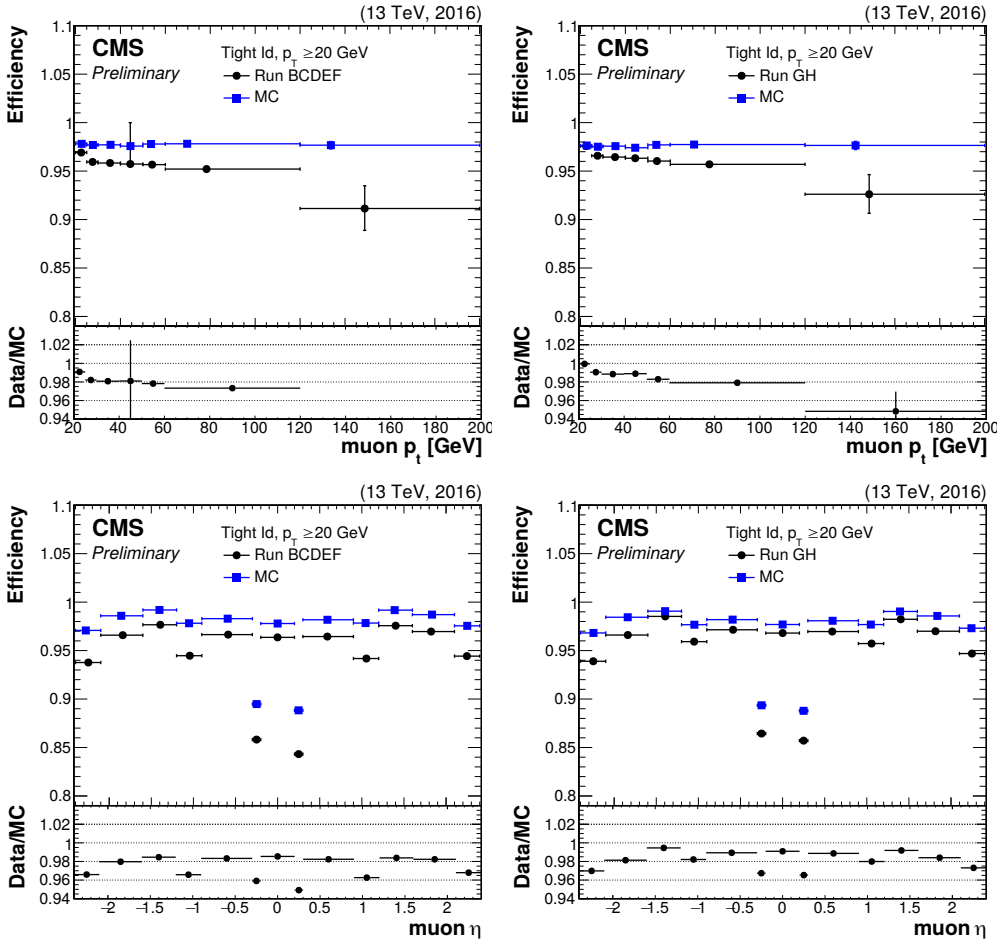


Figure 7.4: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the tight identification. The uncertainties are statistical uncertainties from the tag and probe fit. **upper row:** efficiency distribution in bins of p_T ($|\eta| < 2.4$). **lower row:** efficiency distribution in bins of η ($p_T > 20\,\text{GeV}$). **right column:** efficiency distributions for the runs B, C, D, E and F. **left column:** efficiency distributions for the run G and H.

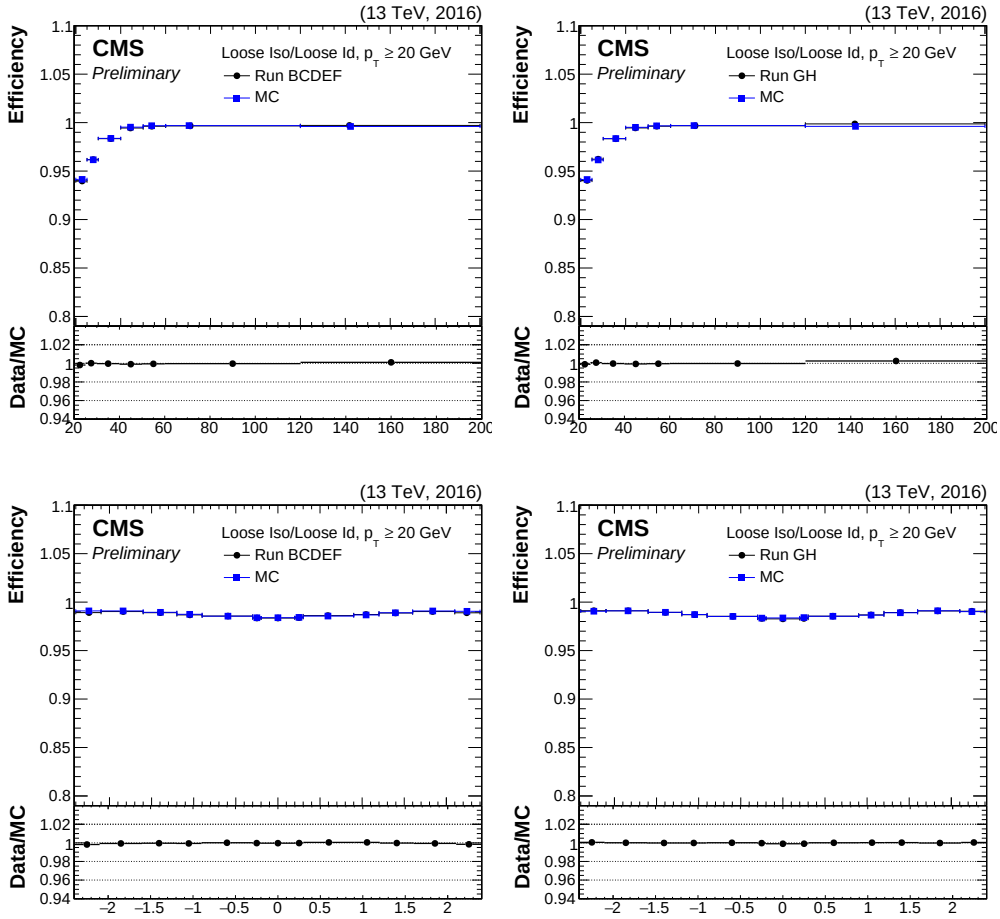


Figure 7.5: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the loose isolation identification. The uncertainties are statistical uncertainties from the tag and probe fit. **upper row:** efficiency distribution in bins of p_T ($|\eta| < 2.4$). **lower row:** efficiency distribution in bins of η ($p_T > 20$ GeV). **right column:** efficiency distributions for the runs B, C, D, E and F. **left column:** efficiency distributions for the run G and H.

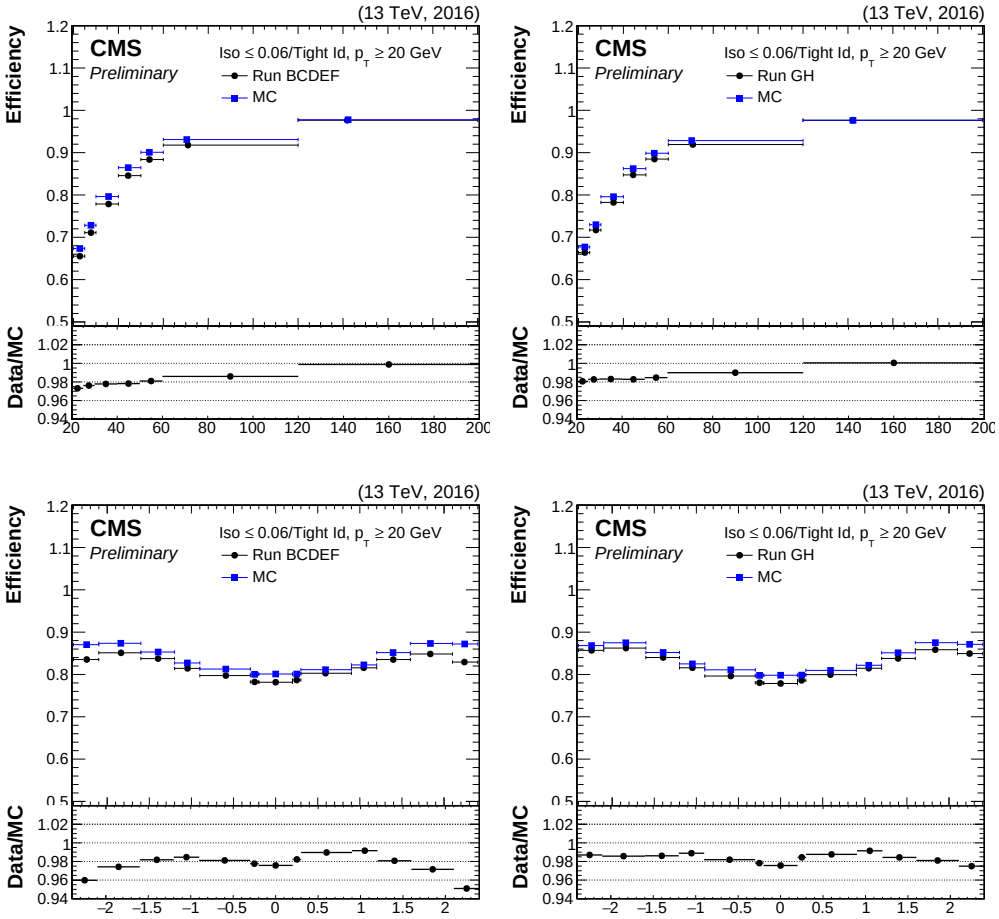


Figure 7.6: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the tight isolation identification. The uncertainties are statistical uncertainties from the tag and probe fit. **upper row:** efficiency distribution in bins of p_T ($|\eta| < 2.4$). **lower row:** efficiency distribution in bins of η ($p_T > 20\,\text{GeV}$). **right column:** efficiency distributions for the runs B, C, D, E and F. **left column:** efficiency distributions for the run G and H.

leg, the efficiency scale factors are flat in p_T around 0.96, except in the $2.1 < |\eta| < 2.4$ region, where a turn-on from 0.9 to 0.96 is visible in the p_T distribution.

For the run H, the efficiency distributions for both the $17\,\mathrm{GeV}$ and $8\,\mathrm{GeV}$ muon can be found in appendix A.11 and A.12, respectively. The efficiency scale factors are flat in p_T at a value within $0.98 - 0.99$ for the $8\,\mathrm{GeV}$ muon. For the $17\,\mathrm{GeV}$ leg, the muon scale factors are flat in p_T around 0.98, expect in the $2.1 < |\eta| < 2.4$ region, where a turn-on from 0.95 to 0.96 is visible in the p_T distribution.

The efficiency of the d_Z are measured on run H. The events in the denominator selection are required to pass the double-muon trigger without the d_Z filter. The average efficiency for data and MC is around 1 and 0.99, respectively, and it is flat as a function of p_T and η of each muon leg.

8
Analysis Strategy

The strategy and details of the VH(bb) analysis performed on the 2016 dataset are described in this chapter, focusing on the Z(ll)H(bb) channel. Section 8.1 gives a brief overview of the analysis strategy. Section 8.2 is dedicated to signal and background selection. It also includes a description of the signal classification and background modeling. Section 8.3 is dedicated to an overview of the systematic uncertainties included in the analysis. The results of the VH(bb) measurement are discussed in Chapter 9.

8.1 Analysis strategy in a nutshell

The VH(bb) analysis is based on a loose *event preselection* specific to each channel, defined to have a high signal efficiency. This selection follows the topology of typical VH(bb) events. In the Z(ll)H(bb) case, it corresponds to one Z boson system, reconstructed from two same-flavor opposite-sign leptons, back-to-back to a dijet system attributed to the Higgs boson $H \to b\bar{b}$ decay.

An additional set of selections is required on top of this preselection to define a *signal region*, enriched in VH(bb) processes. Rather than maximizing the number of signal per background event, it is designed to keep as much signal contributions as possible i.e. to maximize the signal efficiency. The signal-background separation is performed by a BDT trained on Monte-Carlo samples in the signal region. The distribution of the BDT output score is included in the final binned-likelihood fit on data to extract the signal and the corresponding significance. This fit is performed simultaneously on signal and background processes, whose shape predicted by Monte-Carlo simulations are allowed to vary and are determined during the fit. The shape variations are parameterized by nuisance parameters + *background normalization scale factors*, discussed below, for each background process and nuisance parameters + signal strength for the signal process. The signal region and the BDT are described in section 8.2.1.

In parallel, the analysis defines a set of *background control regions* (or *control regions*) to study how the simulated samples model the most relevant physics variables in data. Those control regions are designed to maximize the purity of the main backgrounds: Z + udscg (up, down, strange, charm or gluon) jets, Z + 1 b jet (Z + b), Z + 2 b jets (Z+ 2b) and $t\bar{t}$ multijet ($t\bar{t}$). In the control regions, the lowest CMVAv2 score among the two b jet candidates, labeled as CMVAv2$_{\text{min}}$, provides a discrimination among the background processes. The CMVAv2$_{\text{min}}$ distribution is fit simultaneously with the BDT output score distribution from the signal region in the final binned-likelihood fit, the former bringing signal-background separation and the latter discrimination among the various background processes. The yield of the main backgrounds are allowed to float during the fit through background normalization scale factors which are mostly constrained from the control region CMVAv2$_{\text{min}}$ shapes. Including the control regions in the final fit gives additional information to constrain the systematic uncertainties and the background normalization scale factors.

The analysis strategy is summarized in Figure 8.1. The upper part contains an illustration of a typical VH(bb) event serving as a basis for the analysis preselection. It shows a lepton pair from a Z, W^- or W^+ boson decay, back-to-back in transverse momentum space with respect to two AK04 b-flavor jets from a Higgs boson decay. Below the signal event is a sketch of the dijet invariant mass distribution, $M(jj)$. The signal region is defined around the $125\,\text{GeV}$ mass peak of the VH(bb) process and the control regions in the mass sidebands[1]. The definition of the signal and control region selections are chosen to satisfy the following points: (i) The region selections must be orthogonal (an event cannot be present in more than one region). (ii) The control region definition must be as close as possible to the signal topology. This is to ensure that the systematic uncertainties, in particular the background normalization scale factors, can be extrapolated between the control and signal regions during the binned-likelihood fit. As the closeness in phase-space between the signal and control regions is achieved, no dedicated extrapolation uncertainties are included in the fit . (iii) The background region selections maximize the purity of the main backgrounds and minimize the signal efficiency. (iv) The signal region maximizes the signal efficiency. Examples of CMVAv2$_{\text{min}}$ and BDT output score distributions from control and signal regions, respectively, in the Z(ee)H(bb) sub-channel are shown below the $M(jj)$ plot. Those distributions, as well as other CMVAv2$_{\text{min}}$ and BDT output score distributions from the three channels, are fitted simultaneously in the final signal+background binned-likelihood fit to extract the signal strength, the observed and expected significance, fit the uncertainties and estimated the normalization scale factors. For the three Z(ll)H(bb) channels, this correspond to a total of 7 BDT output score distributions

[1]Some details are omitted from this Figure for simplification purposes. The control regions enriched in Z + udscg and $t\bar{t}$ background processes include the kinematic region around the $125\,\text{GeV}$ mass peak but don't overlap with the signal region due to other selections, detailed in the next two sections.

and 24 CMVAv2$_{\text{min}}$ in the final binned-likelihood fit, performed on the three VH(bb) channels.

The diboson background is similar to the VH(bb) signal, the main difference being that the dijet invariant mass peaks at the Z boson mass, $91.2\,\text{GeV}$ (see section 5.3.1.1). This is exploited to perform a validation of the VH(bb) analysis procedure by extracting the diboson signal in the final binned-likelihood fit and treating the VH(bb) process as background. This analysis is referred to as the *diboson analysis*. The diboson analysis is very similar to the VH(bb) analysis, only a few modifications with respect to the Z(ll)H(bb) analysis are required in the definition of the diboson *Z(ll)Z(bb) channel*. The [90,150] GeV dijet mass window, used as a selection in the signal region and as a veto in the Z + HF control region, is moved to [60-160] GeV to take into account the diboson invariant mass peak (at $91.2\,\text{GeV}$) and resolution. The BDTs are re-trained in the new signal regions, using the diboson process as a signal and replacing the diboson by the VH(bb) process in the background list.

8.2 Event selection

The first loose preselection that follows the signal topology is applied to each analysis channel. It includes leptons and b jets selections applied during the dijet and vector boson reconstruction described in sections 6.4.8 and 6.4.9, respectively. As the vector boson transverse momentum spectrum $p_T(V)$ is harder for signal than for backgrounds, a large value of $p_T(V)$ is required. Events are separated in two regions depending on the $p_T(V)$ value: a low-$p_T(V)$ region defined by $50 < p_T(V) < 150\,\text{GeV}$ and a high-$p_T(V)$ region defined by $p_T(V) > 150\,\text{GeV}$. The high-$p_T(V)$ has more sensitivity than the low-$p_T(V)$ region. The latter is still kept in the analysis due to larger statistics that provide additional information to fit the systematic uncertainties correlated between the two $p_T(V)$ categories. Including the lower $p_T(V)$ category improves the expected sensitivity in the Z(ll)H(bb) channel by 10%.

8.2.1 Signal region

The signal regions of the Z(ll)H(bb) analysis are separated in four categories depending on the decay of the Z boson (Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels) and the Z boson transverse momentum (low- and high-$p_T(V)$ categories). The corresponding selections are listed in Table 8.1, including the preselection mentioned in the previous section.

The dijet invariant mass $M(jj)$ is defined in a window of $[90, 150]\,\text{GeV}$. The selection on the Z mass $M(ll)$, required to be within a $[75, 105]\,\text{GeV}$ mass window, removes background contributions from the $t\bar{t}$ process. Each of the b jet candidate of the event must have a CMVAv2 score higher than the loose CMVAv2 working point, CMVA_{L}. This significantly reduces the Z + udscg background.

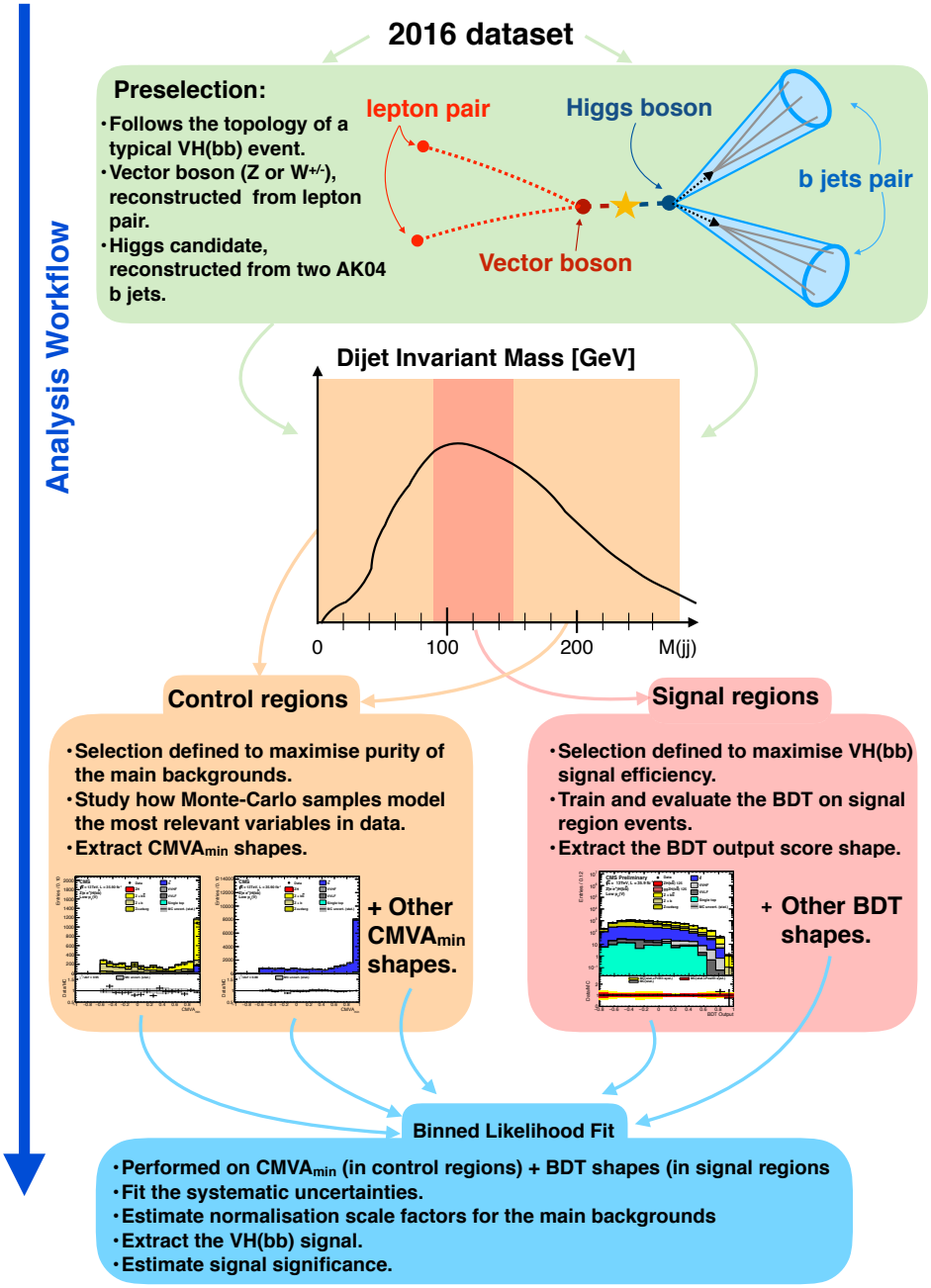


Figure 8.1: Overview of the analysis strategy. From top to bottom: the analysis starts with a event preselection, based on the VH(bb) event depicted in the upper part. Below is an illustration of the dijet invariant mass distribution, $M(jj)$. Signal regions are defined around the Higgs 125 GeV mass peak. The distribution on the left side is the BDT output score in the low-$p_T(V)$ Z(ee)H(bb) sub-channel. Control regions are defined on the dijet invariant mass sidebands. The two distributions on the left side are the CMVAv2$_{\text{min}}$ in two control regions. The BDT and CMVAv2$_{\text{min}}$ shapes from the signal and control regions, respectively, are combined in the final binned likelihood fit to extract the VH(bb) signal strength and the corresponding significance. The nuisances parameters are constrained during the fit.

Variable	Description	Selection
$M(jj)$	Invariant mass of the dijet system	$[90, 150]$
$\Delta\phi(V, jj)$	Angle in the transverse plane between the dijet and vector boson system	> 2.5
$p_T(jj)$	Dijet system transverse momentum	-
$M(ll)$	Invariant mass of the two lepton candidates	$[75, 105]$
$\Delta\phi(Z, jj)$	Angle in tranverse plane between Z boson and dijet system	> 2.5
$\text{CMVAv2}_{\text{max}}$	CMVAv2 score of the leading (highest CMVAv2 score) b jet	$> \text{CMVA}_{\text{L}}$
$\text{CMVAv2}_{\text{min}}$	CMVAv2 score of the sub-leading (second highest CMVAv2 score) b jet	$> \text{CMVA}_{\text{L}}$
BDT score	Score of the BDT trained and evaluate in the signal region	> -0.8
Pre-selection:		
$p_T(Z)$	Z boson transverse momentum	$[50, 150], > 150$
$p_T(l)$	Transverse momentum of the lepton candidates	> 20
$p_T(j_1)$	Transverse momentum of the leading b jet	> 20
$p_T(j_2)$	Transverse momentum of the sub-leading b jet	> 20
Lepton isolation	Isolation of the lepton candidates	$(< 0.25, < 0.15)$

Table 8.1: List of selection for the Z(ll)H(bb) signal region. The first column lists the selection variables, the second column the variable description and the third column the selection cuts applied on the variables. The same signal region selection is used in low ($[50, 150]$ GeV) and high (> 150 GeV) $p_T(V)$ category, and for the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels. The preselection cuts, common to all control and signal regions in the Z(ll)H(bb) channel, are grouped at the bottom of the table. Entries marked with "-" indicate that the variable is not used in that region. The lepton isolation selection is denoted as: (muon isolation selection, electron isolation selection). The selection cuts are given in units of GeV, except for the angles given in radians and the isolation, which is dimensionless.

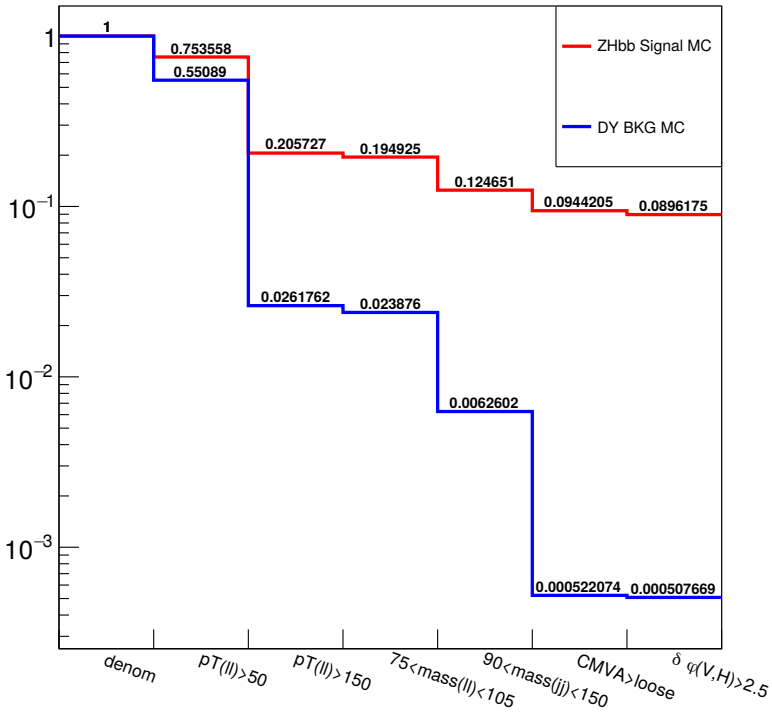


Figure 8.2: List the efficiency reduction of the of the ZH(bb) signal and DY + jets background after each signal region selection applied successively. The denominator corresponds to the Z(ll)H(bb) preselection.

Figure 8.2 lists the efficiency reduction of the ZH(bb) signal and DY + jets background after each signal region selection is applied successively. The denominator corresponds the the Z(ll)H(bb) preselection. The most impactful cut on the signal process is the requirement of the Z boson transverse momentum to be above $150\,\mathrm{GeV}$, reducing the signal contribution by a factor ~ 4 and the $t\bar{t}$ background by a factor ~ 28, which highlights the higher sensitivity of the high $p_T(V)$ category. The $t\bar{t}$ contribution is further reduced by a factor ~ 12 by the dijet invariant mass window of $[90, 150]\,\mathrm{GeV}$.

8.2.1.1 Boosted Decision Tree

A BDT is trained separately on the low- and high-$p_T(V)$ signal regions, merging the Z(ee)H(bb) and Z($\mu\mu$)H(bb) categories. The variables used for the training are listed in Table 8.2. The signal category for the training uses both the qqZH and ggZH processes and the background category includes all samples from Table 6.4.

Variable	Description
$M(jj)$	Invariant mass of the dijet system
$p_T(jj)$	Dijet system transverse momentum
$M(ll)$	Invariant mass of the two lepton candidates
$p_T(Z)$	Z boson transverse momentum
$\Delta\phi(Z,jj)$	Z boson transverse momentum
$p_{T,balance}$	Ratio between the dijet and vector boson transverse momentum
$p_T(j_1)$, $p_T(j_2)$	Momentum of the leading (highest CMVAv2 score) b jet
MET	Transverse momentum of missing energy
$\Delta R(jj)$	Distance in η-ϕ between the two b jet candidates
$\Delta\eta(jj)$	Difference in η between the two b jet candidates
$\text{CMVAv2}_{\text{max}}$	CMVAv2 score of the leading (highest CMVAv2 score) b jet
$\text{CMVAv2}_{\text{min}}$	CMVAv2 score of the sub-leading (lowest CMVAv2 score) b jet
N_{aj}	Number of additional jets
SA5	Number of soft hadronic activity jet with a transverse momentum above $5\,\text{GeV}$

Table 8.2: List of variables used for the BDT training.

The training is performed on half of the Monte-Carlo generated events, referred to as the *training sample*. The second half, the *test sample*, is used to test for eventual cases of overtraining. The degree of overtraining is evaluated by performing a two sample Kolmogorov-Smirnov test on the training and test samples, separately for the background and signal BDT output score distributions. The threshold on the p-value of the test has been set to 0.05, BDTs with a Kolmogorov-Smirnov test score below that value being considered as overtrained. The test values for the signal (background) distribution is 1 (0.49) in the low-$Z(p_T)$ and 1 (0.08) in the high-$Z(p_T)$ category.

The discriminating power of the BDT input variables is evaluated during the training process. The ranking is derived by counting how often a variable is used to split decision tree nodes, and by weighting each split occurrence by the separation gain squared it has achieved and by the number of events in the node. This measurement of the variable importance can be used for a single decision tree as well as for a forest [42]. The most discriminating variables are: the CMVAv2 score of each of the two b-jet candidate ($\text{CMVAv2}_{\text{min}}$ and $\text{CMVAv2}_{\text{max}}$), the transverse momentum $p_T(V)$ of the vector boson, the difference in η between the two b jet candidates and the dijet invariant mass.

8.2.2 Background control regions

Three control regions are defined, corresponding to the three main backgrounds of the Z(ll)H(bb) analysis. The *Z + light* region is enriched in Z + udcsg processes, the *Z + heavy flavor* (Z + HF) in Z + 0b and Z + 1b processes and the $t\bar{t}$ region in $t\bar{t}$ multijet process. Like for the signal region, the control regions are separated in four categories

depending on the Z boson decay and transverse momentum, for a total amount of 12 control regions. The control regions selections are listed in Table 8.3.

- **The Z + light control region** requires a large transverse momentum of the dijet system ($> 100\,\mathrm{GeV}$) and vetos jets with a large CMVAv2 score to exclude contributions from signal, Z + 1 or 2 b jets and $t\bar{t}$. This region is very pure, with more than 99% (95%) Z + udscg events in the low (high) $p_T(V)$ category.

- **The Z + HF control region** vetoes events whose b jet candidates have a low CMVAv2 score and requires a MET value below $60\,\mathrm{GeV}$ to remove contributions from the $t\bar{t}$ process. The dijet invariant mass is vetoed within a range of $\notin [90, 150]\,\mathrm{GeV}$ to be orthogonal to the signal region. The Z + HF region is not pure in Z + 1 b and Z + 2 b, both processes contributing to 82% (71%) of the events in the low (high) $p_T(V)$ category. Other contributions are mainly coming from Z + udscg and $t\bar{t}$ processes.

- **The $t\bar{t}$ control region** vetoes events whose b jet candidates have a low CMVAv2 score, as for the Z + HF. The dilepton invariant mass is vetoed within a range of $[75, 120]\,\mathrm{GeV}$, to exclude contribution from Z + jets, and $[0, 10]\,\mathrm{GeV}$, as Monte-Carlo DY + jets samples are not available in this low mass range. The $t\bar{t}$ region in the low-$p_T(V)$ category has 94% contribution from $t\bar{t}$ processes. In the high-$p_T(V)$ category, the $t\bar{t}$ contribution drops to $\sim 82\%$ due to contamination of single-top and Z + 1 or 2 b backgrounds.

The comparison between data and Monte-Carlo simulations for the most important variables are shown in Figures 8.3-8.14 for all control regions. The background normalization scale factors, whose values have been extracted from the final binned-likelihood fit, are included to re-normalize the yields of the main background contributions. The Monte-Carlo samples include the normalization scale factor corrections for the Z + udscg, Z + 1 b, Z + 2 b and $t\bar{t}$ background processes, extracted in the final binned-likelihood fit. The scale factor values can be found in the next chapter, see section 9.5.

8.3 Systematic uncertainties

This section describes the systematic uncertainties included in the final binned-likelihood fit.

Variable	**Z + light**	**Z + HF**	**$t\bar{t}$**
$M(jj)$	-	$\notin$ [90,150]	-
$\Delta\phi(V, jj)$	-	> 2.5	-
$p_T(jj)$	> 100	-	> 100
$M(ll)$	[75,105]	[85,97]	$\notin$ [0, 10], $\notin$ [75,120]
$\Delta\phi(Z, jj)$	-	> 2.5	-
CMVAv2$_{\text{max}}$	$<$ CMVA$_{\text{T}}$	$>$ CMVA$_{\text{T}}$	$>$ CMVA$_{\text{T}}$
CMVAv2$_{\text{min}}$	$<$ CMVA$_{\text{L}}$	$>$ CMVA$_{\text{L}}$	$>$ CMVA$_{\text{L}}$
MET	-	< 60	
Pre-selection:			
$p_T(Z)$	$[50, 150], > 150$	$[50, 150], > 150$	$[50, 150], > 150$
$p_T(l)$	> 20	> 20	> 20
$p_T(j_1)$	> 20	> 20	> 20
$p_T(j_2)$	> 20	> 20	> 20
Lepton isolation	$(< 0.25, < 0.15)$	$(< 0.25, < 0.15)$	$(< 0.25, < 0.15)$

Table 8.3: The selection of the three control regions used in the Z(ll)H(bb) channel. The first column lists the selection variables, whose definition are given in Table 8.2. The second, third and fourth column list the selection cuts applied on the variables in each control region. The same control region selection is used in low ($[50, 150]$ GeV) and high-$p_T(V)$ ($>$ 150 GeV) category, and for the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels. The preselection cuts are grouped at the bottom of the table. Entries marked with "-" indicate that the variable is not used in that region. The lepton isolation selection is denoted as: (muon isolation selection, electron isolation selection). The selection cuts are given in units of GeV, except for the angles given in radians and the isolation which is dimensionless.

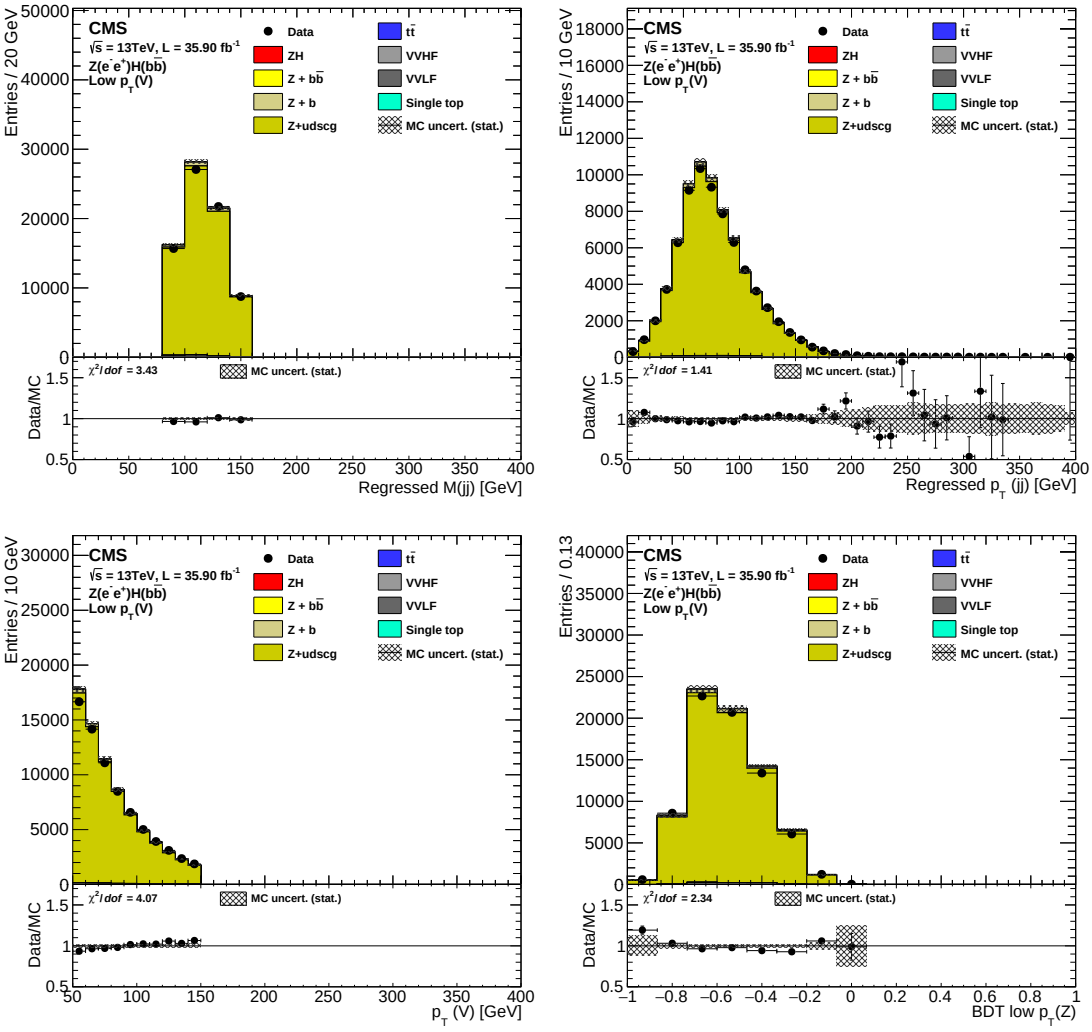


Figure 8.3: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ Z + light control region for the Z(ee)H(bb) sub-channel. The distributions are, from top to bottom and left to right: the regressed dijet invariant mass $M(jj)$, regressed dijet transverse momentum $p_T(jj)$, Z boson transverse momentum $p_T(V)$ and the BDT output score.
The upper part of each figure shows the data (black dot histogram) and Monte-Carlo (colored histograms) distributions. The Monte-Carlo distribution consists of multiple stacked histograms, each histogram corresponding to a particular process listed in the legend. The uncertainty bands for data and Monte-Carlo correspond to statistical Poisson uncertainties. The lower part of each figure is the ratio between the data and the Monte-Carlo distribution. The black dots correspond to the values of the data/Monte-Carlo ratio. The hatching bands around 1 are the Poisson uncertainties of the Monte-Carlo samples propagated to this ratio. A χ^2 test for comparison between the Monte-Carlo and the data distribution is performed, and the corresponding $\chi^2/ndof$, where $ndof$ is the number of degree of freedom, is quoted in the ratio plot [53].

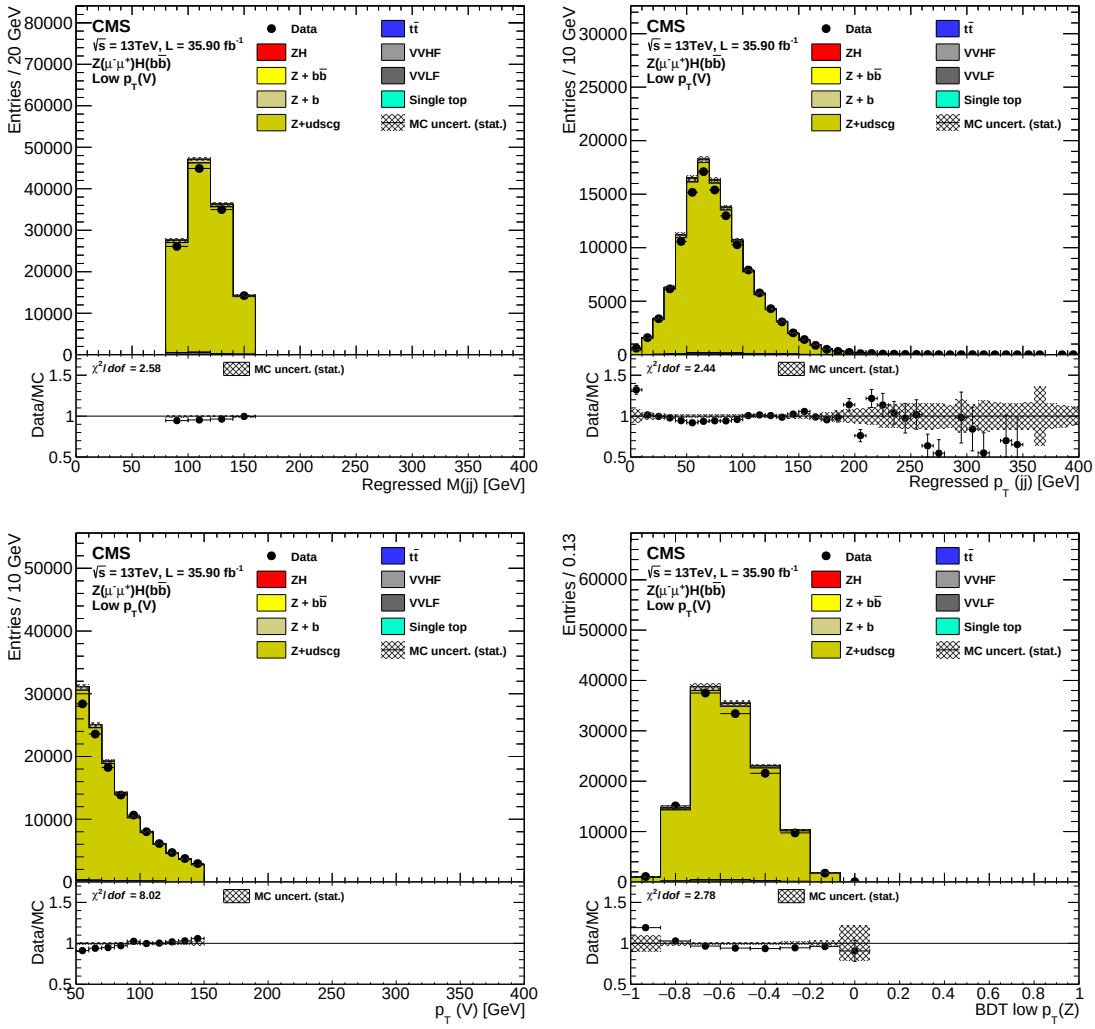


Figure 8.4: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ Z + light control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

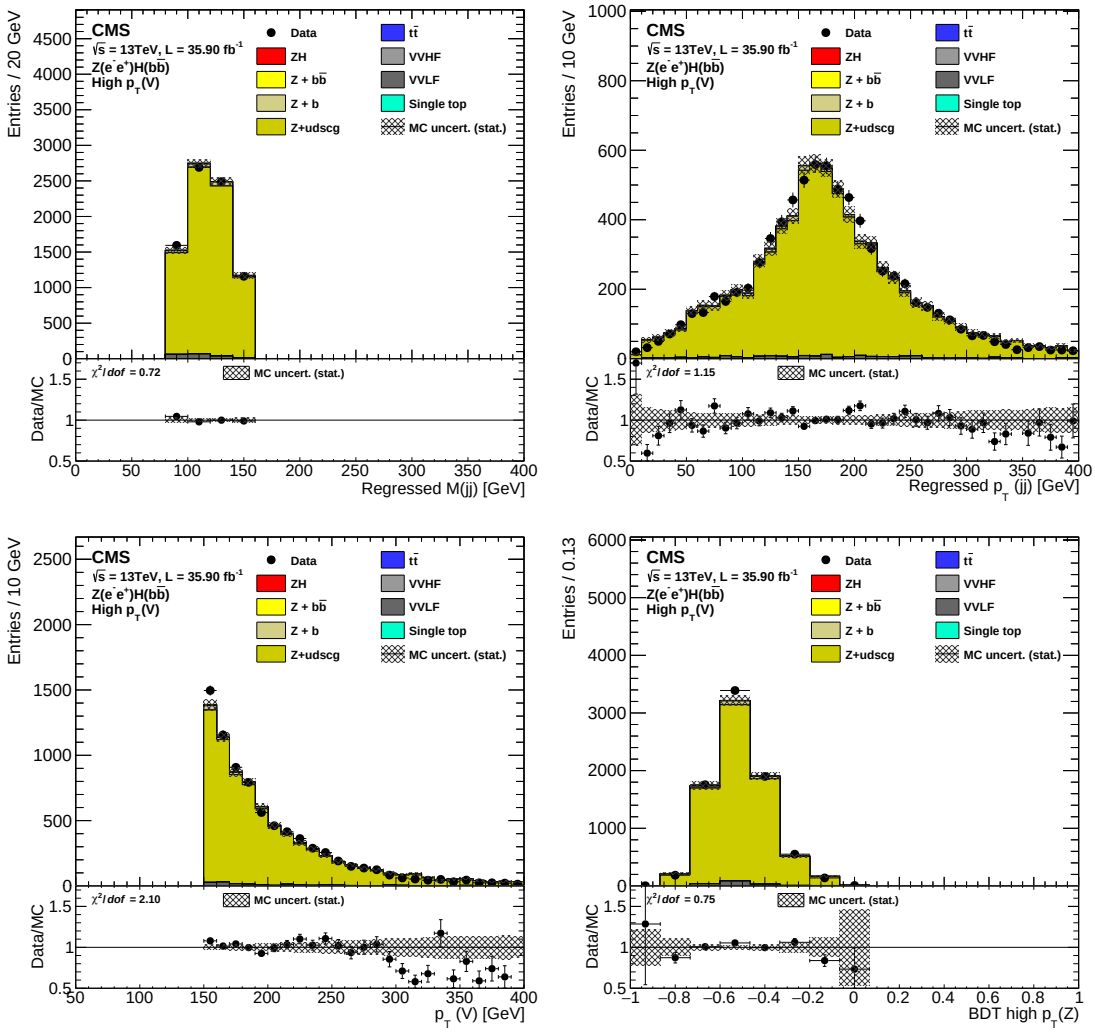


Figure 8.5: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ Z + light control region for the Z(ee)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

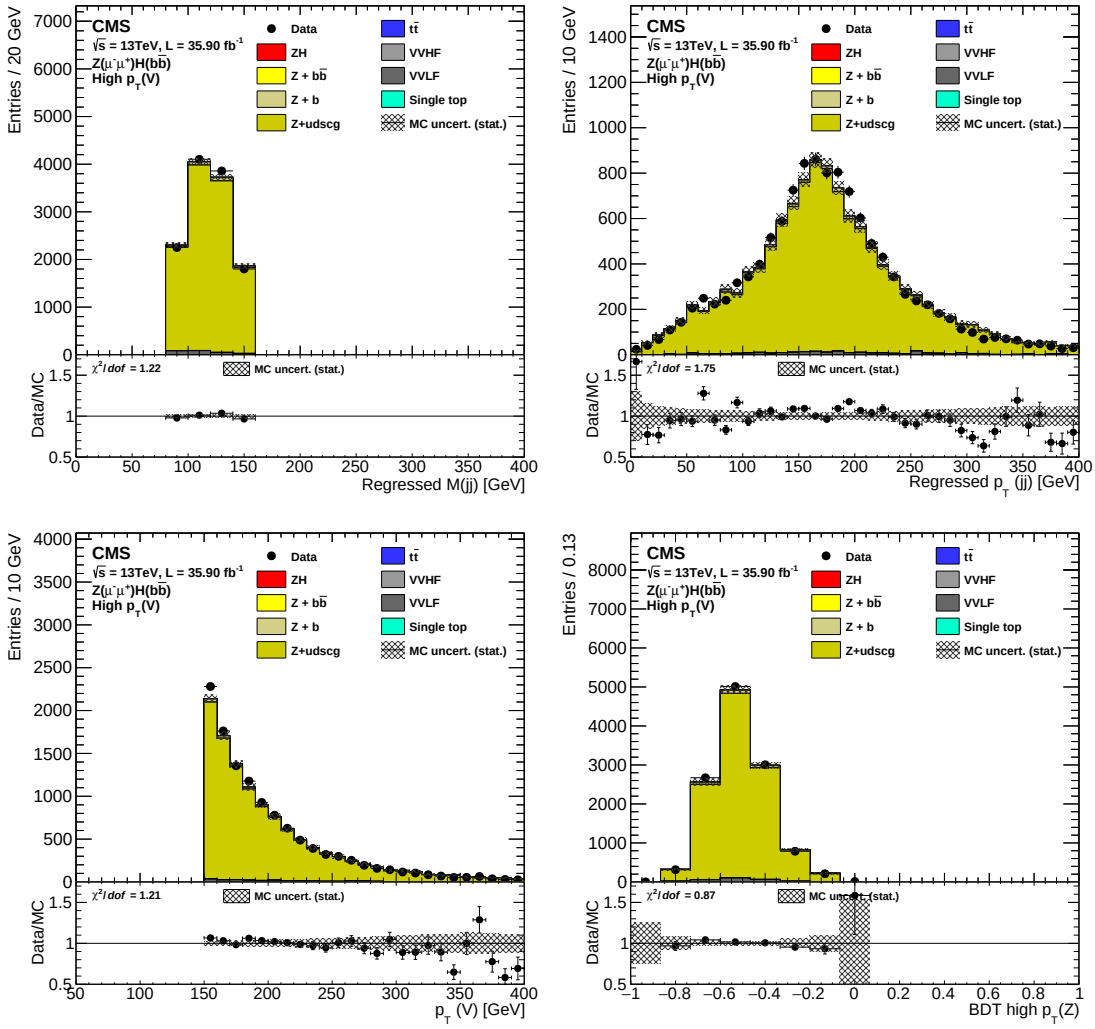


Figure 8.6: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ Z + light control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

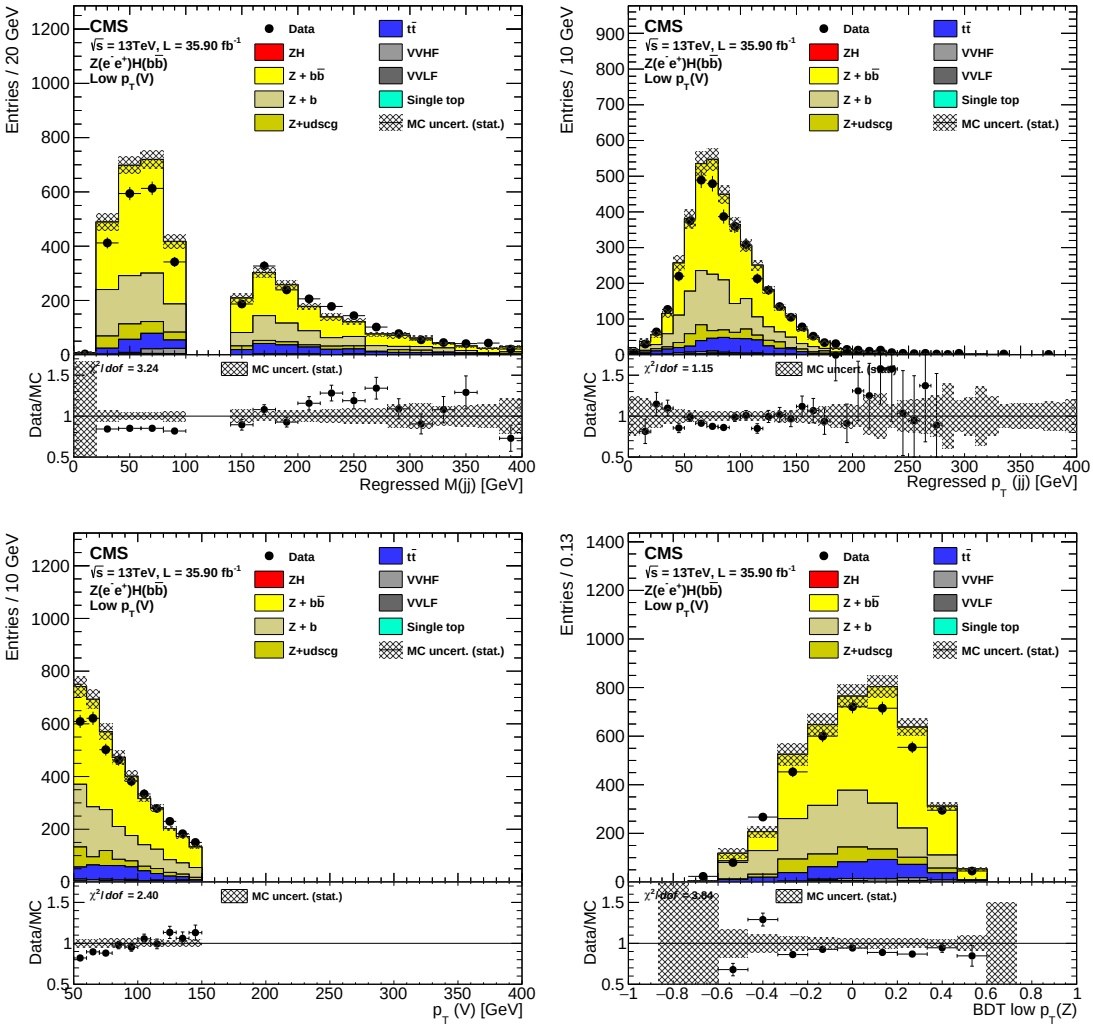


Figure 8.7: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ Z + HF control region for the Z(ee)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

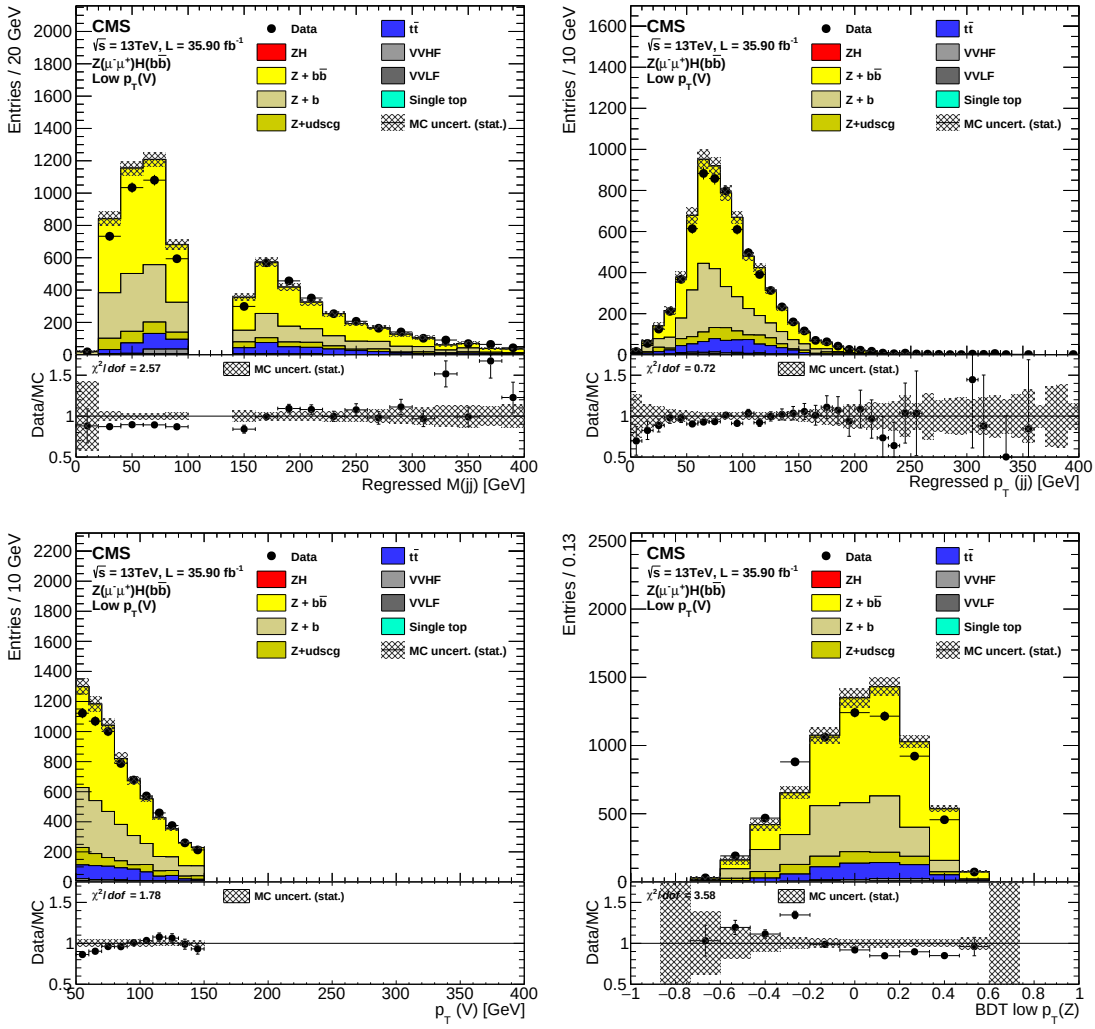


Figure 8.8: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ Z + HF control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

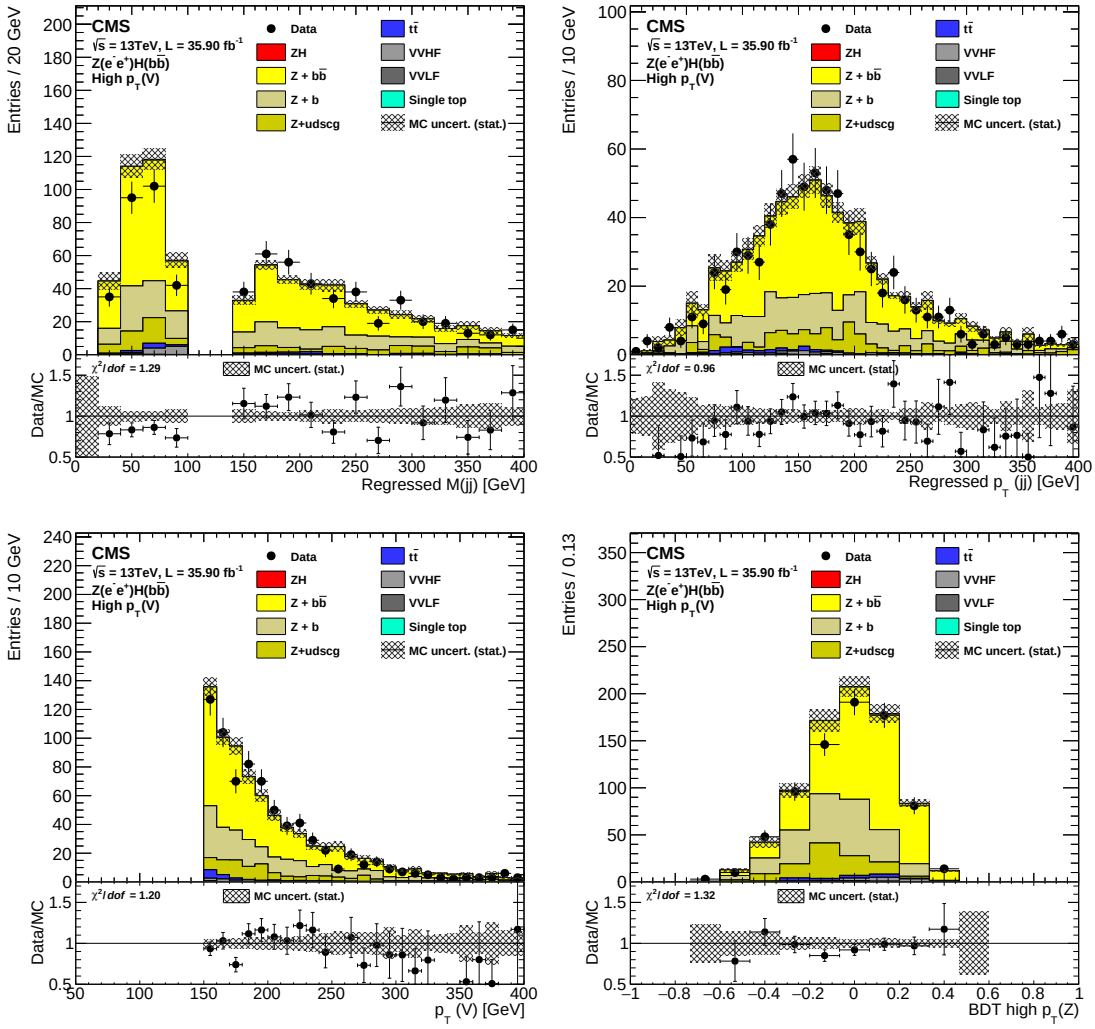


Figure 8.9: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ Z + HF control region for the Z(ee)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

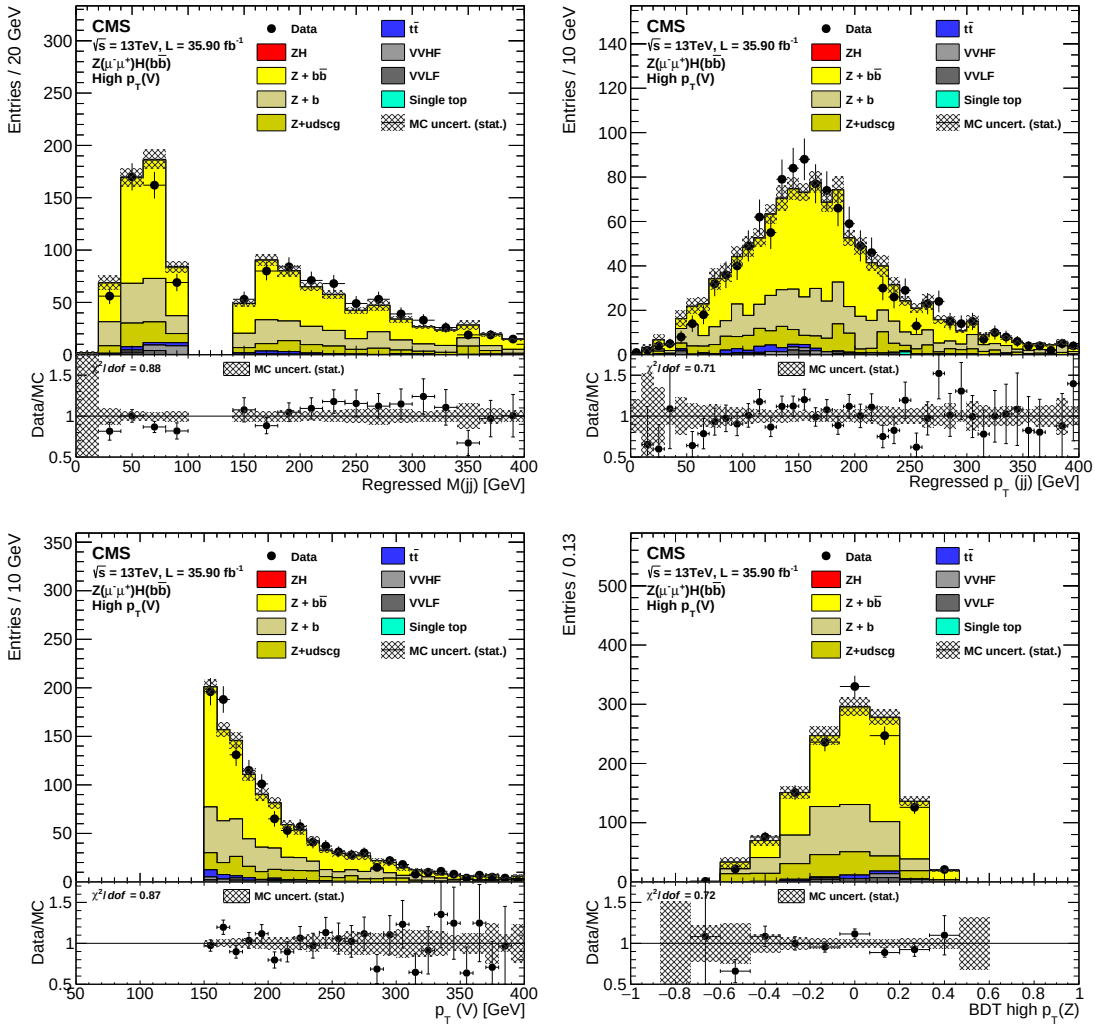


Figure 8.10: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ Z + HF control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

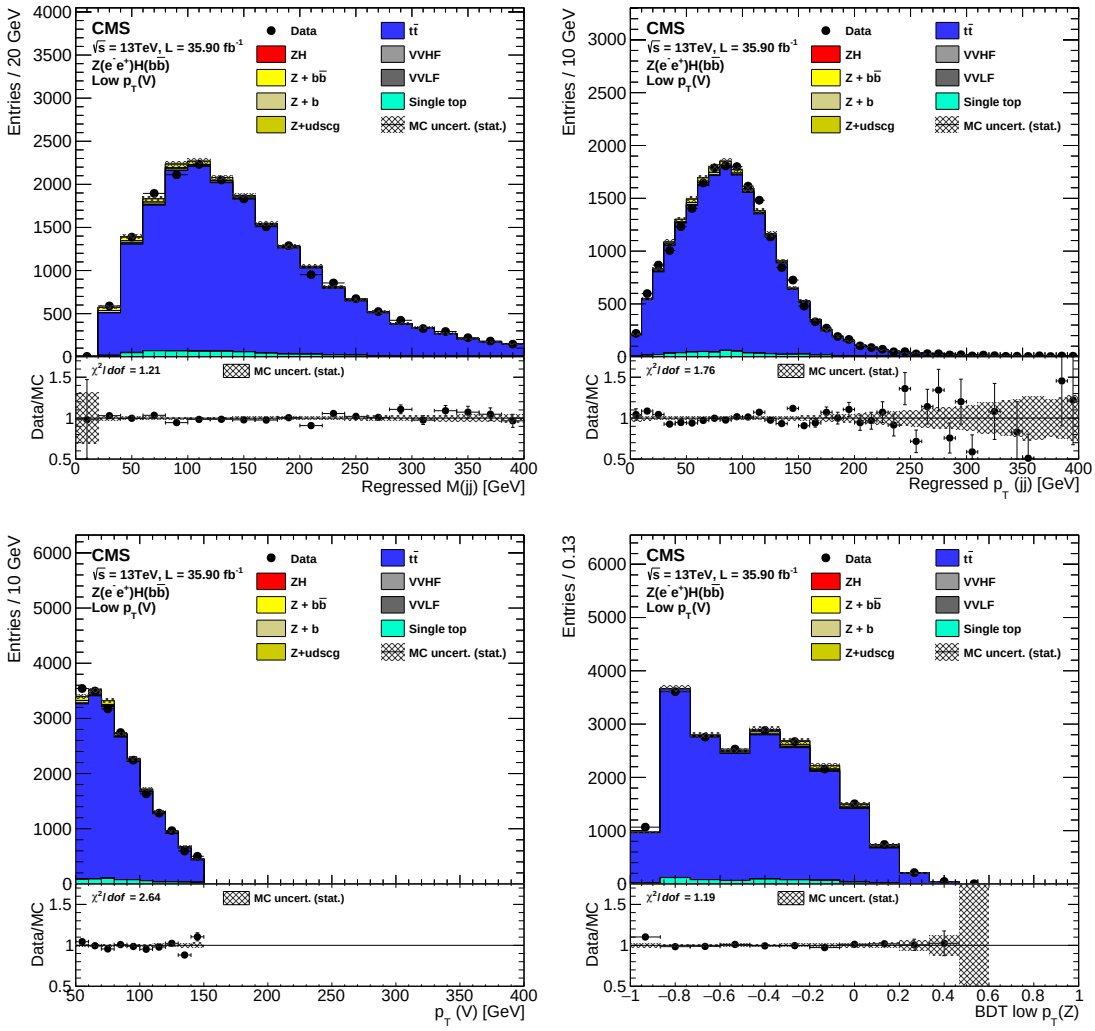


Figure 8.11: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ $t\bar{t}$ control region for the Z(ee)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

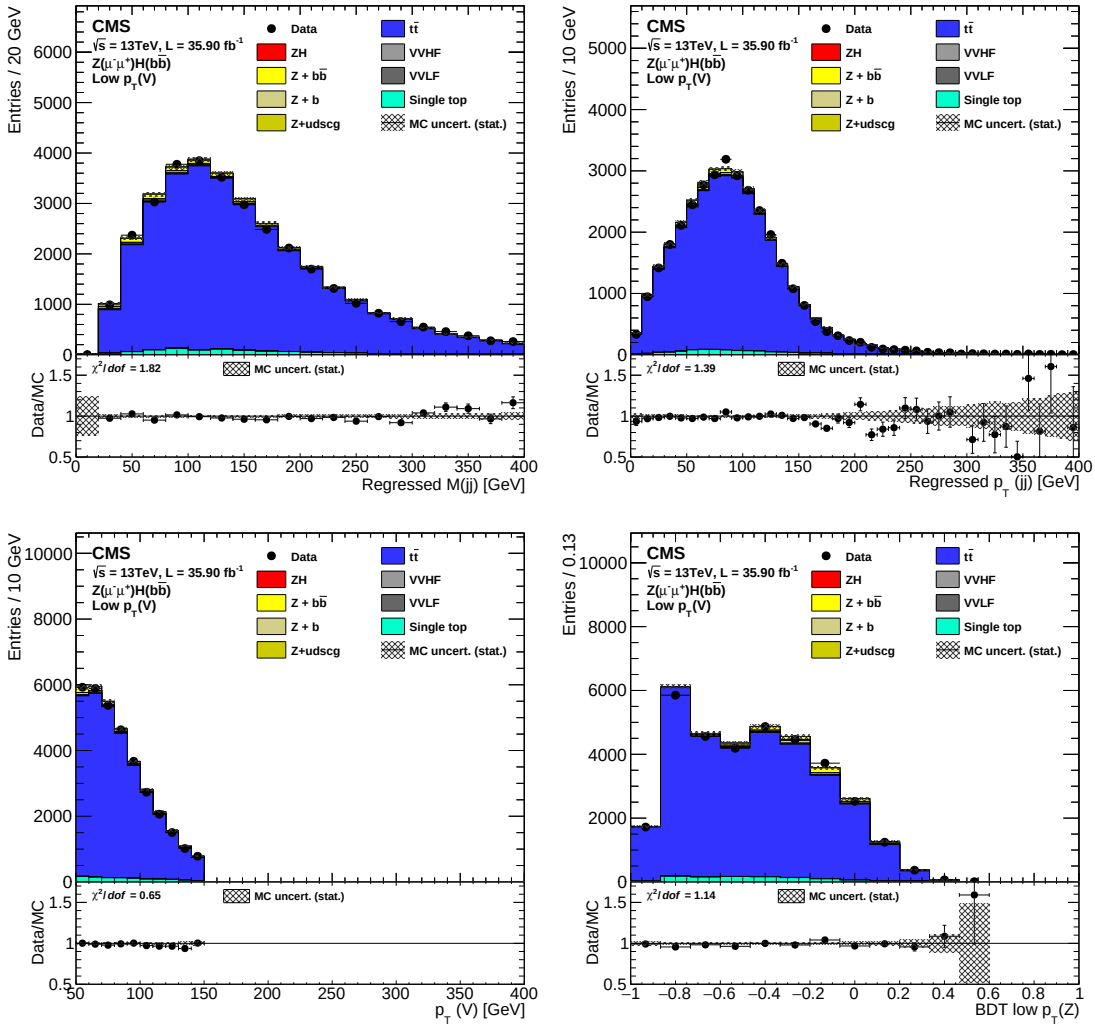


Figure 8.12: Distribution of variables in data and Monte-Carlo samples in the low-$p_T(V)$ $t\bar{t}$ control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

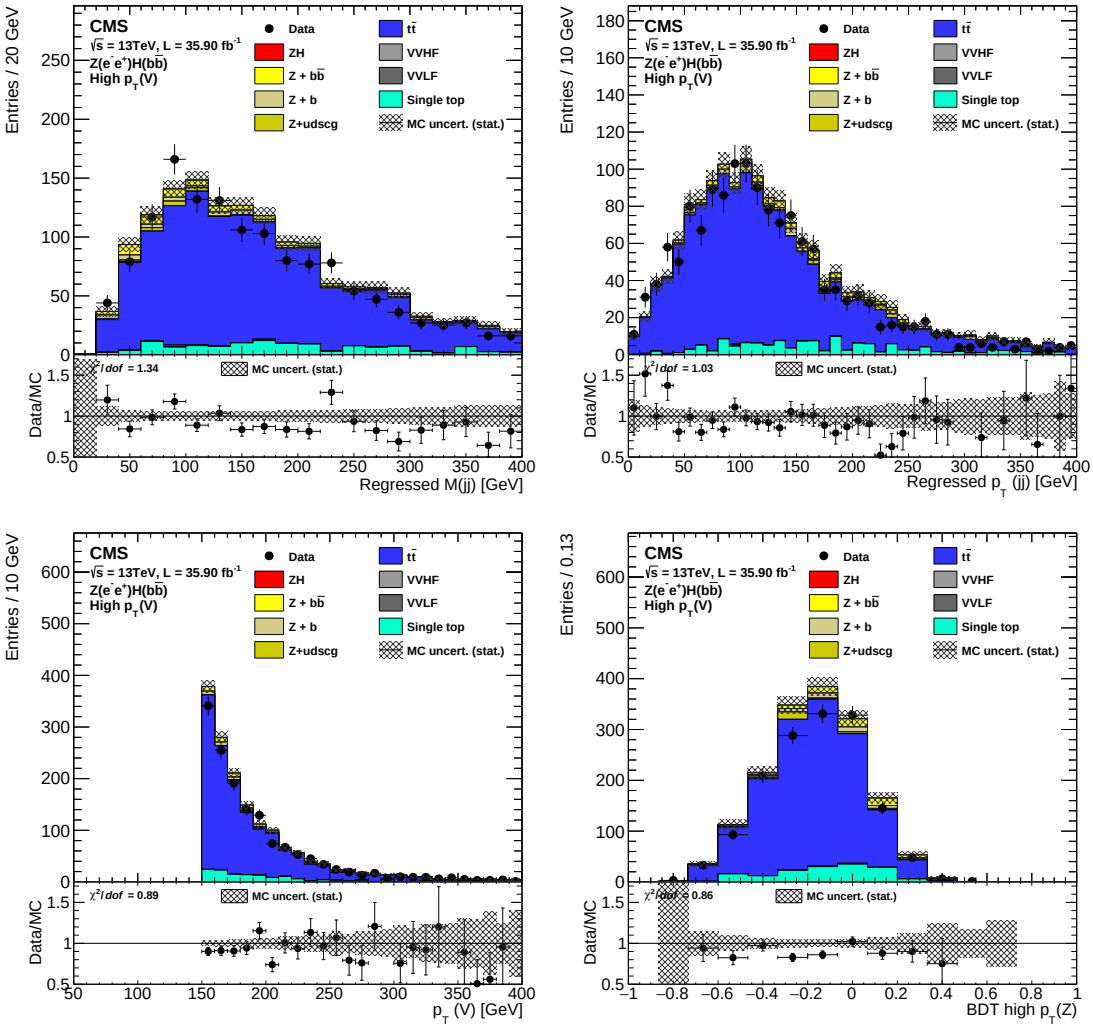


Figure 8.13: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ $t\bar{t}$ control region for the Z(ee)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

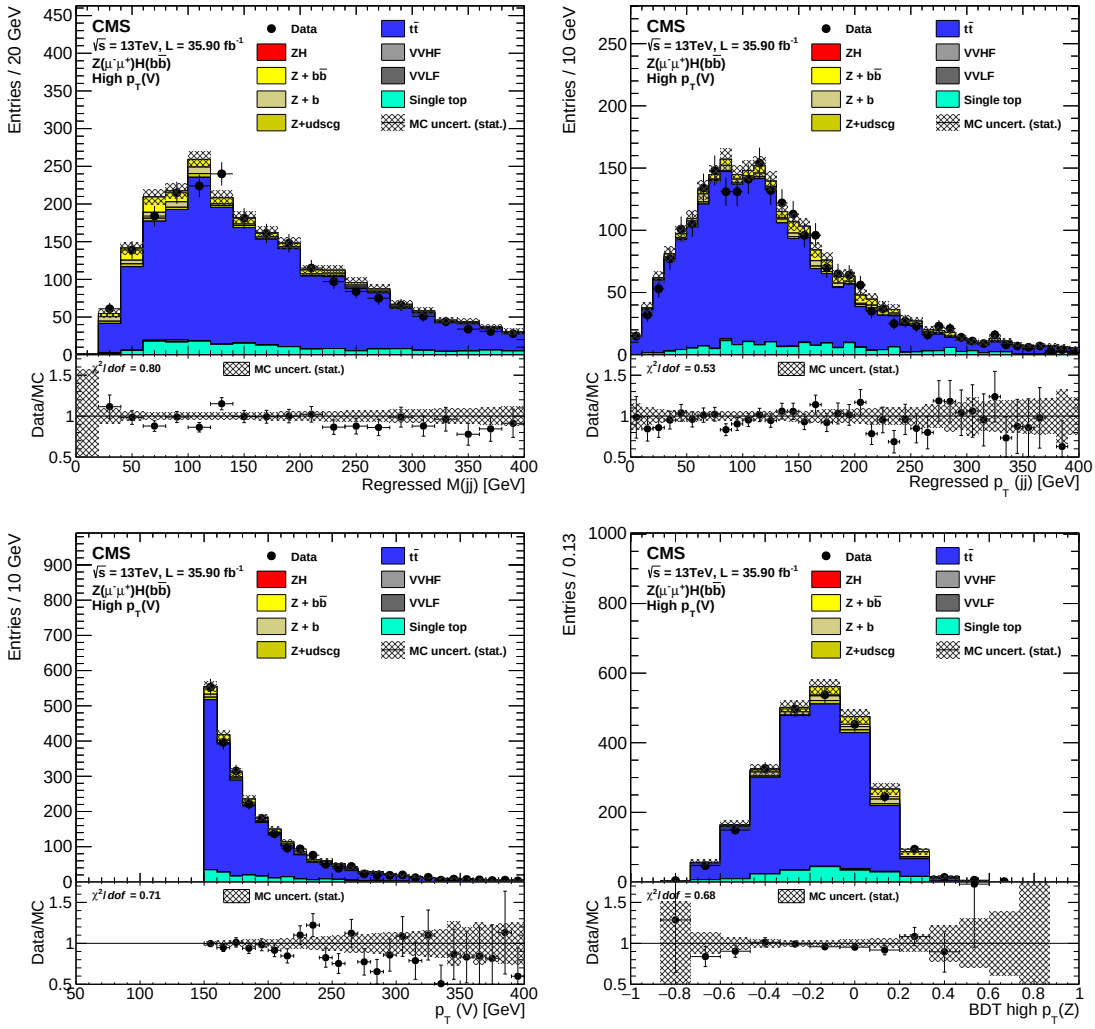


Figure 8.14: Distribution of variables in data and Monte-Carlo samples in the high-$p_T(V)$ $t\bar{t}$ control region for the Z($\mu\mu$)H(bb) sub-channel. The layout and composition of the four figures is similar to what is shown in Figure 8.3 and is described in the legend.

8.3.1 Treatment of the systematic uncertainties

Two cases are considered in the treatment of the systematic uncertainties, depending on how they affect the input distributions of the final binned-likelihood fit for the control and signal regions (CMVAv2$_{\text{min}}$ and BDT). A *normalization uncertainty* uniformly affects the normalization of one or multiple processes. It is parameterized by a log-normal distribution, which affects the event yield in a multiplicative way. An example of such a systematic is the 2.5% uncertainty on the luminosity, affecting all processes. The uncertainty is parameterized by a parameter k, where a deviation of $+1\,\sigma$ ($-1\,\sigma$) corresponds to a yield scaling by a factor k ($1/k$). If $\delta x/x$ is the relative uncertainty on the yield, k is set to $1 + \delta x/x$. The log-normal distribution is

$$p(\widetilde{\theta}|\theta) = \frac{1}{\sqrt{2\pi\theta}\ln k}\exp\left(-\frac{(\ln\theta - \ln\widetilde{\theta})^2}{2\ln^2 k}\right),$$

where θ and $\tilde{\theta}$ is the nuisance parameters and it's best estimate, respectively. The term $p(\widetilde{\theta}|\theta)$ is the likelihood to measure the best estimate $\tilde{\theta}$ prior to the nuisance parameter θ, see section 6.1.1.

A *template uncertainty* affects both the normalization and shape of the input distributions. The initial distribution, referred to as the *nominal* distribution, is modified according to a up and down $\pm1\sigma$ variation of the nuisance parameter. The nominal shape, up and down shapes are then included in the final binned-likelihood fit. Examples of template uncertainties are the JECs (Jet Energy Corrections), which affect the BDT output score distribution, as the value of some BDT input variables, the number of additional jets in the event, dijet transverse momentum, invariant mass and the MET, depends on the JEC.

The up and down variations of the BDT output score shape are derived by re-evaluating the BDT on signal region events, varying the relevant BDT input variables by the $\pm1\sigma$ JECs systematic uncertainties. This shape variation doesn't affect the normalization of the distribution. The effect on the normalization is obtained by simultaneously modifying the signal region selection by the $\pm1\sigma$ JECs systematic uncertainty during the evaluation. As the b jet candidates transverse momentum and other variables affected by the JECs are used for the signal region event selection, this performs a new up and down event selection. An example of a template uncertainty can be found in Figure 8.15. The nuisance parameters for template uncertainties follows a Gaussian distribution

$$p(\widetilde{\theta}|\theta) = \frac{1}{\sqrt{2\pi\theta}\sigma}\exp\left(-\frac{(\theta - \widetilde{\theta})^2}{2\sigma^2}\right).$$

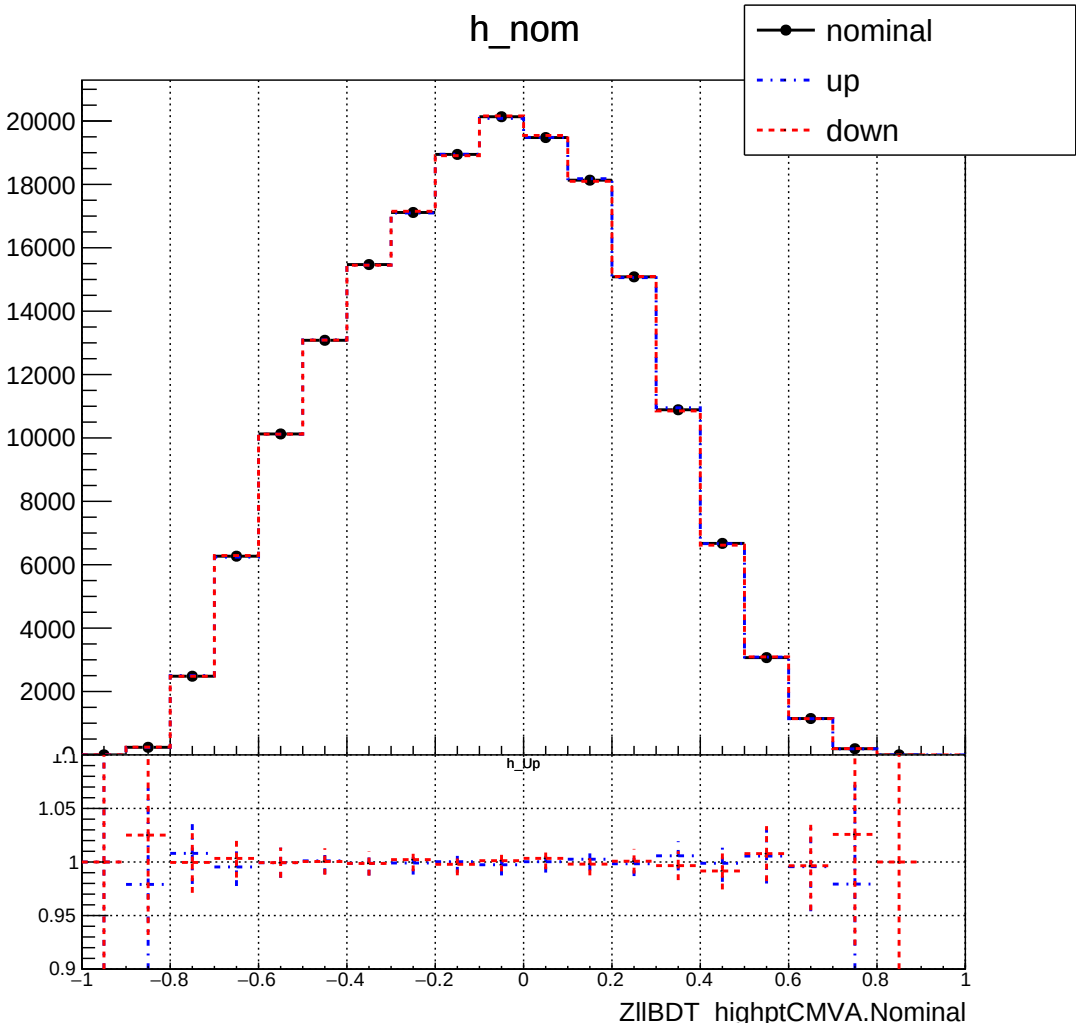


Figure 8.15: Nominal and up/down template variations of the BDT output score in the high-$V(p_T)$ control region. The shape variations correspond to a JEC systematic uncertainty. The template uncertainty affects both the normalization and shape of the BDT output distribution, as described in the text.

8.3.2 Sources of systematic uncertainties

The VH(bb) analysis includes multiple sources of systematic uncertainties, listed in this section. They are treated either as fully correlated, fully anticorrelated or decorrelated. Decorrelated systematic uncertainties are included in the binned-likelihood fit by a one-to-one corresponding multiplicative factors (for normalization uncertainties) or shape templates (for template uncertainty) to model the $\pm 1\sigma$ variations. Two fully correlated systematic uncertainties are obtained by simultaneously varying the nuisance parameters by $\pm 1\sigma$ during the final fit. Two fully anticorrelated systematic uncertainties are obtained by varying one nuisance parameter by $\pm 1\sigma$ and the other by $\mp 1\sigma$. The nuisance parameters in the list below are treated as decorrelated, except for the jet energy scale uncertainties, whose effect on the b-tagging efficiency and jet energy correction are fully correlated.

- **Size of the simulated samples:**
  The limited size of the Monte-Carlo simulation can lead to large statistical uncertainties in the BDT output score bins. This is taken into account by individually varying the shape by Poisson errors in the bins where the statistical uncertainty is large.

- **Simulated sample modeling:**
  To estimate the systematic uncertainties related a particular choice of the Monte-Carlo generator, the BDT and $\text{CMVAv2}_{\text{min}}$ shapes have been simulated with another event generator. The difference in shapes between the two generators is symmetrized and used to define the shape variation. For the Z + jets, the difference in shape between MG5AMC@NLO Monte-Carlo generator at LO (used as the nominal shape) and NLO have been used. For the $t\bar{t}$ process, the differences in shape between POWHEG (used as the nominal shape) and MC@NLO have been used.

- **Signal cross-section:**
  The ZH(bb) signal cross-section has been calculated to NNLO QCD + NLO EW accuracy for the qqZH process and at NNLO + NNLL QCD accuracy for the ggZH process. The associated systematic uncertainties include the effect of the factorization and normalization scale, as well as the PDF uncertainties. The NNLO QCD correction uncertainties are about 4% on the ZH(bb) signal. They are included as normalization uncertainties. The up and down variations from the NLO electroweak correction are applied on the qqZH process as a function of the vector boson transverse momentum to derive the shape uncertainty on

the BDT and CMVAv2$_{\text{min}}$ distribution and result in an average 4% uncertainty.

- **Background cross-section:**
  A normalization uncertainty of 15% is assigned to the single-top and diboson production cross-sections. These uncertainties are about 25% larger than those from the CMS measurements of these processes, to account for the different kinematic regime in which those measurements are performed [72–74].

- **QCD scale** The uncertainty related to the normalization and factorization scale, μ_R and μ_F, are taken into account as template uncertainties for background and signal processes. The templates are generated by independently varying the scales between 0.5 and 2 in the event generator and re-calculating the corresponding CMVAv2$_{\text{min}}$ and BDT output score distributions. This source of systematic uncertainties impacts the shape of the final binned-likelihood fit input distributions and is complementary to the QCD scale uncertainties on the sample cross-section, which are normalization uncertainties.

- **Parton distribution functions:** The limited knowledge of the proton structure is taken into account by the PDF uncertainties. About 100 PDF variations from the NNPDF [11] set are available. Including the BDT output score shapes for each variation in the final binned-likelihood fit would not have been feasible. Instead, the PDF systematic uncertainties are treated as normalization uncertainties for each background process. The BDT output score of each PDF variation is re-calculated separately in the signal region for each background process. The mean and the root mean squared value (RMS) of all the PDF variations are then estimated in each bin of the BDT score. The RMS is defined as RMS $= \frac{1}{n}\sum_i^n b_i^2$, where the sum runs over all the n PDF variations, and b_i is the yield of the BDT score in the bin under consideration, calculated form the i'th PDF shape. The RMS divided by the mean is calculated in each bin of the BDT score distributions for the given process, and the largest variations is selected to be used as the parameter k of the normalization uncertainty.

- **b-tagging efficiency and mistag rate**
  The systematic uncertainties related to the fixed-cut operating point of the b-tagger are not considered, as the full CMVAv2$_{\text{min}}$ distribution is used and the systematic uncertainties related to the shape corrections are included. Those systematic uncertainties consist of 9 uncorrelated sources to cover possible shape

discrepancies between data and Monte-Carlo samples for the CMVAv2$_{\text{min}}$ distributions[2]. They are also propagated to the BDT output score shape.

The systematic uncertainties are separated in four categories. In the first category, one source addresses uncertainties in the jet energy scale. The second category is related to the sample purity when deriving the data-to-simulation scale factors for light (u, d, s or g) and b jets (see section 6.4.5.3). It includes two sources, one for the b jet and the other for light jets. The third category addresses the statistical uncertainties and is also separated for b and light jets. For each of the two jet categories, the statistical fluctuations are parameterize by linear and quadratic functions, corresponding to four sources of systematic uncertainties in total. The fourth category contains two sources evaluating the uncertainties of the c jet data-to-simulation scale factors, one parameterized with the linear function and the other with the quadratic function.

The nuisance parameters associated to jet energy scale and resolution as well as those associated to b-tagging were prone to overconstraining in the signal extraction fit. This was attributed to the fact that only one nuisance parameter was used for all jets regardless of p_T and η, although the size of the nuisance varies depending on p_T and η. The sources of b-tagging efficiency uncertainties are evaluated separately in various p_T-η bins of the b jet candidates. The choice of binning is selected to match the binning in which these systematics are estimated. The average uncertainties on the data-to-simulation scale factors are: 1.5% per b jet, 5% per c jet, and 10% per light jet.

- **Jet energy correction**
  The uncertainties of the jet energy correction and jet energy scale have an effect on the BDT output score shape, as the up/down variation modify multiple input variables to the BDT (see example in section 8.3.1). A total of 26 sources of uncertainties, taking into account the three level of JEC corrections (see section 6.4.4), are included for the JEC[3]. All those sources are fully correlated with the jet energy scale nuisances for the b-tagging efficiencies mentioned in the previous bullet. A single source is used for the jet energy resolution. The uncertainties are evaluated as functions of the p_T-η of each jet in the event.

- **Missing transverse energy:** Another source of systematic uncertainties that affects the MET reconstruction is the estimate of the energy that is not clustered in jets [76]. This affects only the Z(vv)H(bb) and W(lv)H(bb) channels, with an individual contribution to the signal strength uncertainty of 1.3%.

[2]For a more detailed review on how those uncertainties are estimated, see section 8.5 from [65]
[3]The full list of JEC systematics is documented in [75]. See Table 1, page 65.

- **Luminosity:** An total uncertainty of 2.5% is assessed for the luminosity on the 2016 data-taking period [77].

- **Lepton Efficiency:** The efficiencies for the electron and muon trigger, reconstruction, identifications and isolation are estimated with a tag and probe method (see chapter 7). Those uncertainties include the statistical uncertainties in each bin where the tag and probe fit is performed: η bins for the reconstruction efficiencies, p_T-η bins for the trigger, identification, and isolation efficiencies. The uncertainties are propagated following the same binning as for the tag and probe derivation.

9 Results of the VH(bb) analysis

This chapter is dedicated to the results of the VH(bb) analysis. The outcome of the VH(bb) and diboson measurements on the 2016 dataset are reviewed in section 9.1 and 9.2, respectively. In section 9.3, the results on the 2016 dataset are updated with the latest iteration of the CMS VH(bb) analysis, performed on the data recorded during the 2017 data taking period. The setup of this 2017 VH(bb) analysis is outside of the contribution presented in this dissertation and is mentioned for reference.

9.1 VH(bb) analysis on 2016 dataset

In this section, the performances of two b-tagging methods considered for the VH(bb) analysis are studied in the Z(ll)H(bb) channel. The outcomes of the final binned-likelihood fit relative to the shapes, nuisance parameters and background normalization scale factors are described in section 9.1.2. The results of the signal extraction are discussed in section 9.1.3. Section 9.1.4 is dedicated to the combination of the 2016 VH(bb) analysis with previous VH(bb) searches performed on datasets recorded by CMS during the Run 1 data-taking period.

9.1.1 B-tagger discriminator studies

As mentioned in section 6.4.9, the two jets with the highest b-tagging score are selected as the b jet candidates. The two b-tagging algorithms considered for the VH(bb) analysis are CSVv2, previously used in the Run 1 version of the analysis, and CMVAv2. The CMVAv2 algorithm shows a better performance than CSVv2, increasing the b jet identification efficiencies by 5%, 3% and 3% in the tight, medium, and loose working point, respectively, and therefore is the natural choice for the b-tagging method.

An additional study is conducted in the Z(ll)H(bb) channel to compare the analysis sensitivity with each b-tagging algorithm. Two versions of the full analysis are performed, one for each b-tagging algorithm. The expected significance is evaluated in both cases from the combination of the BDT output score distributions in the signal regions and the CMVA_{min} distributions in the control regions used in the final binned-likelihood fit. The main differences between the two versions are the b jet candidates selection, as the b jets have been selected using the highest score of the b-tagging algorithm under consideration, and the signal and control regions definitions, which use the same b-tagging working points as listed in table 8.1 and 8.3 in both versions. As the b-tagging score of both jets candidates is an input variable of the signal region BDT discriminant, the training has been performed separately in both cases. Table 9.1 lists the event yields in 3 out of 15 most sensitive bins of the BDT output score distribution in the high-$V(p_T)$ signal region category for the CMVAv2 and CSVv2 case. A Poisson significance, defined as $S/\sqrt{B}$, points to a better sensitivity for the analysis using the CMVAv2 b-tagger.

All the systematic uncertainties, with the exception of the b-tagging efficiency uncertainties related to the b-tagger shape corrections, are included in the final binned-likelihood fit. No comparisons between the CMVAv2 and CSVv2 uncertainties have been performed. The background normalization scale factors, whose values have been determined by the binned-likelihood fit, are applied. Moving from the CSVv2 to the CMVAv2 increases the expected significance by 7%. Based on this result, the CMVAv2 discriminator is used in all three VH(bb) channels.

9.1.2 Postfit distributions in the signal and control regions

The contribution of this dissertation concerns the VH(bb) analysis conducted in the Z(ll)H(bb) channel on the 2016 dataset. The tools, as well as measurements and studies necessary to setup the analysis in this channel are reviewed in Chapter 6 to 8. The final binned-likelihood fit performed to extract the results on the Z(ll)H(bb) channel and VH(bb) analysis is performed simultaneously on the Z(ll)H(bb), W(lv)H(bb) and Z(vv)H(bb) channels. Therefore, the result mentioned in the rest of this chapter is a combination of separate studies. The analysis in the Z(vv)H(bb) and W(lv)H(bb) channels, as well as the performance of the final binned-likelihood fit, have been performed by other members of the CMS collaboration and are not part of the contribution presented in this thesis.

The final binned-likelihood fit is performed simultaneously on the $\text{CMVAv2}_{\text{min}}$ distributions from the control regions and the BDT output score distributions from the signal regions from the three VH(bb) channels, that is the Z(ll)H(bb), W(lv)H(bb) and Z(vv)H(bb) channel. This corresponds to a total of 7 BDT output score distributions (or, equivalently, signal regions) and 24 $\text{CMVAv2}_{\text{min}}$ distributions (control regions) fitted together. In addition, 635 nuisance parameters are included in the fit to account

Process	CMVAv2		CSVv2	
	electron	muon	electron	muon
Z + 2 b jets	0.19	0.31	0.19	0.42
Z + 1 b jet	4.93	2.57	5.70	3.80
Z + light jets	22.59	27.25	17.35	30.40
$t\bar{t}$	1.98	0.71	1.61	0.73
single top	0	0.33	0	0
diboson + light jets	0.05	0.23	0.01	0.28
diboson + 2 b jets	0.91	1.09	0.61	1.40
Total background	30.64	32.47	25.47	37.01
VH(bb)	6.58	9.76	5.79	9.78
$S/\sqrt{B}$	1.19	1.71	1.14	1.61

Table 9.1: Comparison of the signal and background event yields for the CMVAv2 and CSVv2-based Z(ll)H(bb) analysis in the 3 out of 15 most sensitive bins of the BDT output score distributions. The BDT output score distributions in the high-$V(p_T)$ signal region are considered. The signal regions are separated in the Z(ee)H(bb) (electron) and Z($\mu\mu$)H(bb) (muon) sub-channels.

for all the systematic uncertainties.

The shape and normalization of all distributions are allowed to vary within the uncertainties defined in section 8.3.2, treated as independent nuisance parameters during the fit. The signal strength μ, nuisance parameters and normalization scale factors (see section 8.2.2) are allowed to float freely and are adjusted by the fit.

The postfit (after the final binned-likelihood fit) BDT output score distributions in the four signal regions of the Z(ll)H(bb) channel can be found in Figure 9.1. The first (second) row corresponds to the low (high) -$V(p_T)$ region, and the first (second) column to the Z(ee)H(bb) (Z$\mu\mu$H(bb)) sub-channel. The upper part of each figure shows the data (black dot histogram) and Monte-Carlo (colored histograms) distributions. The Monte-Carlo distribution consists of multiple stacked histograms, each histogram corresponding to a particular process listed in the legend. The signal processes (qqZH and ggZH) are both in red and their presence can be seen in the rightmost bins of the distributions, as the signal-background separation increases with the BDT output score. This is mostly apparent in the figures corresponding to the high-$V(p_T)$ category (lower row) than the low-$V(p_T)$ category (upper row), as the high-$V(p_T)$ signal regions have a higher signal sensitivity. The lower part of each figure shows the ratio plot between the data and the Monte-Carlo histograms. The black dots correspond to the values of the data/Monte-Carlo ratio. The hatching bands around 1 are the systematic uncertainties of the Monte-Carlo samples propagated to this ratio. The grey hatching bands correspond to the statistical (Poisson) uncertainties, the yellow

Process	Low-$V(p_T)$	High-$V(p_T)$
Z + 2 b jets	617.5	113.9
Z + 1 b jet	141.1	17.2
Z + light jets	58.4	4.1
$t\bar{t}$	157.7	3.2
single top	2.3	0.0
diboson + light jets	6.6	0.5
diboson + 2 b jets	22.9	3.8
Total background	1006.5	142.7
VH(bb)	33.7	22.1
Data	1030	179
S/B	0.033	0.15

Table 9.2: The total number of events in the three last (i.e. most sensitive) bins of the BDT output score distribution in the Z(ll)H(bb) channel signal regions. The signal regions are separated in the low and high-$V(p_T)$ categories.

bands to the prefit statistical + systematic uncertainties and the red bands the postfit statistical + systematic uncertainties. By comparing the width of the two bands in both distributions, it can be seen that the total uncertainties are reduced after the fit. As it can be seen in the lower ratio plot, the agreement between data and Monte-Carlo simulation is good for all the distributions, the data points being within the red uncertainty bands in almost all the bins. The total number of Monte-Carlo simulated events in the three rightmost bins of the BDT output score distribution, which have the highest signal sensitivity, are shown in table 9.2.

The postfit CMVAv2$_{\text{min}}$ distributions in the control regions can be found in Figures 9.2 and 9.3 for the low and high-$V(p_T)$ category, respectively. In both figures, the left and right column corresponds to the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels, respectively, and the first, second and third row corresponds to the Z + light, Z + heavy flavor and $t\bar{t}$ control region, respectively. The arrangement of each figure is similar to the BDT output score distributions in the signal regions shown in Figure 9.1. As in the BDT output score distributions, the total uncertainties on the CMVAv2$_{\text{min}}$ distributions are reduced after the fit.

A standard test to evaluate the validity of the 635 nuisance parameters included in the fit is to compare the *pulls*, $(\hat{\theta} - \theta)/\sigma(\theta)$, and the *constrains* given by the *ratios* of the variances, $\sigma(\hat{\theta})/\sigma(\theta)$, where θ ($\hat{\theta}$) is the value of the nuisance before (after) the fit and $\sigma(\theta)$ ($\sigma(\hat{\theta})$) is the nuisance uncertainty before (after) the fit. Large pulls ($>> 0$) can indicate a wrong assumption on the correction related to the nuisance parameter's central value, leading to residual Monte-Carlo modeling of the data after the fit. Ratios

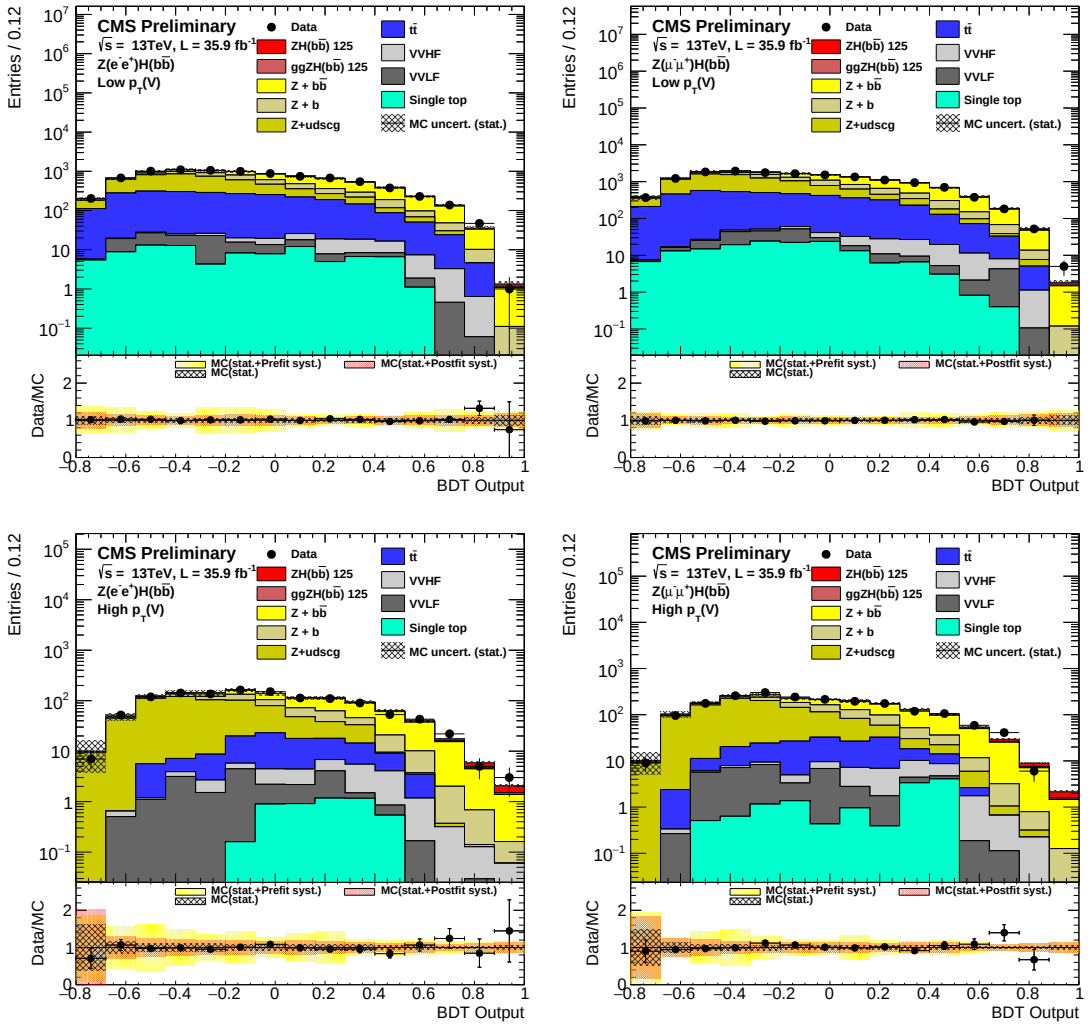


Figure 9.1: Postfit BDT output score distributions in the four signal regions in the Z(ll)H(bb) channel. The binned-likelihood fit is performed on both the control and signal regions from the three VH(bb) analysis channels. The signal regions are, from left to right, top to bottom: Z(ee)H(bb) low-$p_T(Z)$, Z($\mu\mu$)H(bb) low-$p_T(Z)$, Z(ee)H(bb) high-$p_T(Z)$, Z($\mu\mu$)H(bb) high-$p_T(Z)$.
The upper part of each figure shows the data (black dot histogram) and Monte-Carlo (colored histograms) distributions. The Monte-Carlo distribution consists of multiple stacked histograms, each histogram corresponding to a particular process listed in the legend. The lower part of each figure shows the ratio plot between the data and the Monte-Carlo histograms. The black dots correspond to the values of the data/Monte-Carlo ratio. The hatching bands around 1 are the systematic uncertainties of the Monte-Carlo samples propagated to this ratio. The grey hatching bands correspond to the statistical (Poisson) uncertainties, the yellow bands to the prefit statistical + systematic uncertainties and the red bands to the postfit statistical + systematic uncertainties.

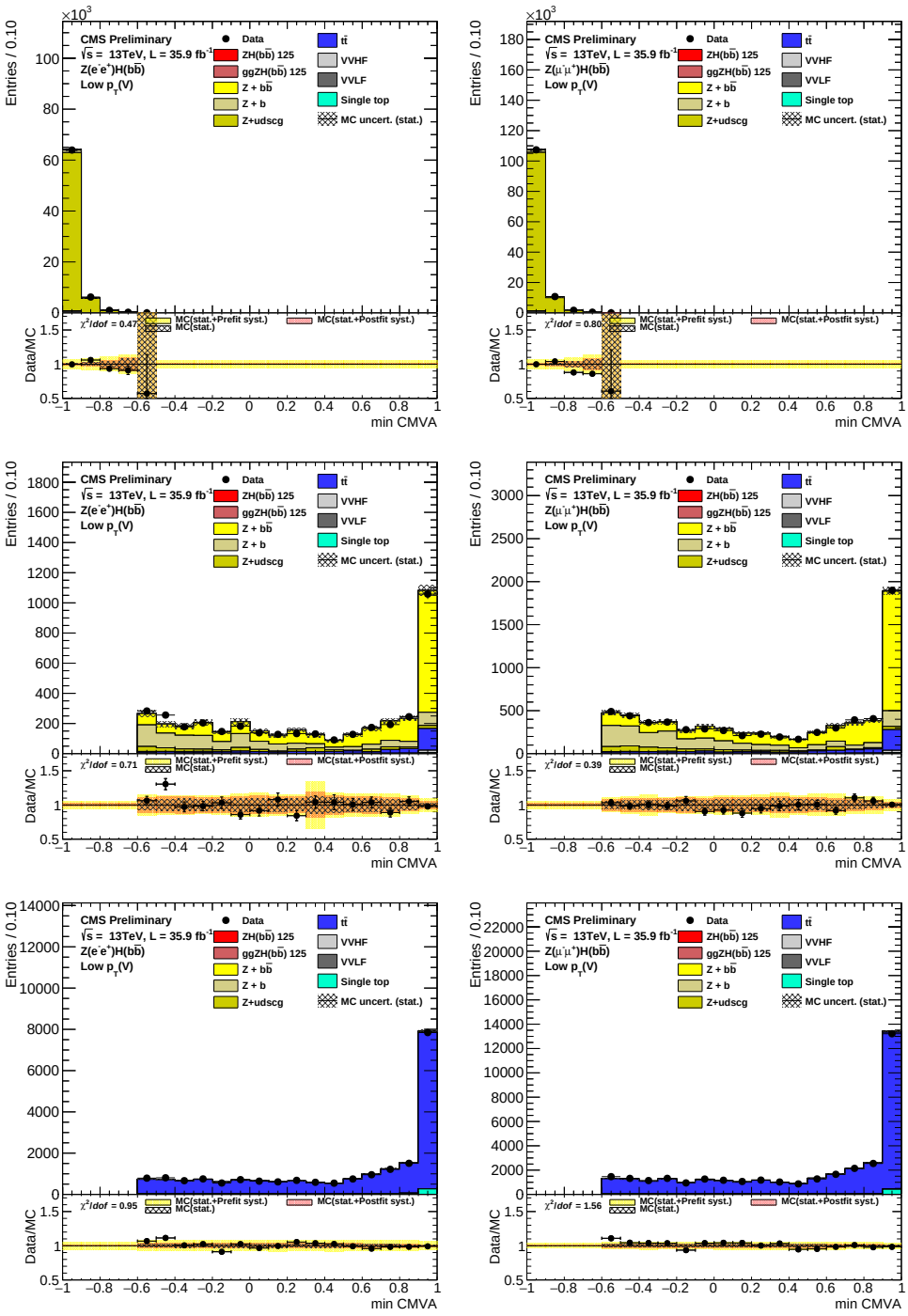


Figure 9.2: Postfit $\mathrm{CMVAv2}_{\mathrm{min}}$ distributions in the six low-$p_T(Z)$ control regions in the Z(ll)H(bb) channel. The binned-likelihood fit is performed on both the control and signal regions from the three VH(bb) analysis channels. The left and right column correspond to the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels, respectively. The first, second and third row correspond to the Z + light, Z + heavy flavor and $t\bar{t}$ control regions, respectively. The arrangement of each figure is similar to the BDT output score plots in the signal region and is described in the legend of Figure 9.1.

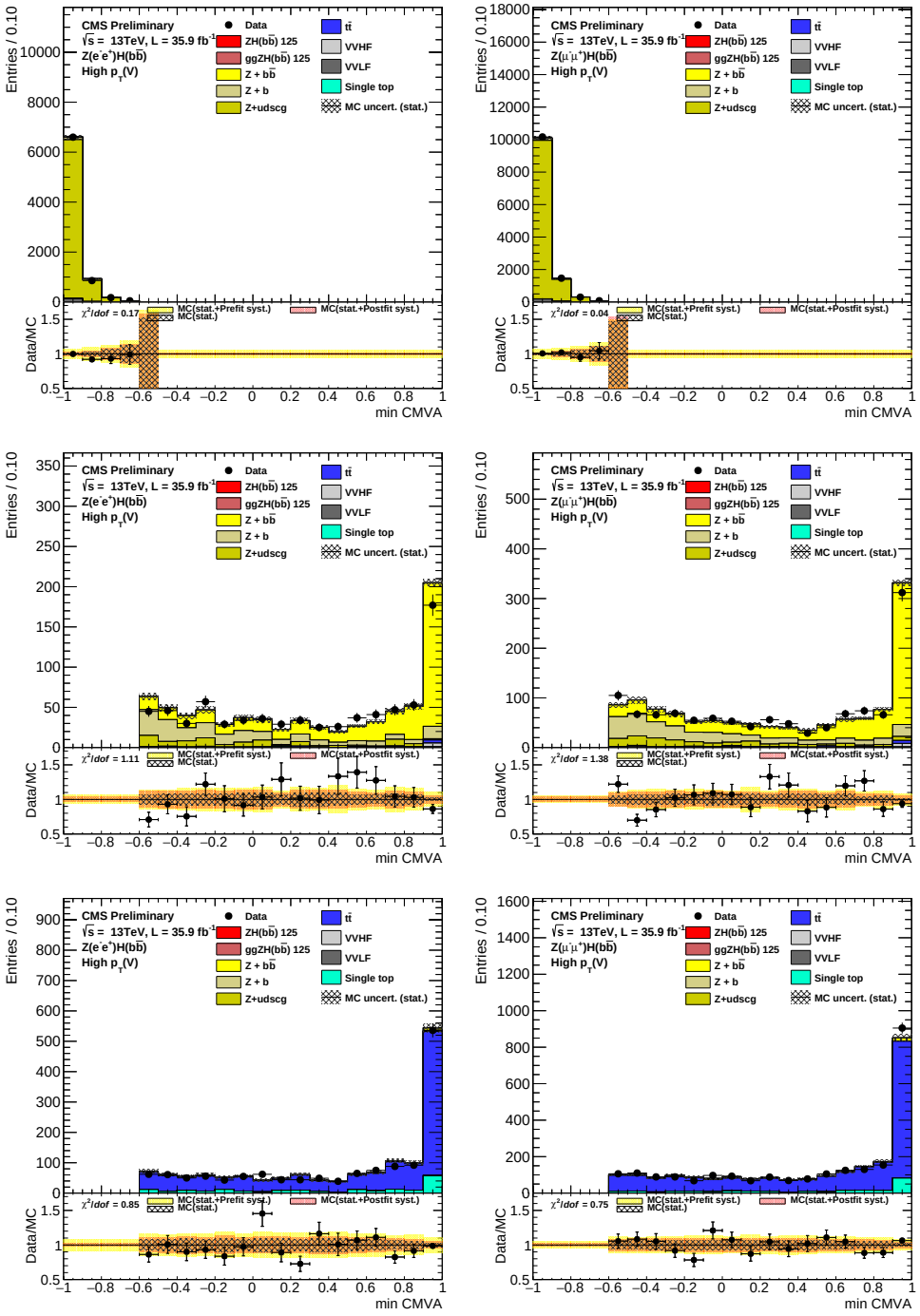


Figure 9.3: Postfit $\mathrm{CMVAv2}_{\mathrm{min}}$ distributions in the six high-$p_T(Z)$ control regions in the Z(ll)H(bb) channel. The binned-likelihood fit is performed on both the control and signal regions from the three VH(bb) analysis channels. The left and right column correspond to the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels, respectively. The first, second and third row correspond to the Z + light, Z + heavy flavor and $t\bar{t}$ control regions, respectively. The arrangement of each figure is similar to the BDT output score plots in the signal region and is described in the legend of Figure 9.1.

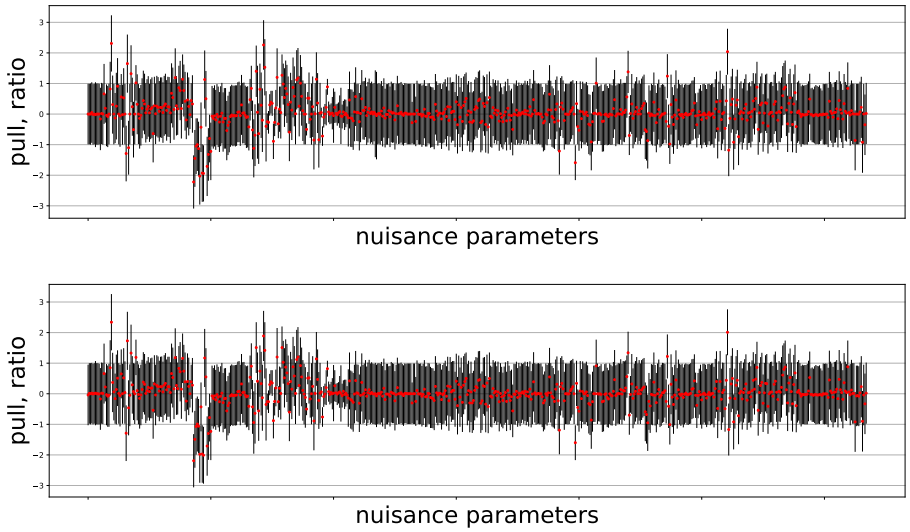


Figure 9.4: Pulls $(\hat{\theta} - \theta)/\sigma(\theta)$ and constrains ratios $\sigma(\hat{\theta})/\sigma(\theta)$ for all nuisance parameters in the Z(ll)H(bb) channel. The red points correspond to the pull values and the error bars to the ratios. **Upper plot:** all signal and background processes are included in the fit. **Lower plot:** only background processes are included in the fit.

$<< 1$ can indicate an overconstraint of the systematic uncertainty value before the fit, pointing to an over-estimation of the nuisance parameter's uncertainty prior to the fit.

The pulls and constrains for all nuisance parameters can be found in Figure 9.4. The red points correspond to the pull values and the error bars to the ratios. The lower plot includes all Monte-Carlo processes while the upper plot doesn't include the signal contribution. The drop on the left sides of both plots (with pulls between approximately -2 and -1) is due to a dozen of nuisances corresponding to the JEC sources of b-tagging efficiency uncertainties. A total of 34 nuisance parameters are considered as *overconstrained* by the fit, with an absolute pull value above 2 and/or a ratio below 0.5. The list of all overconstrained nuisances is shown in Figure 9.5.

9.1.3 Signal extraction

To help visualizing a potential signal excess, the BDT output score distributions adjusted by the fit from all channels are combined into a single distribution shown in Figure 9.6. Events in bins of the BDT output score having a similar signal-to-background ratio (S/B) are merged into the same $\log_{10}(S/B)$ bin in this distribution. In the upper part, the grey contribution corresponds to the stacked events of all the Monte-Carlo background processes and the red line to VH(bb) signal contributions. The data distribution is shown by the black dot histogram. It the lower part, the ratio plot depicts

nuisance	background fit $\Delta x/\sigma_{in}$, σ_{out}/σ_{in}	signal fit $\Delta x/\sigma_{in}$, σ_{out}/σ_{in}	$\rho(\mu, \theta)$
CMS_vhbb_LHE_weights_scale_muF_TT	**+0.19, 0.47**	**+0.22, 0.47**	+0.02
CMS_vhbb_LHE_weights_scale_muF_Wj1b	**+2.31, 0.92**	**+2.34, 0.92**	+0.01
CMS_vhbb_bTagWeightHF_pt4_eta1	**-0.02, 0.47**	**-0.00, 0.45**	+0.01
CMS_vhbb_bTagWeightJES_pt0_eta1	**-2.22, 0.87**	**-2.19, 0.87**	+0.01
CMS_vhbb_bTagWeightJES_pt1_eta3	**-2.03, 0.93**	-1.98, 0.93	+0.02
CMS_vhbb_bTagWeightJES_pt2_eta3	-1.94, 0.92	**-2.01, 0.93**	-0.02
CMS_vhbb_bTagWeightLF_pt2_eta2	**+0.06, 1.52**	+0.41, 0.94	+0.09
CMS_vhbb_bTagWeightLF_pt4_eta1	**+2.26, 0.81**	+1.89, 0.82	-0.16
CMS_vhbb_bTagWeightcErr1_pt0_eta1	**-0.25, 0.27**	**-0.26, 0.26**	-0.01
CMS_vhbb_bTagWeightcErr1_pt0_eta2	**-0.27, 0.37**	**-0.26, 0.37**	+0.00
CMS_vhbb_bTagWeightcErr1_pt0_eta3	**+0.30, 0.35**	**+0.30, 0.35**	-0.00
CMS_vhbb_bTagWeightcErr1_pt1_eta1	**+0.39, 0.37**	**+0.39, 0.38**	+0.00
CMS_vhbb_bTagWeightcErr1_pt1_eta2	**+0.35, 0.41**	**+0.35, 0.42**	-0.01
CMS_vhbb_bTagWeightcErr1_pt2_eta1	**-0.27, 0.45**	**-0.25, 0.46**	+0.01
CMS_vhbb_bTagWeightcErr1_pt3_eta1	**-0.32, 0.47**	-0.26, 0.51	+0.04
CMS_vhbb_eff_e_MVAID_Zll_13TeV	**-0.50, 0.20**	**-0.50, 0.20**	-0.00
CMS_vhbb_eff_e_Wln_13TeV	**+0.52, 0.20**	**+0.51, 0.20**	-0.02
CMS_vhbb_puWeight	**+0.05, 0.16**	**+0.06, 0.16**	+0.00
CMS_vhbb_res_j_13TeV	**-0.72, 0.32**	**-0.76, 0.32**	-0.02
CMS_vhbb_scale_j_AbsoluteMPFBias_13TeV	**-0.10, 0.34**	**-0.09, 0.34**	+0.00
CMS_vhbb_scale_j_AbsoluteScale_13TeV	**+0.00, 0.33**	**-0.01, 0.34**	-0.01
CMS_vhbb_scale_j_FlavorQCD_13TeV	**+0.89, 0.26**	**+0.82, 0.26**	-0.08
CMS_vhbb_scale_j_Fragmentation_13TeV	**+0.07, 0.34**	**+0.07, 0.33**	-0.00
CMS_vhbb_scale_j_PileUpDataMC_13TeV	**-0.01, 0.32**	**-0.02, 0.33**	-0.01
CMS_vhbb_scale_j_PileUpPtBB_13TeV	**+0.02, 0.33**	**+0.04, 0.35**	+0.01
CMS_vhbb_scale_j_RelativeFSR_13TeV	**+0.02, 0.30**	**+0.03, 0.30**	+0.01
CMS_vhbb_scale_j_RelativeJEREC1_13TeV	**-0.02, 0.36**	**-0.01, 0.36**	+0.00
CMS_vhbb_scale_j_RelativePtBB_13TeV	**+0.04, 0.34**	**+0.03, 0.35**	-0.02
CMS_vhbb_scale_j_RelativeStatEC_13TeV	**-0.08, 0.39**	**-0.06, 0.39**	+0.01
CMS_vhbb_scale_j_RelativeStatFSR_13TeV	**+0.07, 0.36**	**+0.08, 0.36**	+0.01
CMS_vhbb_scale_j_SinglePionECAL_13TeV	-0.02, 0.50	**-0.02, 0.48**	-0.00
CMS_vhbb_scale_j_SinglePionHCAL_13TeV	**+0.02, 0.45**	**+0.02, 0.44**	-0.00
CMS_vhbb_scale_j_TimePtEta_13TeV	**-0.04, 0.29**	**-0.04, 0.29**	-0.00
CMS_vhbb_stats_bin5_Zj0b_ZuuHighPt_13TeV	**+2.04, 0.75**	**+2.01, 0.75**	-0.01

Figure 9.5: List of the 34 overconstrained nuisance parameters in the final binned-likelihood fit, with an absolute pull value above 2 and/or a ratio below 0.5. The first column lists the name of the nuisances. The second and third column correspond to the fit with and without the signal contribution, respectively. The values in those two columns are organized as: (pull, constrain ratio). The fourth column shows the linear correlation coefficient between the given nuisance parameter and the signal strength μ.

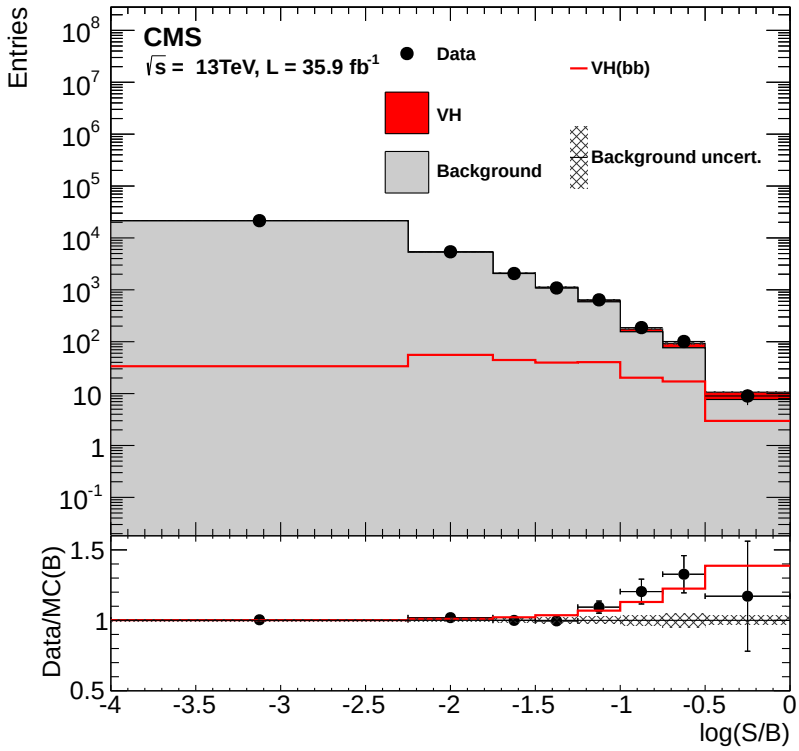


Figure 9.6: Combination of all channels into a single event BDT distribution. Events are sorted in bins of similar expected signal-to-background ratio, as given by the value of the output of their corresponding BDT discriminant (trained with a Higgs boson mass hypothesis of 125 GeV). The bottom plot shows the ratio of the data to the background-only prediction. Looking at the bins with the largest signal-to-background ratio in the ratio plot, an excess in data with respect to the background only distribution is observed. This excess is compatible with the VH(bb) signal simulation of a 125 GeV standard model Higgs boson.

the ratio between data and Monte-Carlo background histograms (black data points) and the red line the ratio between data and the sum of the Monte-Carlo signal + background contributions. Looking at the bins with the largest signal-to-background ratio in the rightmost side of the distribution, an excess in data with respect to the background-only distribution is observed. This excess is compatible with the VH(bb) signal simulation of a $125.09\,\mathrm{GeV}$ standard model Higgs boson.

The local significance of this observed excess of events on 2016 data in the signal extraction fit correspond to a $3.3\,\sigma$ deviation from the background-only hypothesis, for a Higgs boson mass of $m_H = 125.9\,\mathrm{GeV}$. This excess is consistent with the standard model prediction for a Higgs boson production with signal strength μ of

$$\mu = 1.19^{+0.21}_{-0.20}(\mathrm{stat.})^{+0.34}_{-0.32}(\mathrm{syst.}).$$

Channel	Expected significance	Observed significance
Z(vv)H(bb)	1.5	0.0
W(lv)H(bb)	1.5	3.2
Z(ll)H(bb)	1.8	3.1
Combined	2.8	3.3

Table 9.3: The expected and observed significances for the VH(bb) processes, with a Higgs boson mass of $m_H = 125.09$ GeV, for each individual channel as well as for the combination of all three channels.

The "stat." refers to the statistical uncertainty and "syst." to all the other systematic uncertainties. The impact of each individual source of systematic uncertainty, listed in section 8.3.2, is discussed below. The postfit expected significance is $2.8\,\sigma$ (the observed and expected significance definition are given in section 6.1). Together with this result, the lists of the expected and observed significances from the individual fits of each channel are listed in Table 9.3. The $0\,\sigma$ for the observed significance in the Z(vv)H(bb) channel is due to no signal excess above the background-only hypothesis, corresponding to a signal strength of $\mu = 0 \pm 0.5$. For the Z(ll)H(bb) analysis, the observed and expected significance is $3.2\,\sigma$ and $1.5\,\sigma$, respectively, showing the first evidence for the Z(ll)H(bb) process observed at CMS.

The signal strength μ is shown in the lower portion of Figure 9.7 for each individual channel. The observed signal strengths of the three channels are consistent with the combined best fit signal strength with a probability of 5%. The upper portion of Figure 9.7 shows the signal strengths for the ZH(bb) and WH(bb) processes separately. The contributions of the signal processes to the analysis channel are predicted by Monte-Carlo simulations. The ZH(bb) includes the qqZH and ggZH processes contributing to the Z(ll)H(bb) and Z(vv)H(bb) channel, and less than 3% in the W(lv)H(bb) channel, when one of the lepton is out of the detector acceptance. The WH(bb) measurement is driven by the W(lv)H(bb) channel and has a 15% contribution in the Z(vv)H(bb) channel.

The importance of each systematic uncertainty is estimated by its impact on the expected signal strength uncertainty, listed in Table 9.4. The first and second column list the sources and treatments of the systematic uncertainties, which can be implemented as a normalization (*norm.*) or template (*temp.*) uncertainties. The third column shows the uncertainty in % on μ for each source of systematic uncertainty when only that source is considered. The last column shows the percentage decrease on the uncertainty of μ when removing the considered systematic while applying all the others. This takes into account correlations between the sources. The uncertainties associated to the Monte-Carlo normalization scale factors have the largest impact on the

Source	Type	Individual contribution the to μ uncertainty (%)	Effect of removal to the μ uncertainty (%)
Statistical			
Scale factors ($t\bar{t}$, V+jets)	norm.	9.4	3.5
MC sample size			
Size of simulated samples	temp.	8.1	3.1
Theory			
Simulated samples modeling	temp.	4.1	2.9
Signal cross sections	norm.	5.3	1.1
Cross section uncertainties (single-top, VV)	norm.	4.7	1.1
Experimental			
b tagging efficiency	temp.	7.9	1.8
Jet energy scale	temp.	4.2	1.8
Jet energy resolution	temp.	5.6	0.9
b tagging mistage rate	temp.	4.6	0.9
Integrated luminosity	norm.	2.2	0.9
Missing transverse energy	temp.	1.3	0.2
Lepton efficiency and trigger	temp.	1.9	0.1

Table 9.4: List of the impact of the various sources of systematic uncertainties on the signal strength μ. The first column lists the source of systematic uncertainties which are described in section 8.3.2. The second column shows if it is treated as a normalization (norm.) or template (temp.) uncertainty. The third column shows the uncertainty on μ when only that particular source of systematic uncertainty is considered. The fourth column shows the percentage decrease of the uncertainty on μ when removing that specific source of systematic uncertainty while applying all other sources. The approach used in this last column accounts for correlations between the various sources of systematic uncertainties. The dominant sources of systematic uncertainties are: the size of the dataset, which propagates into the uncertainty of the Monte-Carlo normalization scale factors, the statistical size of the Monte-Carlo samples, the shape correction of the Monte-Carlo simulations, the b-tagging efficiency corrections and the JEC and JER uncertainties.

fitted signal strength uncertainty. The second largest effect comes from the size of the simulated samples, particularly due to the size of the Z + 1 b and Z + 2b background processes. The addition of b-enriched and b gen. filter Z + jets samples (see Table 6.4) considerably increases the statistical power of those processes. The third largest group includes the shape correction of the Monte-Carlo simulated background processes and the next one the b-tagging efficiency and JEC and JER uncertainties.

The fitted values and uncertainties of the background normalization scale factors for the Z(ll)H(bb) channel are shown in Table 9.5. The determination of the scale factors are mostly driven by control regions, designed to optimize the discrimination and purity of the main background processes. To validate the extrapolation to signal regions, a new set of normalization scale factors are extracted from a binned-likelihood fit on the control regions $\text{CMVAv2}_{\text{min}}$ shapes, without including the signal region's BDT output score distributions. Both sets of normalization scale factors are compatible within their uncertainties. In a ideal scenario, that is when the background Monte-Carlo simulated samples perfectly model the data distribution before the fit, the scale

Process	low $p_T(Z)$	high $p_T(Z)$
Z + light jets	1.01 ± 0.06	1.02 ± 0.06
Z + 1 b jet	0.98 ± 0.06	1.02 ± 0.11
Z + 2 b jets	1.09 ± 0.07	1.28 ± 0.09
$t\bar{t}$	1.00 ± 0.03	1.04 ± 0.05

Table 9.5: Background normalization scale factors for the main background processes in the Z(ll)H(bb) channel. The same scale factors are used in the Z(ee)H(bb) and Z($\mu\mu$)H(bb) sub-channels. Those values are estimated from the final binned-likelihood fit, performed on all three VH(bb) channels.

factors would have a value a 1.00.

The dijet invariant mass distribution of the VH(bb) and diboson processes can be found in Figure 9.8, where all other processes have been subtracted from the data. It includes all the events from the signal region distributions used in the final VH(bb) binned-likelihood fit. The postfit nuisances and Monte-Carlo normalization scale factors are included in this distribution. In order to better visualize the contributions from the VH(bb) signal, all the events are reweighed by a S/(S+B) weight, where S and B are the numbers of expected post-fit signal and background events, respectively, in the bin of the BDT output score distribution in which each event is contained. The blue lines correspond to the uncertainty from the subtracted Monte-Carlo processes. An excess of data with respect to the subtracted background processes can be observed in the peak, which is well modeled by the VH(bb) + diboson processes. This excess is consistent with a standard model Higgs boson having a mass of $125.09\,\mathrm{GeV}$.

9.1.4 Combination with Run 1

The results presented above are combined with previous VH(bb) searches performed on data recorded by the CMS experiment during the Run 1 data-taking period of the LHC, using proton-proton collisions at a center-of-mass energy $\sqrt{s}$ of 7 and $8\,\mathrm{TeV}$. The size of each dataset corresponds to an integrated luminosity of 5.1 and $18.9\,\mathrm{fb}^{-1}$, respectively [78, 79]. The observed (expected) significance of the combination is $3.8\,\sigma$ ($3.8\,\sigma$) for a Higgs boson with a mass of $125.09\,\mathrm{GeV}$. The signal strength of this excess is $\mu = 1.06^{+0.31}_{-0.29}$. The cross-section uncertainties derived from theory are assumed to be correlated in the combination. All other systematic uncertainties are treated as uncorrelated due to the different conditions between Run 1 and Run 2. The combination results are summarized in Table 9.6.

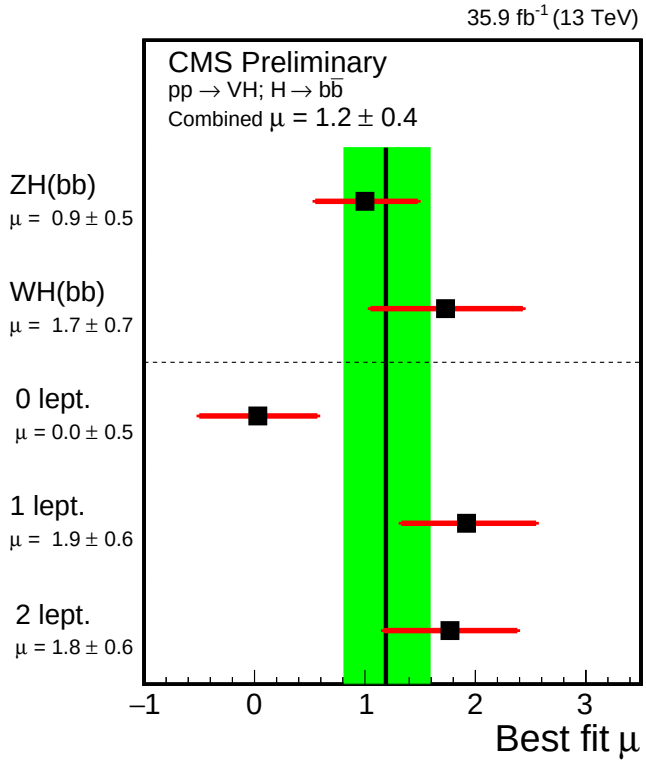


Figure 9.7: Signal strengths estimated from the final binned-likelihood fit. Two binned-likelihood fits have been performed, one for the upper part and one for the lower part of this figure. The upper part shows the signal strength μ estimated separately for the ZH(bb) and WH(bb) processes, where the signal strength of each process is treated independently during the fit. The lower part shows the signal strength separately in the three VH(bb) channels, where the signal strength of each channel is varied independently during the fit. The corresponding combined signal strength from all the channels is $\mu = 1.2 \pm 0.4$. The green band corresponds to the uncertainty of the combined signal strength. The black squares and red lines correspond to central value and uncertainty of the signal strength, respectively, in each category mentioned above.

Dataset	Integrated luminosity	Center-of-mass energy	Observed significance	Observed significance	Signal strength
Run 1	$5.10\,\text{fb}^{-1}$ $18.9\,\text{fb}^{-1}$	7 TeV 8 TeV	2.5	2.1	$0.89^{+0.44}_{-0.42}$
Run 2 (2016 dataset)	$35.9\,\text{fb}^{-1}$	13 TeV	2.8	3.3	$1.19^{+0.40}_{-0.38}$

Table 9.6: Integrated luminosity, center-of-mass energy, expected and observed significance and the observed signal strengths for the VH(bb) processes for Run 1 data, Run 2 (2016) data and for the combination of both datasets. The Higgs boson mass used for the VH(bb) process simulation correspond to 125.9 GeV. Significance values are given in numbers of standard deviations.

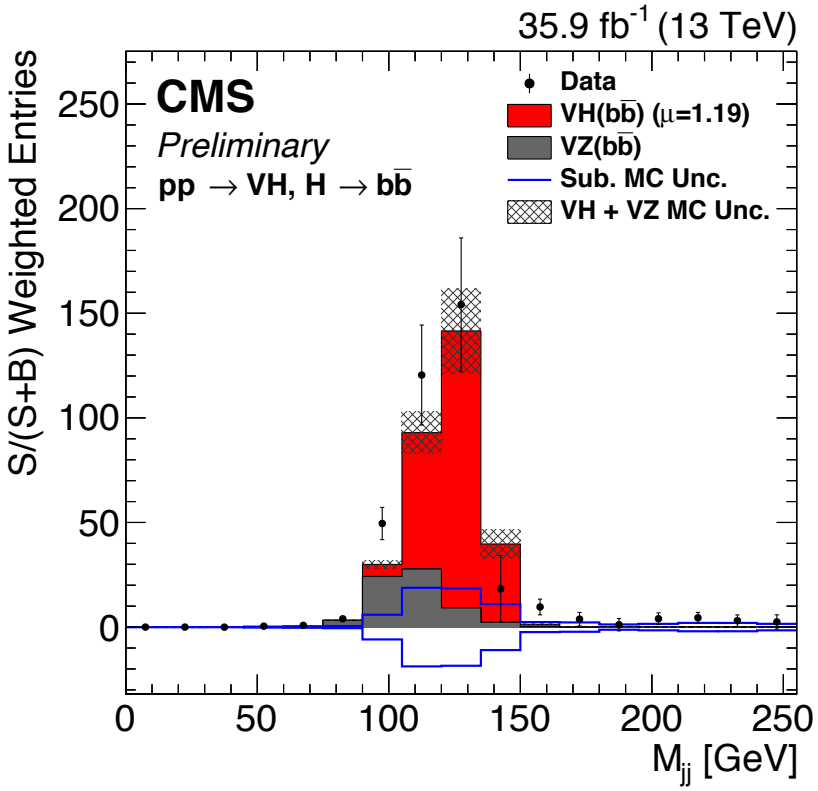


Figure 9.8: Dijet invariant mass distribution after subtracting all Monte-Carlo processes except the VH(bb) signal and the VZ($b\bar{b}$) diboson processes. All events are reweighted by a S/(S+B) weight, where S and B are the numbers of expected signal and postfit background events, respectively, in the bin of the BDT output score distribution in which each event is contained. The blue lines correspond to the uncertainty from the subtracted Monte-Carlo processes. The stacked Monte-Carlo histogram includes the VH(bb) and VZ($b\bar{b}$) processes. An excess is observed in the data peak, well modeled by the VH(bb) + diboson processes.

Channel	Expected signal strength	Observed signal strength	Expected significance	Observed significance
Z(vv)Z(bb)	1.00 ± 0.33	0.57 ± 0.32	3.1	2.0
W(lv)Z(bb)	1.00 ± 0.38	1.67 ± 0.47	2.6	3.7
Z(ll)Z(bb)	1.00 ± 0.31	1.33 ± 0.34	3.2	4.5
Combined	1.00 ± 0.22	1.02 ± 0.22	4.9	5.0

Table 9.7: Results of the diboson cross-check analysis, including expected and observed significances and the signal strengths. Significance values are given in numbers of standard deviations.

9.2 Diboson analysis

The final binned-likelihood fit in the cross-check diboson analysis is performed the same way as in the VH(bb) analysis. The main differences are a modification of the dijet invariant mass selection in the signal and control regions and a retraining of the BDT in the signal region.

The postfit BDT output score distributions in the signal regions for the Z(ll)Z(bb) channel can be found in Figure 9.9. The layout and composition of the four figures is similar to what is shown in Figure 9.1 and described in its caption. The diboson signal process, labeled as VVHF in the legend (diboson + 2 jets), is in light grey and populates the rightmost bins of the distributions, as the signal-background separation increases with the BDT output score. As in the Z(ll)H(bb) analysis, it can be seen in the ratio plot that the agreement between data and Monte-Carlo simulation is good for all the BDT distributions. The fitted control region $\text{CMVAv2}_{\text{min}}$ distributions can be found in Appendix C.

As for the VH(bb) analysis, the BDT distributions adjusted by the fit from all the channels are combined into a single distribution in Figure 9.10. An excess in data with respect to the background only distribution is observed, compatible with the diboson signal simulation.

The observed excess of events in the diboson process extraction performed in the three analysis channels has an observed significance corresponding to $5.0\,\sigma$ deviation from the background-only hypothesis. The expected significance is $4.9\,\sigma$. The corresponding diboson signal strength is $\mu_{VV} = 1.02^{+0.22}_{-0.23}$, and is compatible with the standard model predictions for the diboson signal. Those results and the observed, expected significance and signal strength from the individual fits of each channel are listed in Table 9.7.

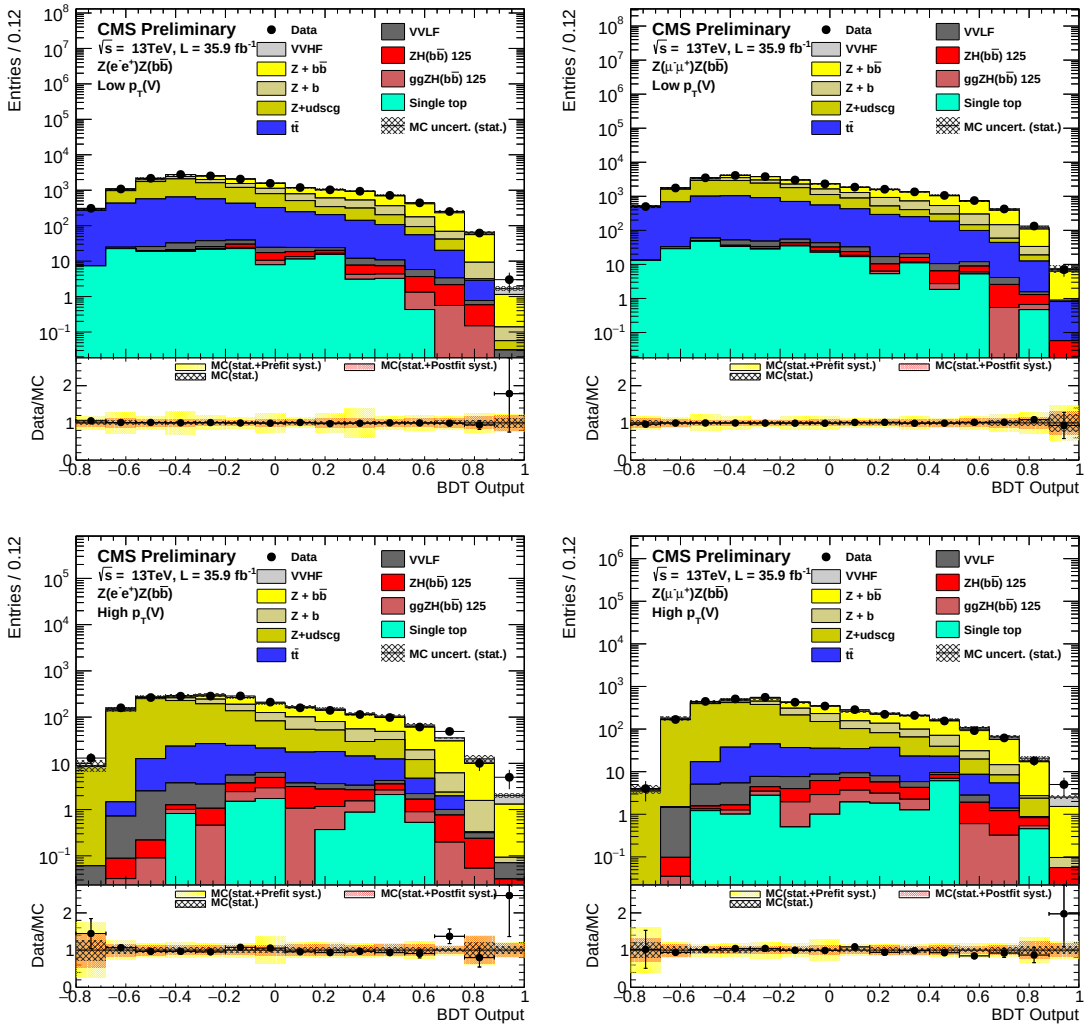


Figure 9.9: Postfit BDT output score distributions in the four signal region in the Z(ll)Z(bb) channel. The signal regions are, from left to right, top to bottom: Z(ee)Z(bb) low-$p_T(Z)$, Z($\mu\mu$)Z(bb) low-$p_T(Z)$, Z(ee)Z(bb) high-$p_T(Z)$, Z($\mu\mu$)Z(bb) high-$p_T(Z)$. The layout and composition of the four figures is similar to what is shown in Figure 9.1 and is described in the legend.

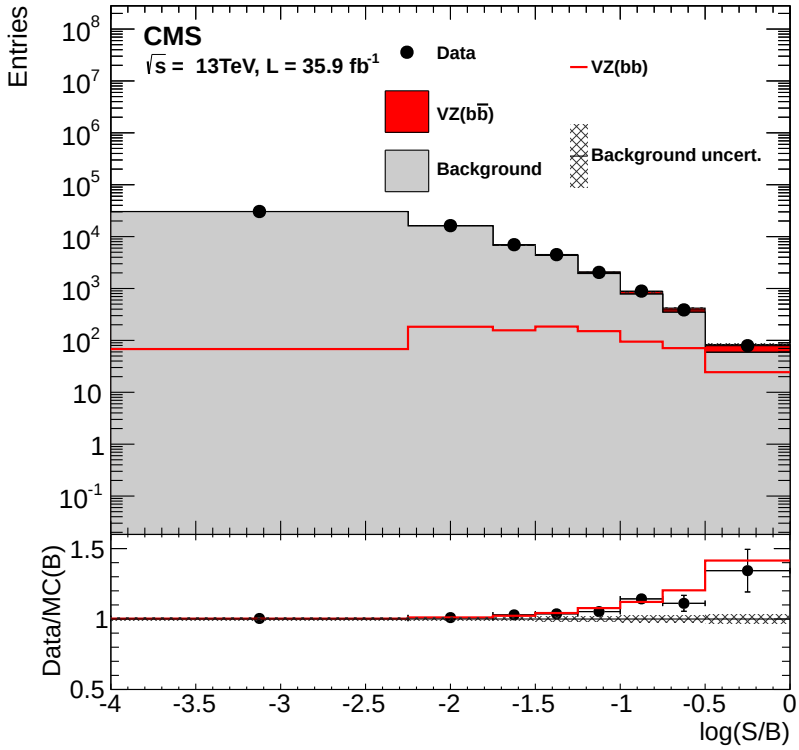


Figure 9.10: Combination of all channels in the diboson search into a single event BDT distribution. Events are sorted in bins of similar expected signal-to-background ratio, as given by the value of the output of their corresponding BDT discriminant. The bottom inset shows the ratio of the data to the predicted background, with a red line overlaying the expected SM contribution from the disboson with $Z \to b\bar{b}$.

9.3 Update with the 2017 dataset

A new version of the CMS VH(bb) analysis was performed on a data sample from the 2017 data-taking period, corresponding to $41.3\,\mathrm{fb}^{-1}$ at a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$.

New techniques have been included to improve the analysis sensitivity. Also, some of the methods described in section 6.4 have been updated with new technologies. The main improvements are listed below. Additional information can be found in [80].

- **Deep neural network:** The signal-background classification in the signal region was performed with a Deep Neural Network (DNN), implemented in TensorFlow [81] and Keras [82].
- **Final state radiation recovery:** Final state radiations can be emitted from the final state jets. Additional jets close to the b jet candidate are recovered to improve the estimation of the dijet invariant mass.
- **Jet energy regression:** A new regression technique making use of a DNN architecture has been used for the 2017 analysis. It improves the dijet invariant mass resolution by 18.2%, versus 15% for the 2016 regression [83].
- **Kinematic Fit:** A simultaneous fit is performed in the Z(ll)H(bb) channel to estimate the energy scale of the two b jet candidates. It uses as constraints the Z boson mass of the two lepton candidates and the sum of the transverse momentum of the leptons and jets in the event.
- **b-tagger algorithm:** A new b-tagging algorithm, DeepCSV [67], is used for the b jet identification. It improves the b jet identification efficiency by a gain of 4% for the tight working point with respect to the CMVAv2 algorithm.
- **DNN multiclassifier:** A multi-output DNN, trained to distinguish among the background components, is used for the heavy flavor control region in the W(lv)H(bb) and Z(vv)H(bb) channels to increase the shape discrimination.

The results of this measurement, as well as its combination with other CMS searches for a $H \to b\bar{b}$ decay, are summarized below.

9.3.1 Results

All the results reported here are obtained for a Higgs boson mass of $m_H = 125.09\,\mathrm{GeV}$. For the VH(bb) analysis on the 2017 dataset, the observed significance is $3.3\,\sigma$ above the background-only hypothesis, while $3.1\,\sigma$ is expected from a Standard Model Higgs

boson. The corresponding measured signal strength is $\mu = 1.08 \pm 0.34$, where the uncertainty is a combination of statistical and systematic components.

Those results are combined with the 2016 VH(bb) analysis presented in this dissertation. All systematic uncertainties are assumed to be uncorrelated in the combination fit, except for theory uncertainties and the dominant JEC uncertainties, which are assumed to be fully correlated. The reason for the JEC uncertainties correlation is that their evaluation is performed with the same methodology and the relevant subdetector part (pixel detector, tracker, ECal and HCal) are identical between the two years. The significance of this Run 2 (2016 and 2017 dataset) combination yields an observed signal significance of $4.4\,\sigma$, where $4.2\,\sigma$ is expected, and a signal strength $\mu = 1.06 \pm 0.20\,\text{stat} \pm 0.17\,\text{syst}$.

The VH(bb) results from Run 2 are combined with the previous VH(bb) searches performed by the CMS Collaboration during Run 1, mentioned in section 9.1.4. Systematic uncertainties in this fit are assumed to be uncorrelated, except for the cross-section uncertainties derived from theory. The combination yields an observed signal significance of $4.8\,\sigma$, where $4.9\,\sigma$ are expected. The measured signal strength is $\mu = 1.01 \pm 0.22[0.17(\text{stat}) \pm 0.09(\text{exp}) \pm 0.06(\text{MC}) \pm 0.08(\text{theo})]$, where the decomposition of the total uncertainty in the bracket follows the description in Table 9.4. When merging the various systematic uncertainties into one category (syst), the signal strength is $1.01 \pm 0.17(\text{stat}) \pm 0.14(\text{syst})$. The statistical uncertainty is therefore still the dominant source of uncertainty of the VH(bb) analysis, primarily affecting the uncertainties of the background normalization scale factors. The dominant sources of systematic uncertainties are the size of the Monte-Carlo samples, the b-tagging efficiencies and the simulated samples modeling.

The left plot in Figure 9.11 shows the distribution of the signal region events in all channels for the Run 1 and the Run 2 combination, similarly to Figure 9.6. An excess of events with respect to the background-only hypothesis is visible in the lower ratio plot and follows the distribution of the red VH(bb) signal. The right plot on Figure 9.11 summarizes the signal strengths of the VH(bb) measurement for the different data sets and the combination. The measurement of the signal strength of the VH(bb) analysis on the Run 1, 2016, 2017 dataset and the Run 1 + Run 2 combination are all compatible with a Standard Model Higgs boson with a mass of $m_H = 125.09\,\text{GeV}$.

A combination of CMS measurements of the $H \to b\bar{b}$ decay is performed, including dedicated analysis for the other Higgs boson production mechanism described in section 5.2: gluon fusion (*ggF*), vector boson fusion (*VBF*), associated production with top quarks (*ttH*) and the VH(bb) search, separating the WH and ZH vector boson contributions. These analysis use data collected at 7, 8 or $13\,\text{TeV}$, depending on the process. The energy scale uncertainties are treated as correlated between processes at the same collision energy, while the theory uncertainties are correlated between all processes and datasets. The observed (expected) signal significance is 5.6 ($5.5\,\sigma$), and the meas-

Dataset	Expected	Observed	Signal strength
Run 1 (2011 + 2012)			
Z($\nu\nu$)H(bb)	1.3	1.3	1.0 ± 0.8
W($l\nu$)H(bb)	1.3	1.4	1.1 ± 0.9
Z(ll)H(bb)	1.1	1.8	0.8 ± 1.0
Combined	2.1	2.1	1.0 ± 0.5
2016			
Z($\nu\nu$)H(bb)	1.5	0.0	0.0 ± 0.5
W($l\nu$)H(bb)	1.5	3.2	1.9 ± 0.6
Z(ll)H(bb)	1.8	3.1	1.8 ± 0.6
Combined	2.8	3.3	1.2 ± 0.4
2017			
Z($\nu\nu$)H(bb)	1.9	1.3	0.73 ± 0.65
W($l\nu$)H(bb)	1.8	2.6	1.32 ± 0.55
Z(ll)H(bb)	1.9	1.9	1.05 ± 0.59
Combined	3.1	3.3	1.08 ± 0.34
Run 2 (2016 + 2017)	4.2	4.4	1.06 ± 0.26
Run 1 + Run 2	4.9	4.8	1.01 ± 0.23
$H \to b\bar{b}$ combination	5.6	5.5	$1.04 \pm 0.14(\text{stat.}) \pm 0.14(\text{sys.})$

Table 9.8: Expected and observed significances, in number of standard deviations, and observed signal strengths for the VH(bb) analysis. Results are shown separately for Run 1 (2011 + 2012 dataset), 2016 and 2017 datasets, Run 2 combination (2016 and 2017 dataset), Run 1 and Run 2 combination and the combination of the $H \to b\bar{b}$ searches at CMS. For the analysis on the Run 1, 2016 and 2017 dataset, results are shown separately for each of the three channels and for a combined simultaneous fit of channels. All results are obtained for a Higgs boson with a mass of 125.09 GeV combining statistical and systematic uncertainties.

ured signal strength is $\mu = 1.04 \pm 0.20$. In addition of the overall signal strength for the $H \to b\bar{b}$ decay, the signal strengths for the individual production processes are also determined in this combination. All results are summarized in Figure 9.12.

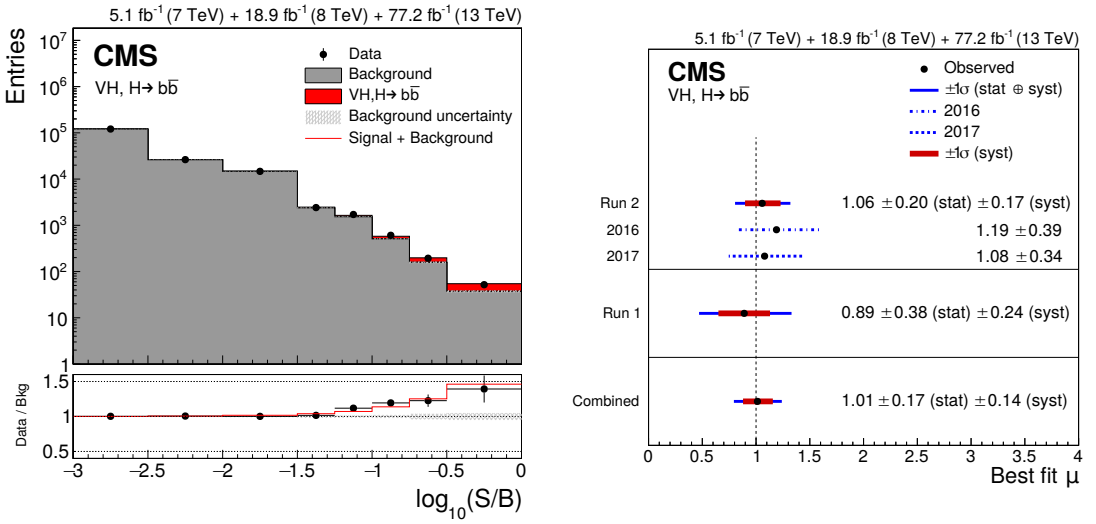


Figure 9.11: Left figure: distributions of signal, background, and data event yields sorted into bins of similar signal-to-background ratio, as given by the result of the fit to their corresponding multivariate discriminant. All events in the VH(bb) signal regions of the combined Run 1 and Run 2 data sets are included. The layout is similar to Figure 9.6 and is described in the legend.
Right figure: best-fit value of the signal strength μ at $m_H = 125.09\,\text{GeV}$ for the fit of all VH(bb) channels in the Run 1 and Run 2 datasets. The individual results of the 2016 and 2017 measurements, the Run 2 combination, and the Run 1 result are also shown. Horizontal error bars indicate the $1\,\sigma$ systematic (red) and $1\,\sigma$ total (blue) uncertainties, and the vertical dashed line indicates the Standard Model prediction.

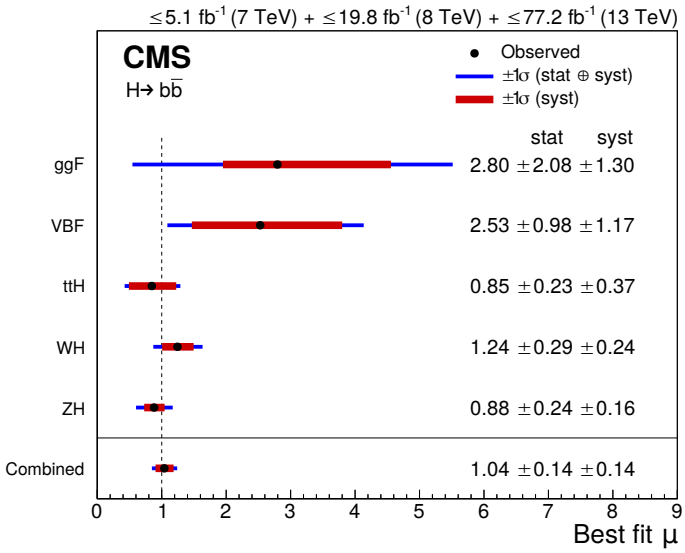


Figure 9.12: Best-fit value of the $H \to b\bar{b}$ signal strength with its 1σ systematic (red) and total (blue) uncertainties for the five individual production modes considered, as well as the overall combined result. The vertical dashed line indicates the standard model expectation. All results are extracted from a single fit combining all input analyses, with $m_H = 125.09\,\text{GeV}$.

Part III

Boosted Studies

10
Boosted W(lv)H(bb) Analysis

The vector and Higgs boson transverse momentum tend to be harder for the VH(bb) signal than for background processes. It is observed in the Z(ll)H(bb) channel that most of the analysis sensitivity is driven by the high-$p_T(Z)$ category, the low-$p_T(Z)$ category contributing to $\sim 10\%$ of the overall sensitivity (see section 8.2). This encourages to study topologies with a *boosted* (i.e. large transverse momentum) vector and Higgs boson in the laboratory frame. This chapter gives an overview of a VH(bb) analysis conducted in the W(lv)H(bb) channel and performed on the 2016 dataset in a boosted topology, where both the vector and Higgs boson are required to have a transverse momentum greater than $250\,\mathrm{GeV}$. This analysis will be referred to as the *boosted W(lv)H(bb) analysis*.

The boosted W(lv)H(bb) analysis shares some similarities with the W(lv)H(bb) channel included in the final binned-likelihood fit of the VH(bb) measurement, described in Chapter 9. This is illustrated in Figure 10.1, which separates possible VH(bb) analysis depending on the nature of the vector boson decay and the approach used to reconstruct the Higgs boson candidate. The upper part represents a VH(bb) process at particle level, without considering detector-related reconstruction of the vector and Higgs boson. The lower part lists various approaches of conducting the VH(bb) analysis, depending on the channel related to the vector boson decay and the reconstruction method used for the Higgs candidate. The rows differentiate between the Z(ll)H(bb) and the W(lv)H(bb) channel corresponding to the upper and the lower row, respectively.

The columns differentiate two approaches of reconstructing the b hadron pair from the Higgs boson decay. In the left column, a dijet system composed of two particle-flow AK04 jets is employed to reconstruct the Higgs boson candidate. The analysis relying on this approach will be referred to as *resolved analysis*, as it is mostly suited

for events with two angularly separated b jets[1]. The three analysis channels of the VH(bb) analysis mentioned in Section 5.3.0.1 are examples of resolved analysis and will be referred as such during the rest of this thesis to distinguish them from *boosted analysis*, described below.

In the right column, a single particle-flow AK08 jet, referred to as a *fat jet* due to its large cone size, is employed to reconstruct the decay from the Higgs boson candidate. This approach is suited for studies of boosted topologies where the transverse momentum of the Higgs boson is large and the two AK04 b jets would start to overlap and no longer be resolved. The analysis employing that approach will be referred to as boosted analysis and are performed in a phase-space where the vector and Higgs boson transverse momentum is larger than $250\,\mathrm{GeV}$.

Each cell of the lower table in Figure 10.1 corresponds to a particular analysis represented by a sketch of the event reconstruction and referred to by a number. They are described below.

1 is the VH(bb) resolved analysis in the Z(ll)H(bb) channel, or *resolved Z(ll)H(bb) analysis*, documented in Chapters 6 to 9.

2 is the VH(bb) resolved analysis in the W(lv)H(bb) channel, or *resolved W(lv)H(bb) analysis*. It is combined with the Z(ll)H(bb) and Z(vv)H(bb) channels in the final binned-likelihood fit to extract the results of the VH(bb) analysis as described in Chapter 9.

It differs from the resolved Z(ll)H(bb) analysis due to the nature and decay of the W vector boson. It however uses the same approach in the reconstruction of the jet candidates attributed to the dijet mass system. An overview of this analysis is given in Section 10.1.

3 is the *VH(bb) boosted analysis* in the Z(ll)H(bb) channel. It is shown for illustration purposes only.

4 is the VH(bb) boosted analysis in the W(lv)H(bb) channel, or boosted W(lv)H(bb) analysis, which is the object of the studies presented in the rest of this dissertation.

An overview of this analysis is given in Section 10.2. The analysis strategy is described in Section 10.3. If differs from the resolved W(lv)H(bb) analysis in terms of the reconstruction of the jet candidates attributed to the dijet mass system. It

[1]We consider the two b jets as angularly separated when $\Delta R_{\mathrm{j1,j2}} > 0.4$, where $\Delta R_{\mathrm{j1,j2}}$ is the angular distance between the two b jet axis.

Channel	L1 Seed	HLT Paths
W(lμ)H(bb)	SingleMu20	HLT_IsoMu24 OR HLT_IsoTkMu24
W(le)H(bb)	SingleIsoEG22er OR SingleEG25	HLT_Ele27_WPTight_Gsf

Table 10.1: List of L1 and HLT triggers used for the resolved W(lv)H(bb) analysis 2016 dataset.

however uses the same approach in the reconstruction of the vector boson.

As the resolved W(lv)H(bb) channel is very similar to the boosted W(lv)H(bb) channel, as mention in the previous bullets, the former analysis is reviewed in the next section. A comparison between both analysis in terms of sensitivity is shown in Chapter 11.

10.1 Resolved W(lv)H(bb) analysis

This section reviews the resolved W(lv)H(bb) analysis performed on the 2016 dataset, which is included in the final binned-likelihood fit of the VH(bb) analysis as described in Chapter 9.

10.1.1 Data Taking

The data taking conditions, described Section 6.2 for the resolved Z(ll)H(bb) analysis, are the same for the boosted W(lv)H(bb) analysis. The dataset used in this analysis has been recorded during the year 2016 by the CMS detector at a center-of-mass energy of $13\,\mathrm{TeV}$, with a minimum bunch spacing time of $25\,\mathrm{ns}$.

The resolved W(lv)H(bb) channel is separated in two sub-channels, *W(ev)H(bb)* and *W(μv)H(bb)*, depending on the decay products of the vector boson. The W(ev)H(bb) sub-channel uses the *SingleElectron* dataset and W(μv)H(bb) sub-channel the *SingleMuon* dataset. The trigger conditions related to those datasets are discussed below.

10.1.1.1 High level trigger definitions

The resolved W(lv)H(bb) channel uses a mix of single-lepton triggers. The triggers corresponding to each dataset are listed in Table 10.1.

The single-muon HLT has a p_T threshold of $24\,\mathrm{GeV}$ and an additional isolation requirement. The single-electron HLT has a p_T threshold of $27\,\mathrm{GeV}$ and additional isolation and identification selections to improves the purity of signal electrons while keeping a low p_T threshold.

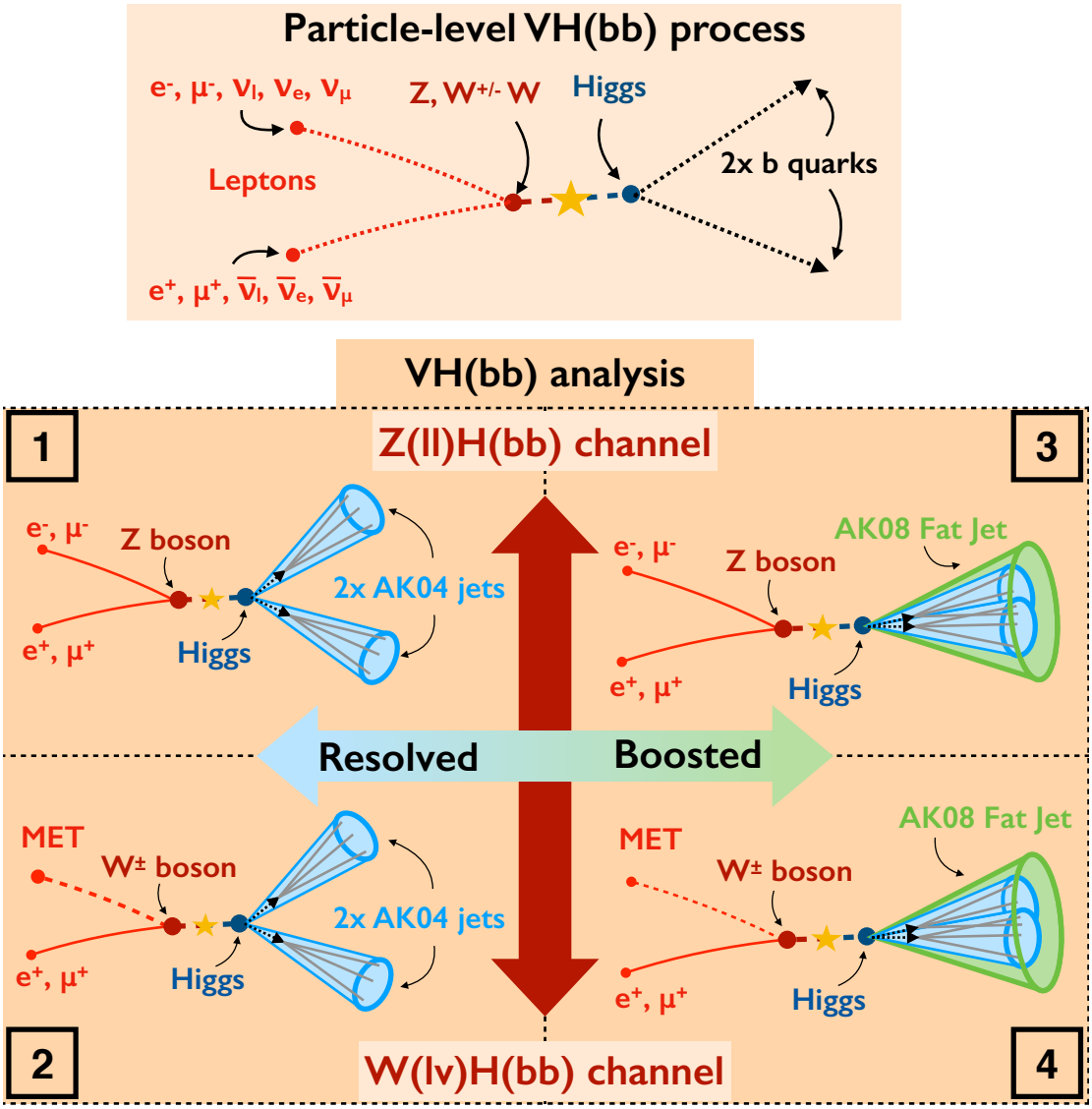


Figure 10.1: Representation of resolved and boosted analysis in the Z(ll)H(bb) and W(lν)H(bb) channels. The upper part is a general representation of the VH(bb) process at particle level. The table in the lower part represents the event reconstruction in the, from left to right and top to bottom, resolved Z(ll)H(bb) analysis, boosted Z(ll)H(bb) analysis, resolved W($l\nu$)H(bb) analysis and boosted W($l\nu$)H(bb) analysis. The Higgs candidate is reconstructed with two AK04 jets in the left column and a single AK08 fat jet in the right column.

Process	Generator	Order	Cross-section
W^+	POWHEG (v2) [18–20] + MinLO [51, 51] + PYTHIA 8	NLO, rescaled to NLO+NLL QCD	$0.840 \times 0.108535 \times 0.5824$
W^-	"	"	$0.533 \times 0.108535 \times 0.5824$

Table 10.2: List of Monte-Carlo simulations for signal processes in the W(lv)H(bb) channel. The cross-sections are denoted as *WH production* $\times\ W \rightarrow l\nu \times H \rightarrow b\bar{b}$. The cross-sections values are taken from [34].

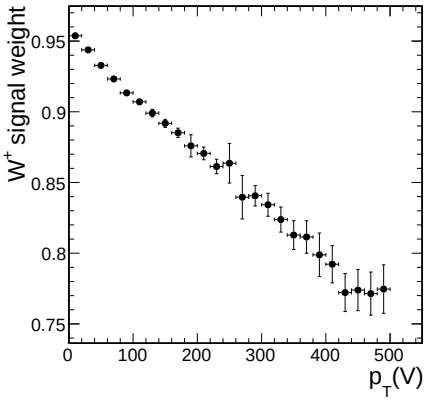


Figure 10.2: Multiplicative signal NLO EW cross-section correction applied in bins of W boson p_T distribution.

10.1.2 Monte-Carlo simulation samples

The production of Monte-Carlo simulated signal and background events for the resolved W(lv)H(bb) analysis is performed with the same tools as described in Section 6.3.

10.1.2.1 Signal simulation

The signal samples for the resolved W(lv)H(bb) channel are simulated for a Higgs boson with a mass of $125\,\mathrm{GeV}$ and are listed in Table 10.2. There are two contributions depending on the charge of the W boson, W^+ and W^-. The cross-sections are denoted as $\sigma(pp \rightarrow WH) \times BR(W \rightarrow l\nu) \times BR(H \rightarrow b\bar{b})$, where *BR*(*decay*) stands for the branching ratio of the corresponding decay.

The signal cross-section is rescaled to NNLO QCD accuracy. This rescaling factorizes with additional NLO EW correction applied differentially in the vector boson p_T [34]. The corresponding relative weight is displayed in Figure 10.2.

10.1.2.2 Background simulation

The list of Monte-Carlo simulated background samples includes the samples already mentioned in Section 6.3.2 for the resolved Z(ll)H(bb) channel, with the addition of

W + jets, QCD multijet and new diboson processes (with one W boson in addition to the $Z \to b\bar{b}$ decay), listed in Table 10.3. The production of W + jets samples is split in different phase-space to optimize the computation efficiency, similarly to the DY + jets sample. The stiching procedure described in Section 6.3.2 is performed to account for overlapping events among the W + jets samples.

10.1.3 Physics Objects

Most of the physics objects used in the resolved Z(ll)H(bb) analysis are also used in the W(lν)H(bb) analysis. The reconstruction of the primary vertex, missing transverse energy, soft activity and the pileup treatment are the same in both analyses and are described in Section 6.4. The reconstruction of the Higgs boson candidate employs the same method for both analyses, described in Section 6.4.8. Two particle-flow AK04 jets with the highest CMVAv2 score are selected as the b-jet candidates. The reconstruction, pileup mitigation, JEC and JER corrections, b-tagging algorithm and bottom jets identification efficiency corrections are common to both analysis.

As the Z(ll)H(bb) and W(lν)H(bb) signal processes differ in the nature and decay of the vector boson, the only different objects between the two analysis are the vector boson and the electrons and muons, whose reconstruction, selection and related Monte-Carlo simulations corrections for the resolved W(ll)H(bb) channel are described below.

10.1.3.1 Vector boson

In the resolved W(lv)H(bb) analysis, the vector boson candidates are identified by the presence of a lepton and additional missing transverse energy. The transverse momentum $p_T(W)$ and transverse mass M_T of the vector boson are reconstructed as

$$p_T(W) = \sqrt{(\text{MET}_x + p_x^l)^2 + (\text{MET}_y + p_y^l)^2},$$

$$M_T = \sqrt{(\text{MET}_x + p_T^l)^2 - p_T^2(W)},$$

where p^l refers to the momentum of the lepton candidate. The tight identification and isolation selections are requiered on both electrons and muons to remove possible contamination from the QCD multijet processes. It is also required that the azimuthal angle between the MET direction and the lepton is less than 2.0 radians.

Muon candidates are required to have a transverse momentum above $25\,\text{GeV}$ and pass the tight identification and isolation cut described in section 6.4.3.2 and 7.3.2, respectively. The electron candidates must have a transverse momentum above $30\,\text{GeV}$ and pass the tight version of the selection described in 6.4.3.3, corresponding to a signal efficiency of 80%.

Process	Generator	Order	Cross-section
Diboson (WW)	MG5AMC@NLO [47] + FxFx[54] + PYTHIA 8	NLO	118.7
Diboson (WZ)	MG5AMC@NLO [47] + FxFx[54] + PYTHIA 8	NLO	47.13
DY + Jets, Z_{mass} = [10, 50]:	MADGRAPH 5 [55] + MLM[16]+ PYTHIA 8	LO, rescaled to NLO	-
DY + 1 Jet	"	"	725
DY + 2 Jet	"	"	394.5
DY + 3 Jet	"	"	96.47
W + Jets	"	"	
H_T inclusive	"	"	61526.7×1.23
H_T = [100, 200]	"	"	1345×1.23
H_T = [200, 400]	"	"	359.7×1.23
H_T = [400, 600]	"	"	48.91×1.23
H_T = [600, 800]	"	"	12.05×1.23
H_T = [800, 1200]	"	"	5.501×1.23
H_T = [1200, 2500]	"	"	1.329×1.23
H_T = [2500, inf]	"	"	0.03216×1.23
W + Jets, b-enriched	"	"	
p_T(W) = [100, 200]	"	"	6.004×1.23
p_T(W) = [200, inf]	"	"	0.8524×1.23
W + Jets, b gen. filter	"	"	
p_T(Z) = [100, 200]	"	"	26.1×1.23
p_T(Z) = [200, inf]	"	"	3.545×1.23
QCD multijet	MADGRAPH 5 [55] + MLM[16]+ PYTHIA 8	LO	-
H_T = [100, 200]	"	"	27990000
H_T = [200, 300]	"	"	1712000
H_T = [300, 500]	"	"	347700
H_T = [500, 700]	"	"	29400
H_T = [700, 1000]	"	"	6831
H_T = [1000, 1500]	"	"	1207
H_T = [1500, 2000]	"	"	119.9
H_T = [2000, inf]	"	"	25.42

Table 10.3: List of Monte-Carlo simulations for the background processes in the boosted W(lv)H(bb) analysis. The analysis also includes all the samples listed in Table 6.4. The process name is given in the first column. The production of the DY + jets is split in different regions of phase-space, depending on the number of jets, the jet transverse momentum H_T and the W vector boson transverse momentum. The second column describes the corresponding event generator. The third column contains the order of the event generator, as well as an eventual rescaling of the cross-section. The fourth column lists the cross-sections for each process. For the W + jets background, whose cross-section has been rescaled from LO to NLO with a 1.23 k-factor, the cross-section is written as "cross-section $\times$ k-factor". A quote in one of the column (") indicates that this parameter is the same as for the sample above.

10.1.3.2 QCD and electroweak corrections

As for the Z vector boson case in the resolved Z(ll)H(bb) analysis, a difference in shape has been observed between data and the Monte-Carlo simulations, the latter showing a harder W boson transverse momentum spectrum than the data.

This difference in shape is recovered after applying higher order (NLO) QCD and electroweak corrections on the Monte-Carlo sample [68]. The QCD corrections are derived in different bins of H_T. They correspond to a k-factor of 1.31, 1.19, 1.23, 0.94 for the H_T bins $100-200$, $200-400$, $400-600$ and $>600\,\mathrm{GeV}$, respectively. The electroweak correction is a function of the W boson transverse momentum p_T, defined as

$$f(p_T) = -0.83 + 7.94 \cdot (p_T + 877)^{-0.31}, \quad \text{if } p_T > 100\,\mathrm{GeV},$$
$$f(p_T) = 1, \quad \text{if } p_T < 100\,\mathrm{GeV}.$$

10.1.3.3 Transverse momentum corrections

The spectrum of the W boson transverse momentum, $p_T(W)$, is observed to be softer in data than in the Monte-Carlo simulated samples after all the NLO and QCD corrections mentioned in the previous section are applied.

Linear re-weighting functions are derived to correct the $p_T(W)$ slope for $t\bar{t}$, W + udscg, and the combination of W + 2 b and single top processes via a simultaneous fit of the reconstructed $p_T(W)$ in the W(lv)H(bb) control regions to data. The control region selections are defined in section 10.1.4. The input template for the fit in each control region is a sum of the Monte-Carlo predictions for each process corrected by a linear function of the reconstructed $p_T(W)$ with a slope that is adjusted by the fit. The relative composition of the fitted processes in each control region is fixed as the background composition is adjusted by background normalization scale factors during the final binned-likelihood fit of the analysis, described below.

The result of this simultaneous fit in the resolved W(lv)H(bb) analysis is summarized in Table 10.4, which lists the values of the linear correction slopes after the fit. The uncertainties correspond to the statistical uncertainties from the fit. The values of the fitted slope have been validated on the control regions, where they cover the residual differences in $p_T(W)$ distributions between data and Monte-Carlo simulated samples after being applied to the processes mentioned above.

10.1.4 Analysis strategy

The analysis strategy for the resolved W(lv)H(bb) analysis is very similar to the one of the resolved Z(ll)H(bb) analysis, described in Chapter 8. It follows the exact same steps as described by the workflow in Figure 8.1, except the plots illustrating the

Process	Fitted Slop $[\mathrm{GeV}^{-1}]$
$t\bar{t}$	0.000380 ± 0.000089
W + udscg	0.000575 ± 0.000046
W + 2b + single top	0.001670 ± 0.000130

Table 10.4: Linear correction factors and statistical uncertainties obtained from a simultaneous fit to the $p_T(V)$ distribution in data in the resolved W(lv)H(bb) analysis.

$\mathrm{CMVA}_{\mathrm{min}}$ and BDT output score distributions in that figure are taken from the resolved Z(ll)H(bb) channel.

The resolved W(lv)H(bb) analysis starts with loose pre-selection requirements, summarized in the lower part of Table 10.5. This table also lists the selection of signal and control regions, described below. The variables N_{aj} and σ_{MET} correspond to the number of additional jets in the event and the MET significance, respectively.

An additional set of selections is applied on top of the preselection to define a signal region designed to keep as much signal contributions as possible. The signal region selections are listed in the second column of Table 10.5. The signal-background separation is performed by a BDT trained on Monte-Carlo samples in the signal region.

The BDT discriminant is trained simultaneously on the W(ev)H(bb) and W(μv)H(bb) sub-channels. The signal category of this training corresponds to the W^+, W^-, qqZH and ggZH processes. The background category includes all processes from Tables 6.4 and 10.3, with the exception of the QCD multijet background, as the small size of this sample causes peaks (a single bin with a very large weight) in the distributions of the discriminating variables used in the BDT training. Half of the Monte-Carlo samples are used as a training sample and the other half to evaluate the BDT score. A two-sample Kolmogorov-Smirnov test is performed to check that there is no overtraining. The inputs of the BDT training include variables as the dijet invariant mass and W boson transverse mass, the transverse momentum of the dijet and the W boson, the difference of the ϕ between the dijet and the W boson, $\mathrm{CMVAv2}_{\mathrm{min}}$, the number of additional jets, SA5 (soft activity) and MET-related variables.

In parallel, the boosted W(lv)H(bb) analysis defines four background control regions to study how the simulated samples models the most relevant physics variables in data. Those control regions are designed to maximize the purity of the main backgrounds: W + udscg (up, down, strange, charm or gluon) jets, W + 1 b jet (W + b), W + 2 b jets (W+ 2b) and $t\bar{t}$ multijet ($t\bar{t}$). The *W + light control control region* is enriched in W + udscg processes, the two heavy flavor *W + HF control region* enriched in W + 1 or 2 b jets and the $t\bar{t}$ *control region* in $t\bar{t}$ processes. The selections of the control regions are listed in the last three columns of Table 10.5.

In the control regions, the $\mathrm{CMVAv2}_{\mathrm{min}}$ variable provides a discrimination among the background processes. The $\mathrm{CMVAv2}_{\mathrm{min}}$ distributions from the control regions and the

BDT output score distributions from the signal regions are combined in a final binned-likelihood fit to measure the signal strength and extract the observed and expected significance with respect to the background-only hypothesis. The BDT output score distributions brings a signal-background separation and the CMVAv2$_{\text{min}}$ discriminates between the various background processes. A total of ten shapes are included in the fit, as the signal region and the four control regions are separated in the W(ev)H(bb) and W(μv)H(bb) sub-channels. Including the control regions in the final fit gives additional information to constrain the systematic uncertainties. In the control and signal regions, the shapes predicted by Monte-Carlo simulations are allowed to vary and are determined during the fit. The yield of the main backgrounds are allowed to float during the fit through the background normalization scale factors which are mostly constrained from the control region CMVAv2$_{\text{min}}$ shapes. Other shape variations are parameterized by nuisance parameters, which account for the systematic uncertainties, for each background process and nuisance parameters + signal strength for the signal processes. The systematic uncertainties mentioned in section 8.3.2 are common to the resolved Z(ll)H(bb) and resolved W(lv)H(bb) analysis.

10.2 Boosted W(lv)H(bb) analysis

This section reviews the boosted W(lv)H(bb) analysis. It has many similarities with the resolved W(lv)H(bb) analysis in terms of dataset, reconstructions and corrections of the Monte-Carlo samples.

Table 10.6 gives a comparison between the boosted and resolved W(lv)H(bb) analysis. It summarizes the dataset, physics objects and analysis strategy for both analysis, organized following the first column. The descriptions relative to the resolved and boosted W(lv)H(bb) are listed in the second and third column, respectively. The fourth column refers to the related section of this dissertation. More details on the physics objects for the boosted analysis case are given in the following sections.

Both analysis are performed on the 2016 CMS SingleElectron and SingleMuon dataset, recorded at a center-of-mass energy of $13\,\text{TeV}$. The same single-lepton high level triggers are used by both analyses. The primary vertex, missing transverse energy, soft activity reconstructions and the pileup treatment are the same in both analysis. The same reconstruction procedure is performed for the W boson, with a selection of $W(p_T) > 100\,\text{GeV}$ for the resolved case and $W(p_T) > 250\,\text{GeV}$ for the boosted case. The same corrections on Monte-Carlo simulations are applied for both analysis but the linear $W(p_T)$ corrections are derived separately for the resolved and boosted case.

The reconstruction of the Higgs candidate uses another approach in the boosted W(lv)H(bb) analysis, as it relies on a single AK08 fat jet. This approach relies on various tools mentioned below and described in more detail in the following sections of this dissertation.

Variable	Signal region	W + light	W + HF	$t\bar{t}$
$M(jj)$	[90,150]	< 250	< 90, [150,250]	< 250
CMVA_{max}	$> \text{CMVA}_{\text{T}}$	$> \text{CMVA}_{\text{T}}$	$> \text{CMVA}_{\text{T}}$	$> \text{CMVA}_{\text{T}}$
CMVA_{min}	$> \text{CMVA}_{\text{L}}$	$> \text{CMVA}_{\text{T}}$	$> \text{CMVA}_{\text{L}}$	$> \text{CMVA}_{\text{T}}$
N_{aj}	$= 0$	-	$= 0$	> 1
$\sigma(MET)$	-	> 2	> 2	-
Pre-selection:				
$\Delta\Phi(MET, l)$	< 0.2	< 0.2	< 0.2	< 0.2
$\Delta\phi(W, jj)$	> 2.5	> 2.5	> 2.5	> 2.5
$p_T(jj)$	> 100	> 100	> 100	> 100
$p_T(W)$	> 100	> 100	> 100	> 100
$p_T(l)$	(> 25,> 30)	(> 25,> 30)	(> 25,> 30)	(> 25,> 30)
$p_T(j_1)$	> 25	> 25	> 25	> 25
$p_T(j_2)$	> 25	> 25	> 25	> 25
Lepton isolation	(< 0.06, -)	(< 0.06, -)	(< 0.06, -)	(< 0.06, -)
BDT score	> 0.5	-	-	-

Table 10.5: Selections for the signal and controls regions in the resolved W(lv)H(bb) analysis. The first column lists the variables used for the selection. The variable N_{aj} corresponds to the number of additional jets in the event and σ_{MET} to the MET significance. The second column correspond to the signal region selection. The third, fourth and fifth column to the W + light, W + HF and $t\bar{t}$ control region, respectively. In the signal region, events with a BDT output score below 0.5 are removed due to the low sensitivity in that region. Two W + HF are defined, one with a low dijet invariant mass ($M(jj) < 90$ GeV) and one with a high dijet invariant mass ($M(jj) \in [150, 250]$ GeV). Entries marked with "-" indicate that the variable is not used in that region. The lepton isolation selection is denoted as: (muon isolation selection, electron isolation selection). The selection cuts are given in units of GeV, except for the angles given in radians and the isolation which is dimensionless.

The invariant mass of this fat jet is referred to as the *PUPPI soft drop mass* as it relies on a combination between the *PUPPI* algorithm [84], for pileup-mitigation, and the *soft drop* algorithm [85], to remove soft and high angle radiations within the jet. The selection of the fat jet is performed through the *double-b tagger algorithm* [86], which discriminates between jets with two b hadron decay and jets originating from light (up, down, charm, strange, gluon) quarks or a top quark. The main variables related to the fat jet system are the PUPPI soft drop mass, the transverse momentum of the fat jet and the *N-jettiness ratios* τ_{21} and τ_{32} that provide information about the inner structure of the fat jet [87]. The JEC and JER are applied on additional jets in the event, reconstructed from particle flow candidate with AK04 as in the resolved analysis, see section 6.4.4. The JEC and JER are also propagated to the PUPPI soft drop mass. Specific corrections for the PUPPI soft drop mass are also applied as a function of the fat jet transverse momentum and $|\eta|$. Corrections related to the difference in efficiency on data and Monte-Carlo event for a selections using the N-jettiness ratio variable τ_{21} are also considered. The analysis strategy is similar between both analysis. An overview of the boosted W(lv)H(bb) analysis strategy is given in section 10.3.

10.2.1 Data Taking and high level trigger

The same dataset and HLT path described in section 10.1.1 and 10.1.1.1, respectively, has been used for the boosted and resolved W(lv)H(bb) analysis.

10.2.2 Monte-Carlo simulation samples

The same Monte-Carlo samples for background and signal processes described in section 10.1.2 have been used for the boosted and resolved W(lv)H(bb) analysis.

10.2.3 Physics Objects

The physics objects and the corresponding corrections common to the resolved W(ll)H(bb) and boosted W(ll)H(bb) analysis are summarized in Table 10.6 and are not repeated in this section. The only differences in the W boson reconstruction is the high transverse momentum requirement, $W(p_T) > 250\,\mathrm{GeV}$, and the vector boson transverse momentum corrections applied on the Monte-Carlo simulated samples, which are rederived in the boosted phase space for the boosted W(lv)H(bb) analysis, as described below.

The main difference between the resolved and boosted W(lv)H(bb) is related to the reconstruction of the Higgs boson candidate, referred to as the *fat jet system* for the boosted case, described in section 10.2.3.2.

Analysis overview	**Resolved W(lv)H(bb) analysis**	**Boosted W(lv)H(bb) analysis**	**Section**
Dataset	SingleElectron and SingleMuon 2016 CMS dataset, recorded at $\sqrt{s} = 13\,\text{TeV}$		10.1.1
High level trigger	Single-electron and single-Muon trigger		10.1.1.1
Physics objects			
Primary vertex	Mentioned in previous section		6.4.1
Pileup treatment	"		6.4.2
Soft Activity	"		6.4.7
Tracks and vertex	"		6.4.5.1
W boson	Reconstructed from MET and electron/muon candidate		10.1.3.1
• Selection	$W(p_T) > 100\,\text{GeV}$	$W(p_T) > 250\,\text{GeV}$	-
• Monte-Carlo corrections	linear $W(p_T)$ reweighting	linear $W(p_T)$ reweighting	10.1.3.3, 10.2.3.1
	• NLO EWK QCD in $W(p_T)$		10.1.3.2
	• Electron and muon efficiency		6.4.3.2 and 6.4.3.3
Higgs candidate	Two AK04 b-tagged jets	One AK08 bb-tagger fat jet	6.4.9, 10.2.3.2
• Selection	dijet$(p_T) > 100\,\text{GeV}$	AK08fat jet$(p_T) > 100\,\text{GeV}$	-
• Invariant mass	Regressed dijet invariant mass	PUPPI soft drop mass	6.4.9.1, 10.2.3.3
• Tagging	CMVAv2 b-tagger	double-b tagger	6.4.5.2, 10.2.3.7
• Structure variables	• regressed dijet mass • regressed dijet p_T	• PUPPI soft drop mass • fat jet p_T • N-jettiness ratios (τ_{21}, τ_{32})	10.3, 10.2.3.5
• Data and Monte-Carlo corrections	• JEC and JER on all jets		6.4.4
	• LO to NLO weight on W + jet samples		6.4.9.2
	• CMVAv2 efficiency	• double-b tagger linear corrections	10.2.3.8
		• PUPPI soft drop mass corrections	10.2.3.4
		• JER and JES on PUPPI soft drop mass	10.3.3
		• τ_{21} corrections	10.2.3.6
Analysis strategy			10.1.4, 10.3
• Signal region • Control regions • Binned likelihood fit	BDT for signal-background discrimination Maximize background purity Performed simultaneously on control and signal regions		

Table 10.6: Comparison between the boosted and resolved W(lv)H(bb) analysis. Dataset, physics objects and analysis strategy for both analysis are organized following the first column. The descriptions relative to the resolved and boosted W(lv)H(bb) and listed in the second and third column, respectively. The fourth column refers to the related section of this dissertation. For items where the resolved and boosted analysis differ, it is organized as: "reference for resolved case", "reference for boosted case".

10.2.3.1 W boson corrections

Similarly to the resolved W(lv)H(bb) analysis, the spectrum of the W boson transverse momentum, $p_T(W)$, is observed to be harder in the Monte-Carlo simulated samples than in data after all the NLO and QCD corrections are applied on the W + jets Monte-Carlo simulated samples (see section 10.1.3.2).

A first attempt to address this discrepancy was explored by using the same linear $p_T(W)$ corrections as for the resolved W(lv)H(bb) analysis, see section 10.1.3.3. Those corrections are observed to overcorrect the Monte-Carlo distributions, the $p_T(W)$ spectrum being harder in the data than in the Monte-Carlo distributions after the linear corrections have been applied. Therefore, the independent linear re-weighting was re-derived in the phase-space of the boosted W(lv)H(bb) analysis following the same procedure as in the resolved case.

Independent linear re-weighting functions are derived to correct the $p_T(W)$ slope for $t\bar{t}$, W + udscg, and the combination of W + 2 b and single top processes via a simultaneous fit of the reconstructed $p_T(W)$ in the boosted W(lv)H(bb) control regions, described below, to data. The input templates for the fit in each control region is a sum of the Monte-Carlo prediction for each process corrected by a linear function of the reconstructed $p_T(W)$ with a slope that is allowed to float in the fit.

The result of this simultaneous fit in the boosted W(lv)H(bb) analysis is summarized in Table 10.7. The uncertainties correspond to the statistical uncertainties from the fit. The corrections essentially concerns the W + udscg process, the best fitted value of the slope being 0 for the two other processes. The central value for the W + udscg correction is compatible with what is obtained in the resolved W(lv)H(bb) analysis, unlike the correction for the W + 2 b or single top processes and the $t\bar{t}$ process that are not reweighted by the central value. Those differences have not been the object of additional studies and are attributed to the difference between the phase-space of both analysis.

The effect of the corrections in a regions enriched in W + udscg jets can be seen in Figure 10.3, showing two $p_T(W)$ distributions in the W + light control region of the boosted W(lv)H(bb) analysis. In the lower ratio plot, where the black points correspond to the values of the data divided by the stacked Monte-Carlo histogram in each bin, a linear trend can be observed in the left figure, before the linear corrections are applied, and is no longer visible in the right figure, where the Monte-Carlo processes mentioned above are reweighted by the linear corrections derived in the boosted W(lv)H(bb) control regions. The application of this linear $p_T(W)$ re-weighting removes the trend observed in the $p_T(W)$ spectrum.

Process Boosted analysis	Fitted Slop $[\mathrm{GeV}^{-1}]$
$t\bar{t}$	0 ± 0.0000896
W + udscg	0.000548 ± 0.0000870
W + 2b + single top	0 ± 0.0003500

Table 10.7: Linear correction factors and statistical uncertainties obtained from a simultaneous fit to the $p_T(W)$ distribution in data in the boosted W(lv)H(bb) analysis.

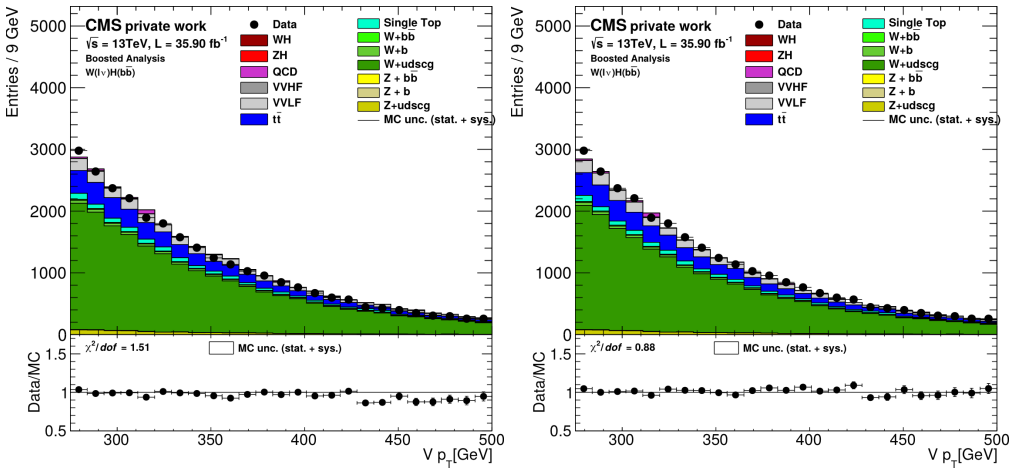


Figure 10.3: W boson transverse momentum distributions in the W + light control region of the boosted W(lv)H(bb) analysis. Left figure: no linear W boson p_T corrections are applied. Right figure: linear W boson p_T corrections derived in the boosted W(lv)H(bb) control regions are applied. The layout and composition of the two figures is similar to what is shown in Figure 8.3 and is described in the legend.

10.2.3.2 The fat jet system

The angular separation between the two b jets from the Higgs boson decay decreases as the Higgs boson momentum increases. For a transverse momentum larger than $250\,\mathrm{GeV}$, the ΔR separation between the two b jets is expected[2] to be lower than 1. Looking for an event signature involving two b jets reconstructed with an anti-k_T algorithm with a cone radius parameter of $\Delta R = 0.4$ (AK04 jet) is not suited for this scenario. Overlap between the two b jets reduces the reconstruction quality and in some cases both jets are merged into one single AK04 jets. The boosted analysis therefore reconstructs the pair of b hadrons from the Higgs candidate decay using a single anti-k_T fat jet with a larger cone radius parameter of $\Delta R = 0.8$ (AK08).

The fat jet candidate is selected with a double-b tagging algorithm, which is designed to identify fat jets containing a pair of b hadrons. The double-b tagger algorithm is described in section 10.2.3.7. In case of multiple fat jets in the event, the fat jet with the highest double-b tagger score (i.e. which is the most likely to contain two b hadrons) is selected as the fat jet candidate. In addition, the fat jet candidate must have a transverse momentum larger than $250\,\mathrm{GeV}$ and be within the tracker acceptance ($|\eta| < 2.5$).

The pileup treatment for the fat jet is performed via the *pileup per particle identification* (PUPPI) algorithm [84]. It assigns a weight to each jet particle or jet constituent for how likely it is to originate from pileup or the hard scattering vertex prior to the jet clustering, then rescales the four momentum of each jet particle by that weight. The weight is built based on the event pileup properties, local shape and tracking information. The jet energy corrections (JEC), described in the previous section 6.4.4, computed for AK08 jets with the PUPPI pileup mitigation, are applied on top of the PUPPI corrections for the fat jet.

10.2.3.3 PUPPI Softdrop mass

Due to contributions from initial state radiations, underlying events and pileup, the reconstructed jet mass can be far higher than the mass of the initial parton. This effect is increased by using a large distance parameter, here $\Delta R = 0.8$, for the jet reconstruction.

Contamination from pileup, initial state radiations and underlying events are removed from the fat jet candidate through a *grooming* technique combined to the PUPPI algorithm, referred to as the *PUPPI soft drop* algorithm [85]. The PUPPI soft drop algorithm removes the soft wide-angle radiations through a *declustering* procedure. Starting with an initial jet j of radius R_0 (here $\Delta R_0 = 0.8$), it breaks the jet j into two sub-jets j_1 and j_2. The softer constituent is removed unless

[2] This approximation uses the relation $\Delta R_{jj} = M_H / Hp_T$, where ΔR_{jj} is the angular separation between both b jets, M_H the Higgs boson mass and Hp_T the Higgs boson transverse momentum.

$$\frac{\min(p_T^{j_1}, p_T^{j_2})}{p_T^{j_1} + p_T^{j_2}} > z_{cut} \left(\frac{\Delta_{12}}{R_0} \right)^{\beta},$$

where $p_T^{j_{1/2}}$ is the transverse momentum of the jet $j_{1/2}$. The soft wide-angle radiation is controlled by the soft radiation fraction threshold z_{cut} and an angular exponent parameter β. The default CMS parameters of $\beta = 0$ and $z_{cut} = 1$, which remove soft and collinear contributions, are used. If the jets j_1 and j_2 satisfy the equation above, the jet j is considered as the final soft drop jet. Otherwise, the jet j is redefined as the sub-jet with the largest p_T from j_1 and j_2 and the procedure is repeated. If the sub-jet j is a singleton and cannot longer be declustered, it is considered as the final soft drop jet.

The invariant mass after the application of the PUPPI soft drop algorithm is referred to as the PUPPI soft drop mass, M_{SD}, and is one of the main variables of the boosted analysis.

The softdrop algorithm improves the pileup rejection when combined with the PUPPI algorithm. In addition, the softdrop algorithm contributes to the separation between fat jets from heavy objects (Higgs boson, W boson) and light quark or gluon jets. The signal and background normalized invariant mass distribution of the fat jet candidate are shown in figure 10.4, before and after the PUPPI soft drop algorithm is applied. The events must pass the boosted analysis pre-selection described in section 10.4. The mass of QCD jets is significantly shifted to lower values by the grooming procedure. A shift to lower values is also observed for the mass of the Z + jets and W + jets, where the fat jet mostly consists of light hadrons, though less significantly than for the QCD process. For all the processes having a dijet mass resonance, the PUPPI soft drop grooming procedure moves the mass peak closer to the real value of the heavy particle. The WH(bb) mass peak is moved closer to the Higgs mass $m_H = 125.18\,\mathrm{GeV}$, the diboson closer to the Z mass $m_Z = 91.2\,\mathrm{GeV}$ and the single top and $t\bar{t}$ closer to the W boson mass $m_W = 80.38\,\mathrm{GeV}$ (coming from the top quark decay).

The JECs from the fat jet are propagated to the PUPPI soft drop mass. In addition, corrections to account for differences in scale and resolution between the data and Monte-Carlo generated PUPPI soft drop mass are applied to the simulation. The resolution corrections are applied with the stochastic smearing method, described in section 6.4.4, and have an uncertainty of 20%. The scale corrections are applied as a multiplicative factor to the PUPPI soft drop mass distribution and have an uncertainty of 9.4%.

10.2.3.4 PUPPI soft drop mass corrections

The value of the PUPPI soft drop mass is observed to be dependent of the fat jet p_T and η. Dedicated corrections are applied in order to reduce this dependency. They

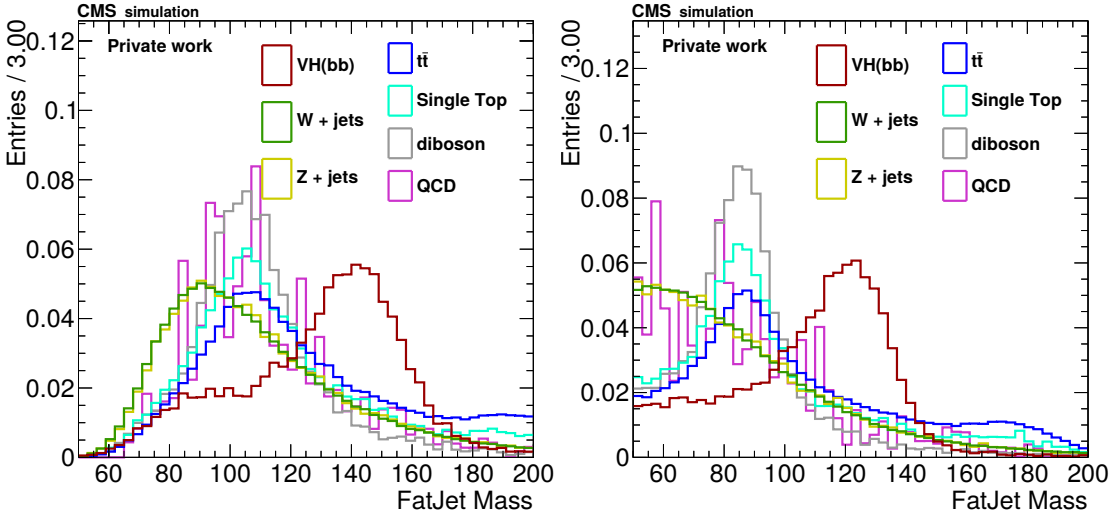


Figure 10.4: Invariant mass distribution of the fat jet candidate. Left: no PUPPI and soft drop corrections applied. Right: after PUPPI soft drop grooming is applied.

consist of two sets of p_T-dependent corrections, derived and recommended by the CMS JMAR (Jet MET Algorithms Reconstruction) working group. The first correction accounts for small shifts in the generated vector boson soft drop mass. The second correction accounts for shifts in the reconstructed jet PUPPI soft drop mass, and is applied separately for jets in the barrel and the endcaps regions. Both corrections are derived on simulated boosted W jets.

The mass shift introduced at generator level is corrected by a fit of M_{PDG}/M_{GEN} as a function of the jet p_T, where $M_{PDG} = 80.4\,\text{GeV}$ and M_{GEN} is the fitted mean of the generator level soft drop mass. To correct for the residual shift between generator and reconstruction level, a fit to $(M_{RECO} - M_{GEN})/M_{RECO}$ is performed, where M_{RECO} is the reconstructed fat jet PUPPI soft drop mass. The corrected soft drop mass M_{SD} is obtain by applying the generator and reconstruction level correction w_{GEN} and w_{RECO}, respectively, on the uncorrected PUPPI soft drop mass $M_{SD,\text{uncorr}}$ on both the data and Monte-Carlo simulations:

$$M_{SD} = M_{SD,\text{uncorr}} \times w_{RECO} \times w_{GEN}.$$

The distribution and corresponding fits for both corrections are shown in Figure 10.5 for the generator and reconstruction level, showing the size of these corrections and the associated uncertainty.

As mentioned above, the generator and reconstruction level corrections are both derived from simulation studies based on W vector bosons. In the boosted W(lv)H(bb) analysis, those corrections are assumed to be independent of the nature of the boson (W, Z or Higgs) and are applied to the PUPPI soft drop mass of all fat jet candidate.

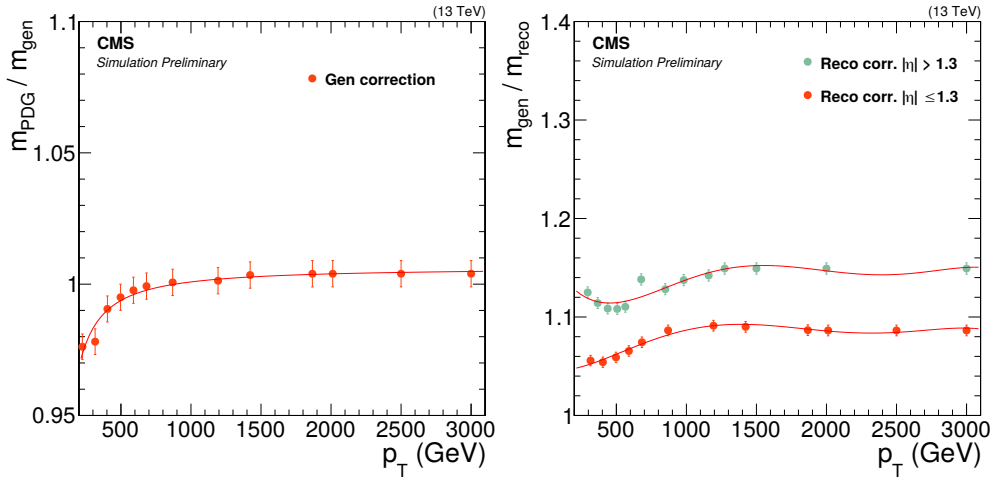


Figure 10.5: PUPPI soft drop mass corrections. Left figure: M_{PDG}/M_{GEN} as a function of the generated W boson transverse momentum. Right figure: $(M_{RECO} - M_{GEN})/M_{RECO}$ as a function of the reconstructed W boson transverse momentum in two η bins. Both figures are produces by the CMS JMAR working group.

Simulation studies of the PUPPI soft drop mass corrections for a Z or Higgs boson have not been considered in this analysis.

10.2.3.5 Substructure variable

The main backgrounds of the boosted W(lv)H(bb) analysis are the $t\bar{t}$ and W + jets processes. As illustrated in Figure 10.6, the number of sub-jet contained by the fat jet candidate can differ between the WH(bb) signal and the $t\bar{t}$ and W + jets background processes. This figure represents the sub-jet structure of a fat jet candidate for the WH(bb) signal process (left), $t\bar{t}$ process (middle) and the W + jets process (right). For the WH(bb) signal process, the fat jet candidate contains two b-flavoured sub-jets from the $H \rightarrow b\bar{b}$ decay. In the $t\bar{t}$ case, one of the top quarks is reconstructed as the fat jet candidate with a three sub-jet structure: one b sub-jet (in blue), and two light sub-jets from the W hadronic decay (in red). The three jets from the top quark decay are not necessarily all contained in the fat jet, such cases are discussed below. For the W + jets case, one or more additional jets are reconstructed in the fat jet candidate, the case corresponding to only one additional jet case is illustrated in the figure.

The different number of sub-jets reconstructed by the fat jet among the processes can be exploited with the use of *jet substructure variables*. An observable that is widely used in jet substructure is the N-subjettiness, denoted by τ_N [87], that estimates how likely the jet constituent consists of N numbers of sub-jets. The estimation of τ_N takes as input a reconstructed jet. In the case of the W(lv)H(bb) analysis, the fat jet candidate is considered. A number N of sub-jet candidates are then identified to define

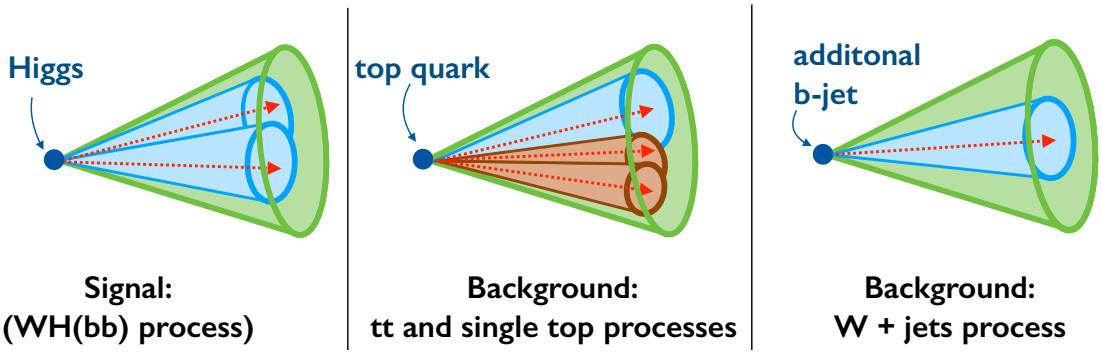


Figure 10.6: Illustration of the fat jet substructure for the WH(bb) signal process (left), tt or single top background processes (middle) and W + jets process (right). The b jets are illustrated by the blue cone in each example, the fat jet correspond to the large green cone and the two red cone to jets (not necessarily from b quarks).

N *subjettiness axis*. The value of τ_N is then calculated as

$$\tau_N = \frac{1}{\Sigma_k p_{T,K} \Delta R_0} \Sigma_k p_{T,k} \min\left(\Delta R_{k1}, \cdots, \Delta R_{kN}\right),$$

where k runs over the constituent particles in a given jets, $p_{T,k}$ is the transverse momentum of the constituent particle k, ΔR_{kl} is the separation between the constituent k and the subjettiness axis l and ΔR_0 is the jet radius. Jets with $\tau_N \approx 0$ are likely to have N or fewer sub-jets and jets with $\tau_N >> 0$ are likely to have at least $N+1$ sub-jets.

The N subjettiness axes are chosen to minimize the value of τ_N. The iterative procedure to find this minimum is described in [88, 89]. The N initial subjettiness axis to start the minimization (or *seeds*) are obtained by reclustering the jet with the *exclusive k_T algorithm*. It is similar to the anti-k_T algorithm defined in section 4.4.5.1, only that the distances d_{ij} between two jet elements i and j and the distance d_{iB} between an element i and the beam are defined as

$$\begin{aligned} d_{ij} &= \min\left(p_{\mathrm{T}_i}^{2p}, p_{\mathrm{T}_j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2} \\ d_{i\mathrm{B}} &= p_{\mathrm{T}_i}^{2p}, \end{aligned}$$

where $p = 1$, and that the clustering stops once N jets are found. The parameter R_{ij} is the distance in the $\eta - \phi$ plane between the components i and j, R defines the size of the jets and $p_{\mathrm{T}_i}^2$ is the transverse momentum of the element i. The axis of jets clustered this way are then used as seeds for the minimization procedure [89]. The k_T algorithm is more sensitive to the soft radiation than the anti-k_T algorithm, the former (latter) starting by clustering the soft (hard) components.

The N-subjettiness can be combined to define the *N-jettiness ratio* $\tau_{N+1,N} = \tau_{N+1}/\tau_N$. A high (low) τ_{N+1}/τ_N value indicates that the jet is more (less) likely to be N-prong than (N+1)-prong. The distributions of the $\tau_{21} = \tau_2/\tau_1$ and $\tau_{32} = \tau_2/\tau_1$ N-jettiness ratios can be found in the left and right side of Figure 10.7, respectively, for signal and background processes. The event selection corresponds to the boosted analysis pre-selection described in section 10.3.2.

The τ_{21} distribution shows a separation between the 2-prong processes such as VH(bb) or diboson and the 1-prong processes such as QCD, Z + jets or W + jets, as τ_{21} distribution corresponding to the VH(bb) or diboson processes is shifted to the left with respect to the QCD, Z + jets or W + jets processes, which populate the right side of the distribution.

The τ_{32} distribution shows a discrimination between the 3-prong process such as $t\bar{t}$ or single top and the other processes. Most of the separation can be seen in the left tail of the τ_{32} distributions (below $\tau_{32} \approx 0.5$). The τ_{32} distribution peaks around $\tau_{32} = 0.85$ for all the processes, including $t\bar{t}$ and single top. If all the top quarks from the $t\bar{t}$ and single top reconstructed by the fat jet candidate have a three prong structure as illustrated by the middle of Figure 10.6, the τ_{32} distribution is expected to be shifted to the right with respect to the other processes, which is not what is observed. This is due to part of the sub-jets from the top quark decay being out of the AK08 fat jet cone and therefor not being reconstructed by the fat jet candidate. This reconstruction feature can observed in the PUPPI soft drop mass distribution in the right side of Figure 10.4, where the largest peak of the PUPPI soft drop mass is around $\approx 85\,\mathrm{GeV}$, which shows that part of the energy from the top quark decay was not recovered by the fat jet reconstruction.

The discrimination of the two jettiness ratios τ_{21} and τ_{32} have been considered in the W(lv)H(bb) boosted analysis.

10.2.3.6 N-jettiness efficiency corrections

The τ_{21} N-jettiness ratio is usually used to identify W boson jets. To take into account differences between data and simulations, inclusive scale factors are extracted in three working points, tight ($\tau_{21} < 0.35$), medium ($\tau_{21} < 0.4$) and loose ($\tau_{21} < 0.55$), to correct the Monte-Carlo normalization to the data [90].

Those scale factors have been evaluated on the W boson with a tag and probe technique using a selection of $t\bar{t}$ events. A W boson with a semileptonic decay is required to select the event. A second W boson decaying fully hadronically, reconstructed with an AK08 fat jet with PUPPI pileup mitigation, is required and used to perform the measurement. The efficiencies are measured by fitting the W boson mass distribution by a double Crystall Ball to remove the background contributions. The values of the scale factors and the corresponding uncertainties are: 0.99 ± 0.11 (tight), 1.00 ± 0.06 (medium) and 1.03 ± 0.14 (loose). The uncertainties include statistical and systematic

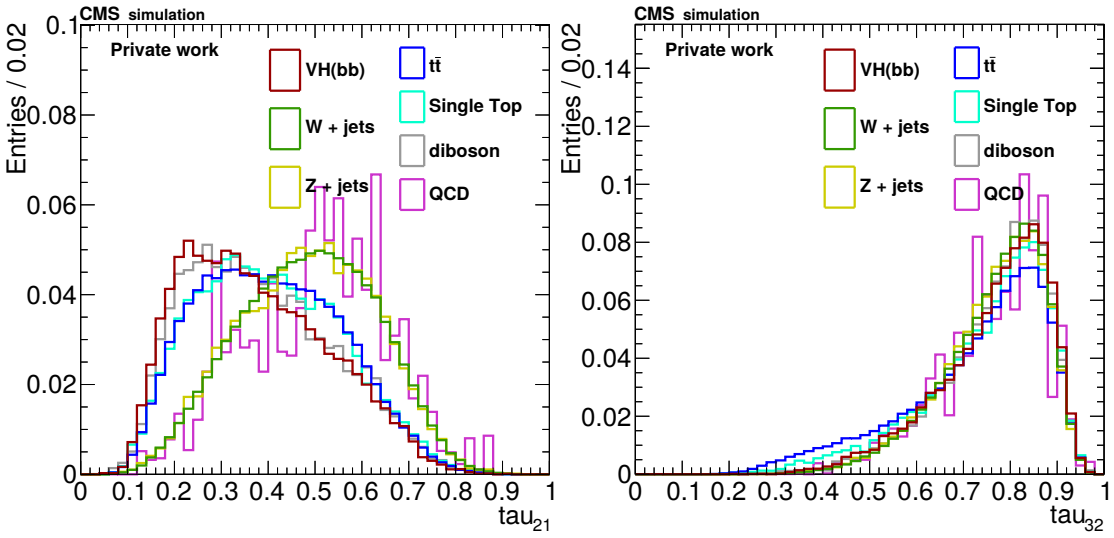


Figure 10.7: The N-jettiness ratio for normalized signal and background distributions in the boosted W(lv)H(bb) channel. Left: τ_{21} distributions. Right: τ_{32} distributions.

uncertainties. The systematic uncertainties include the bias from the choice of the Monte-Carlo generator and are evaluated by comparing the uncertainties from Pythia and HERWIG Monte-Carlo samples. The second source of systematics is connected to the choice of the fit model, replacing the Crytal Ball by a Gaussian function. Additional uncertainties due to differences between Higgs and W fat jets are determined from simulations for the loose working point. Two parton shower + hadronization algorithms are used, PYTHIA and HERWIG, to measure the differences in efficiency between the Higgs and the W boson. The corresponding uncertainties are 6% for per fat jet.

10.2.3.7 Double-b tagger

The decay products of boosted objects such as the Higgs boson candidate are highly collimated such that they merge into a single AK08 fat jet. Such boosted objects represent a challenge for the jet identification algorithms.

This can be illustrated by two different approaches used during the Run 1 period to identify boosted $H \rightarrow b\bar{b}$ candidates: the *fat jet b-tagging* and *subjet b-tagging* [86]. Both approaches are based on standard b-tagging algorithms which take advantage of the tracking and vertexing information and are designed to identify jets originating from a single b quark. In the first approach, the standard b-tagging algorithms are applied to the fat jet but with the track and vertex association criteria are relaxed due to a larger jet cone size. In the second approach, the sub-jets are first defined by declustering the fat jet with the *Cambrige/Aachen algorithm* [91, 92], obtained by setting the $p = 0$ in equation 10.2.3.5, combined with the *pruned jet clustering algorithm*

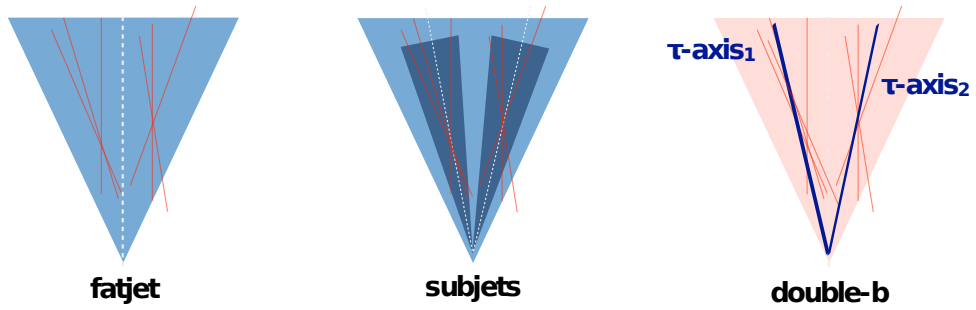


Figure 10.8: Schematic comparison of the fat jet (left), subjet b tagger (middle) and double-b tagger (right) methods. The figure is taken from [87].

to remove soft wide-angle radiations [93]. The standard b tagging is then applied to each of the sub-jet. The performance of the fat jet b-tagging is inherently limited by the fact that the algorithm is not designed to profit from the 2-prong structure of the fat jet, the tracks and secondary vertices selection being performed with respect the the fat jet axis . On the other hand, the sub-jet b tagging, with its focus on individual sub-jets, does not fully profit from the global properties of the fat jets containing two b hadrons. The two approaches are therefore complementary.

The *double-b tagger* algorithm combines information from both approaches. To identify boosted $H \rightarrow b\bar{b}$ candidates, it exploits the presence of two b quarks inside a fat jet and their topology in relation to the jet substructure, namely the fact that the b hadron flight directions are strongly correlated with the energy flows of the two sub-jets. The secondary vertices from the b hadron decay are identified with the IVF algorithm, described in section 6.4.5.1. The decay chain of the two b hadrons is reconstructed by associating the secondary vertices observables to the two 2-subjetinness axes, referred to as τ-axis. The *double-b tagger algorithm* combines 27 observables in a multivariate discriminant implemented in the TMVA package [42], which are reconstructed from the tracks, τ-axis, secondary and primary vertex information. The mass, transverse momentum and substructure variables such as the N-jetinness are not included in the multivariate discriminant, such that the double-b tagger can be used on a wide range of analyses to select jets originating from a W, Z or H boson. The output of the double-b tagger is a single value within a range $[-1, 1]$. A fat jet with a high (low) double-b tagger score is more likely (unlikely) to contain a 2-prong structure originating from a pair of b hadrons. A detailed description of the double-b tagger algorithm can be found in reference [86].

The difference between the fat jet, subjet b and double-b tagger algorithm are illustrated in Figure 10.8. The right side of the figure represents the fat jet b-tagger algorithm, where the tracks (in red) and secondary vertices (not shown on the figure) are associated to the fat jet based on their angular distance with respect to the jet

axis. The middle of the figure represents the subjet b tagger. The algorithm starts by undoing the last fat jet clustering step to define two sub-jets. The tracks and secondary vertices are then associated to each sub-jet, and the CSVv2 b-tagging algorithm is applied on each sub-jet. The right part of the figure represent the double-b tagger algorithm. It starts by defining the two τ-axis; the tracks and the secondary vertices are then associated to closest τ-axis.

The distribution of the double-b tagger value of the fat jet candidate for the signal and background processes can be found in Figure 10.9. Backgrounds with a similar double-b distributions have been merged in a same category, as described below.

- $t\bar{t}$ + *single top category* (TT + s. top): for $t\bar{t}$ and single top processes.
- *Diboson category* VZ(bb) + light-flavour: for diboson processes without b jets in the final state.
- *W + heavy flavor category* (W + hf): for W + 1 or 2 b jets processes.
- *W + light category*: for W + udscg jets
- *Remaining background category.*

As it can be seen in Figure 10.9, the double-b tagger variable has a good signal-background discrimination, the VH(bb) signal peaking above 0.9. This is the most discriminating variable after the PUPPI soft drop mass, which is the most discriminating variable for the boosted VH(bb) analysis discussed in this dissertation (see the variable ranking in section 10.3.2.2).

10.2.3.8 Double b-tagger efficiency corrections

The efficiencies related to a double-b tagger selection are different for data and Monte-Carlo simulations. In order to fully exploit the double-b tagger shape by including it in a BDT discriminator, flavor-dependent Monte-Carlo shape corrections and related systematic uncertainties must be derived, similarly to CMVAv2 b-tagger efficiency corrections included in the resolved analysis (see section 6.4.5.3).

Efficiency corrections are evaluated by the CMS b-tagging group, as described in section 6 and 7 of [86], for signal and background ($t\bar{t}$ and single top) processes in four double-b tagger fixed-cut working points: loose (double-b tagger score > 0.3), medium 1 (> 0.6), medium 2 (> 0.8) and tight (> 0.9). Those corrections are given as 2-3 scale factors in bins of the fat jet transverse momentum. They correct the scale of Monte-Carlo process distributions for a given double-b selection but not the overall double-b tagger shape.

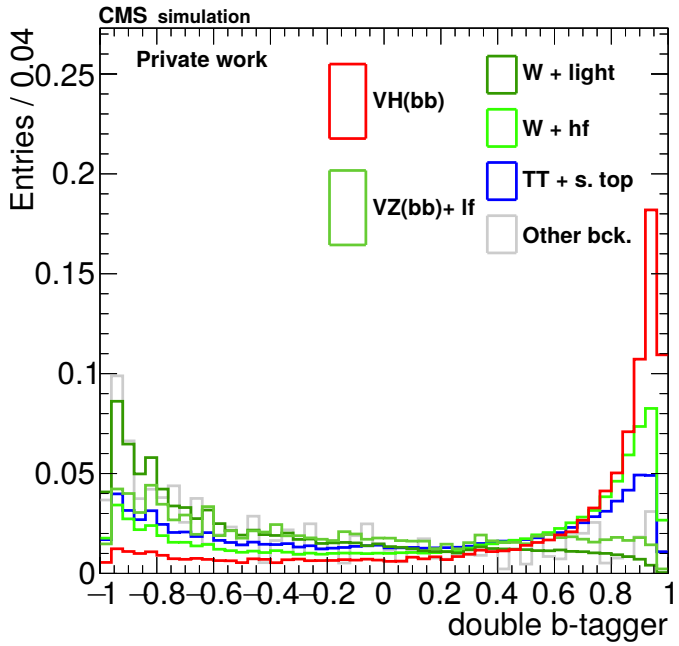


Figure 10.9: The normalized distribution of the fat jet candidate double-b tagger value for signal and background processes. The background categories are described in the text.

Dedicated double-b tagger shape corrections are therefor extracted for the boosted W(lv)H(bb) analysis. They are derived for the main background processes by a simultaneous fit performed in the boosted W(lv)H(bb) phase-space to data. Two regions of phase-space are considered for the fit, one *light flavor region* enriched in W + udscg jets processes and one *heavy flavor region* enriched in $t\bar{t}$, single top and W + 1/2 b jets processes. The light flavor region is selected by requiring a fat jet with a double-b tagger score below 0.8. The heavy flavor region is selected by requiring at least two jets outside of the fat jet cone having a CMVAv2 score above the loose working point. In both regions, events with a double-b value below -0.8 are removed to match the selection of the boosted analysis (see section 10.3.2.3).

The input templates for the fit are the background double-b tagger distributions obtained from the Monte-Carlo predictions corrected by a linear function of the double-b tagger score with a slope that is allowed to float in the fit. Background processes with a similar double-b tagger shapes are grouped in a same template category: a $t\bar{t}$, diboson + light flavor, W + light flavor and W + heavy flavor category, described above. The remaining backgrounds, having a negligible contribution to the Monte-Carlo predicted yield with respect to four aforementioned background categories, is subtracted from data and not considered in the fit template. The yield of the sum of all four templates is scaled by a same factor to normalize it to the data prior to the fit, such that the double-b tag linear shape corrections do not affect the total Monte-Carlo normal-

ization.

The linear corrections and related statistical uncertainties from the fit in the four categories are listed in Table 10.8. For processes with a b-flavour final state, such as $t\bar{t}$ and W + heavy flavor, the linear corrections have a slope of 0.062 ± 0.00015 and 0.07 ± 0.15706, respectively. The double-b tagger shape corrections are larger for processes involving light quarks in the final state, such as the W + light and diboson + light processes, which have a linear slope of 0.201 ± 0.01629 and 0.118 ± 0.01245, respectively. The uncertainty is the largest for the W + heavy flavor, with a linear correction slope compatible with 0. This is due to the shape and the low events yield of the W + heavy flavor template category that mainly populates the last four bins of the double-b tagger distributions, as it can be seen in the lower row of Figure 10.11, described below.

The shape of the templates before and after the fit are compared in Figure 10.10. The first row and the second row correspond to the prefit and postfit template distributions, respectively. The left column corresponds to the light flavor region and the right column to the heavy flavor region. The four template categories are illustrated separately by the colored dashed histograms described in the legend. The blue histogram represents the stack of the four template categories, and the black dot histogram represents the data. By comparing the upper and the lower row, it can be seen that the double-b tagger linear shape corrections improve the Monte-Carlo modeling of the data distribution. A χ^2 test for comparison between the stacked Monte-Carlo histogram and the data distribution is performed. The corresponding $\chi^2/ndof$, where $ndof$ is the number of degree of freedom, goes from 5.77 (heavy flavor region) and 16.65 (light flavor region) to 2.1 (heavy flavor region) and 2.48 (light flavor region) after the double-b tagger linear shape corrections are applied.

The double-b tag distributions before and after the application of the linear corrections in the heavy flavor and light flavor region are compared in Figure 10.11, arranged to a similar way to Figure 10.10. The linear trend visible in lower the ratio plot in the two upper row figures (no double-b tage shape correction applied). This trend is reduced after the double-b tag corrections are applied, as it can be seen in the lower ratio plot in the two lower row figures (with double-b tage shape correction applied). The remaining discrepancies are covered by the systematic uncertainties of the analysis, see the post-fit double-b tag distributions in the next chapter.

10.3 Boosted W(lv)H(bb) analysis strategy

The strategy and details of the boosted W(lv)H(bb) analysis performed on the 2016 dataset are described in this section. An overview of the analysis strategy is given in section 10.3.1. Section 10.3.2 is dedicated to signal and background selection. It also includes a description of the signal classification and background modeling. Sec-

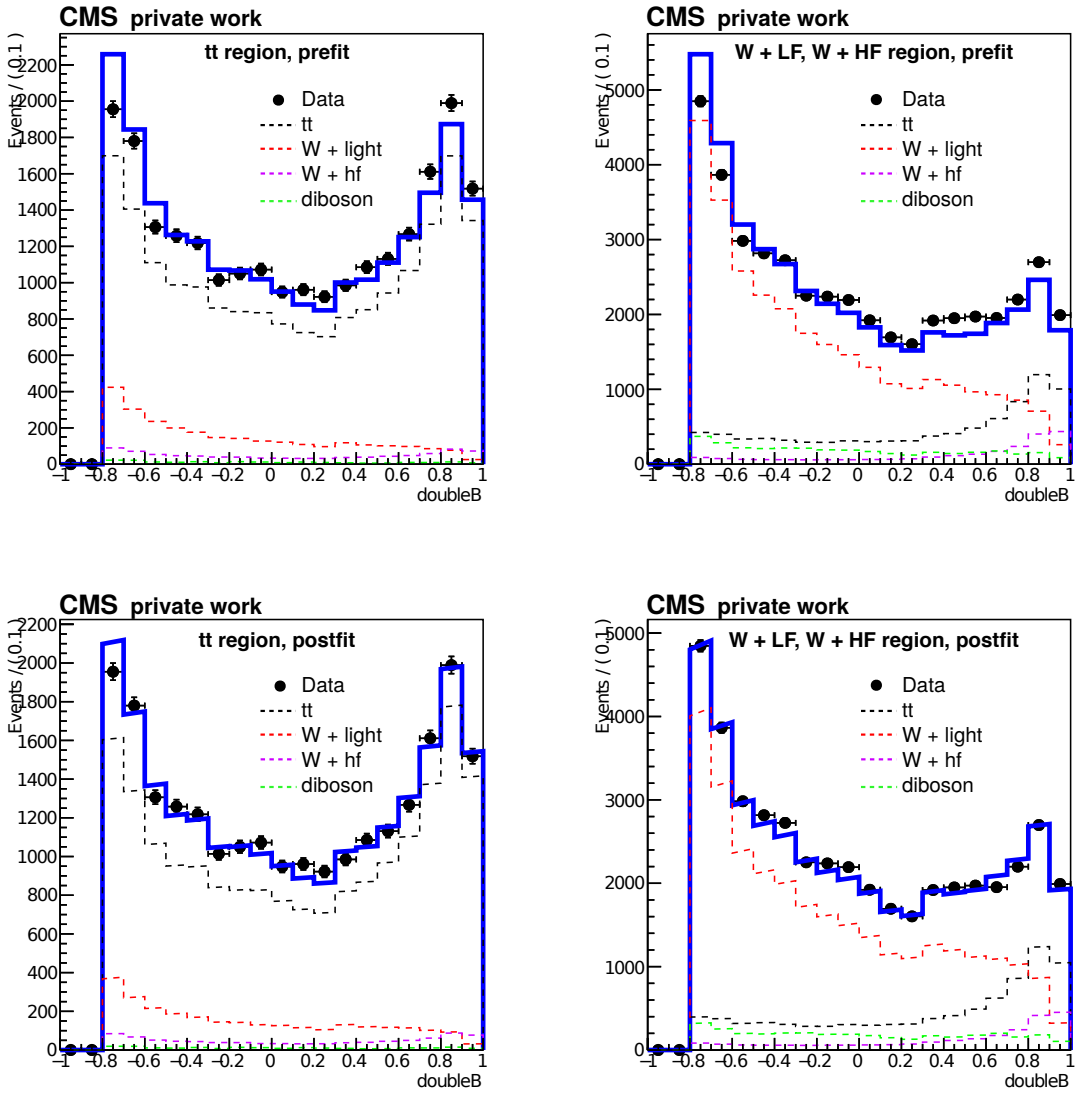


Figure 10.10: Template shapes for the different background categories as described in the text, before (upper row) and after (lower row) the fit. The heavy flavor and light flavor region corresponds to the left and right column, respectively.

Background category	Linear double-b tagger correction slope
$t\bar{t}$	0.062 ± 0.00015
W + heavy flavor	0.07 ± 0.15706
W + light	0.201 ± 0.01629
diboson + light	0.188 ± 0.01245

Table 10.8: Value and fit uncertainties on the slope of the linear corrections on the double-b tagger distribution for the different background categories.

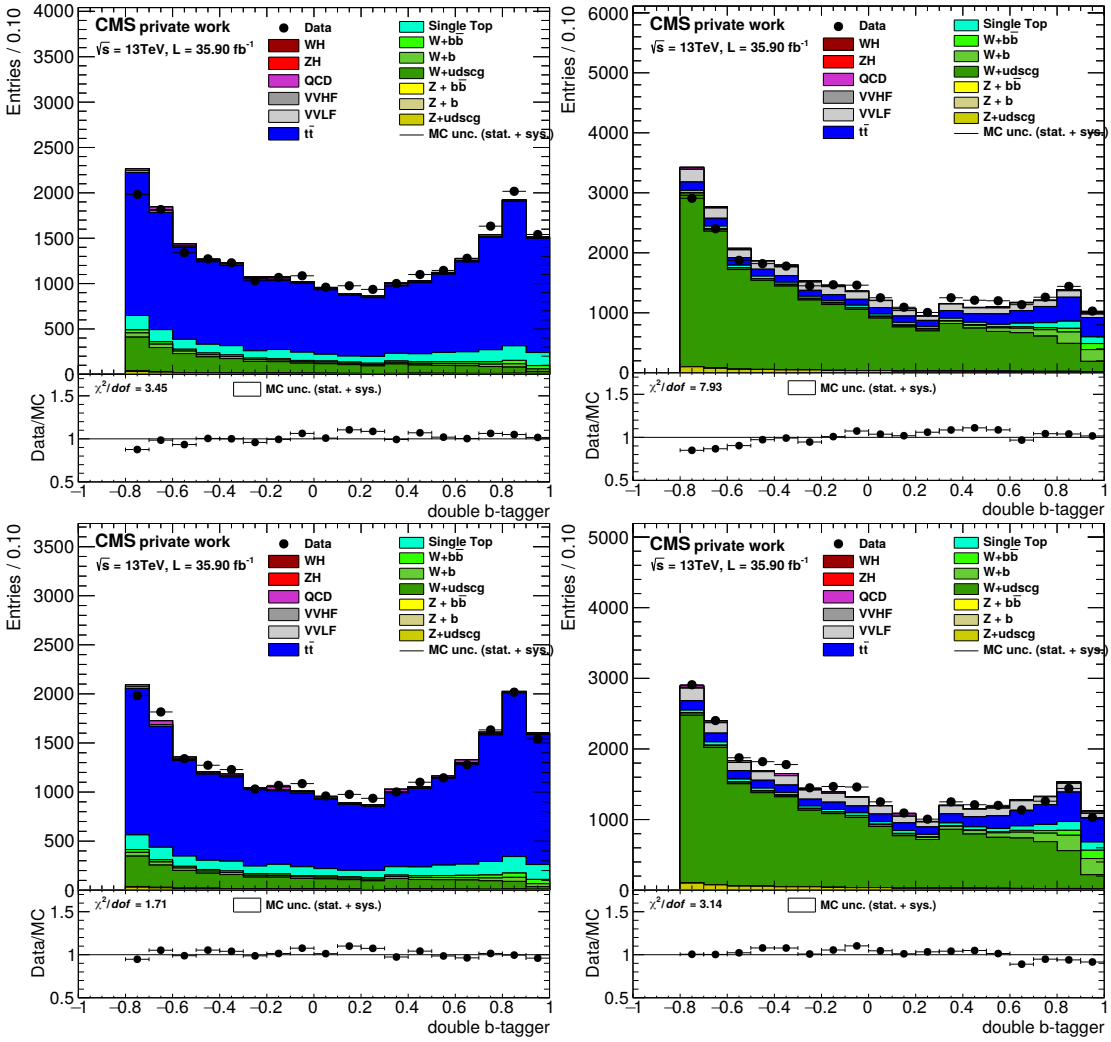


Figure 10.11: Double-b tag distributions in the two regions used for the fit of the linear double-b correction. From left to right, top to bottom: region enriched in $t\bar{t}$ and single without the linear corrections, the same region single after applying the linear corrections, region enriched in W + light jets without the linear correction, the same region after applying the linear corrections.

tion 10.3.3 is dedicated to an overview of the systematic uncertainties included in the analysis.

10.3.1 Analysis strategy in a nutshell

The boosted W(lv)H(bb) analysis starts with a loose event preselection. This selection follows the topology of a typical WH(bb) event in a phase-space with large transverse momentum of the W and Higgs boson. The W boson system is reconstructed from one lepton and MET, back-to-back to a fat jet system attributed to the Higgs boson $H \rightarrow b\bar{b}$ decay. Both the W and fat jet system are required to have a transverse momentum above $250\,\mathrm{GeV}$.

An additional set of selections is defined on top of this preselection to define the signal region, enriched in VH(bb) processes. The signal-background separation is performed by a BDT trained on Monte-Carlo samples in the signal region. The distribution of the BDT output score is included in the final binned-likelihood fit on data to extract the signal and the corresponding significance. This fit is performed simultaneously on signal and background processes, whose shape predicted by Monte-Carlo simulations are allowed to vary and are determined during the fit. The shape variations are parameterized by nuisance parameters for each background process, and nuisance parameters + signal strength for the signal process. The signal region and the BDT are described in section 10.3.2.

In parallel, the analysis defines a set of control regions to study how the simulated samples model the most relevant physics variables in data. Those control regions are designed to maximize the purity of the main backgrounds: W + udscg (up, down, strange, charm or gluon) jets, W + 1 b jet (W + b), W + 2 b jets (W+ $b\bar{b}$) and $t\bar{t}$ multijet ($t\bar{t}$). In the control regions, the lowest double-b tagger score of the fat jet provides discrimination among the background processes. The double-b tagger score distribution is fitted simultaneously with the BDT output score distribution from the signal region in the final binned-likelihood fit, the former bringing signal-background separation and the latter discrimination among the various background processes. In addition, including the control regions in the final fit gives additional information to constrain the systematic uncertainties. The yield of the main backgrounds are allowed to float during the fit through background normalization scale factors, which are mostly constrained from double-b tagger score shapes in the control regions.

The analysis strategy is summarized in Figure 10.12. The upper part contains an illustration of a typical W(lv)H(bb) signal event serving as a basis for the analysis preselection. It shows a lepton pair from a W^+ boson decay, back-to-back in transverse momentum space with respect to the double-b tagged AK08 fat jet from a Higgs boson decay. Below the signal event is a sketch of the PUPPI soft drop mass distribution, M_{SD}. The signal region is defined around the $125\,\mathrm{GeV}$ mass peak of the

W(lv)H(bb) process and the control regions in the mass sidebands[3]. The definition of the signal and control region selections are chosen to satisfy the following points: (i) The region selections must be orthogonal (an event cannot be present in more than one region). (ii) The control region definition must be as close as possible to the signal topology. This is to ensure that the systematic uncertainties, in particular the background normalization scale factors, can be extrapolated between the control and signal regions during the binned-likelihood fit. No dedicated extrapolation uncertainties are included in the fit. (iii) The background region selections maximize the purity of the main backgrounds and minimize the signal efficiency. (iv) The signal region maximizes the signal efficiency. Examples of double-b tagger score and BDT output score distributions from control and signal regions, respectively, in the W(ev)H(bb) sub-channel are shown below the M_{SD} plot. Those distributions, as well as other double-b tagger score and BDT output score distributions from all the W(lv)H(bb) signal + control regions, are fitted simultaneously in the final signal+background binned-likelihood fit to extract the signal strength, the observed and expected significance, fit the uncertainties and estimated the normalization scale factors. This correspond to a total of 2 BDT output score and 6 double-b tagger score distributions in the final binned-likelihood fit, performed on the boosted W(lv)H(bb) channel.

10.3.1.1 Combination with the resolved W(lv)H(bb) analysis

The boosted analysis can be combined in the final binned-likelihood with the resolved W(lv)H(bb) analysis on the 2016 dataset. The expected and observed sensitivity from this combination is an estimate of the potential improvement brought by the boosted technologies (fat jet, double-b tagger, ...) to the VH(bb) analysis conducted in the W(lv)H(bb) channel.

Performing this combination requires to make a choice concerning the treatment of the *overlapping events*, present in both the resolved and boosted W(lv)H(bb) analysis. This is due to the selections of both analysis not being fully orthogonal. An event with a boosted W boson candidate back-to-back to a boosted Higgs boson candidate, reconstructed from two AK04 b-tagger jets, can also be selected by the boosted analysis, as the b jet pair can be reconstructed by the double-b tagged fat jet.

Two approaches are considered for the treatment of the overlaping events. In the *overlap boosted* approach, all the overlapping events are selected in the boosted analysis and removed from the resolved analysis. The opposite is done in the *overlap resolved* approach, where the overlapping events are selected in the resolved analysis and removed from the boosted analysis.

[3]Some details are omitted from this figure for simplification purposes. The control regions enriched in W + udscg and $t\bar{t}$ background processes include the kinematic region around the $125\,\mathrm{GeV}$ mass peak but don't overlap with the signal region due to other selections, detailed in the next sections.

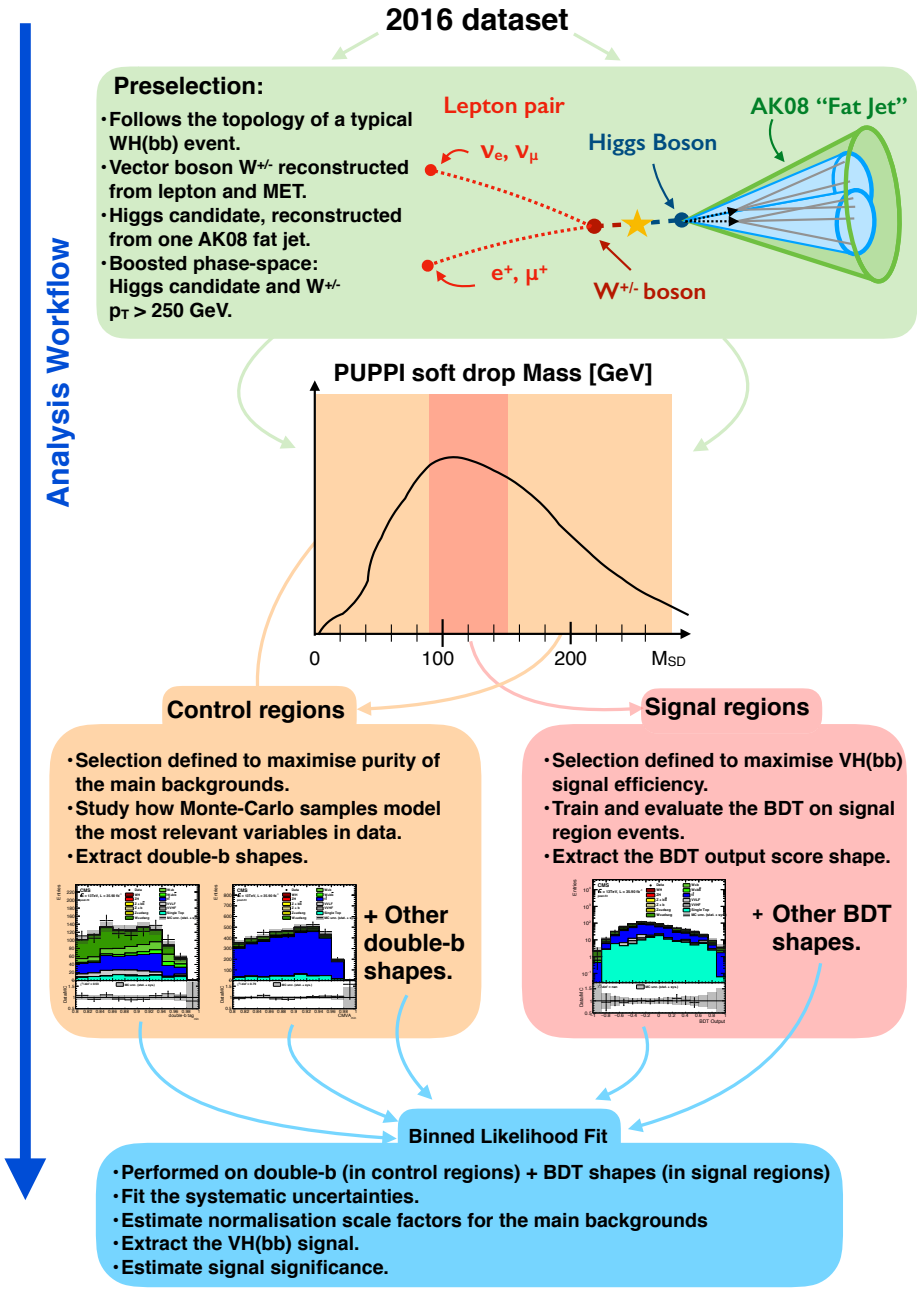


Figure 10.12: Overview of the analysis strategy for the boosted W(lv)H(bb) analysis. From top to bottom: the analysis starts with a event preselection, based on the W(lv)H(bb) event depicted in the upper part. Below is an illustration of the PUPPI soft drop mass, M_{SD}. Signal regions are defined around the Higgs 125 GeV mass peak. The distribution on the left side is the BDT output score in the W(le)H(bb) sub-channel. Control regions are defined on the dijet invariant mass sidebands. The two distributions on the left side are the double-b tagger score in two control regions. The BDT output score and double-b tagger score shapes from the signal and control regions, respectively, are combined in the final binned likelihood fit to extract the VH(bb) signal strength and the corresponding significance. The nuisances parameters are constrained during the fit.

This can be visualized in Figure 10.13 that illustrates four scenarios, referred to by a number in each corner. The layout for each scenario contains an illustration of the transverse momentum of the W boson in the upper part, and a Venn diagram representing the event selection in the lower part. The events selected by the resolved W(lv)H(bb) analysis are in grey, the ones selected by the boosted W(lv)H(bb) analysis in blue. Each scenario is described below. For each case, a binned-likelihood fit is performed by combining all the events represented by the lower Venn diagram.

- 1 corresponds to the resolved W(lv)H(bb) analysis, ignoring the existence of a possible boosted W(lv)H(bb) analysis. It is described in section 10.1. The results of this analysis are included in the VH(bb) analysis performed on the 2016 dataset and can be found in Chapter 9.

- 2 corresponds to the boosted W(lv)H(bb) analysis, ignoring the existence of a possible resolved W(lv)H(bb) analysis. This analysis is the object of the current section, all the figure and tables have been produced with the events selected by this analysis

- 3 corresponds to the overlap resolved scenario. Blue events selected from the boosted analysis are added to the $W(p_T)$ distributions above $250\,\mathrm{GeV}$, and the overlapping events are represented by a grey triangle with one side being a dashed black line. Both events from the resolved and boosted W(lv)H(bb) analysis are considered, and the overlapping events are excluded from the control and signal regions of the boosted W(lv)H(bb) analysis. The final binned-likelihood fit is performed simultaneously on the control and signal regions from the resolved + boosted analysis. All the systematic uncertainties are treated as correlated, except the ones unique to each analysis, see section 10.3.3.

- 4 corresponds to the overlap boosted scenario. Blue events selected from the boosted analysis are added to the $W(p_T)$ distributions above $250\,\mathrm{GeV}$, and the overlapping events are represented by a blue triangle with one side being a dashed blue line. Both events from the resolved and boosted W(lv)H(bb) analysis are considered, and the overlapping events are excluded from the control and signal regions of the resolved W(lv)H(bb) analysis. The resolved analysis part is therefore equivalent to the one in the second bullet. The final binned-likelihood fit is performed simultaneously on the control and signal regions from the resolved + boosted analysis. All the systematic uncertainties are treated as correlated, except the ones unique to each analysis, see section 10.3.3.

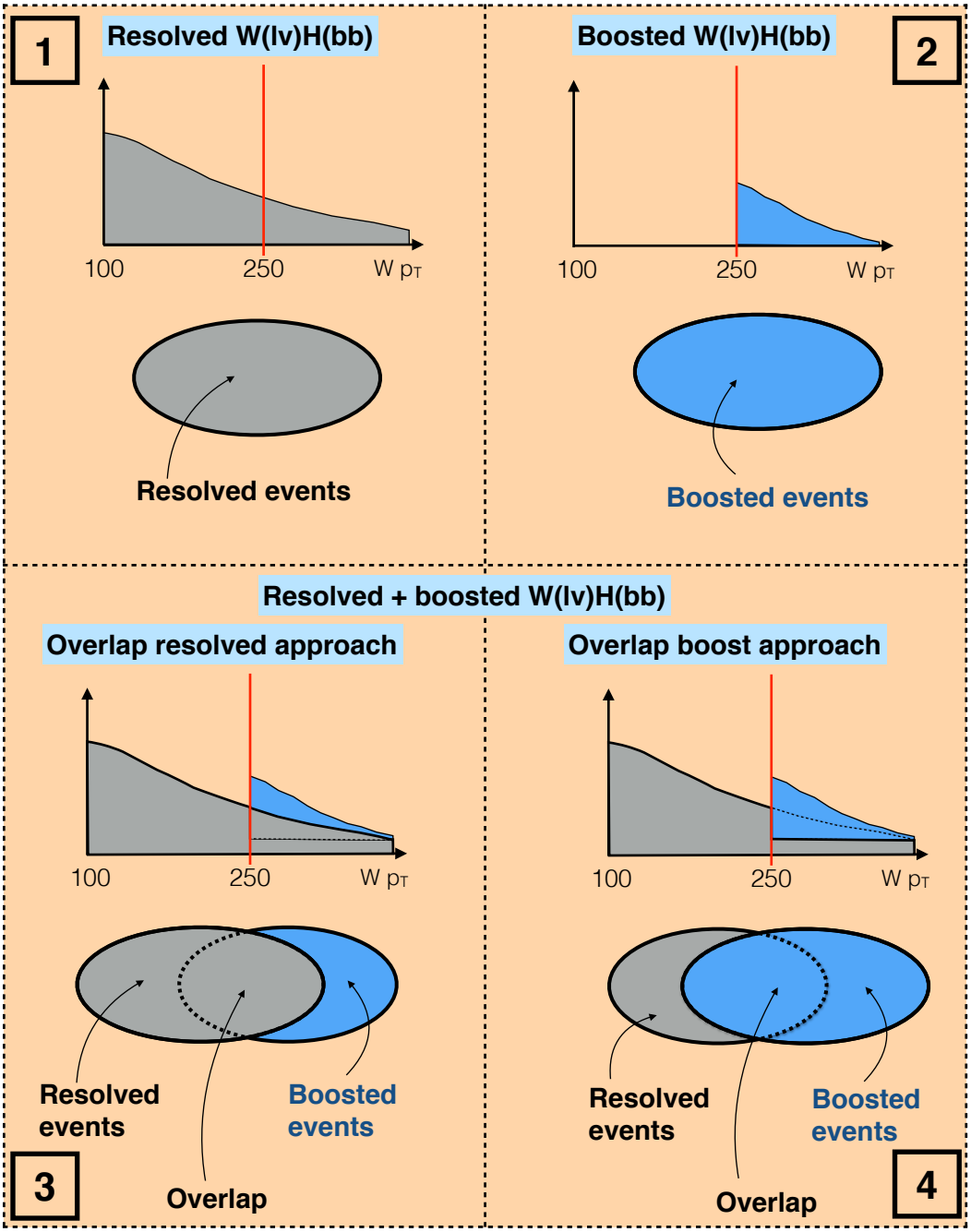


Figure 10.13: Overview of the resolved, boosted, and the boosted + resolved combinations for the W(lv)H(bb) analysis. The layout for each scenario contains an illustration of the transverse momentum of the W boson in the upper part, and a Venn diagram representing the event selection in the lower part. The events selected by the resolved W(lv)H(bb) analysis are in grey, the ones selected by the boosted W(lv)H(bb) analysis in blue. Each scenario is described in the text. For each case, a binned-likelihood fit is performed by combining all the events represented by the lower Venn diagram. The signal strength of the VH(bb) signal process and the expected and observed significance are extracted in each scenario from a binned likelihood fit.

10.3.2 Event selection

The typical signal events targeted by the boosted W(lv)H(bb) analysis contains a boosted fat jet candidate, back-to-back to a W vector boson candidate. Both the W boson and fat jet are required to have a transverse momentum above $250\,\mathrm{GeV}$. Additional selections on the vector boson candidate and the fat jet are mentioned in section 10.2.3. In addition, the softdrop mass is required to be above $50\,\mathrm{GeV}$. All the selections are listed in the lower part of Table 10.9.

10.3.2.1 Signal regions

The signal region consists of a sequence of three selections on top of the event pre-selection to remove contributions from the main backgrounds of the analysis.

The first selection is applied on the double-b tagger score of the fat jet candidate. Among the four working points listed in section 10.2.3.7, the double-b tagger score above 0.9 selection gives the highest sensitivity, estimated as $S/\sqrt{S+B}$, where S is the total number of signal events and B the total number of background events after the selection. As the double-b tagger score is an input variable of the BDT (described in the following section), the looser double-b tagger > 0.8 working point is used in the final signal region selections to increase the signal efficiency (0.5 for the 0.8 cut and 0.35 for the 0.9 cut). This selection mainly suppresses background contributions from W + udscg processes.

The second selection is a requirement on the number of b jets outside the fat jet, $N_{bjet \notin fatjet}$. Such jets are reconstructed as particle flow AK04 jets and are required to have a CMVAv2 score above the loose working point and have a cone radius distance above 0.8 ($\Delta R_{bjet,fatjet} > 0.8$). The selection that maximizes the $S/\sqrt{S+B}$ sensitivity correspond to no additional jet in the event, so $N_{bjet \notin fatjet} = 0$. It mostly suppresses contribution from the $t\bar{t}$ background.

The third selection concerns the PUPPI soft drop mass of the fat jet candidate, M_{SD}, that is required to be within the 90-$150\,\mathrm{GeV}$ range to reduce the contamination from the W + 1 b jet and W + 2b jets processes.

The event efficiency for the VH(bb) signal and the main background sources for each successive cuts are summarized in Figure 10.14. The denominator corresponds to the event pre-selection mentioned in the previous section. The double-b tagger score of the fat jet is referred to as $double - b_{max}$. The signal region selections are listed in Table 10.9.

10.3.2.2 Boosted decision tree

A BDT is trained in the signal region commonly on the W(ev)H(bb) and W(μv)H(bb) sub-channels. The BDT input variables have been selected by starting with a list of

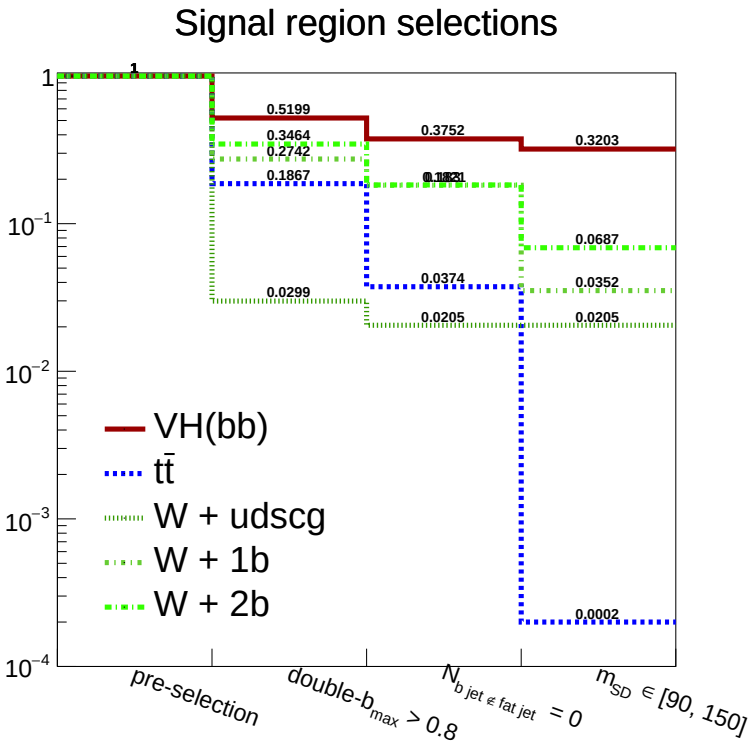


Figure 10.14: Efficiency for each successive control region cuts for the VH(bb) signal and the main W(lv)H(bb) backgrounds. The denominator correspond to the event pre-selections.

Variable	Description	Selection	Efficiency (VH(bb), tt, W + udscg, W + 1b, W + 2b)	Rejected background
$double - b_{max}$	Double-b tagger score of the fat jet candidate	> 0.8	(51 %, 19 %, 3 %, 27 %, 35 %)	W + udscg
$N_{bjet \notin fatjet}$	Number of b jets outside the fat jet cone	$= 0$	(38 %, 4 %, 2 %, 18 %, 18%)	$t\bar{t}$
M_{SD}	PUPPI Soft drop mass of the leading fat jet	$[90, 150]$	(32%, 0.02%, 2%, 3.5%, 7%)	W + 1 or 2 b
Pre-selection:				
$p_T(W)$	W boson transverse momentum	> 250	100%	-
Fat jet $p_T(jj)$	Fat jet transverse momentum	> 250	100%	-
$\Delta\phi(MET, l)$	Difference in ϕ between the MET and the lepton candidate	< 2	100%	-
$p_T(l)$	Transverse momentum of the lepton candidates	(> 25,> 30)	100%	-
Lepton isolation	Isolation of the lepton candidates	(< 0.06, -)	100%	-

Table 10.9: List of selections for the signal region in the boosted W(lv)H(bb) analysis. The first column lists the variables used for the selections, which are described in the second column. The double-b tagger score of the fat jet is referred to as $double - b_{max}$. The third column lists the selection applied on each variables. The fourth column lists the efficiency on the signal and main background processes after successively applying each of the selection variables as in Table 10.14. The pre-selection is used as an denominator in the efficiency estimation (with a default efficiency of 100%). The fifth column lists the background processes that are targeted by the first three selection variables.

discriminating variables and estimate the contribution to the significance of each variable by individually removing this variable from the training. The significance used for this study relies on a Gaussian approximation for the Poisson significance, defined as

$$\text{significance} = \Sigma_k \frac{s_k}{\sqrt{b_k}},$$

where the index k runs on all the bins of the BDT distribution and s_k and b_k are the total signal and background yields of the k'th bin, respectively. This approximation includes the statistical uncertainties from the expected yields of the background contribution in each bin, estimated as $\sqrt{b_k}$, but does not take systematic uncertainties into account.

To estimate the importance of a discriminating variable, this significance is evaluated on a new BDT after the variable under consideration is removed from the training. The list of the input variables, ordered in terms of impact on the approximated significance are listed in Table 10.10. Variables having an impact smaller than 2% are removed from the final BDT training. The N-subjettiness ratio τ_{21} brings a 3.10% improvement on the approximated significance at a cost of a systematical uncertainty of $\sim 10\%$ and $\sim 6\%$ (see section 10.2.3.5). Due to the small expected improvement on the analysis sensitivity and large corresponding systematics, the N-subjettiness ratio τ_{21} and τ_{32} are not used in the boosted analysis BDT and event selection.

The same BDT weights are used for the overlap boosted and overlap resolved approaches of the boosted W(lv)H(bb) analysis (see section 10.3.1). The BDT training includes events corresponding to the overlap boosted selection, i.e. it includes all the events from the boosted W(lv)H(bb) selection, even if they are also selected in by resolved W(lv)H(bb) analysis.

10.3.2.3 Control regions

Three control regions are defined in the boosted W(lv)H(bb) analysis: a *W + light control region*, enriched in W + udscg processes, a $t\bar{t}$ *control region*, enriched in $t\bar{t}$ processes, and a *W + heavy flavor (W+HF) control region*, enriched in the W + 1 b or and W + 2 b processes. The selection relies on the same variables as for the signal region definitions.

- The W + light control region is defined by inverting the cut on the double-b tagger score, which is required to be below 0.8. A difference in modeling of the double-b tagger distributions between data and the Monte-Carlo simulated processes is observed in the region where the double-b value is below -0.8, where the yield of the Monte-Carlo simulated processes is $\sim 30\%$ larger than

Variable	Description	Impact	In BDT
M_{SD}	PUPPI Soft drop mass of the leasing fat jet	15.63	yes
$double - b_{max}$	Double-b tagger score of the fat jet candidate	9.43	yes
SA5	Number of soft hadronic activity jet with a transverse momentum above $5\,\mathrm{GeV}$	7.82	yes
		7.82	yes
$\Delta\phi(MET, lep)$	Difference in ϕ between the MET and lepton	5.12	yes
$p_{T,balance}$	Ratio between the dijet and vector boson transverse momentum	4.18	yes
Fat jet $p_T(jj)$	Fat jet transverse momentum	4.18	yes
$\Delta\eta(W, H)$	Difference in η between the fat jet and W boson	3.91	yes
M_T	Transverse mass of the W boson candidate	3.64	yes
τ_{21}	N-subjettiness ratio	3.10	no
MET	Transverse momentum of missing energy	2.02	yes
τ_{32}	N-subjettiness ratio	1.48	no
$\Delta\phi(W, H)$	Difference in ϕ between the fat jet and W boson	0.94	no
N_{aj}	Number of additional jets	0.40	no
$p_T(W)$	W boson transverse momentum	0.40	no

Table 10.10: List of BDT input variables for the W(lv)H(bb) boosted analysis.

the data[4]. This region is excluded in the W + light control region. An additional selection removes contamination from the $t\bar{t}$ process by requiring no b tagged jets outside the fat jet cone, $N_{bjet \notin fatjet} = 0$.
About 70% of the background yield in the W + light control region comes from the W + light jet process. The main other sources of background contributing to this region are $t\bar{t}$ (13%), diboson + light jets (7%) and single top (3%).

- The $t\bar{t}$ *control region* is required to have a double-b tagger score above 0.8 and the presence of at least one b jet outside the fat jet cone, $N_{bjet \notin fatjet} > 0$.
  It is the purest control region of the three, about 79% of the background yield coming from the $t\bar{t}$ process. The main other sources of background are single top (10%) and W + light (6%).

- The selection of the W + HF control region requires the PUPPI soft drop mass of the fat jet candidate to be below $90\,\mathrm{GeV}$. As in the W + light control region, part of the $t\bar{t}$ is removed with the $N_{bjet \notin fatjet} = 0$ selection.
  The W + HF control region has a low purity, the W + 1 b and W + 2 b jets processes constituting 13% and 5.3% of the total background contribution, re-

[4]This is observed with no double-b tagger linear efficiency corrections applied. This region is excluded and the efficiency corrections derived in section 10.2.3.8 have been extracted for a double-b tagger range of $[-0.8,1]$.

Variable	**W + light**	**W + HF**	**$t\bar{t}$**
M_{SD}	-	$< 90\,\mathrm{GeV}$	-
$N_{bjet \notin fatjet}$	$= 0$	$= 0$	> 0
$double - b_{max}$	> -0.8,< 0.8	> 0.8	> 0.8
Pre-selection:			
$p_T(W)$	> 250	> 250	> 250
Fat jet $p_T(jj)$	> 250	> 250	> 250
$\Delta\phi(MET, l)$	< 2	< 2	< 2
$p_T(l)$	(> 25,> 30)	(> 25,> 30)	(> 25,> 30)
Lepton isolation	(< 0.06, -)	(< 0.06, -)	(< 0.06, -)

Table 10.11: List of selections for the W(lv)H(bb) control regions.

spectively. The main background processes contributing to this region are the W + light (32%), $t\bar{t}$ (30%) and single top (10%).

The selections for the three control regions are listed in Table 10.11. The yields of the different background processes in the three control regions are listed in Table 10.12. The prefit (before the final binned-likelihood fit) data and Monte-Carlo distributions for the main discriminating variables can be found in Figures 10.15, 10.16, 10.17 and 10.18.

10.3.3 Systematic uncertainties

The systematic uncertainties are included as nuisance parameters in the final binned-likelihood fit. All the uncertainties listed in section 8.3 are included, with the exception of the CMVAv2 b-tagging efficiencies, as no CMVAv2 b-tagging on the AK04 jets is used in the boosted analysis. Additional uncertainties specific for the boosted W(lv)H(bb) analysis are as follows.

- **Vector boson transverse momentum reweighting:** The slope of the W boson linear p_T correction are varied within the up and down statistical uncertainties derived in the fit, described in section 10.2.3.1, and propagated to shapes of the corresponding background processes as template uncertainties. In case of a combination with the resolved W(lv)H(bb) analysis, they are treated as decorrelated with respect the vector boson transverse momentum reweighting systematic of the resolved case.

- **Double-b tagger efficiency:** Systematic uncertainties of the linear double-b tagger corrections described in section 10.2.3.8 are taken into account in a similar way to the vector boson corrections mentioned in the previous bullet. The

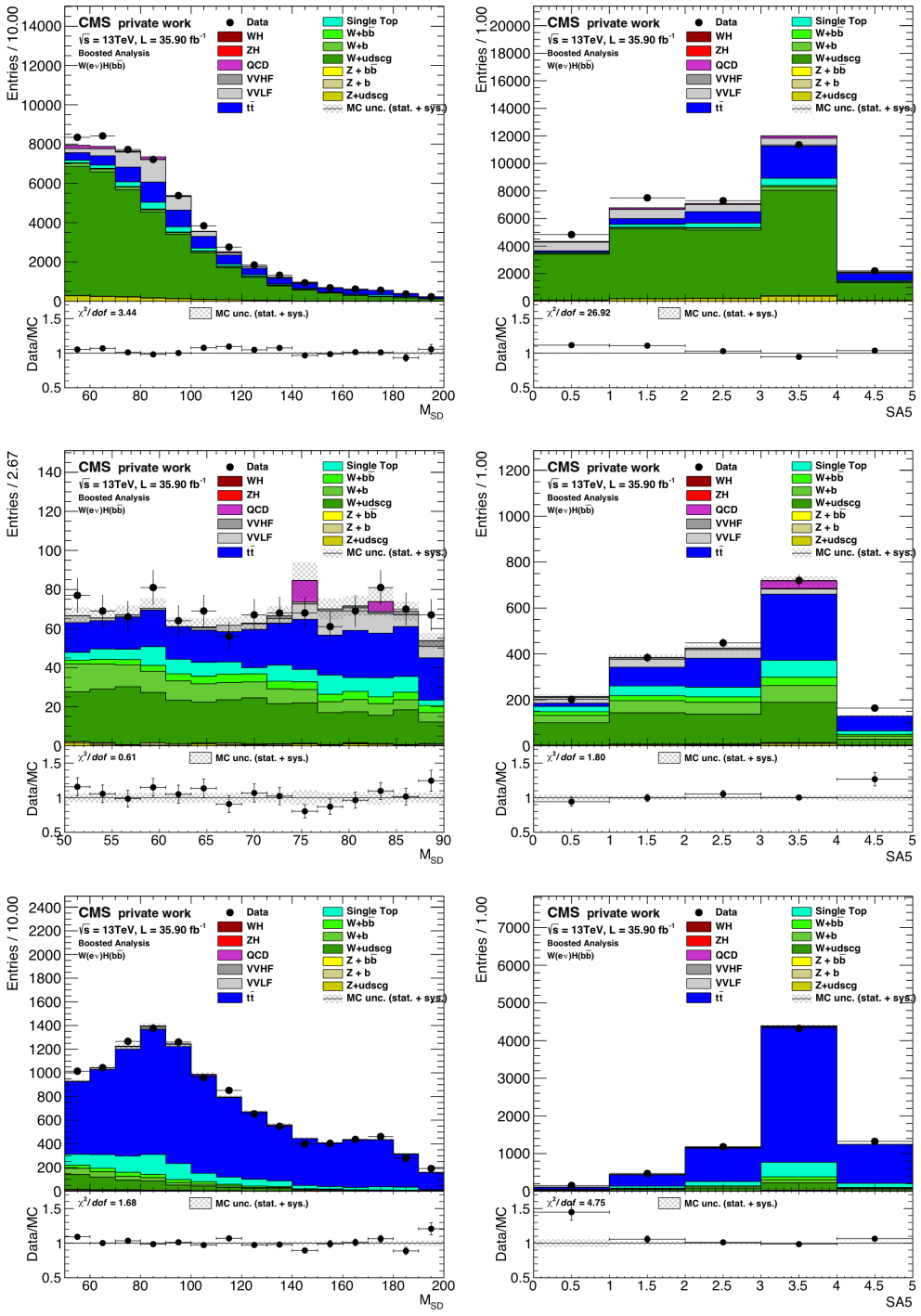


Figure 10.15: Data and Monte-Carlo generated processes prefit distributions of PUPPI soft drop mass (left column) and the soft activity (right column) in the three boosted control regions for the W($e\nu$)H(bb) sub-channel. First row : W + light control region. Second row: W + HF control region. Third row: $t\bar{t}$ control region.
The layout and composition of the six figures is similar to what is shown in Figure 8.3 and is described in the legend.

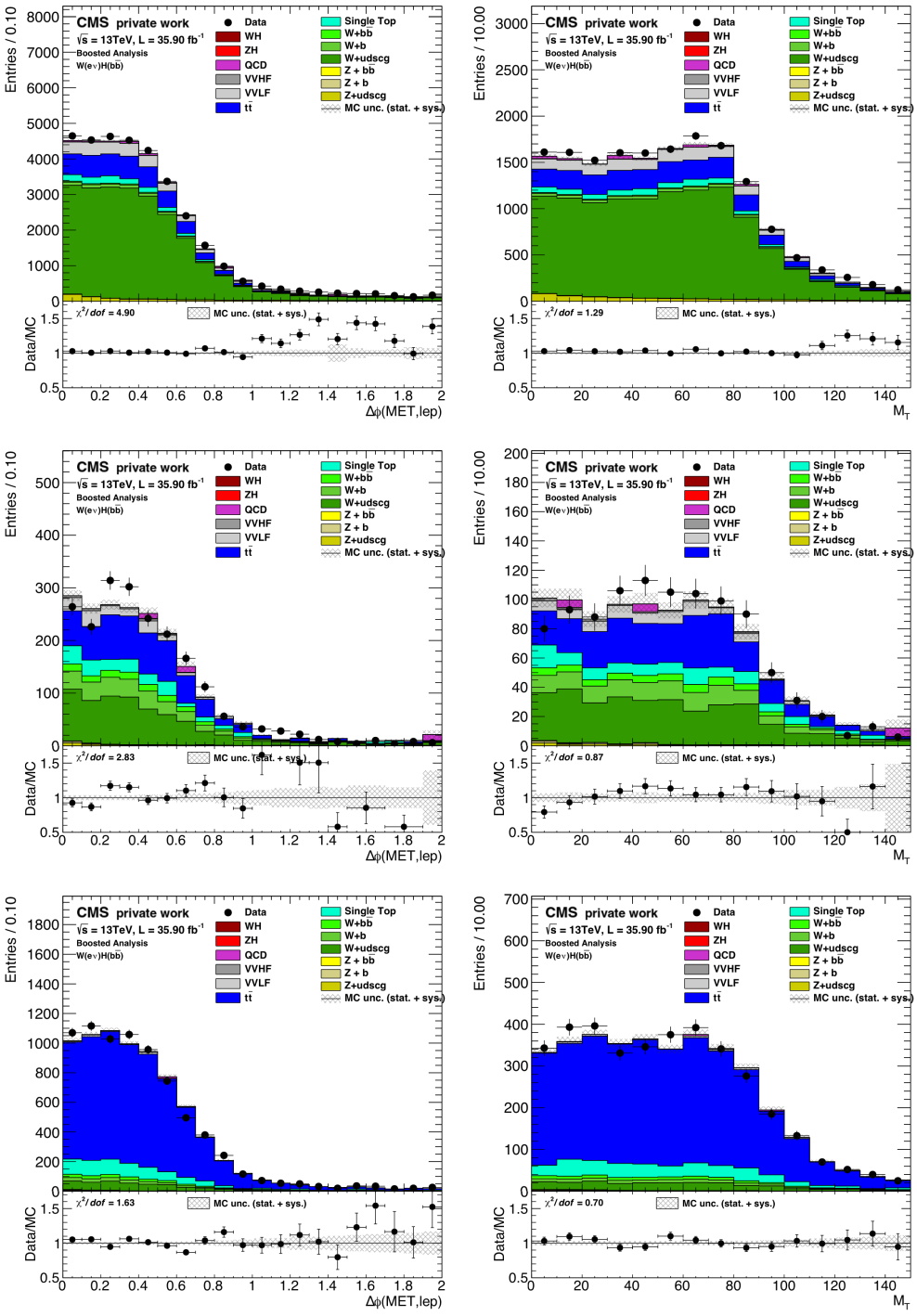


Figure 10.16: Data and Monte-Carlo generated processes prefit distributions of the angular distance $\Delta\phi$ between the MET and the electron (left column) and the W transverse mass (right column) in the three boosted control regions for the W($e\nu$)H(bb) sub-channel. First row : W + light control region. Second row: W + HF control region. Third row: $t\bar{t}$ control region.
The layout and composition of the six figures is similar to what is shown in Figure 8.3 and is described in the legend.

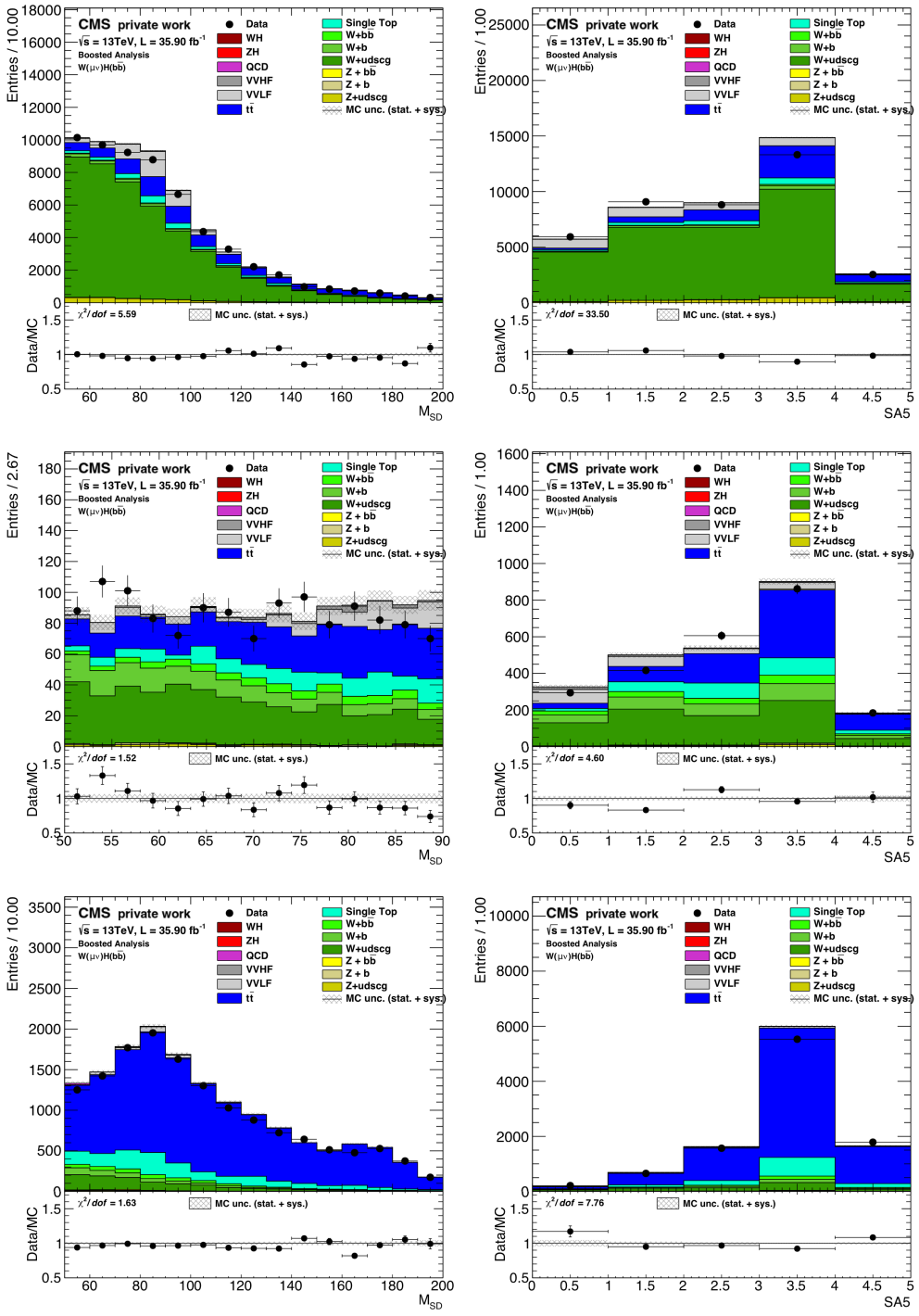


Figure 10.17: Data and Monte-Carlo generated processes prefit distributions of PUPPI soft drop mass (left column) and soft activity (right column) in the three boosted control regions for the W($\mu\nu$)H(bb) sub-channel. First row : W + light control region. Second row: W + HF control region. Third row: $t\bar{t}$ control region.
The layout and composition of the six figures is similar to what is shown in Figure 8.3 and is described in the legend.

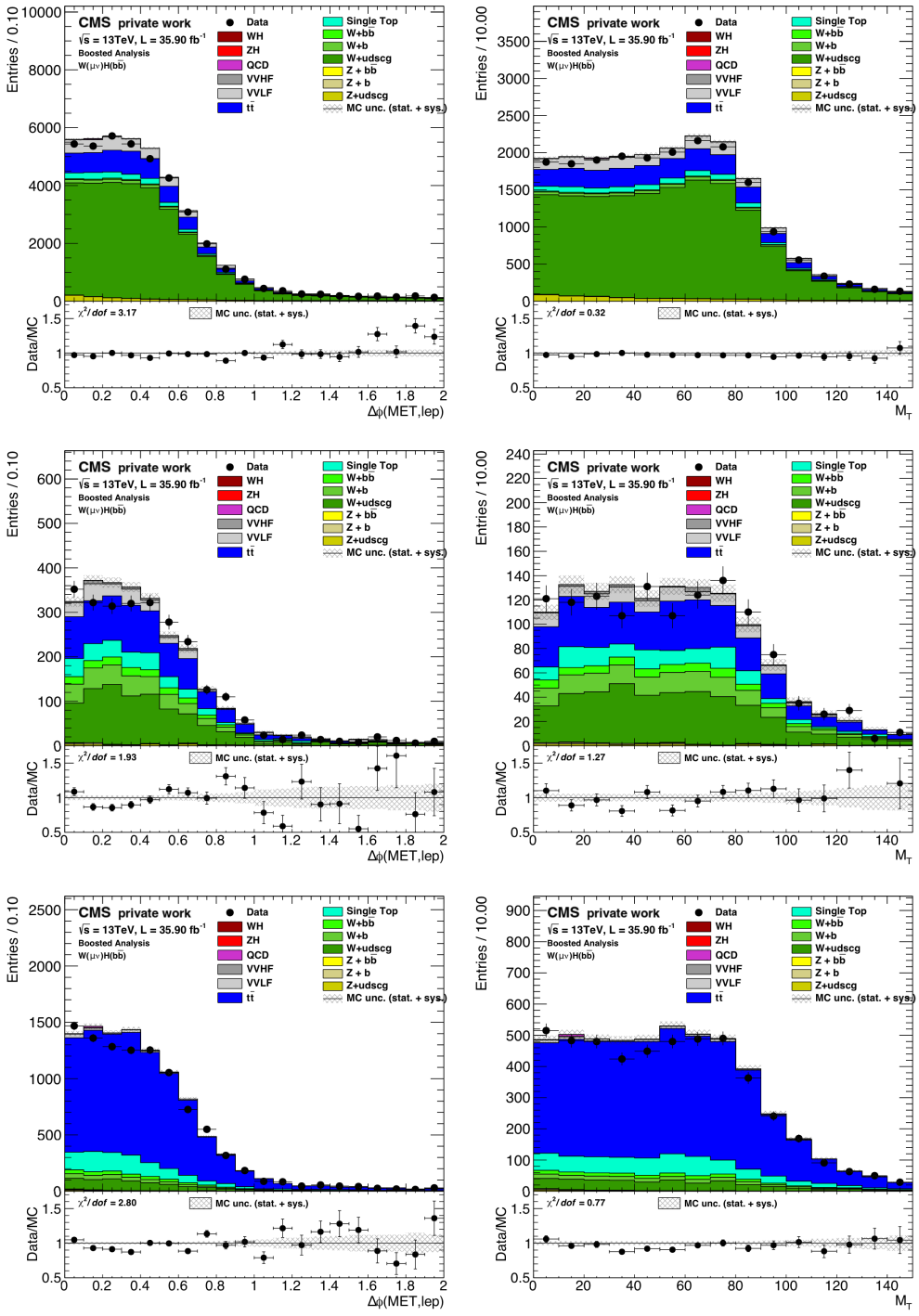


Figure 10.18: Data and Monte-Carlo generated processes prefit distributions of the angular distance $\Delta\phi$ between the MET and the muon (left column) and the W transverse mass (right column) in the three boosted control regions for the W($\mu\nu$)H(bb) sub-channel. First row : W + light control region. Second row: W + HF control region. Third row: $t\bar{t}$ control region.
The layout and composition of the six figures is similar to what is shown in Figure 8.3 and is described in the legend.

Process	W + light	W + HF	$t\bar{t}$
Z + 2 b jets	14.12	5.72	19.5
Z + 1 b jet	44.32	15.63	21.9
Z + light jets	980.5	20.03	25.13
W + 2 b jets	245.37	126.14	200.22
W + 1 b jet	739.89	300.22	221.38
W + light jets	27039.23	763.45	533.59
$t\bar{t}$	5087.79	712.12	7586.63
single top	1390.98	241.76	955.56
diboson + light jets	2912.10	188.13	94.69
diboson + 2 b jets	26.44	28.06	14.23
VH(bb)	8.32	1.53	5.02
Background	38480.74	2373.2	9658.6
Data	39436	2470	9732

Table 10.12: Yields (in number of events) for Monte-Carlo generated processes and data in the three boosted W(lv)H(bb) control regions.

slopes of the linear double-b tagger corrections are varied within the up and down uncertainties from the fit and propagated to the background shapes as a template uncertainties.

No shape corrections are available for the signal process. Instead, efficiency scale factors derived for a double-b tagger cut of > 0.8 are applied on the VH(bb) signal. The systematic uncertainty on the scale factors, described in section 6 from [86], are propagated to the signal shape used in the binned-likelihood fit as template uncertainties.

- **PUPPI soft drop mass scale and resolution:** The PUPPI soft drop energy scale uncertainty is propagated to the BDT shape by scaling the PUPPI soft drop mass with the up and down scale variation ($\pm 9.4\%$) before re-evaluating the BDT. For the PUPPI soft drop mass resolution uncertainty ($\pm 20\%$), a smearing of the PUPPI soft drop mass is performed for the up variation. Since the stochastic smearing method is used (see section 6.4.4), the down variation cannot be included, as the mass resolution cannot be reduced by this method. The down variation is set to be equal to the nominal value. The PUPPI soft drop mass resolution uncertainty is therefore taken into account by a one-sided uncertainty band, as it can be seen on the BDT output score distributions on figure 10.19.

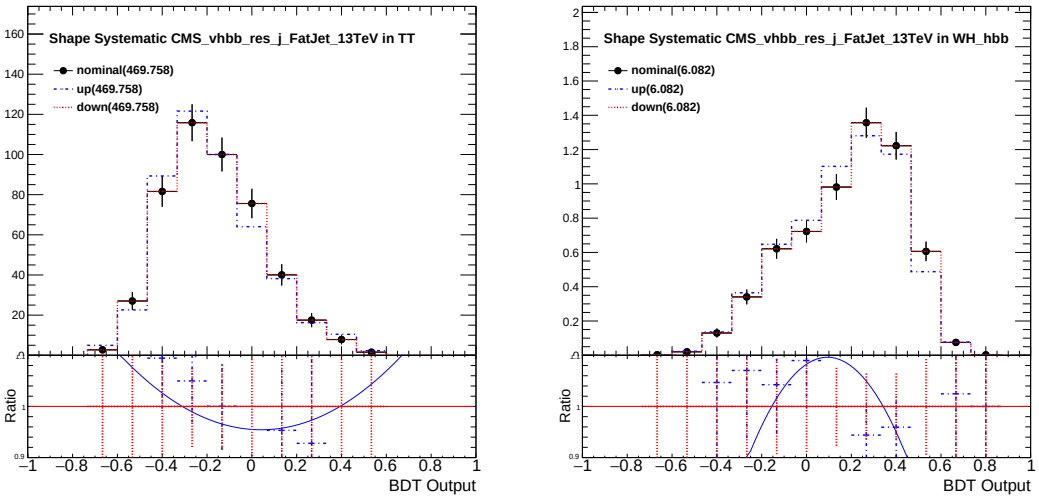


Figure 10.19: PUPPI Soft drop mass resolution uncertainty bands on the BDT output score distributions. Left: VH(bb) signal process. Right: $t\bar{t}$ background process.

11

Boosted W(lv)H(bb) analysis results

The results of the boosted W(lv)H(bb) analysis on the 2016 dataset are reviewed in this chapter. Section 11.1 is dedicated to a study of the discrimination power brought by including the double-b tagger score in the boosted W(lv)H(bb) analysis BDT discriminator. Section 11.2 reviews the impact of the nuisance parameters on the analysis sensitivity and the background normalization scale-factors extracted from the final binned-likelihood fit. The results from the boosted W(lv)H(bb) analysis are given in section 11.3. In section 11.3.2, the expected sensitivity from a combination between the boosted and resolved W(lv)H(bb) 2016 analysis is reviewed.

11.1 Impact of the double-b tagger

To correct and quantify differences in the double-b tagger efficiencies on the main backgrounds in the boosted W(lv)H(bb) channel, linear fit corrections are extracted (see section 10.2.3.8). Additional studies are required to include the double-b tagger shapes in the training of the BDT discriminant with respect to what is done in the preliminary boosted W(lv)H(bb) analysis presented in this thesis. Measuring the double-b tagger efficiency corrections on the shape of the VH(bb) signal process is particularly important, as a large effect on the double-b tagger shape in the signal region could have an important impact on the analysis sensitivity and the measured signal strength.

To estimate the importance of the double-b tagger, two versions of the boosted W(lv)H(bb) analyses have been performed, with and without the double-b tagger score included in the training of the BDT. For both analysis, the events overlapping with the resolved analysis have been included in the boosted category, following the overlap boosted scenario (see scenario 2 in Figure 10.13). In the case where the double-b tagger score isn't included in the training of the BDT, the uncertainties on the linear double-b tagger corrections are not considered in the final binned likelihood fit. Uncertainties

Boosted W(lv)H(bb) analysis	Without double-b tagger shape in the BDT training	With double-b tagger shape in the BDT training
Expected significance	0.37σ	0.42σ

Table 11.1: Asimov significance of the boosted W(lv)H(bb) analysis without and with the double-b tagger score used as a training variable in the BDT. Including the double-b tagger score improves the expected analysis sensitivity by $\sim 10\%$.

related to the signal efficiencies for a double-b tagger selection above 0.8 are included in both cases.

The expected significance from a binned-likelihood fit including all the systematic uncertainties are 0.42σ and 0.37σ for a boosted analysis with and without the double-b tagger shape included in the BDT training, respectively. This corresponds to a $\sim 10\%$ improvement coming from the double-b tagger. For the results discussed below, the double-b tagger shape is included in the BDT training for the boosted W(lv)H(bb) analysis.

11.2 Binned-likelihood fit

The validation of the likelihood fit procedure is performed following the overlap boosted scenario. The final binned-likelihood fit for the boosted W(lv)H(bb) channel is performed simultaneously on the control and signal regions of the boosted W(lv)H(bb) channel, taking as input the double-b tagger shapes for the control regions and the BDT output score shapes for the signal regions. The shape and normalization are allowed to vary within uncertainties, treated as independent nuisance parameters during the fit. The nuisance parameters and background normalization scale factors are allowed to float freely and are adjusted by the fit. The postfit values of the background normalization scale-factors are listed in Table 11.2 and are discussed below. The measured signal strength is discussed in the next section.

Three background normalization scale factors are considered. The $t\bar{t}$ scale factor for the $t\bar{t}$ process, the *W + light* scale factor for the W + udscg process and the *W + 1 or 2 b* scale factor for both the W + 1 b jet and W + 2 b jets processes. This differs from the resolved W(lv)H(bb) analysis, which uses different scale factors for the W + 1 b jet and W + 2 b jets processes. Unlike the CMVAv2$_{min}$ distributions used in the resolved analysis control regions, the double-b tagger shape doesn't provide a good discrimination between the two processes. This affects the performance of the binned likelihood fit, which doesn't converge when using a W + 1 b and W + 2 b scale factor. The two scale factors are therefore merged into the W + 1 or 2 b before the fit in the boosted W(lv)H(bb) analysis.

The fitted values and uncertainties of the background normalization scale factors

Process	Overlap boosted scenario	Overlap resolved scenario
	Background normalization scale-factor	
$t\bar{t}$	0.94 ± 0.09	0.90 ± 0.09
W + light	1.04 ± 0.12	1.06 ± 0.12
W + 1 or 2 b	0.65 ± 0.13	0.72 ± 0.14

Table 11.2: Value and total uncertainties of the background normalization scale factors estimated by the binned-likelihood fit performed on an Asimov dataset.

for the boosted W(lv)H(bb) channel for both the overlap boosted and overlap resolved scenario are listed in Table 11.2. The values of the background normalization scale factors are compatible between the two scenarios.

The pulls and ratios for the 551 nuisance parameters included in the fit are shown in Figure 11.1. The upper part corresponds to a fit including all background + signal processes and the lower part to a fit performed on background processes only. In an ideal scenario, corresponding to no pulls and constrains, the red point corresponding to the pull value is centered at 0 and the ratios represented by the uncertainty bands are between -1 and 1. As it can be seen in both the upper and lower part of Figure 11.1, the majority of the nuisance parameters are compatible with their prefit value with a ratio of 1, which is the sign of a good performance of the likelihood fit. Five nuisances parameters considered as overconstrained by the fit, with an absolute pull value above 2 and/or a ratio below 0.5, are listed in Figure 11.2 for both the background and background + signal fit.

The importance of each systematic uncertainty is estimated by its impact on the expected signal strength uncertainty. To evaluate the impact on the signal strength for a particular nuisance parameter θ, the parameter is fixed at its $+1\sigma$ or -1σ postfit value. The binned-likelihood fit is performed again, with all other parameters profiled to their central value. The corresponding shift $\Delta\mu$ on the signal strength quantifies the impact of the nuisance parameter.

The impacts for the 30'th most important nuisance parameters are listed in Figure 11.3. The first column lists the name of the nuisance parameters, ranked according to their impact listed in the third column. The second column lists the pulls and ratios. In the third column, the impacts on the signal strength corresponding to the $+1\sigma$ (red) and -1σ (blue) shifts of the nuisance parameter values are shown. Most of the highly ranked nuisance parameters correspond to the limited size of the Monte-Carlo simulated samples, particularly for the single top and $t\bar{t}$ processes. The first and second parameters in the ranking are the statistical uncertainties of the single top (referred to by s_Top in the nuisance parameter name) and $t\bar{t}$ (TT) samples in the highest BDT bin in the W(μv)H(bb) (W(ev)H(bb)) sub-channel. Other nuisances having an important impact on the signal strength apart from the size of the Monte-Carlo samples are the resolution and scale uncertainties of the PUPPI soft drop mass in the fourth and fifth

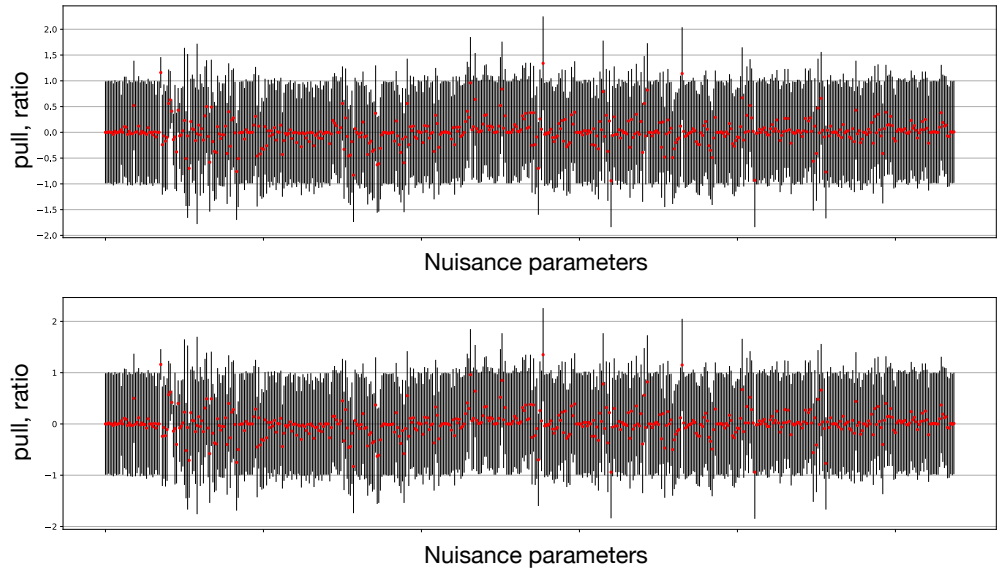


Figure 11.1: Pulls and ratios for all nuisance parameters. The red points correspond to the pull values and the error bars to the ratios. **Upper plot:** all signal and background processes are included in the fit. **Lower plot:** only background processes are included in the fit.

nuisance	background fit $\Delta x/\sigma_{in}$, σ_{out}/σ_{in}	signal fit $\Delta x/\sigma_{in}$, σ_{out}/σ_{in}	$\rho(\mu, \theta)$
CMS_vhbb_eff_e_Wln_13TeV	+1.16, 0.30	+1.16, 0.30	-0.01
CMS_vhbb_res_j_FatJet_13TeV	+0.41, 0.23	+0.42, 0.23	+0.21
CMS_vhbb_scale_j_FatJet_13TeV	+0.43, 0.43	+0.40, 0.45	-0.19
CMS_vhbb_scale_j_PileUpPtEC2_13TeV	-0.07, 1.59	-0.07, 1.60	+0.00
CMS_vhbb_scale_j_RelativePtEC2_13TeV	-0.03, 1.75	-0.03, 1.73	-0.01

Figure 11.2: List of the 5 overconstrained nuisance parameters in the final binned-likelihood fit, with an absolute pull value above 2 and/or a ratio below 0.5. The first column lists the name of the nuisances. The second and third column correspond to the fit with and without the signal contribution, respectively. The values in those two columns are organized as: (pull, constrain ratio). The fourth column shows the linear correlation coefficient between the given nuisance parameter and the signal strength μ.

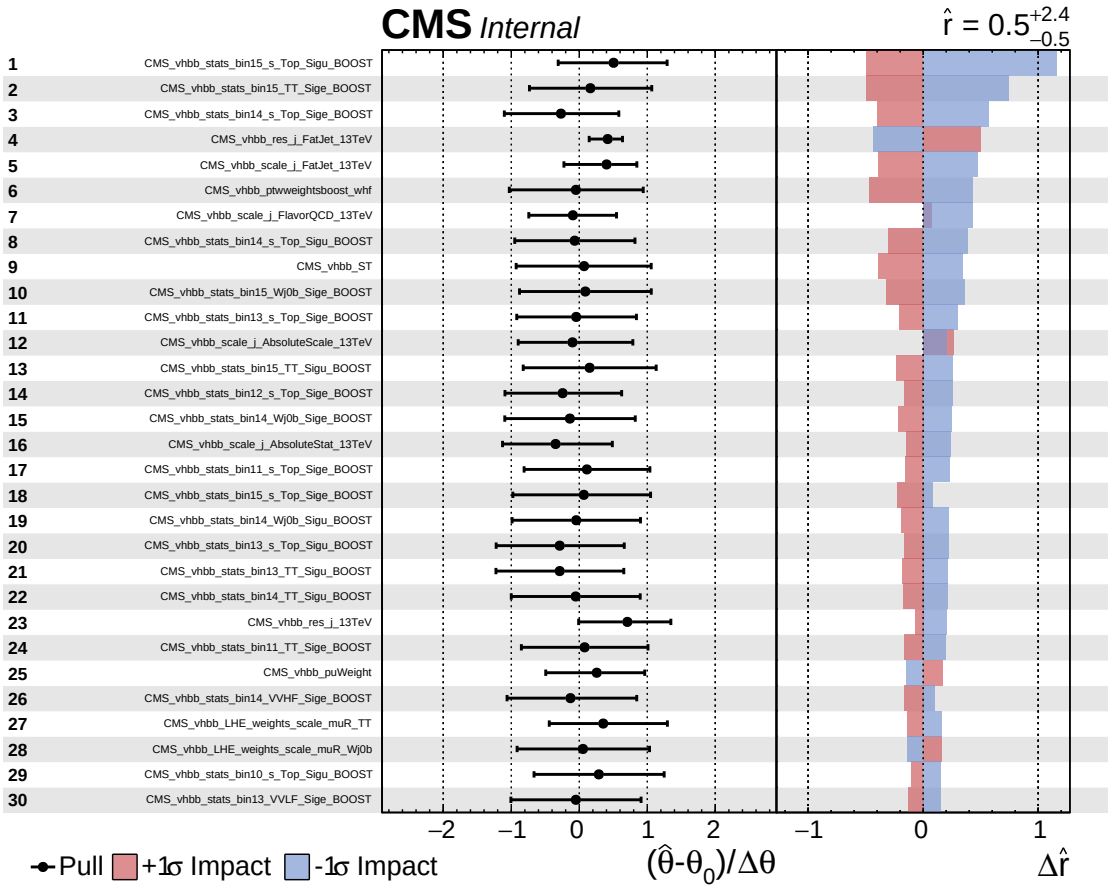


Figure 11.3: Highest 30 impact for the boosted W(lv)H(bb) analysis in the overlap resolved scenario. The first column lists the name of the nuisance parameters, ranked according to the impacts on the signal strength uncertainty (third column). The second column lists the pulls and ratios, represented similarly as in Figure 11.1. In the third column, the impacts on the signal strength corresponding the $+1\sigma$ (red) and -1σ (blue) are shown.

position and the simulated sample modeling from the W boson correction from the W + 2 b and single top category (section 10.2.3.1) in the sixth position. An asymmetry can be observed in the uncertainty of the pull for the PUPPI softdrop mass resolution in the fourth position. This might be caused by the treatment of this systematic in the smearing procedure as described in section 10.3.3, leading to an asymmetry in the nuisance uncertainty, although this effect has not been directly verified. This can be treated by using the hybrid method to perform the smearing procedure (see section 6.4.4).

11.3 Results

The results of a standalone boosted W(lv)H(bb) analysis, that is without combining with the resolved W(lv)H(bb) analysis, are given below in both the overlap resolved and overlap boosted scenario (scenario 1 and 2 in Figure 10.13, respectively). This is followed by an estimation of the sensitivity of a boosted + resolved W(lv)H(bb) analysis combination in both the overlap resolved and overlap boosted scenario (scenario 3 and 4 in Figure 10.13, respectively).

11.3.1 Standalone boosted W(lv)H(bb) analysis

A final binned-likelihood fit is performed on data separately for the overlap resolved and overlap boosted scenario in the boosted W(lv)H(bb) analysis. The double-b tagger score and BDT output score distributions from the boosted W(lv)H(bb) control and signal regions are displayed in Figures 11.4 and 11.5 for the overlap boosted and overlap resolved approach, respectively. Both systematic and statistical components are included in the Monte-Carlo uncertainties. The agreement between data and Monte-Carlo distributions is good for the double-b tagger and the BDT score distributions in both scenarios. As it can be seen by comparing the first three rows between Figure 11.4 and Figure 11.5, the double-b tagger distributions have a lower amount of bins in the overlap resolved case than in the overlap boosted case. This different binning choice motivated by the lower number of events entering the boosted W(lv)H(bb) analysis in the overlap resolved scenario, as about 63% of the background events are attributed to the resolved category in this case, as discussed below. Using a lower amount of bins for the double-b tagger distribution in the boosted W(lv)H(bb) control regions allows to reduce the statistical uncertainty in the boosted resolved scenario.

The total numbers of events in the four most sensitive bins of the BDT output score distributions in the signal regions (overlap boosted scenario) are shown in Table 11.4. The most important background contribution in this sensitive region comes from $t\bar{t}$ processes (36%), followed by W + b (22.9%) and single top (18%) processes.

The signal strength of the boosted W(lv)H(bb) analysis in the overlap boosted scenario is $0.5^{2.4}_{-0.5}$, where the total uncertainty includes both the statistical and systematic contributions. The asymmetry in the signal strength's uncertainty is due to the requirement of signal strength above 0 during the fit. The observed (expected) significance is $0.18\,\sigma$ ($0.42\,\sigma$).

For the overlap resolved scenario, the signal strength is $0.0^{3.4}_{-0.0}$. As in the boosted scenario, the asymmetry in the signal strength's uncertainty is due to the requirement a signal strength above 0 during the fit. The observed and expected significances are 0σ, corresponding to no observed excess in the data, and $0.17\,\sigma$.

Those results are summarized in Table 11.3. The performance of the boosted W(lv)H(bb) analysis is lower in the overlap resolved than in the overlap boosted scenario due to

Strategy	Expected significance	Observed significance	Signal strength
Overlap resolved	$0.18\,\sigma$	$0.00\,\sigma$	$0.0^{+3.4}_{-0.0}$
Overlap boosted	$0.42\,\sigma$	$0.21\,\sigma$	$0.5^{+2.4}_{-0.5}$

Table 11.3: Observed and expected significances and signal strengths in the boosted W(lv)H(bb) analysis shown for the overlap boosted and overlap resolved scenarios.

Process	Event yield
Z + 2 b jets	0.19
Z + 1 b jet	0.15
Z + light jets	0
W + 2 b jets	22.66
W + 1 b jet	8.1
W + light jets	8.96
$t\bar{t}$	34.94
single top	17.39
diboson + light jets	6.66
diboson + 2 b jets	1.35
Total background	98.38
Signal (VH(bb))	5.18
S/B	0.053

Table 11.4: The total number of events in the three most sensitive BDT bins for the boosted W(lv)H(bb) channel in the overlap boosted scenario. The S/B in the lowest row corresponds to the expected VH(bb) signal yield divided by the expected background yield.

the loss of events selected by the resolved W(lv)H(bb) analysis.

11.3.2 Combination with the resolved W(lv)H(bb) analysis

The boosted W(lv)H(bb) analysis can be combined to the resolved W(lv)H(bb) analysis. The expected significance is evaluated in both the overlap resolved and overlap boosted scenario to estimate which treatment of the overlapping events is the most optimal.

The postfit CMVAv2$_{min}$ and BDT output distributions from a resolved W(lv)H(bb) analysis are shown in Figure 11.6 and Figure 11.7 for the overlap boosted combination scenario and in Figures 11.8 and Figure 11.9 for the overlap resolved combination scenario. In each case, the fit is performed on the control and signal regions of the resolved W(lv)H(bb) analysis with the corresponding treatment of overlapping events.

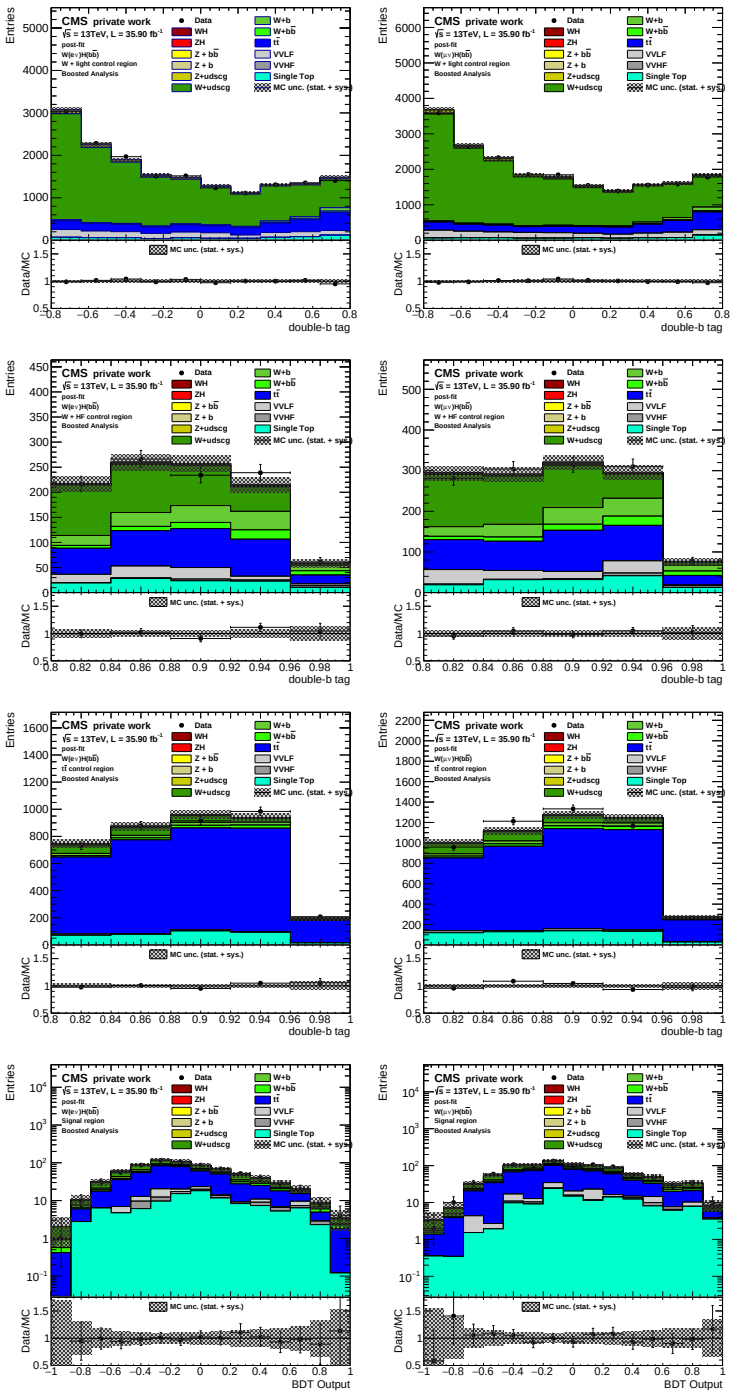


Figure 11.4: Postfit double-b tagger and BDT score distributions in the control and signal region for the boosted W(lv)H(bb) analysis in the overlap boosted scenario. The left and right column correspond to the W(ev)H(bb) and W($\mu\nu$)H(bb) sub-channel, respectively. The double-b tagger distributions in the first, second and third row correspond to the W + light, W + HF and $t\bar{t}$ control region, respectively. The BDT output score distributions in the third row corresponds to the signal region, where data has been removed in the four last BDT bins.

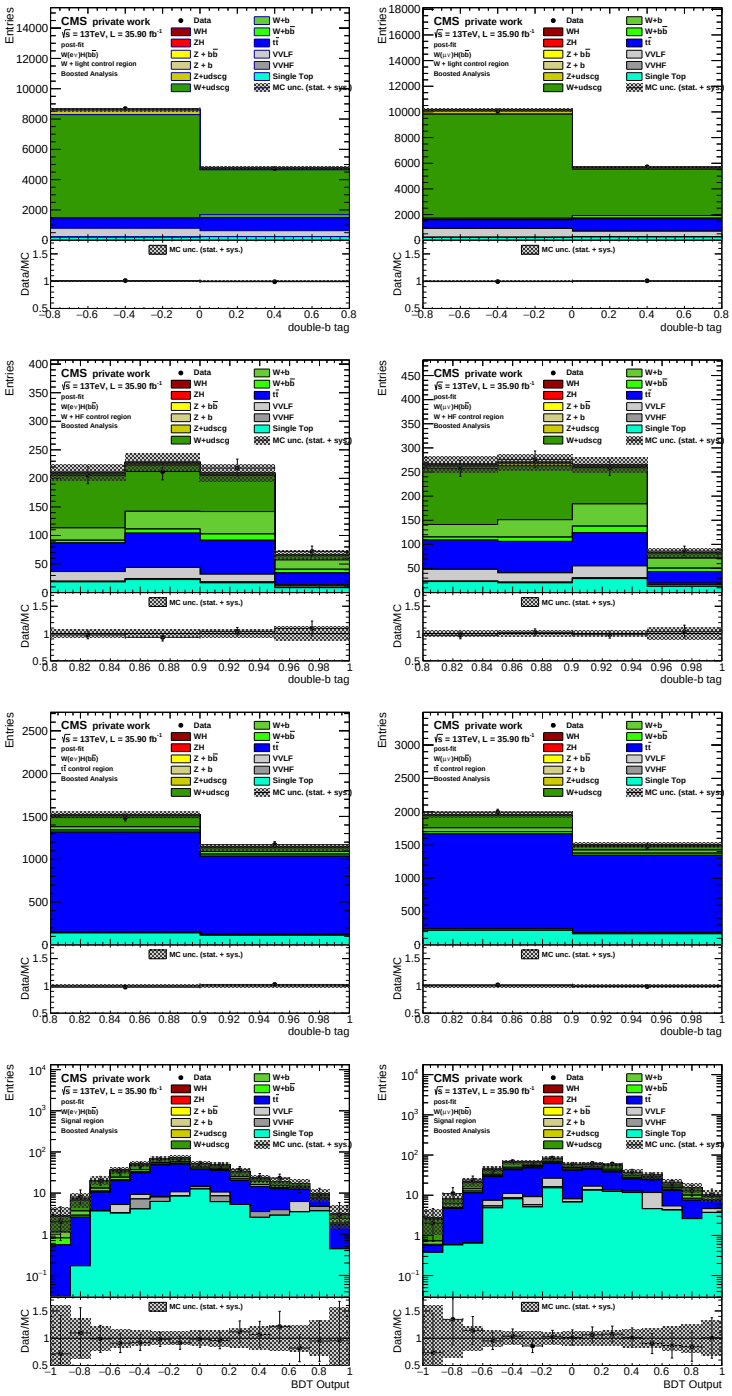


Figure 11.5: Postfit double-b tagger and BDT score distributions in the control and signal region for the resolved W(lv)H(bb) analysis in the overlap resolved scenario. The left and right column correspond to the W(ev)H(bb) and W($\mu\nu$)H(bb) sub-channel, respectively. The double-b tagger distributions in the first, second and third row correspond to the W + light, W + HF and $t\bar{t}$ control region, respectively. The BDT output score distributions in the third row corresponds to the signal region, where data has been removed in the four last BDT bins.

The data has been removed from the last three bins of the BDT score distribution in the signal regions of the resolved W(lv)H(bb) analysis to keep the results *blinded* (without looking at the data in the sensitive part of the signal region), as the goal of this study is to extract the expected Asimov sensitivity only. The BDT output score distributions start at 0.5 due to the signal region BDT cut in the signal region definition (see Table 10.5). Those postfit plots are derived to validate the Monte-Carlo modeling of the data in the resolved W(lv)H(bb) analysis prior to the expected Asimov sensitivity estimation from the boosted + resolved W(lv)H(bb) combination. A good agreement between the data and Monte-Carlo simulated processes can be observed in the CMVAv2$_{min}$ and BDT output score in both the overlap boosted and overlap resolved approach.

The expected events yields for all processes in the last three bins from the BDT output distributions in the signal regions are listed in Table 11.5 in three orthogonal categories: resolved events in the second column, overlapping events in the third column and boosted events in the fourth column. Those three categories can be visualized by the Venn diagrams in the lower part of Figure 10.13. The second column corresponds to all events selected by the resolved analysis in an overlap boosted scenario. The third column corresponds to the overlapping events that could be selected by both the resolved and the boosted analysis. The fourth column corresponds to all events selected by the boosted analysis in an overlap resolved scenario. Removing the boosted events from the resolved W(lv)H(bb) analysis correspond to a 9% and 22% drop of signal and background events, respectively. When removing resolved events from the boosted W(lv)H(bb) analysis, this corresponds to a 58% and a 36% drop of signal and background events. The performance of the boosted W(lv)H(bb) is therefore strongly affected when removing the overlapping events, as it is shown from the Asimov significance in Table 11.3, discussed below.

The Asimov significance of the boosted and resolved W(lv)H(bb) analysis are summarized in Table 11.6 for the overlap boosted and overlap resolved approach. In the overlap resolved scenario, new events brought by the boosted W(lv)H(bb) give a negligible improvement on the expected significance from the resolved W(lv)H(bb) analysis, which goes from $1.41\,\sigma$ to $1.43\,\sigma$. In the overlap boosted scenario, the removal of the overlapping events reduce the expected significance of the resolved analysis from 1.41σ to 1.27σ and improves the performance of the boosted analysis from $0.18\,\sigma$ to $0.42\,\sigma$. The corresponding expected significance from this combination is 1.32σ, which is lower than in the overlap resolved scenario.

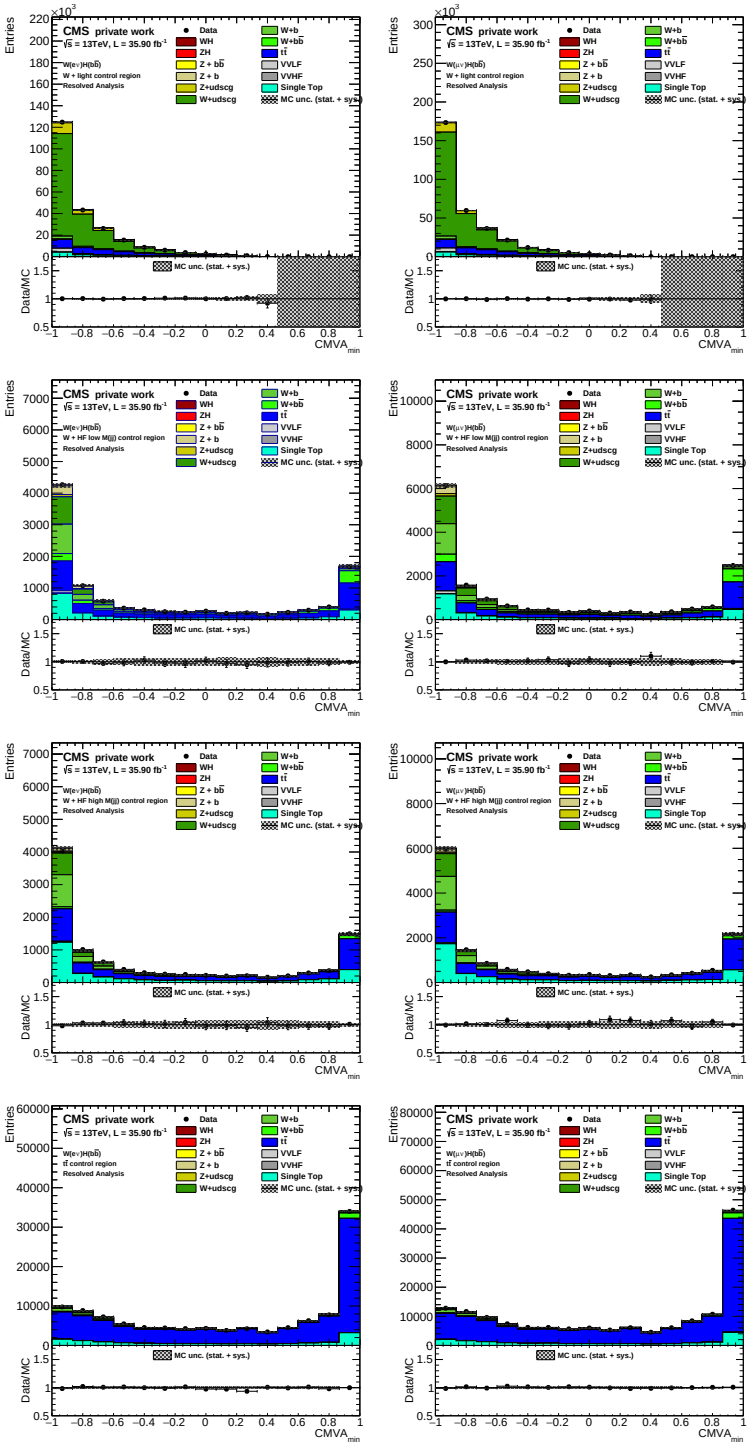


Figure 11.6: Postfit CMVA_{min} distributions from the resolve W(lv)H(bb) analysis in the overlap boosted scenario. The left column and right column correspond to the W(ev)H(bb) and W(μv)H(bb) sub-channel, respectively. The first, second, third and fourth row correspond to the W + light control, W + HF control region (low dijet mass), W + HF control region (high dijet mass) and $t\bar{t}$ control region, respectively.

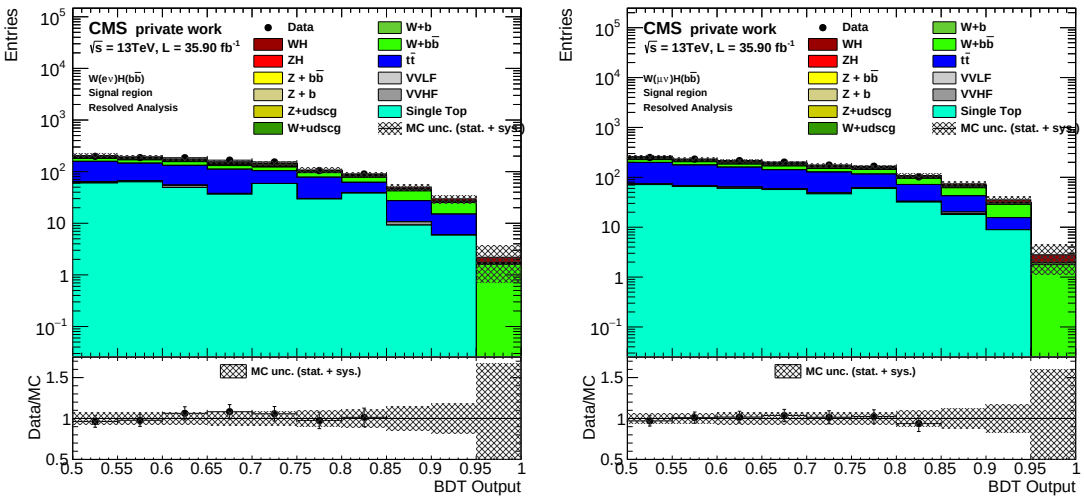


Figure 11.7: Postfit signal region BDT output score distribution from the resolved W(lv)H(bb) analysis in the overlap boosted scenario. The VH(bb) signal shape in the distribution is taken before the fit with a signal strength fixed to 1 and the data is removed in the three last bin. The left and right plot corresponds to the W(ev)H(bb) sub-channel and W(μv)H(bb) sub-channel, respectively.

Process	Resolved	Overlap	Boosted
Z + 2 b jets	44.36	1.37	2.15
Z + 1 b jet	16.34	2.83	3.51
Z + light jets	9.36	0.37	7.14
W + 2 b jets	287.12	77.88	57.81
W + 1 b jet	84.30	46.05	105.62
W + light jets	84.07	106.05	186.64
$t\bar{t}$	1249.64	310.23	542.70
single top	482.91	65.74	163.30
diboson + light jets	7.59	22.93	50.17
diboson + 2 b jets	24.44	9.07	15.53
Total background	2357.47	648.53	1139.41
Signal (VH(bb))	67.33	6.75	4.83
S/B	0.029	0.01	0.004

Table 11.5: Event yields in the resolved, boosted and overlapping categories in the W(lv)H(bb) channel. The three selections in the table are orthogonal, the resolved and boosted category do not contain the overlapping events. The S/B in the lowest row corresponds to the expected VH(bb) signal yield divided by the expected background yield.

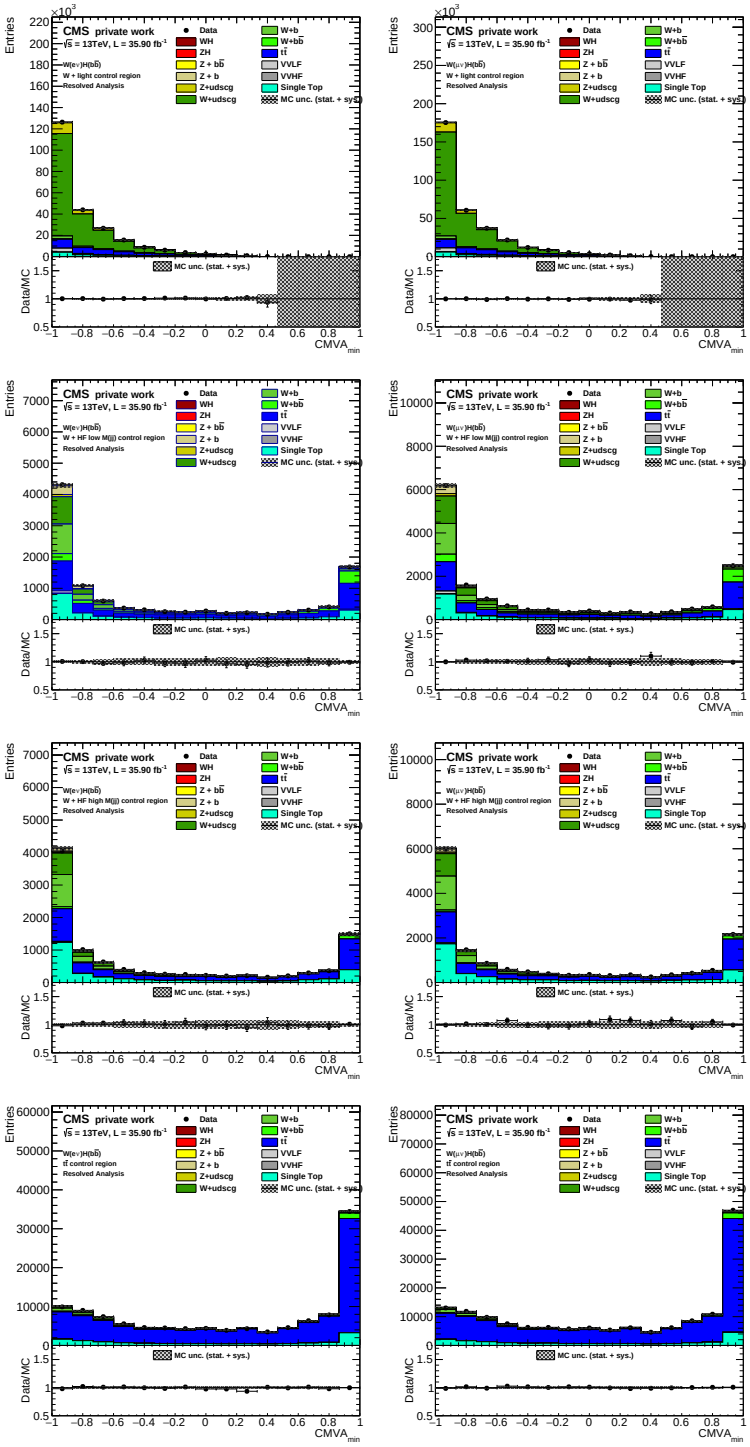


Figure 11.8: Postfit CMVA_{min} distributions from the resolve W(lv)H(bb) analysis in the overlap resolved scenario. The left column and right column correspond to the W(ev)H(bb) and W(μv)H(bb) sub-channel, respectively. The first, second, third and fourth row correspond to the W + light control, W + HF control region (low dijet mass), W + HF control region (high dijet mass) and $t\bar{t}$ control region, respectively.

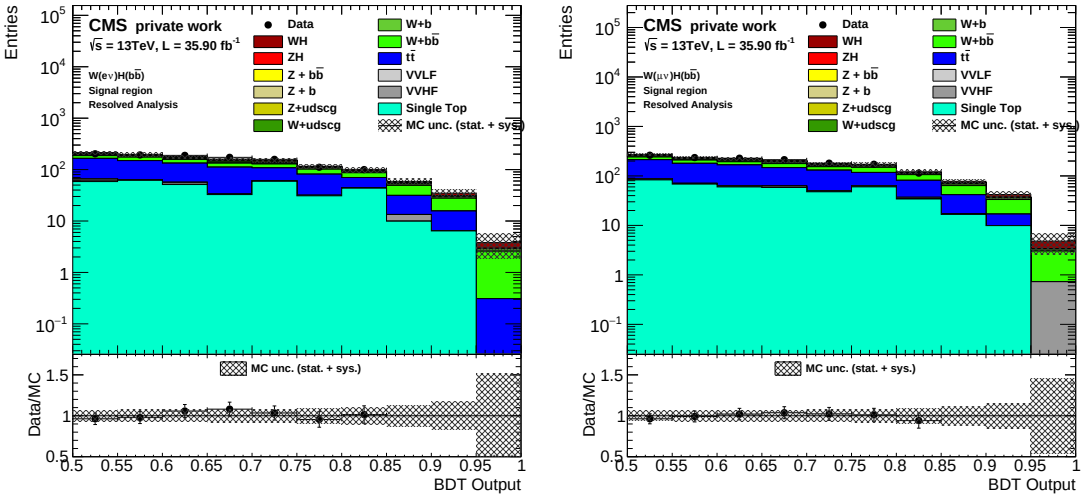


Figure 11.9: Postfit signal region BDT output score distribution from the resolve W(lv)H(bb) analysis in the overlap resolved scenario. The VH(bb) signal shape in the distribution is taken before the fit with a signal strength fixed to 1 and the data is removed in the three last bin. The left and right plot corresponds to the W(ev)H(bb) sub-channel and W($\mu\nu$)H(bb) sub-channel, respectively.

Strategy	Resolved significance	Boosted significance	Boost Analysis
As in 2016 VH(bb) analysis	$1.41\,\sigma$	-	$1.41\,\sigma$
Overlap resolved	$1.41\,\sigma$	$0.18\,\sigma$	$1.43\,\sigma$
Overlap boosted	$1.27\,\sigma$	$0.42\,\sigma$	$1.32\,\sigma$

Table 11.6: Expected significance for the resolved and boosted analysis for both treatment of the overlapping events.

12
Conclusion

A search for a Higgs boson decaying into two bottom quarks associated with a Z or W vector boson decaying leptonically, referred to as the VH(bb) process, was performed on the $35.9\,\mathrm{fb}^{-1}$ 2016 dataset recorded by the CMS detector at a center-of-mass energy of $13\,\mathrm{TeV}$. The analysis presented in this thesis focused on the $Z \to l\bar{l}$ channel, in which the associated vector boson is a Z decaying into an electron or muon pair.

The selection of the lepton pair by a double-lepton trigger and additional isolation and identification requirements on each lepton are used to completely remove the contribution from the QCD multijet background in this channel. Differences in efficiency between the data and Monte-Carlo simulated processes due to those lepton requirements have been addressed by scale factors corrections, whose measurement for the muon case is part of the contribution presented in this thesis. A signal region is defined in a range of [90, 150 GeV] around $125.1\,\mathrm{GeV}$ invariant mass of the Higgs boson candidate by making use of kinematic variables related to the event topology, mass of the vector boson and the b-tagger score from each of the two b-jet candidates attributed to the Higgs boson decay that estimate how likely is the jet to originate from a b quark. In this signal region, the discrimination between the VH(bb) signal and background processes, both simulate by Monte-Carlo event generators, is performed by a Boosted Decision Tree trained on Monte-Carlo simulations. A good understanding of the background processes is crucial for the performance of this analysis. Three control regions defined orthogonally to the signal region, respectively enriched in $t\bar{t}$, W + additional light jets, and W + one or two b-tagged jets, are dedicated to study the modeling of the main background processes and extract normalization corrections for the aforementioned simulated backgrounds. Other backgrounds contributing to this analysis are the diboson (ZZ(bb)) and single top processes. The signal strength and significance with respect to the background-only hypothesis are extracted with a fi-

nal binned-likelihood fit performed simultaneously on the control and signal regions of the $Z \to l\bar{l}$ channel and two other complementary analysis channels contributing to the VH(bb) search: the $W \to l\overline{\nu}$ channel, targeting a W boson decaying to an electron-neutrino or muon-neutrino pair, and the $Z \to \nu\overline{\nu}$ channel, targeting a Z boson decaying into a neutrino pair.

As the diboson background presents the same final state as the VH(bb) process, a measurement of the diboson process is performed with the same approach as for the VH(bb) search to validate the analysis methodology. The result of the this diboson analysis corresponds to an observed (expected) significance of $4.5\,\sigma$ ($3.2\,\sigma$) deviations with respect to the background-only hypothesis in the $Z \to l\bar{l}$ channel. The corresponding signal strength is 1.33 ± 0.34. When combining with the $W \to l\overline{\nu}$ and $Z \to \nu\overline{\nu}$ channel, the observed (expected) significance is $5\,\sigma$ and $4.9\,\sigma$ with a corresponding VZ(bb) diboson signal strength of 1.02 ± 0.22.

For the VH(bb) analysis conducted in the $Z \to l\bar{l}$ channel, the observed (expected) significance is $3.1\,\sigma$ ($1.8\,\sigma$) deviations with respect to the background-only hypothesis. The corresponding signal strength is 1.8 ± 0.6 for a $125.09\,\mathrm{GeV}$ standard model Higgs boson. When combining with the $W \to l\overline{\nu}$ and $Z \to \nu\overline{\nu}$ channels, the observed (expected) significance is $3.3\,\sigma$ ($2.8\,\sigma$). The corresponding signal strength for the VH(bb) process is $\mu = 1.19^{+0.21}_{-0.20}(\mathrm{stat.})^{+0.34}_{-0.32}(\mathrm{syst.})$ for $125.09\,\mathrm{GeV}$ standard model Higgs boson, where stat. and sys. refer to the statistical and systematic uncertainties, respectively. The most important source of systematic uncertainties are coming from the Monte-Carlo normalization corrections and the limited size of the Monte-Carlo samples. The results mentioned above are compatible with a $125.9\,\mathrm{GeV}$ standard model Higgs boson.

The aforementioned VH(bb) measurement in the $Z \to l\bar{l}$ channel has importantly contributed to the Higgs boson search at CMS. It was included in a combination between the VH(bb) searches performed on the Run 1 dataset ($5.1\,\mathrm{fb}^{-1}$ at $7\,\mathrm{TeV}$ + $18.9\,\mathrm{fb}^{-1}$ at $8\,\mathrm{TeV}$), 2016 dataset ($35.9\,\mathrm{fb}^{-1}$ at $13\,\mathrm{TeV}$) and the 2017 dataset ($41.3\,\mathrm{fb}^{-1}$ at $13\,\mathrm{TeV}$) recorded by CMS. This combination results in an observed (expected) significance of $4.8\,\sigma$ ($4.9\,\sigma$) with respect to the background-only hypothesis and a signal strength of 1.01 ± 0.23 for the VH(bb) process with a $125.09\,\mathrm{GeV}$ standard model Higgs boson. A combination between this results and other CMS searches for the $H \to \bar{b}b$ final state has lead to the first observation of the $H \to \bar{b}b$ decay by the CMS Collaboration with an observed and expected significance of $5.5\,\sigma$ and $5.6\,\sigma$ with a corresponding signal strength of $\mu = 1.04 \pm 0.14(\mathrm{stat.}) \pm 0.14(\mathrm{syst.})$. This result is compatible with the standard model predictions for a $125.9\,\mathrm{GeV}$ standard model Higgs boson.

In addition to the ZH(bb) analysis presented in this thesis, an extension of the VH(bb) search was performed in a boosted event topology, in which both the Higgs and vector boson are required to have a transverse momentum above $250\,\mathrm{GeV}$, was conducted in the $W \to l\overline{\nu}$ channel. This boosted W(lv)H(bb) analysis investigated the use of technologies dedicated for such boosted regime, such as using a fat AK08 jet for the Higgs boson candidate reconstruction, the double-b tagger, the PUPPI soft drop mass and jet substructure variables, with an analysis methodology similar to the one VH(bb) mentioned above.

This first boosted analysis was less sensitive with respect to the VH(bb) analysis mentioned previously, with an observed (expected) signal strength of $0.21\,\sigma$ ($0.42\,\sigma$) deviation with respect to the background-only hypothesis. The corresponding signal strength is $0.5^{+2.38}_{-0.50}$ for a WH(bb) process with a $125.09\,\mathrm{GeV}$ standard model Higgs boson. It however sets a first baseline for future VH(bb) analysis investigating the boosted regime and has the potential of benefiting from notable improvements, such as using the most recent double-b tagging methods like the DeepAK08 algorithm [94], which improves the double-b tagger background rejection by 40% with respect to the double-b tagger algorithm.

Part IV

Appendices

A
Appendix A

A.0.1 Efficiency distributions

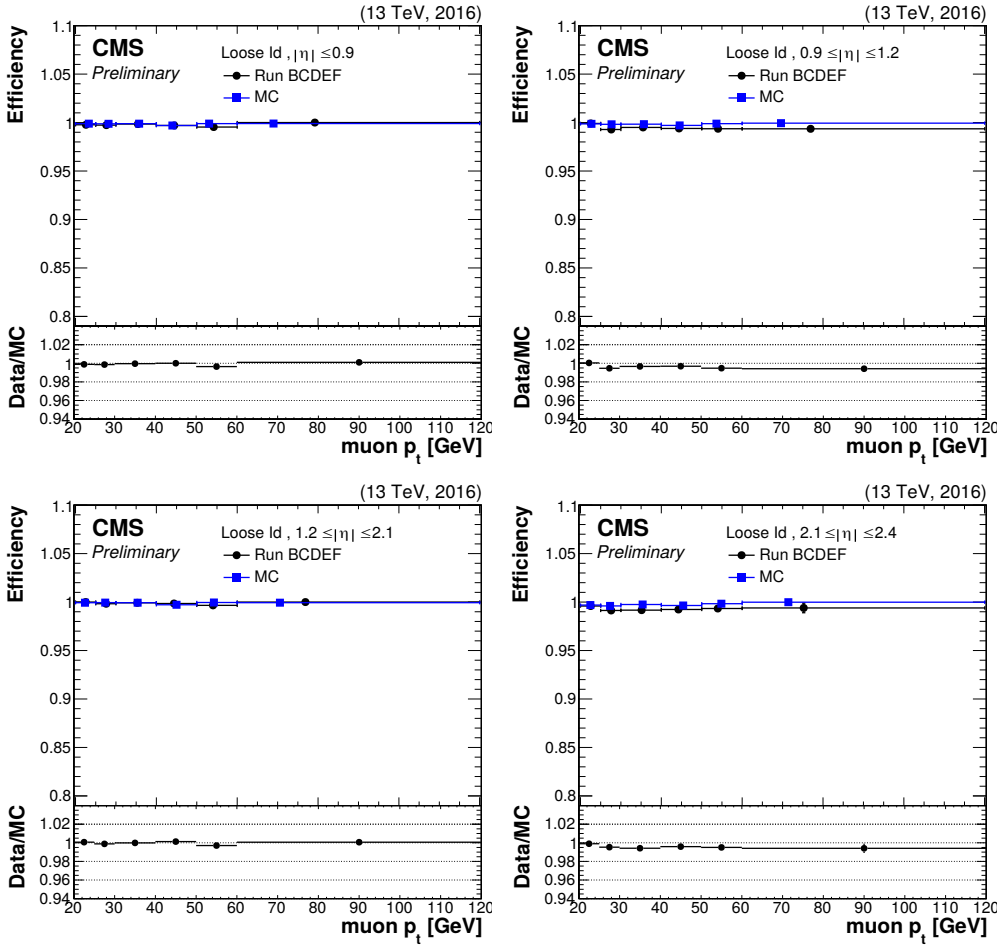


Figure A.1: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

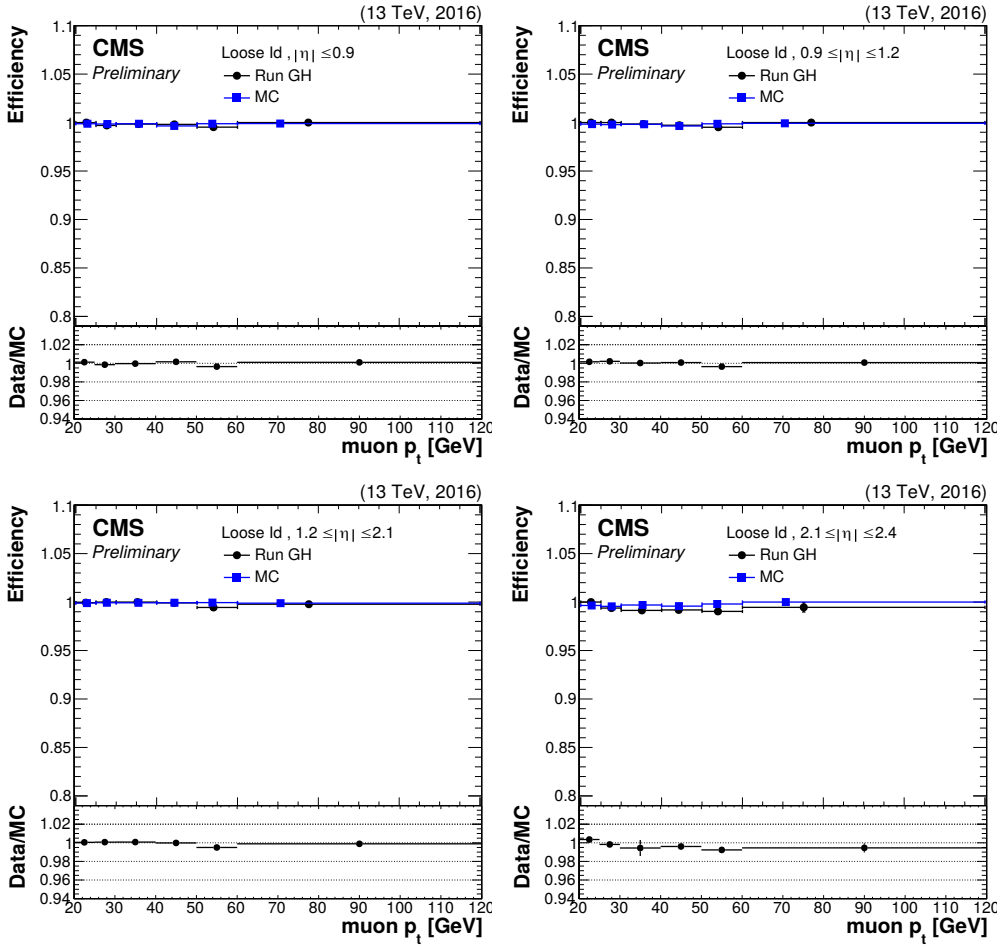


Figure A.2: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs G and H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

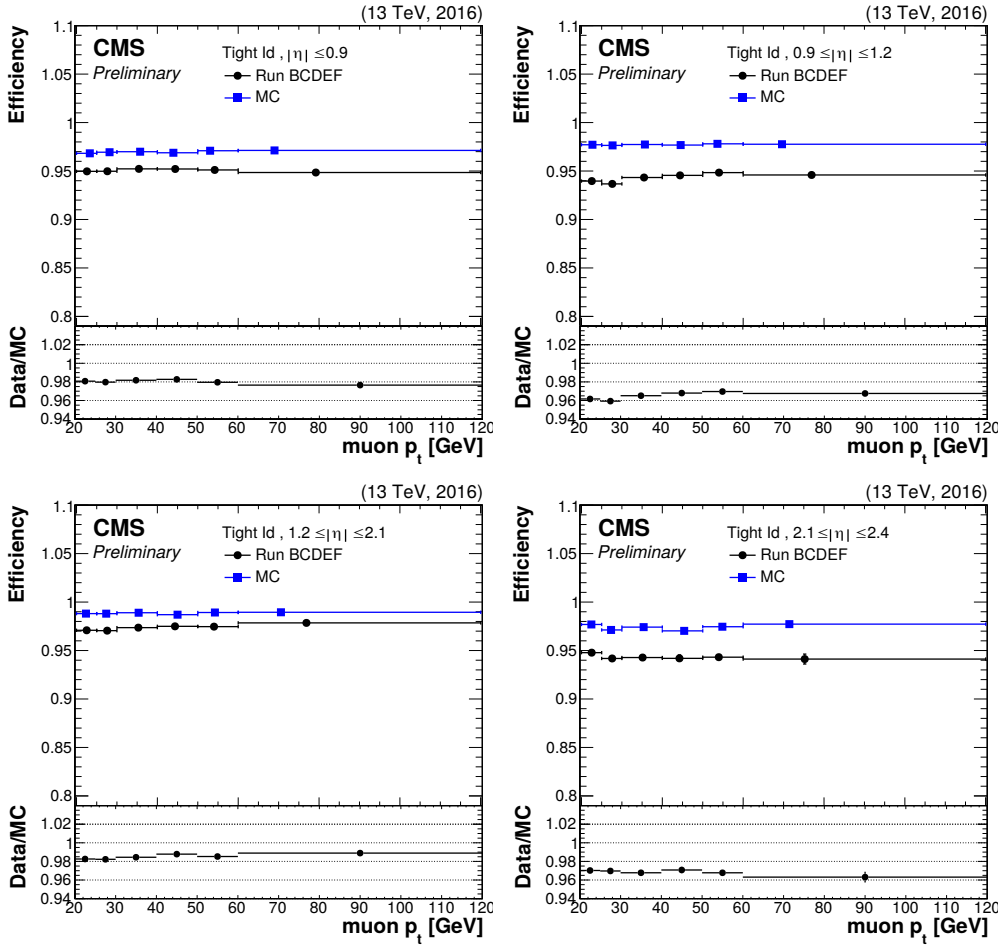


Figure A.3: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

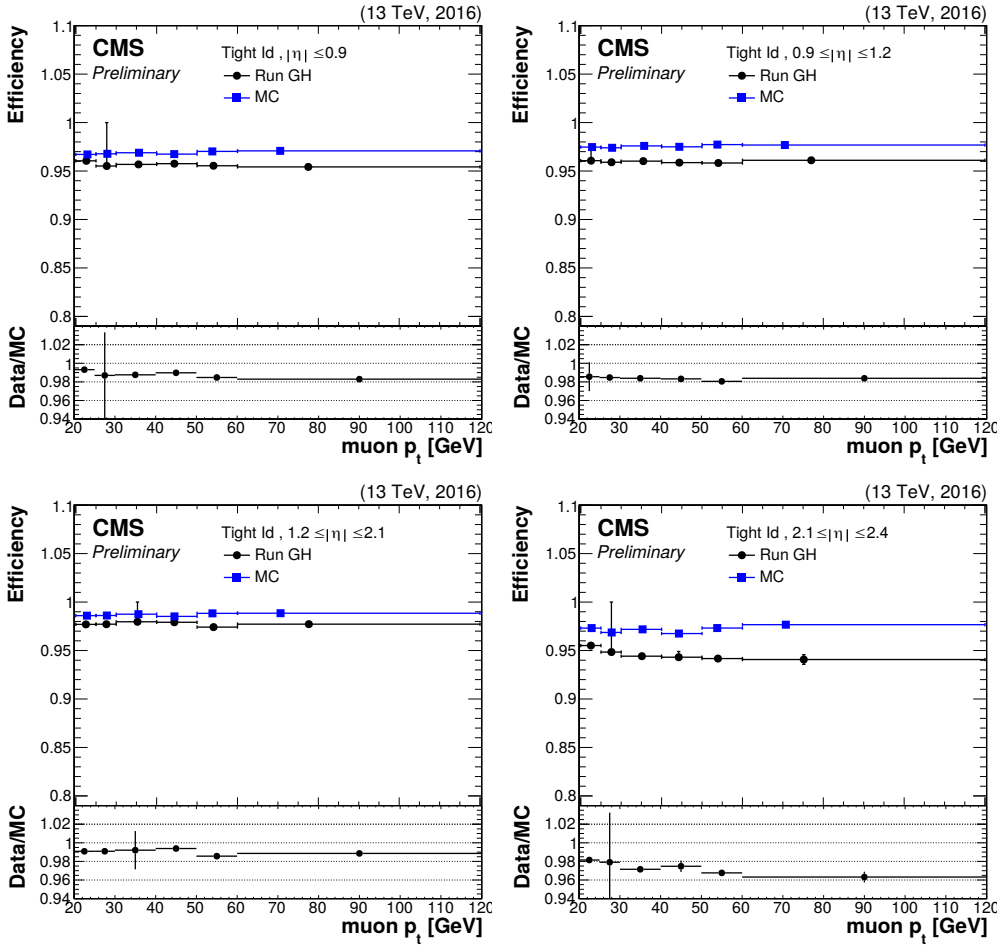


Figure A.4: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs G and H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

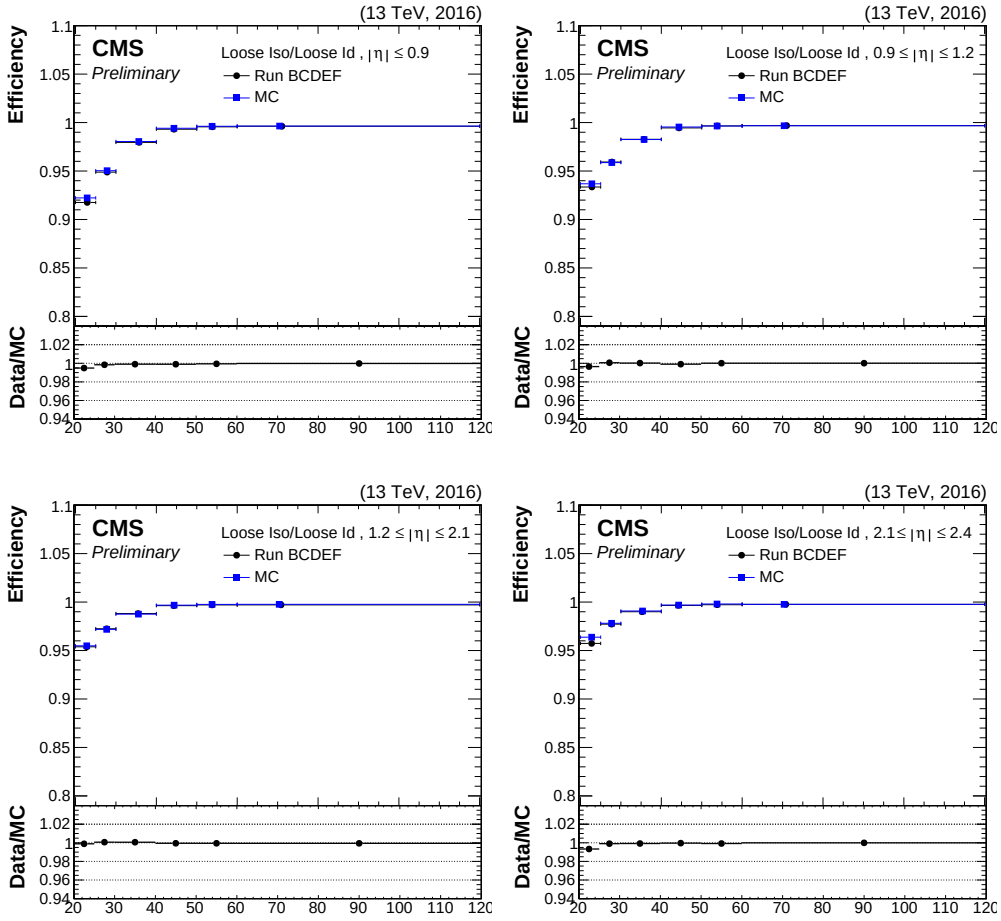


Figure A.5: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

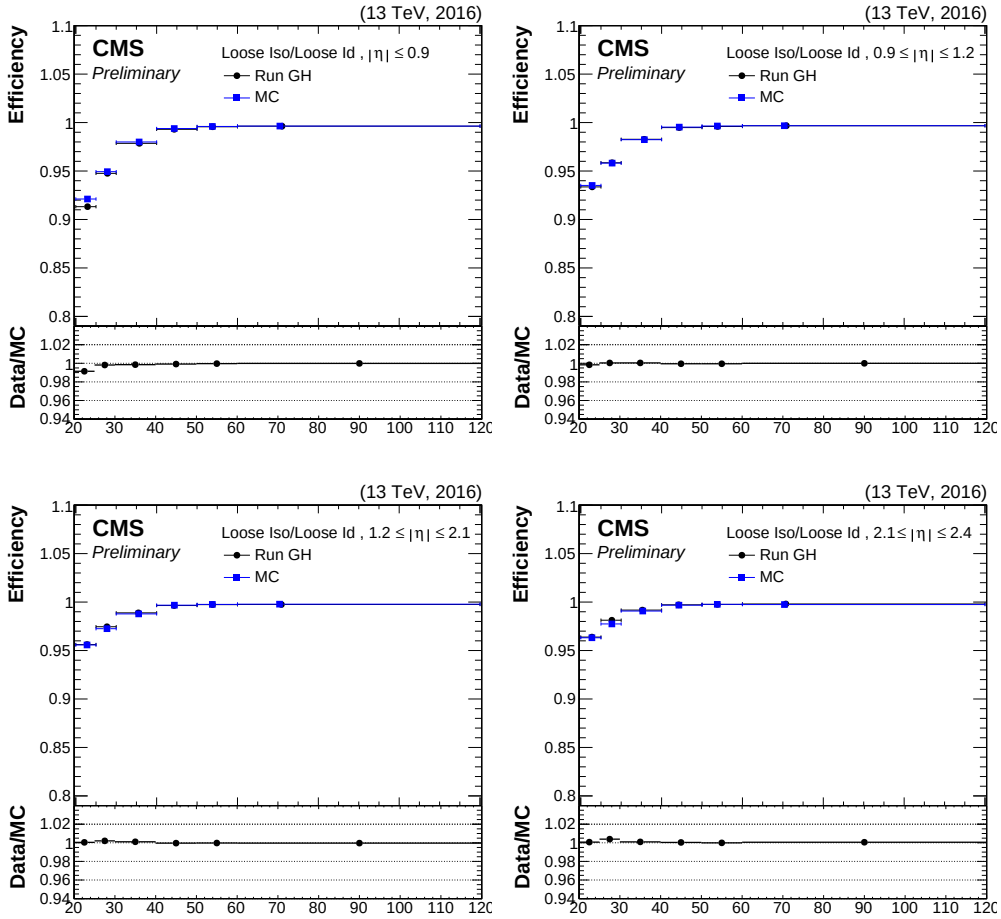


Figure A.6: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs G and H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

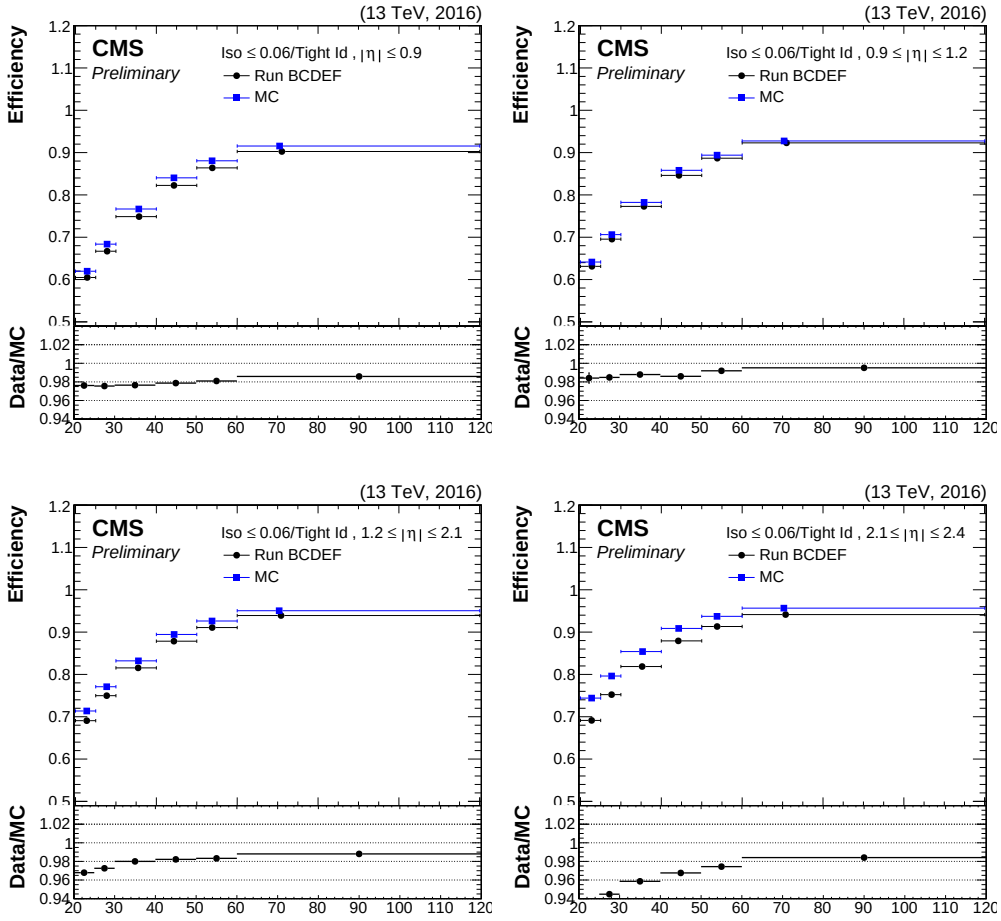


Figure A.7: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

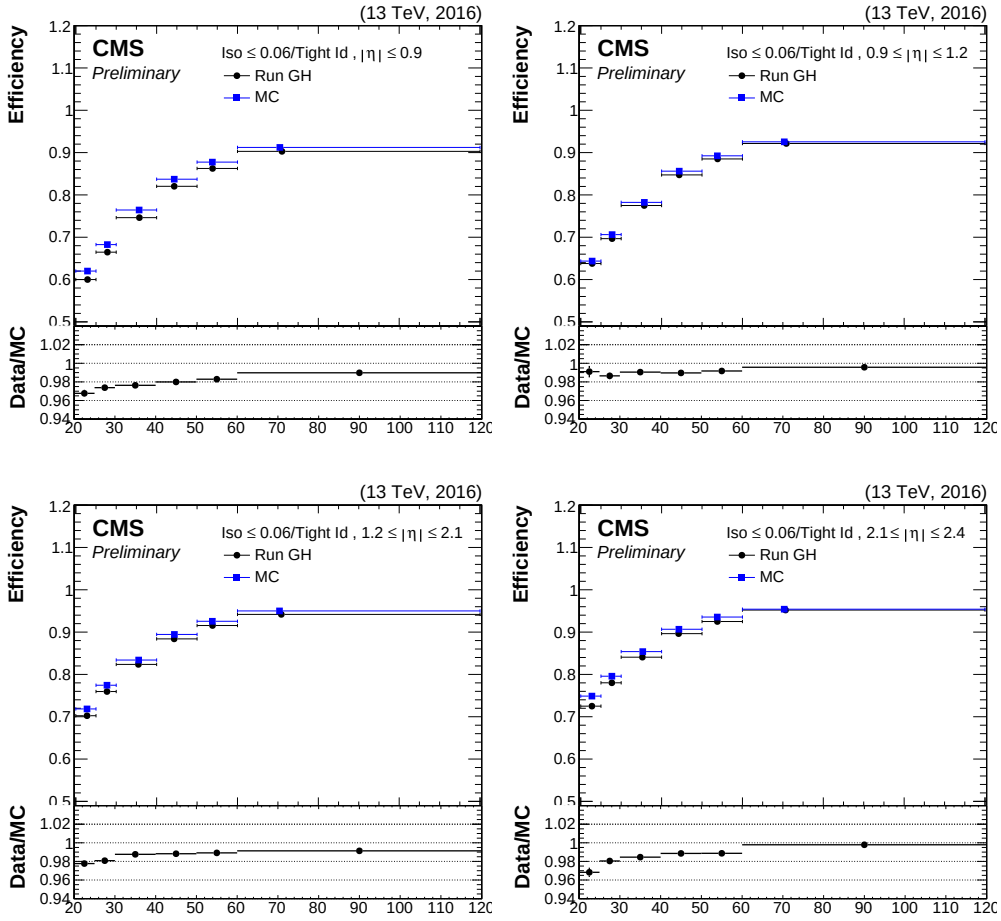


Figure A.8: Efficiency distributions of the data and DY + jets Monte-Carlo sample in the muon p_T for different bins of $|\eta|$. The data include the runs G and H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

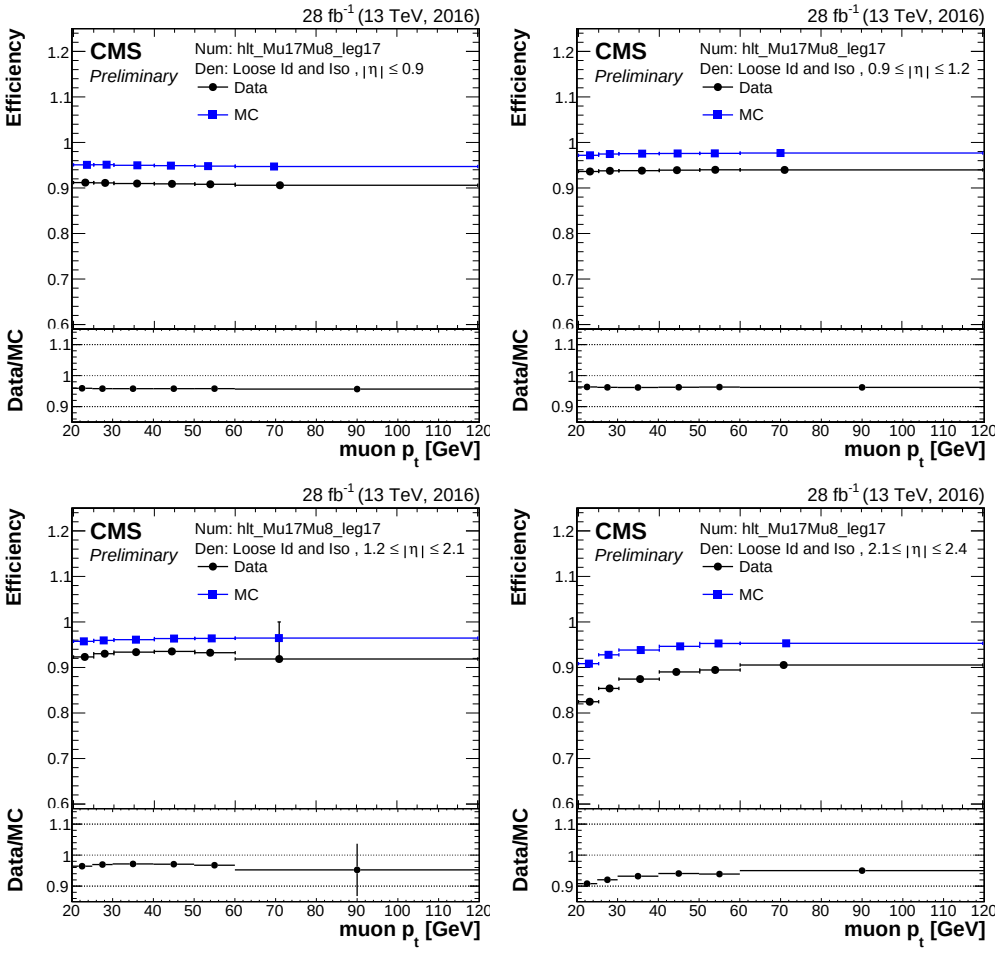


Figure A.9: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the 17 GeV muon in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

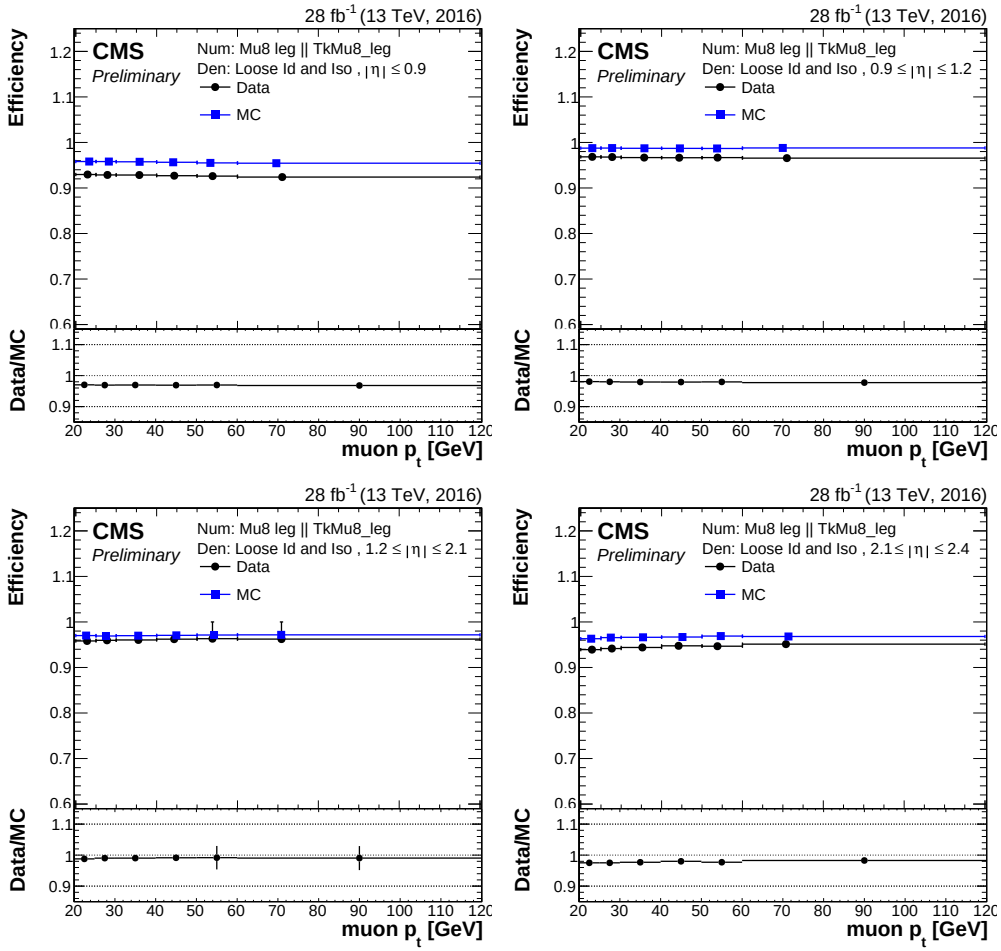


Figure A.10: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the 8 GeV muon in the muon p_T for different bins of $|\eta|$. The data include the runs B, C, D, E and F. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

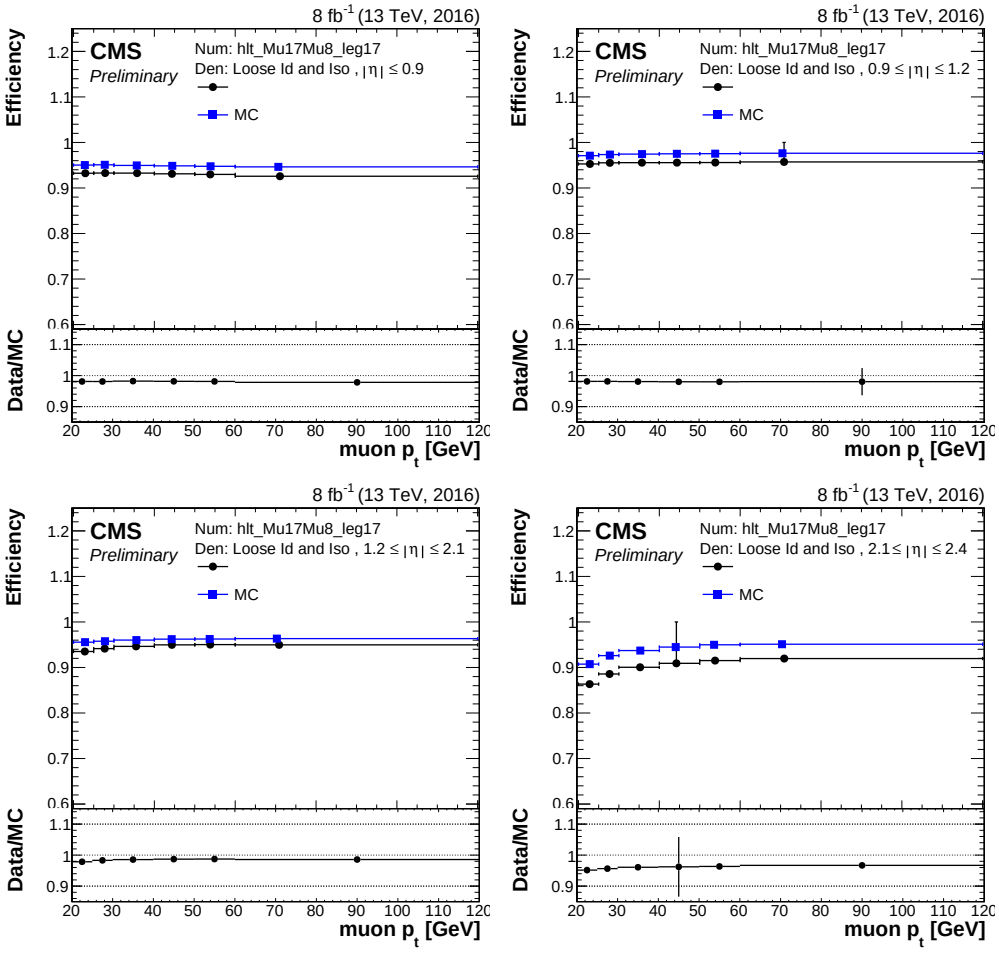


Figure A.11: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the 17 GeV muon in the muon p_T for different bins of $|\eta|$. The data include the run H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

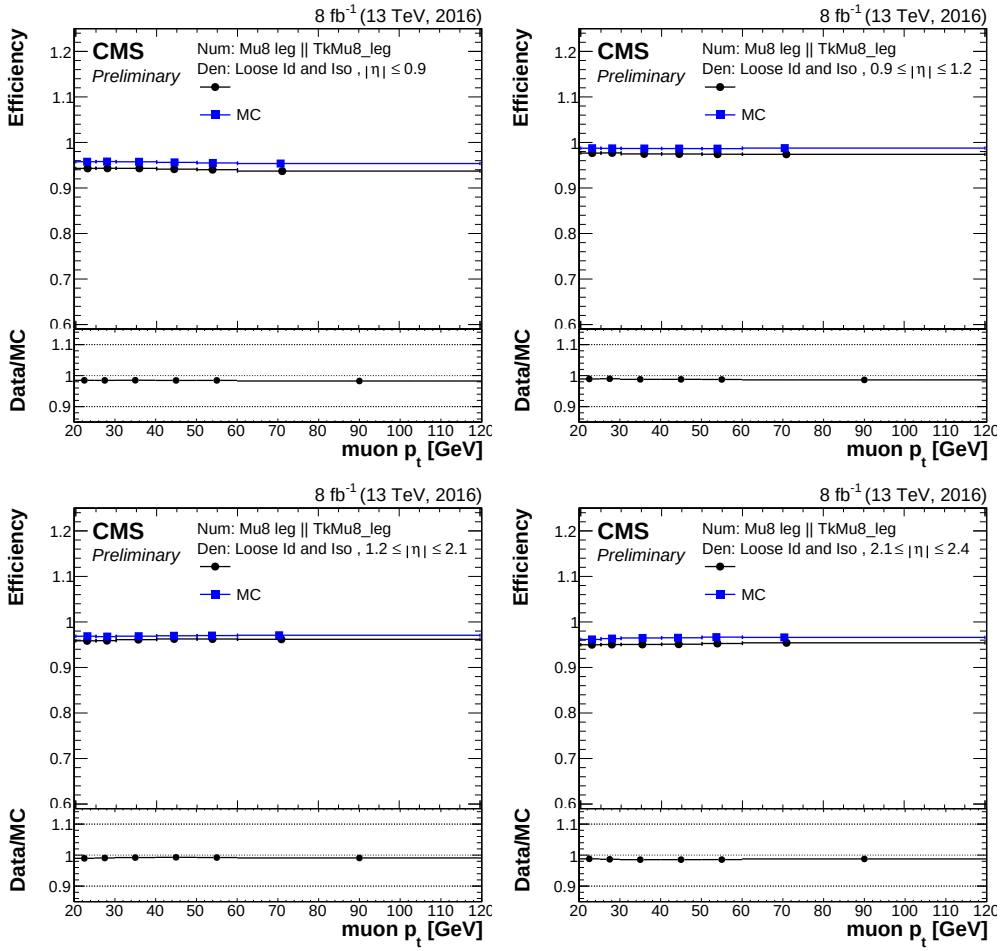


Figure A.12: Efficiency distributions of the data and DY + jets Monte-Carlo sample for the 8 GeV muon in the muon p_T for different bins of $|\eta|$. The data include the run H. The uncertainties are statistical uncertainties from the tag and probe fit. The $|\eta|$ bins are, from top to bottom and left to right, $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

B

Appendix B

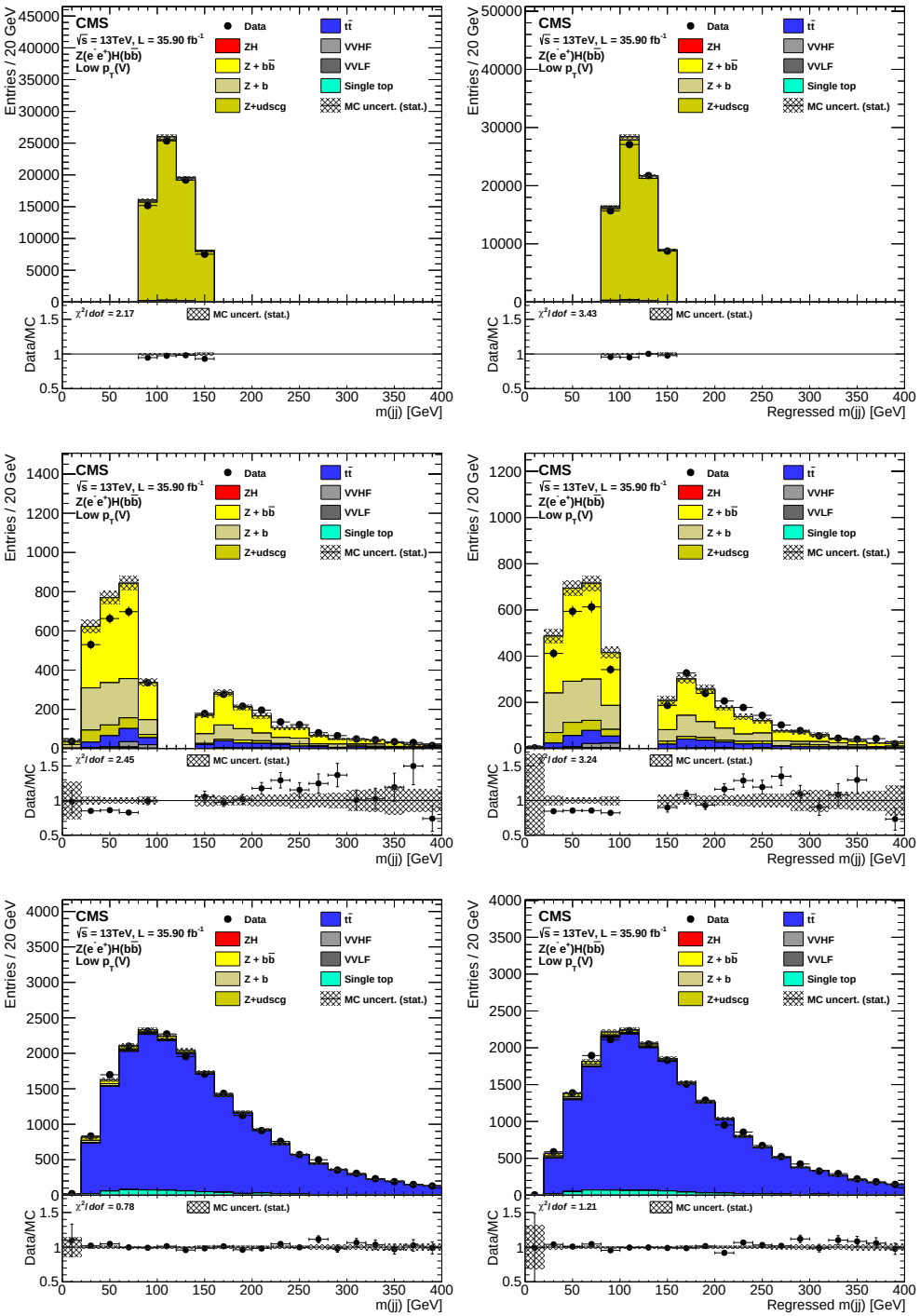


Figure B.1: Dijet invariant mass in region of the Z(ll)H(bb) enriched in background processes. **First row:** region enriched in Z + light jets processes. **Second row:** region enriched in Z + b jets processes. **Third row:** region enriched in $t\bar{t}$ processes. The dijet mass is compared with and without regression. **First column:** no regression applied. **Second column:** with regression applied.

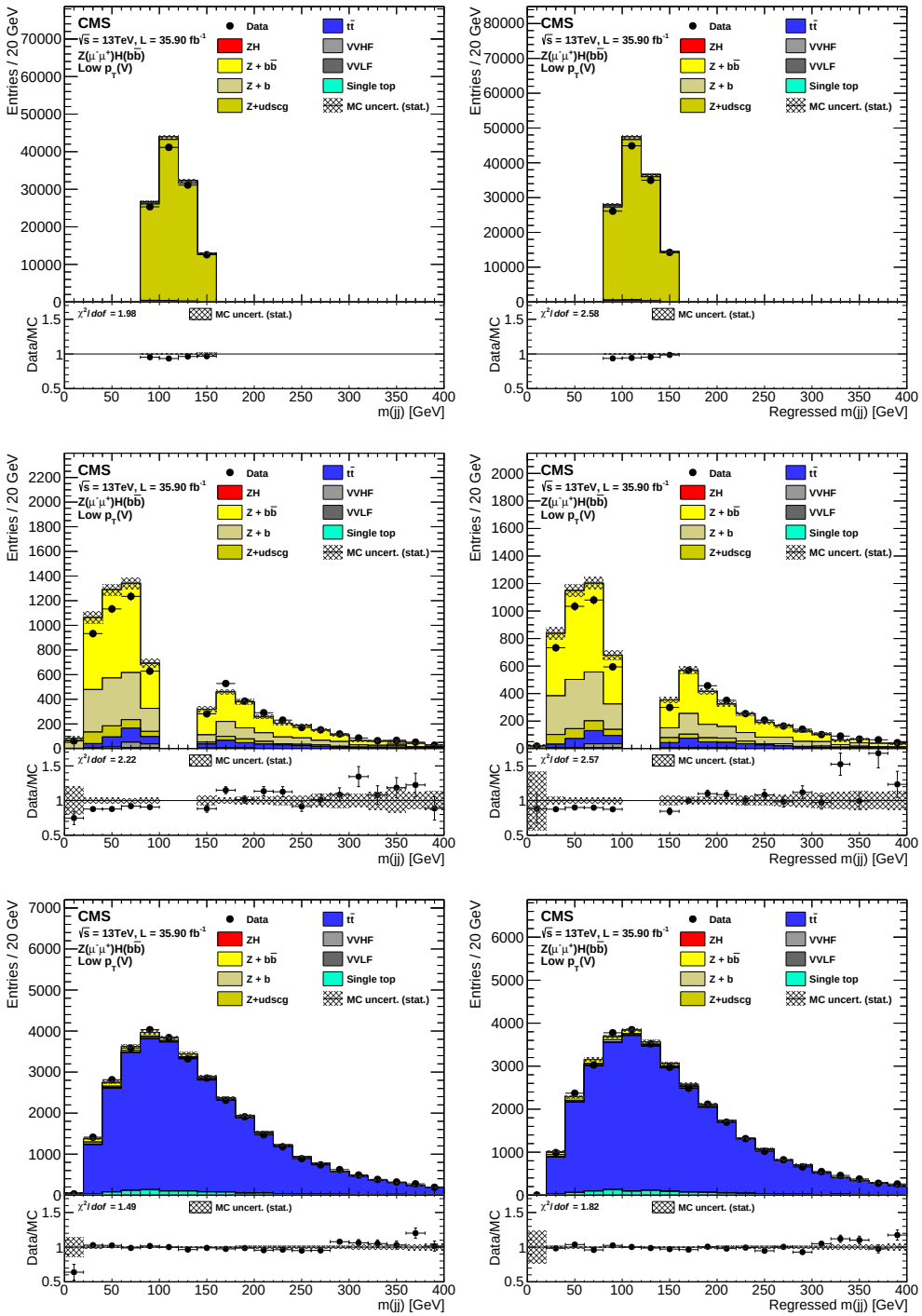


Figure B.2: Dijet invariant mass in region of the Z(ll)H(bb) enriched in background processes. **First row:** region enriched in Z + light jets processes. **Second row:** region enriched in Z + b jets processes. **Third row:** region enriched in $t\bar{t}$ processes. The dijet mass is compared with and without regression. **First column:** no regression applied. **Second column:** with regression applied.

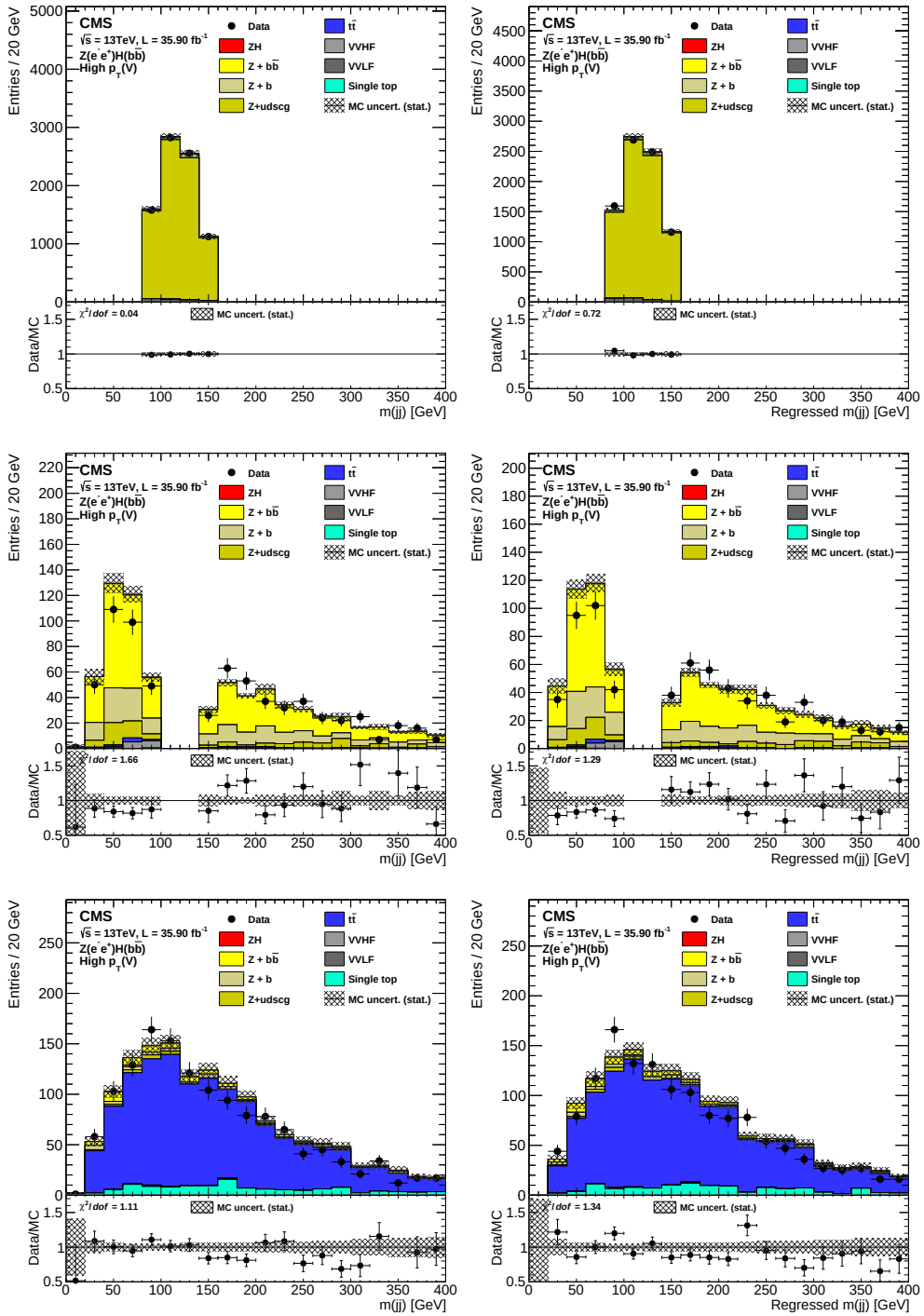


Figure B.3: Dijet invariant mass in region of the Z(ll)H(bb) enriched in background processes. **First row:** region enriched in Z + light jets processes. **Second row:** region enriched in Z + b jets processes. **Third row:** region enriched in $t\bar{t}$ processes. The dijet mass is compared with and without regression. **First column:** no regression applied. **Second column:** with regression applied.

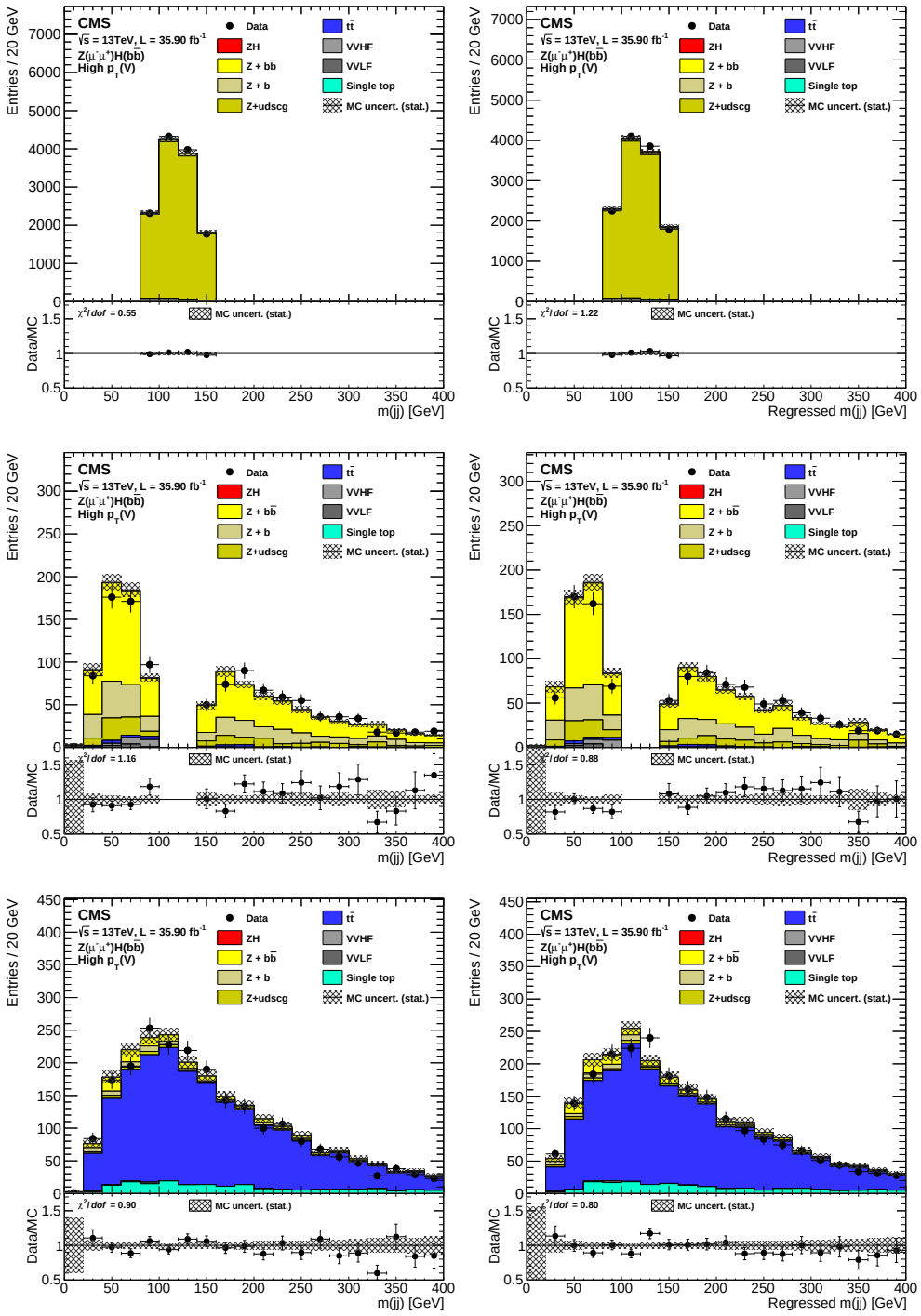


Figure B.4: Dijet invariant mass in region of the Z(ll)H(bb) enriched in background processes. **First row:** region enriched in Z + light jets processes. **Second row:** region enriched in Z + b jets processes. **Third row:** region enriched in $t\bar{t}$ processes. The dijet mass is compared with and without regression. **First column:** no regression applied. **Second column:** with regression applied.

C
Appendix C

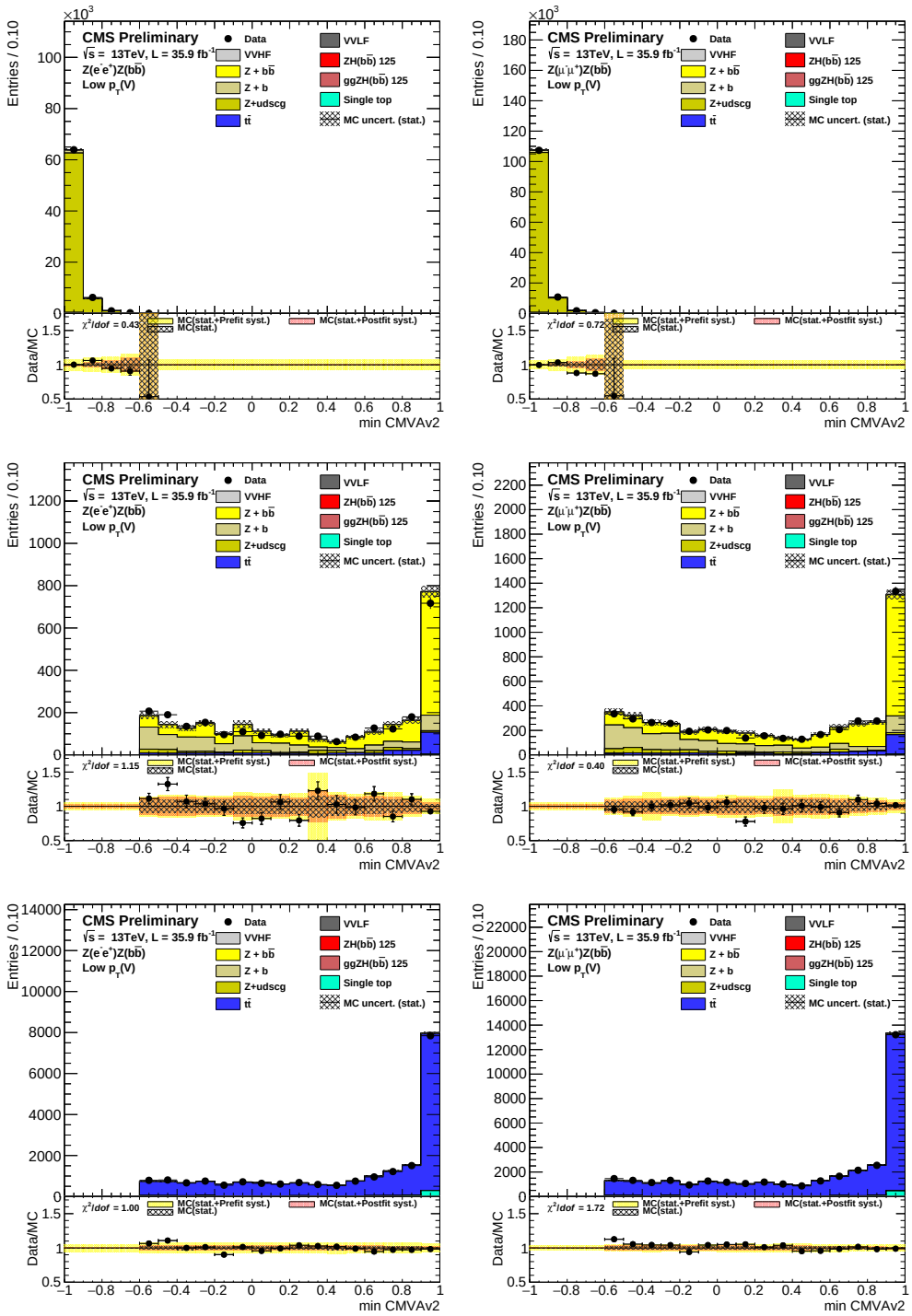


Figure C.1: Postfit $\text{CMVAv2}_{\text{min}}$ distributions in the six low-$p_T(Z)$ control regions in the Z(ll)Z(bb) channel. The left and right column correspond to the Z(ee)Z(bb) and Z($\mu\mu$)Z(bb) sub-channels, respectively. The first, second and third row correspond to the Z + light, Z + heavy flavor and $t\bar{t}$ control regions, respectively. The arrangement of each figure is similar to the BDT output score plots in the signal region and is described in the legend of Figure 9.1.

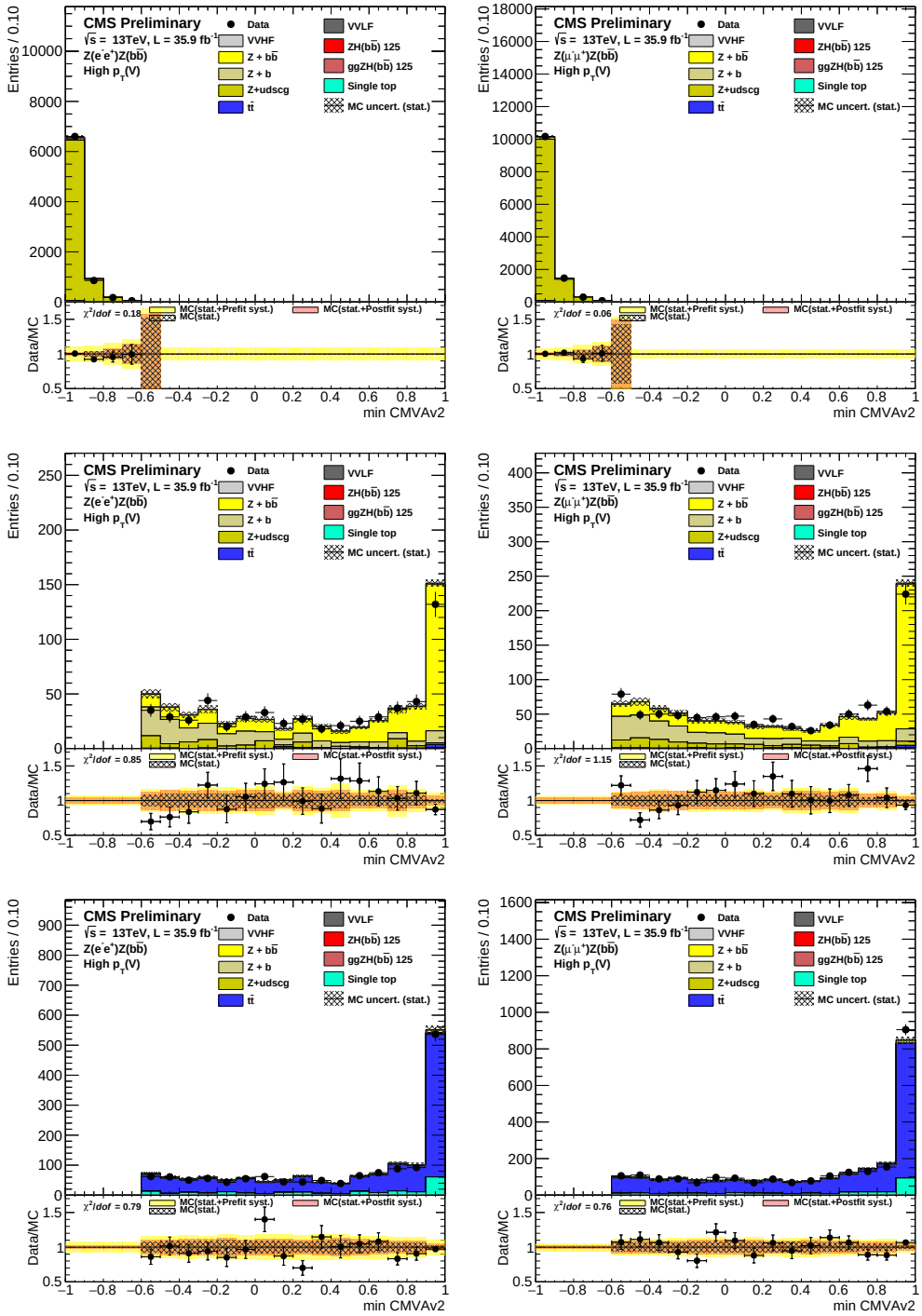


Figure C.2: Postfit $\text{CMVAv2}_{\text{min}}$ distributions in the six high-$p_T(Z)$ control regions in the Z(ll)Z(bb) channel. The left and right column correspond to the Z(ee)Z(bb) and Z($\mu\mu$)Z(bb) sub-channels, respectively. The first, second and third row correspond to the Z + light, Z + heavy flavor and $t\bar{t}$ control regions, respectively. The arrangement of each figure is similar to the BDT output score plots in the signal region and is described in the legend of Figure 9.1.

Acknowledgments

A sincere thank you to Christoph Grab who gave me the opportunity to realize my PhD within his group and the CMS VH(bb) team. I am very grateful for his support, encouragement and valuable feedback during this four and half year journey. I would also like to thank Rainer Wallny for his generous support during the last months of my PhD.

I am also thankful to Luca Perrozzi for his supervision, 24/7 availability, countless feedback and discussions during the ongoing analysis and slides-making. It was a pleasure to work with him. I also appreciate his availability during the first months of his new career path, when I am sure he was very busy.

I would also like to express my thanks to the other members of the ETH VH(bb) team from whom I have greatly benefited. To Alessandro who has carefully read my PhD thesis numerous times and given valuable feedback. Thank you to Pirmin for having improved the Xbb framework by orders of magnitude, and for his expertise and availability. To Krunal for the discussion we had on statistical methods and ML approaches.

Furthermore, I would like to express my thanks to the CMS VH(bb) team members. The VH(bb) search is a complex analysis, and the results presented in this thesis would not have been possible without the work of Michele, Caterina, David, Jacobo, Sean-Jiun, Andrea, Stephane, James, Chris, Lorenzo, Silvio, Leonardo, Rainer, Adinda, Luca M. and other members from the VH(bb) team.

I would also like to thank my colleagues from the IPA group, Lukas, Simon, David, Daniele, Maren, Diego, Vasilije, both Christians, Christina, Anne-Ma, Micha and Maria Giulia who made it very enjoyable to work at HPK and brightened my social life with numerous team-building activities during these four years. A special thank you to the administrative team of IPA: Gabi, Rosa, Bettina, Jennifer and Gabriele who helped with music rooms, neptun orders and many other items relevant to my work at IPA.

Last but not least, I would like to thank all those who have encouraged me and made my life more colorful during the past four and half years. To my parents Marc-André and Nina for believing in me, Delphine and Adam for their help with English corrections from across the Atlantic (including this acknowledgment section), Nathalie, Pascal, Audray and Alix for their support, to my friends from the french side: Léandre, Felix and Laurent, flatmates: Alice, Cedric and Thierry and my piano teacher Julie.

Finally, to 马滕英子 who is the sun in my life.

Bibliography

[1] Michael Edward Peskin and Daniel V. Schroeder, *An Introduction to Quantum Field Theory*. Westview Press Reading (Mass.), 1995.

[2] Francis Halzen, Alan D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. John Wiley & Sons, 1984.

[3] M. Tanabashi et al. (Particle Data Group), "The Review of Particle Physics." *Phys. Rev. D 98* no. 98, (030001 (2018)) .

[4] 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. B* **716** (2012) 1–27.

[5] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett. B* **716** (2012) 30–61.

[6] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV," *JHEP* (2013) .

[7] G. Stagnitto, *Scale dependence of physical observables and theoretical uncertainties*. PhD thesis, Universita Di Pavia, 2018.

[8] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev.* (1957) .

[9] A. Pich, "The Standard Model of Electroweak Interactions," arXiv:0502010 [hep-ph].

[10] T. Gleisberg, S. Hoeche, F. Krauss, M. Schoenherr, S. Schumann, F. Siegert, J. Winter, "Event generation with SHERPA 1.1," *JHEP* **0902:007** (2008) , arXiv:0811.4622.

[11] "Parton distributions for the LHC Run II," *J. High Energ. Phys.* (2015) .

[12] A. Buckley, J. Butterworth, S. Gieseke et al., "General-purpose event generators for LHC physics," *Physics Reports* no. 504, (2011) 145–233.

[13] M. Bahr, S. Gieseke, M. A. Gigg, D. Grellscheid, K. Hamilton, O. Latunde-Dada, S. Platzer, P. Richardson, M. H. Seymour, A. Sherstnev, J. Tully, B. R. Webber, "Herwig++ Physics and Manual," *Eur.Phys.J.* (2008) 639–707.

[14] R. Keith Ellis and W. James Stirling and B.R. Webber, *QCD and collider physics.* No. 8. Cambridge University Press, 1996.

[15] E. Gardi, N. Glover, A. Robson, *LHC Phenomenology.* Scottish Graduate Series.

[16] J. Alwall, S. Hoche, F. Krauss, N. Lavesson, L. Lonnblad, F. Maltoni, M.L. Mangano, M. Moretti, C.G. Papadopoulos, F. Piccinini, et al., "Comparative Study of Various Algorithms for the Merging of Parton Showers and Matrix Elements in Hadronic Collisions," *Eur.Phys.J.C53* (2008) 473–500.

[17] S. Frixione, F. Stoeckli, P. Torrielli et al., "The MC@NLO 4.0 Event Generator," *CERN-TH/2010-216* (2010) , arXiv:1010.0819 [hep-ph].

[18] P. Nason, "A new method for combining NLO QCD with shower Monte Carlo algorithms," *J. High Energ. Phys.* **2004** (2004) .

[19] S. Frixione, P. Nason, C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method," *J. High Energ. Phys.* **2007** (2007) .

[20] S. Alioli, P. Nason, C. Oleari, E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX," *J. High Energ. Phys.* **2010** (2010) .

[21] P. Bortignon, *Search for the standard model Higgs boson produced in association with a Z boson with the CMS detector at the LHC.* PhD thesis, ETH Zurich, 2014. pages 29-32.

[22] O. S. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, P. Proudlock, "LHC Design Report," *Geneva : CERN,* (2004) 548.

[23] P. Bortignon, *Search for the standard model Higgs boson produced in association with a Z boson with the CMS detector at the LHC.* PhD thesis, ETH Zurich, 2014. pages 33-51.

[24] CMS Collaboration, "CMS Physics Technical Design Report Volume I: Detector Performance and Software," *CERN* (2006) .

[25] "Official cms-cern webpage." cms.web.cern.ch/cms.

[26] CMS Collaboration, "CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems," *Technical Design Report CMS* (2000) .

[27] CMS Collaboration, "CMS Physics: Technical Design Report, Volume 1: Detector Performance and Software," *Technical Design Report CMS* (2006) .

[28] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting," *Nucl.Instrum.Meth.* no. A262:444—450, (1987) .

[29] W. Adam and R. Fruhwirth and A. Strandlie and T. Todorov, "Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC," *eConf* no. C0303241:TULT009, (2003) .

[30] Matteo Cacciari and Gavin P. Salam and Gregory Soyez, "The Anti-k(t) jet clustering algorithm," *JHEP* **0804:063** (2008) .

[31] CMS Collaboration, "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS de- tector," *Technical Report CMS-PAS-PFT-10-001* (2010) .

[32] CMS Collaboration, "Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV," *Technical Report CMS- PAS-PFT-10-003* (2010) .

[33] CMS Collaboration, "Commissioning of the Particle-Flow recon- struction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV," *Technical Report CMS-PAS-PFT-10-002* (2010) .

[34] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector," *arXiv:1610.07922* (2017) .

[35] CMS Collaboration, "Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair," *arXiv:1709.05543* (2017) .

[36] CMS Collaboration, "Search for the standard model Higgs boson produced through vector boson fusion and decaying to bb," *Phys. Rev. D92 (2015)* .

[37] CMS Collaboration, "Search for ttH production in the all-jet final state in proton-proton collisions at s= 13 TeV," *arXiv:1803.06986* (2018) .

[38] CMS Collaboration, "Search for ttH production in the H $\longrightarrow$ bb decay channel with leptonic tt decays in proton-proton collisions at s= 13 TeV," *arXiv:1804.03682* (2018) .

[39] Cowan, Glen and Cranmer, Kyle and Gross, Eilam and Vitells, Ofer, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C* **71** no. 2, (2011) 1—19.

[40] The ATLAS and CMS Collaborations and The LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011," *Technical Report CMS-NOTE-2011-005* (2011) .

[41] Roe, Byron P. and Yang, Hai-Jun and Zhu, Ji and Liu, Yong and Stancu, Ion and Mcgregor, Gordon, "Boosted decision trees as an alternative to artificial neural networks for particle identification," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **543** no. 2-3, (2005) 577—584.

[42] Therhaag, Jan and Team, Tmva Core Developer, "TMVA - Toolkit for multivariate data analysis,".

[43] Shrestha, Durga and Solomatine, Dimitri, "Experiments with AdaBoost.RT, an Improved Boosting Scheme for Regression," *Neural Computation* **18** (07, 2006) 1678–1710.

[44] "Official cms luminosity public results webpage." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

[45] CMS Collaboration, "Electron and photon performance in CMS with the full 2016 data sample." *CMS-DP-2017-004,CERN-CMS-DP-2017-004* (2017) .

[46] The CMS Collaboration, "Muon Identification and Isolation efficiency on full 2016 dataset," *CMS Performance Note* **CMS-DP-2017-007** (2017) .

[47] Alwall, J. and Frederix, R. and Frixione, S. and Hirschi, V. and Maltoni, F. and Mattelaer, O. and Shao, H.-S. and Stelzer, T. and Torrielli, P. and Zaro, M. and et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics* **2014** no. 7, (2014) .

[48] T. Sjöstrand and S. Mrenna and P. Skands, "A Brief Introduction to PYTHIA 8.1," *Comput. 819 Phys. Commun.* (2008) .

[49] CMS Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements," *EPJC* (2019) .

[50] GEANT4 Collaboration, "GEANT4 — a simulation toolkit," *arXiv:1804.03682* **506** (2013) 250–303.

[51] K. Hamilton and P. Nason and G. Zanderighi, "MINLO: Multi-Scale Improved NLO," *JHEP* (2012) .

[52] Alwall, J. and Hoche, S. and Krauss, F. and Lavesson, N. and Lonnblad, L. and Maltoni, F. and Mangano, M.l. and Moretti, M. and Papadopoulos, C.g. and Piccinini, F. and et al., "Comparative Study of Various Algorithms for the Merging of Parton Showers and Matrix Elements in Hadronic Collisions,".

[53] N.D. Gagunashvili, "Comparison of weighted and unweighted histograms," *PoS(ACAT)* **054** (2007) , arXiv:0605123.

[54] Alwall, J. and Frederix, R. and Frixione, S. and Hirschi, V. and Maltoni, F. and Mattelaer, O. and Shao, H.-S. and Stelzer, T. and Torrielli, P. and Zaro, M. and et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics* **2014** no. 7, (2014) .

[55] Alwall, J. and Hoche, S. and Krauss, F. and Lavesson, N. and Lonnblad, L. and Maltoni, F. and Mangano, M.l. and Moretti, M. and Papadopoulos, C.g. and Piccinini, F. and et al., "Comparative Study of Various Algorithms for the Merging of Parton Showers and Matrix Elements in Hadronic Collisions,".

[56] "The powheg box v2 framework (draft)." http://th-www.if.uj.edu.pl/~erichter/POWHEG-BOX-V2/Docs/V2-paper.pdf.

[57] CMS Collaboration, "Pileup Removal Algorithms," *CMS Physics Analysis Summaries* no. CMS-PAS-JME-14-001, (2018) .

[58] CMS Collaboration, "The CMS experiment at the CERN LHC," *CERN-ARCH-Series-PH-TH-DivRep* (2011) .

[59] "Search for neutral Higgs bosons decaying to tau pairs in pp collisions at 7 TeV," *Physics Letters B* .

[60] Collaboration, The Cms, "Determination of jet energy calibration and transverse momentum resolution in CMS," *Journal of Instrumentation* **6** no. 11, (2011) .

[61] Collaboration, The Cms, "Identification of b-quark jets with the CMS experiment," *Journal of Instrumentation* **8** no. 04, (2013) .

[62] "Identification of heavy-flavour jets with the CMS detectorin pp collisions at 13 TeV," *JINST* (2017) .

[63] Joosep Pata, *Search for the Production of the Higgs Boson in Association with a Top Quark Pair with CMS at $\sqrt{s}$ = 13 TEV.* PhD thesis, ETH Zurich, 2018.

[64] CMS Collaborationi, "Measurement of $b\bar{b}$ angular correlations based on secondary vertex reconstruction at $s = 7$ TeV," *JHEP* **136** (2011) , arXiv:1102.3194.

[65] Cms, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV," *arXiv.org* (May, 2018) . https://arxiv.org/abs/1712.07158.

[66] "Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector," Aug, 2016. https://cds.cern.ch/record/2205284.

[67] Cms, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV," *arXiv.org* (May, 2018) . https://arxiv.org/abs/1712.07158.

[68] Kallweit, S. and Lindert, J. M. and Maierhöfer, P. and Pozzorini, S. and Schönherr, M., "NLO QCD EW predictions for V jets including off-shell vector-boson decays and multijet merging," *Journal of High Energy Physics* **2016** no. 4, (2016) 1—51.

[69] P. Bortignon, *Search for the standard model Higgs boson produced in association with a Z boson with the CMS detector at the LHC.* PhD thesis, ETH Zurich, 2014.

[70] The CMS Collaboration, "Muon identification and isolation efficiencies with 2017 and 2018 data," *CMS Performance Note* **CMS-DP-2017-049** (2018) .

[71] CMS Collaboration, "Operation and Performance of the CMS outer tracker," *PoS* no. (Vertex 2017), 013, (2017) .

[72] CMS Collaboration, "Cross section measurement of t-channel single top quark production in pp collisions at $\sqrt{s} = 13$ TeV.," *Phys. Lett. B* (2017) .

[73] "Measurement of the WZ production cross section in pp collisions at $\sqrt{s} = 13$ TeV.," *Phys. Lett. B* (2017) .

[74] CMS Collaboration, "Measurement of the ZZ production cross section and $Z \to l^+l^-l'^+l'^-$ branching fraction in pp collisions at $\sqrt{s} = 13$ TeV.," *Phys. Lett. B* (2017) .

[75] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV," *JINST 12 (2017) P02014* (2017) .

[76] CMS Collaboration, "Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector," *CMS Physics Analysis Summary CMS-PAS-JME-16-004* (2016) .

[77] "Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV," *Phys. Rev. D* (2017) .

[78] ATLAS and CMS Collaborations, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysisof the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV," *High Energ. Phys.* (2016) .

[79] CMS Collaboration, "Search for the standard model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$," *Phys. Rev. D* (2015) .

[80] CMS Collaboration, "Observation of Higgs boson decay to bottom quarks." *Phys. Rev. Lett. 121* (2018) .

[81] M. Abadi et al, Google Research, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2015. `https://www.tensorflow.org/`. Software available from tensorflow.org.

[82] Chollet, François and others, "Keras." `https://keras.io`, 2015.

[83] N. Chernavskaya et al, "Deep neural network based simultaneous b-jet energy correction and resolution estimator," 2018.

[84] Daniele Bertolini, Philip Harris, Matthew Low, Nhan Tran, "Pileup Per Particle Identification," *JHEP* (2014) .

[85] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, Jesse Thaler, "Soft Drop," *JHEP* (2014) .

[86] CMS Collaboration, "Identification of double-b quark jets in boosted event topologies," *CMS Physics Analysis Summary* no. CMS-PAS-BTV-15-002, (2016) .

[87] Jesse Thaler, Ken Van Tilburg, "Identifying Boosted Objects with N-subjettiness," *JHEP* (2011) .

[88] S. Marzani, G. Soyez, M. Spannowsky, "Looking inside jets: an introduction to jet substructure and boosted-object phenomenology," `arXiv:0811.4622 [hep-ph]`.

[89] J. Thaler, K. Van Tilburg, "Maximizing boosted top identification by minimizing N-subjettiness," *J. High Energ. Phys.* (2012) .

[90] CMS Collaboration, "Jet algorithms performance in 13 TeV data," *Technical Report CMS-PAS-JME-16-003* (2017) .

[91] Yu. L. Dokshitzer, G. D. Leder, S. Moretti, B. R. Webber, "Better jet clustering algorithms,".

[92] M. Wobisch, T. Wengler, "Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering," arXiv:9907280v1 [hep-ph].

[93] S. D. Ellis, C. K. Vermilion, J. R. Walsh, "Techniques for improved heavy particle searches with jet substructure," arXiv:0903.5081 [hep-ph].

[94] CMS Collaboration, "Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment," CMS Physics Analysis Summaries.