Simplified template cross-section measurement for Higgs boson decay to b-quarks in association with a vector boson with the full Run 2 CMS dataset

Dissertation

zur Erlangung des Doktorgrades

an der Fakultät für Mathematik, Informatik und Naturwissenschaften

Fachbereich Physik

der Universität Hamburg

vorgelegt von

Hessamoddin (Sam) Kaveh

aus Mashhad, Iran

Hamburg

2022

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Die Dissertation wurde in der vorgelegten oder einer ähnlichen Form nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Hamburg, den May 17, 2022.

Hessamoddin Kaveh

Gutachter der Dissertation:	Dr. Rainer Mankel Prof. Dr. Elisabetta Gallo
Zusammensetzung der Prüfungskommission:	Dr. Rainer Mankel Prof. Dr. Elisabetta Gallo Prof. Dr. Freya Blekman Dr. James Ferrando Prof. Dr. Sven-Olaf Moch
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Sven-Olaf Moch
Datum der Disputation:	15 Juni 2022
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Wolfgang J. Parak
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Günter H. W. Sigl
Dekan der MIN-Fakultät:	Prof. Dr. Heinrich Graener

Abstract

The Simplified Template Cross-Section (STXS) measurement of the 125 GeV Higgs boson decaying to a pair of b-quarks, produced in association with a vector boson is presented. The analysis is performed on the full Run 2 data collected at the Compact Muon Solenoid detector (CMS) at Large Hadron Collider (LHC) corresponding to the integrated luminosity of 138 fb^{-1}. The analysis is performed considering the leptonic decays of the associated vector boson, while the Higgs candidate is reconstructed from either a pair of b-quark jets or one fat b-quark jet; both corresponding to a pair of final state b-quarks in different kinematic regions. The inclusive signal strength and an instructive mass-based cross-check analysis where the mass of the Higgs candidate is considered as the observable to extract the signal strength are performed and results for each data taking era and for the full Run 2 are reported.

Zusammenfassung

Es wird die Messung des Simplified Template Cross-Section (STXS) des 125 GeV Higgs-Bosons vorgestellt, das in ein Paar von b-Quarks zerfällt, die in Verbindung mit einem Vektorboson erzeugt werden. Die Analyse wird mit den vollständigen Daten des Run 2 durchgeführt, die am Compact Muon Solenoid Detektor (CMS) am Large Hadron Collider (LHC) gesammelt wurden, was einer integrierten Luminosität von 138 fb^{-1} entspricht. Die Analyse wird unter Berücksichtigung der leptonischen Zerfälle des zugehörigen Vektorbosons durchgeführt, während der Higgs-Kandidat entweder aus zwei b-Quark-Jets oder einem fetten b-Quark-Jet rekonstruiert wird, der einem Paar von b-Quarks im Endzustand in verschiedenen kinematischen Regionen entspricht. Die inklusive Signalstärke und eine aufschlussreiche massenbasierte Cross-Check-Analyse werden auch für den vollständigen Run 2 durchgeführt und die Ergebnisse werden berichtet.

Contents

Introduction

The study presented in this thesis is an effort to understand the properties of the Higgs boson and its Yukawa couplings to b-quarks. In Chapter 1 the theoretical basis needed to perform this study and the framework of the Standard Model (SM) of particle physics is presented. Chapter 2 is an overview of the Compact Muon Solenoid (CMS) apparatus, while chapter 3 explains the techniques employed to reconstruct the physics objects from the detector responses. The physics analysis to measure the Simplified Template Cross-Section (STXS) for the Higgs boson decaying to a pair of b-quarks, produced in association with a vector boson is described in detail in chapter 4 and the results gained from this study is found in chapter 6. In chapter 5, the mass-based cross-check analysis is explained, where a parallel strategy is employed to ensure the robustness of the main approach.

Chapter 1

The Standard Model of particle physics

The objective of this thesis is to study matter and its interactions. Specifically, the Higgs boson and its decay into b-quarks is studied. In this chapter, the theoretical formulation and the matter and its interactions will be introduced. In the context of high energy physics, matter and its interactions follow the laws of quantum mechanics and need to be treated in the relativistic regime. These are very challenging premises to mathematically formulate into mathematical models and successfully predict the outcomes. The Standard Model of Particle Physics (SM) [58, 93, 109] is a quantum field theory that to date has successfully described the result of experiments. To demonstrate the challenge faced, the time between conceptualizing a concept/particle in the SM and the experimental discovery is shown in Fig. 1.1, it can be seen that the current formalism is the result of more than a century of theoretical and experimental work.

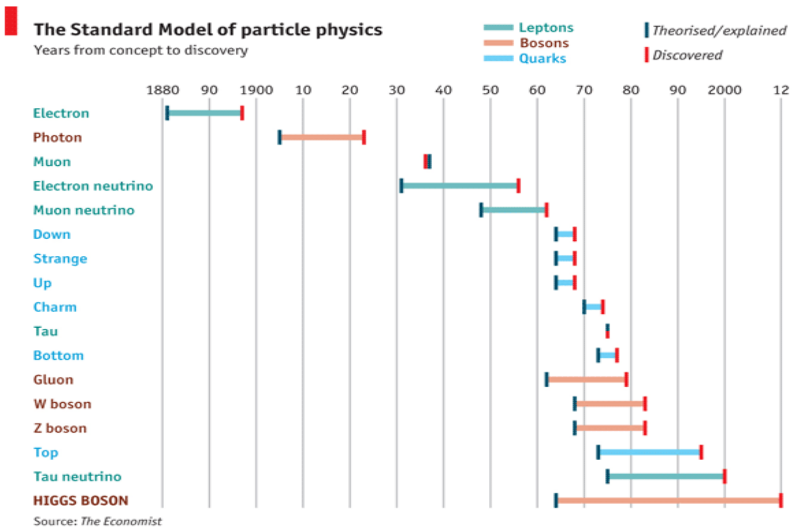


Figure 1.1: The formulation of a concept in particle physics and years to its discovery [source:The Economist]

1.1 Natural units

Before starting to introduce the elements of the SM, it is needed to introduce the natural units. Natural units is the redefinition of units so that Planck's constant and the speed of light are both equal to one ($\hbar = c = 1$). This allows one to express all quantities (mass, length, area, time, rate, momentum) as powers of GeV. Furthermore, the electric charge is presented as integer multiples of the fundamental charge e = $1.602176634 \times 10^{-19}$C [63].

1.2 Matter in the Standard Model

To introduce the elementary components of matter in the SM, it is best to categorize them based on their quantum numbers. Considering the spin quantum numbers, which is the intrinsic angular momentum of a particle, integer spin particles are called bosons and half-integer-spin particles are categorized as fermions.

Fermions

The fermions are in turn grouped into three generations of leptons and quarks. The difference between the leptons and quarks is that quarks exhibit strong interactions and have the color quantum number. The difference between generations is only in the mass property, which increases per generation. In each generation, there are two particles, forming an isospin doublet. For leptons, these doublets are formed with a charged lepton and its corresponding neutral neutrino. The three generations of leptons are as follows

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix} \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}, \tag{1.1}$$

where the electron is the lightest of the leptons with a mass of 0.511 MeV while muons are almost 200 times more massive and finally tau leptons have a mass of 1.777 GeV [63]. Neutrinos have negligible masses and were originally assumed massless in the SM.

Quark doublets are made of two quarks, where the first element of each doublet has $+(2/3)$ electric charge and the second element $-(1/3)$ electric charge,

$$\begin{pmatrix} u \\ d \end{pmatrix} \begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} t \\ b \end{pmatrix}. \tag{1.2}$$

The quark masses range from few MeV for the up and down quarks and to 172.76 ± 0.30 GeV [63] for the top quark. Quarks carry the color quantum number (red, green, blue, anti-red, anti-greed and anti-blue), which is introduced to preserve the Pauli exclusion principle in multi-quark bound states [61].

Gauge bosons

In the SM, interactions between fermions are mediated by gauge bosons. The electromagnetic interaction, which describes the interactions between objects with electric charge, is mediated by an exchange of a massless boson called photon (γ) according to the formalism of SM.

The strong interaction has 8 corresponding massless bosons that are called gluons. These for example interact with the up and down quarks to form protons and neutrons and bind them to form the atomic nuclei.

The weak interaction, which manifest itself in radioactivity and decay of unstable subatomic particles, is mediated with massive $W^{\pm}$ and the Z bosons, 80.379 ± 0.012 and 91.1876 ± 0.0021 GeV [63], respectively. The masses of the mediator reflects the fact that this force is short ranged.

The Higgs boson

Finally, the only scalar boson (0-spin) of the SM, is the Higgs boson. The interaction with the Higgs field is the mechanism for particles to acquire mass in the SM. It was discovered on July 4, 2012 by ATLAS and CMS, and the study of the properties of this particle in its vector-boson associated decays to two b-quarks is the main topic of this thesis.

A summary of the SM constituents is given in Fig. 1.2.

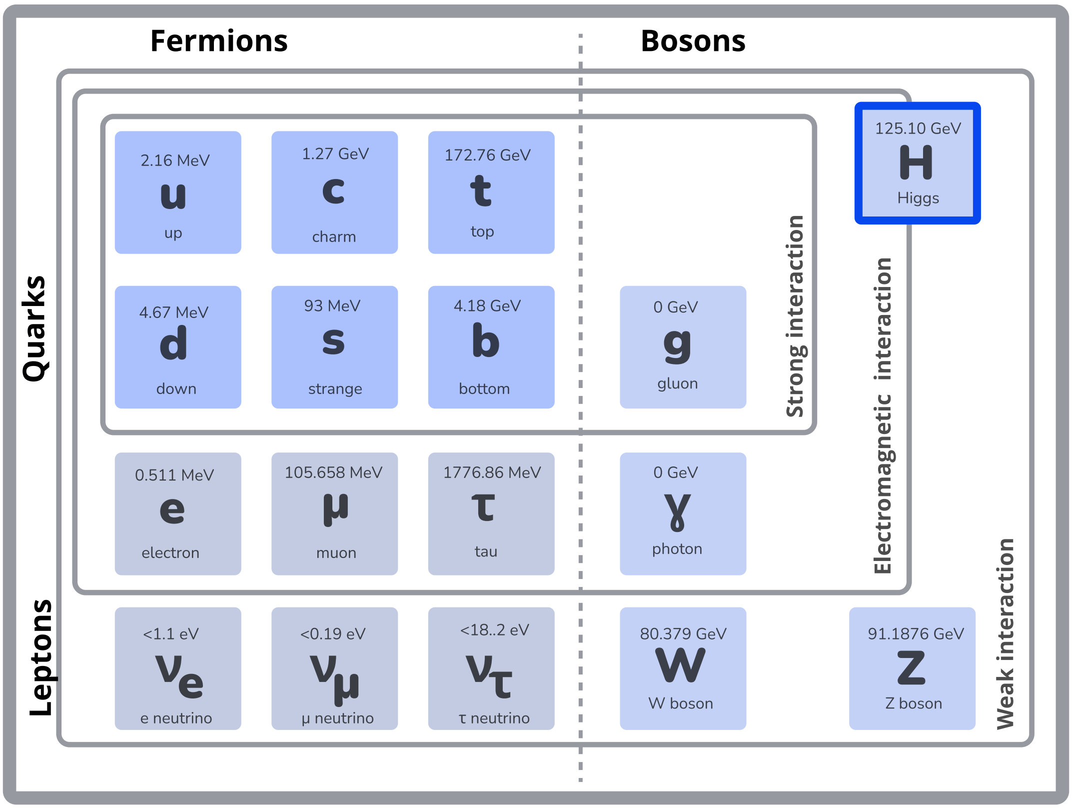


Figure 1.2: Constituents of the SM

1.3 The framework of the Standard Model

In this section, the theoretical framework of the SM is presented. Many theories in physics have been formulated with the principle of least action, which the SM also follows. In this approach, the path that minimizes the action is found and is independent of the coordinates chosen to parametrize the problem. Furthermore, the dynamics are formulated with the Lagrangian, which allows us to use mathematical tools such as the Noether theorem [85], in which symmetries result in conserved quantities.

In the SM, to find the equations of motion, the action S which has the following form is minimized:

$$S = \int \mathcal{L}(\psi(x), \partial_\mu \psi(x))\, \mathrm{d}t,\ x = (\vec{x}, t). \tag{1.3}$$

The quantity in the integration, the Lagrangian density, is a function of $\psi(x)$, where x is the four vector of space-time and $\psi(x)$, the space-time dependent field. Particles and their interactions are conceptualized as fields, the derivation of components of the SM from these fields is elaborated in the following sections.

1.3.1 Quantum electrodynamics

The interactions of electrically charged fermions with the photon (γ) are formulated as quantum electrodynamics. It expands the invariant formulation of Maxwell equations, in which $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the field strength tensor for the vacuum. The vector potential A^μ describes the photon field and the particle fields (leptons) are denoted as Dirac spinors $\psi(x)$.

In classical electrodynamics, the field strength is invariant under the addition of the gradient of an arbitrary function $\theta(x)$ to A^μ. This invariance introduces new degrees of freedoms which are redundant in the sense that they don't always have physical reality, any mathematical formalism that regulates this redundancy is called a gauge (Lorentz gauge for example is expressed by $\partial_\mu A^\mu = 0$). Consequently, a transformation between these gauges is called a gauge transformation. The gauge transformations which regulate local degrees of freedom are local gauge transformations and result in conserved quantities according to Noether theorem (Noether theorem states how local and global symmetries result in conserved quantities in the Lagrangian formalism).

The Lagrangian density of quantum electrodynamics (QED) in its covariant form is as follows,

$$\mathcal{L}_{QED} = i\bar{\psi}\gamma^\mu D_\mu \psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - m\bar{\psi}\psi, \tag{1.4}$$

where $D_\mu = \partial_\mu - ieA_\mu$ is called the covariant derivative and γ^μ are Dirac γ-matrices. This Lagrangian is invariant under the following gauge-transformation,

$$A^\mu(x) \to A^{\mu\prime}(x) = A^\mu(x) - \partial^\mu \theta(x) \tag{1.5}$$

$$\psi(x) \rightarrow \psi'(x) = \psi(x)e^{-ie\theta(x)}. \tag{1.6}$$

The transformations 1.6 form an Abelian group U(1), Abelian since the order of applying the consecutive transformations does not change the outcome. Therefore, QED is invariant under U(1) local gauge symmetry and the corresponding conserved quantity is the electric charge.

1.3.2 Quantum chromodynamics

Quantum chromodynamics (QCD), that describes the strong interaction in the SM, has the following properties: the strength of interaction decreases asymptotically as the energy scale increases and the corresponding length scale decreases (asymptotic freedom [62, 90]), the property that no free quarks or gluons can be detected (confinement [111]) and that the quanta of this model appears as bound states called hadrons. The formulation is based on a local SU(3) color symmetry, resulting in a theory that phenomenologically exhibits these properties.

Similar to the covariant Lagrangian of QED, the QCD Lagrangian for a quark of flavor q is written as

$$\mathcal{L}_{QCD} = i\bar{q}_i(\gamma^\mu D_\mu)_{ij}q_j - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - m\delta_{ij}\bar{q}_i q_j. \tag{1.7}$$

with i, j as color-indices with discrete values from 0 to 3 and the covariant derivative

$$D_\mu = \partial_\mu + ig_s\frac{\lambda_a}{2}\mathcal{A}^a_\mu, , a = 1, ..., 8, \tag{1.8}$$

where λ_a are the Gell-Mann matrices and the field strength tensor

$$G^a_{\mu\nu} = \partial_\mu\mathcal{A}^a_\nu - \partial_\nu\mathcal{A}^a_\mu + g_s f^{abc}\mathcal{A}^b_\mu\mathcal{A}^c_\nu, \tag{1.9}$$

where f^{abc} are the structure constant of the SU(3) group. Finally, the Lagrangian is invariant under the simultaneous transformation of gluons and quark fields as follows,

$$\mathcal{A}^a_\mu \rightarrow \mathcal{A}'^a_\mu = \mathcal{A}^a_\mu - \frac{1}{g_s}\partial_\mu\alpha_a(x) - f_{abc}\alpha^b(x)\mathcal{A}^c_\mu \tag{1.10}$$

and

$$q(x) \rightarrow q'(x) = q(x)e^{-ig_s\frac{1}{2}\alpha_a(x)\lambda_a}. \tag{1.11}$$

Transformations 1.11 form a non-Abelian $SU(3)_{color}$ group. The non-Abelian nature dictates that the gluons themselves will carry color charges, therefore gluons are self-interacting, this results in confinement. The other consequence of the non-Abelian gauge symmetries is that the coupling appears in the field strength tensor, this implies that all particle fields should couple with the same coupling constant to gauge fields.

The coupling constant for the strong interaction, $\alpha_s = \frac{g_s^2}{4\pi}$, is expressed as $\alpha_s(\mu_R)$ in perturbative regime where, α_s is small, performing perturbative calculations are possible. The perturbative regime is often used for calculating the rates of observable quantities, where μ_R is an arbitrary (unphysical) renormalization scale. Studying the properties of the renormalization at different momentum transfer scales Q shows that the strong coupling α_s is small for large momentum transfers (hard processes), this phenomenon is asymptotic freedom. For example, α_s is around 0.1 for Q in the 0.1-1 TeV range, while α_s is large for Q around 1 GeV or below [63].

1.3.3 Weak interaction and the electroweak theory

The disintegration of nuclei through emission or fission, motivated the formulation of weak interaction. The electroweak theory, describes the electromagnetic force and the weak force in a unified theory based on a local $SU(2)_L \times U(1)_Y$ gauge symmetry. This structure implies weak isospin (associated with $SU(2)_L$) and hypercharge (associated with $U(1)_Y$) conserved quantities, it states that the left-handed and right-handed fermions have different weak interactions. The chiral operators $(1-\gamma_5)/2$ and $(1+\gamma_5)/2$ projects fermions into their respective left/right-handed components.

Left-handed fermions transform as weak isospin-doublets under SU(2)

$$\mathcal{X}_L = L_L \begin{pmatrix} \nu_e \\ e_L \end{pmatrix} \text{ or } Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \tag{1.12}$$

and the right-handed fermions

$$\psi_R = e_R \text{ or } u_R \text{ or } d_R, \tag{1.13}$$

as singlets (e and ν_e for leptons and u, d for quarks).

The corresponding local transformations are

$$\begin{aligned} \mathcal{X}_L \to \mathcal{X}_L' &= e^{-ig\vec{\alpha}\cdot\vec{T} - ig'\beta\frac{Y}{2}}\mathcal{X}_L, \\ \psi_R \to \psi_R' &= e^{-ig'\beta\frac{Y}{2}}\psi_R, \end{aligned} \tag{1.14}$$

where $\vec{T} = \frac{1}{2}\vec{\sigma}$ are the three generators of the $SU(2)_L$ group with $\vec{\sigma}$ being the Pauli matrices and Y is the hypercharge operator. The couplings g' and g are gauge couplings of $U(1)_Y$ and $SU(2)_L$ respectively.

The field strength tensors are defined as,

$$\begin{aligned} B^{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu \\ \mathcal{W}^a_{\mu\nu} &= \partial_\mu \mathcal{W}^a_\nu - \partial_\nu \mathcal{W}^a_\mu - g\epsilon^{abc}\mathcal{W}^b_\mu \mathcal{W}^a_\nu, \end{aligned} \tag{1.15}$$

where ϵ^{abc} is the Levi-Civita [60] tensor and $\vec{\mathcal{W}}_\mu = \left(W^1_\mu, W^2_\mu, W^3_\mu\right)$ and B_μ represent the gauge fields.

The covariant derivative for the electroweak theory is defined as,

$$D_\mu = \partial_\mu + ig\vec{T}\cdot\vec{\mathcal{W}}_\mu + ig'\frac{Y}{2}B_\mu. \tag{1.16}$$

Finally, the Lagrangian density invariant under $SU(3)_c \times SU(2)_L \times U(1)_Y$ is expressed as follows,

$$\begin{aligned}\mathcal{L} &= i\bar{L}_{iL}\not{D}L_{iL} + i\bar{Q}_{iL}\not{D}Q_{iL} + i\bar{e}_{iR}\not{D}e_{iR} + i\bar{u}_{iR}\not{D}u_{iR} + i\bar{d}_{iR}\not{D}d_{iR} \\ &\quad -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - \frac{1}{4}\vec{\mathcal{W}}^{\mu\nu}\cdot\vec{\mathcal{W}}_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu}, \\ D_\mu &= \partial_\mu + ig_s\frac{\lambda_a}{2}\mathcal{A}^a_\mu + ig\vec{T}\cdot\vec{\mathcal{W}}_\mu + ig'\frac{Y}{2}B_\mu.\end{aligned} \tag{1.17}$$

which describes a self-consistent massless theory of strong and electroweak interactions.

The Higgs mechanism and EWSB

As already stated in Section 1.2, the matter of the standard model have different masses while Eq. (1.17) describes massless matter, as including mass terms breaks the local gauge invariance. The electroweak symmetry breaking (EWSB) which is rooted in the ideas arising from condensed matter physics on global symmetries, introduces mass terms for local symmetries.

A Lagrangian density, is added:

$$\mathcal{L}_{Higgs} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi^\dagger\phi), \tag{1.18}$$

where D_μ,

$$D_\mu = \partial_\mu + ig\vec{T}\cdot\vec{\mathcal{W}}_\mu + ig'\frac{Y}{2}B_\mu, \tag{1.19}$$

which is designed to break the $SU(2)_L \times U(1)_Y$ to produce the physical symmetries and masses for W and Z bosons and leave the photon massless. The covariant derivative of a Higgs scalar complex field ϕ, which is a $SU(2)_L$ doublet of the following form,

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix} = \begin{pmatrix}\phi^+ \\ \phi^0\end{pmatrix}, \tag{1.20}$$

describes the couplings to gauge fields. The procedure of EWSB, requires the potential $V(\phi^\dagger\phi)$ to have an infinite number of equivalent minima. Here the following form is considered,

$$V(\phi^\dagger\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2, \text{ with } \mu^2 < 0. \tag{1.21}$$

The shape of this potential is depicted in Fig. 1.3. Picking one minimum, i.e. $\phi^+_{vac} = 0$ and $\phi_{vac} = v$ where $v = \sqrt{\frac{-\mu^2}{2\lambda}} = 246.22$ [63] the symmetry of the vacuum is simultaneously broken and v is the vacuum expectation value. Transforming the fields as follows,

$$\begin{aligned} W^{\pm\mu} &= \frac{1}{\sqrt{2}}(W^{1\mu} \mp iW^{2\mu}) && \longrightarrow W^\pm \quad bosons \\ Z^\mu &= -B^\mu\sin\theta_w + W^{3\mu}\cos\theta_w && \longrightarrow Z \quad boson \\ A^\mu &= B^\mu\cos\theta_w - W^{3\mu}\sin\theta_w && \longrightarrow \gamma \quad photon, \end{aligned} \tag{1.22}$$

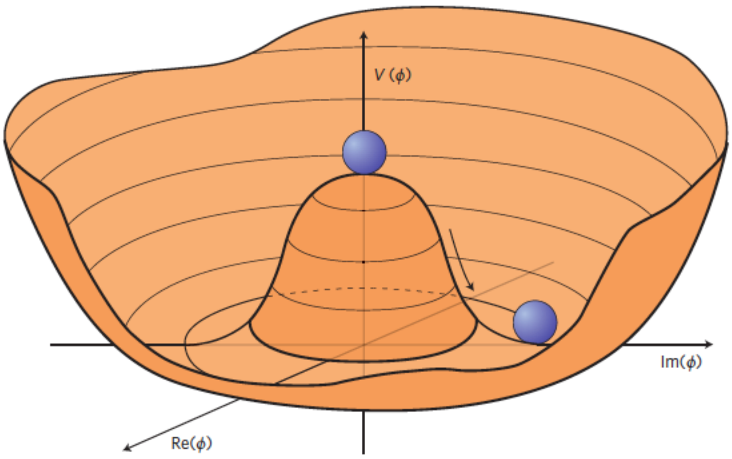


Figure 1.3: Shape of the Higgs potential, figure take from [45]

where θ_w is the weak mixing angle, the fields observed in nature are described in the theory. The angle is determined experimentally, $sin^2\theta_w = 0.23121 \pm 0.00004$ [63], the boson masses are related by $M_W = \frac{1}{2}vg, M_Z = \frac{1}{2}\sqrt{g^2 + g'^2}$.

An excitation of the Higgs field is defined as,

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix}; \tag{1.23}$$

replacing it in Eq. (1.18) gives rise to the Higgs boson mass term with $m_H = \sqrt{2\lambda}v$ and the trilinear and quartic self-coupling of the Higgs boson.

To add the fermionic masses, the following Lagrangian density terms are added,

$$\mathcal{L}_{HF} = -y^u_{ij}\bar{Q}_{iL}\tilde{\phi}u_{jR} - y^d_{ij}\bar{Q}_{iL}\phi d_{jR} - y^e_{ij}\bar{L}_{LiL}\phi e_{jR} + h.c., i, j = 1, ..., 3; \tag{1.24}$$

where y_{ij} are Yukawa-couplings and $\tilde{\phi} = i\sigma_2\phi^*$. The resulting fermionic masses, in the fermion mass eigenstate basis, have the following form,

$$m^u_i = \frac{y^u_i v}{\sqrt{2}}, m^d_i = \frac{y^d_i v}{\sqrt{2}}, m^e_i = \frac{y^e_i v}{\sqrt{2}}. \tag{1.25}$$

This demonstrates that for various fermions, the ratio of interaction strengths (to Higgs) equals the ratio of their masses.

1.4 Higgs phenomenology

In the following, the Higgs production channels at the Large Hadron Collider (LHC) with some of their decay channels are introduced to express the importance of the channel studied in this thesis and its relation with other Higgs studies. The mass of the Higgs boson is $m_H = 125.35 \pm 0.15$ GeV [33].

1.4.1 Higgs production modes

The leading order diagrams that depict how the Higgs production happens from the quark and gluon initial states (as is the case at LHC since protons collide) are shown in Fig. 1.4.

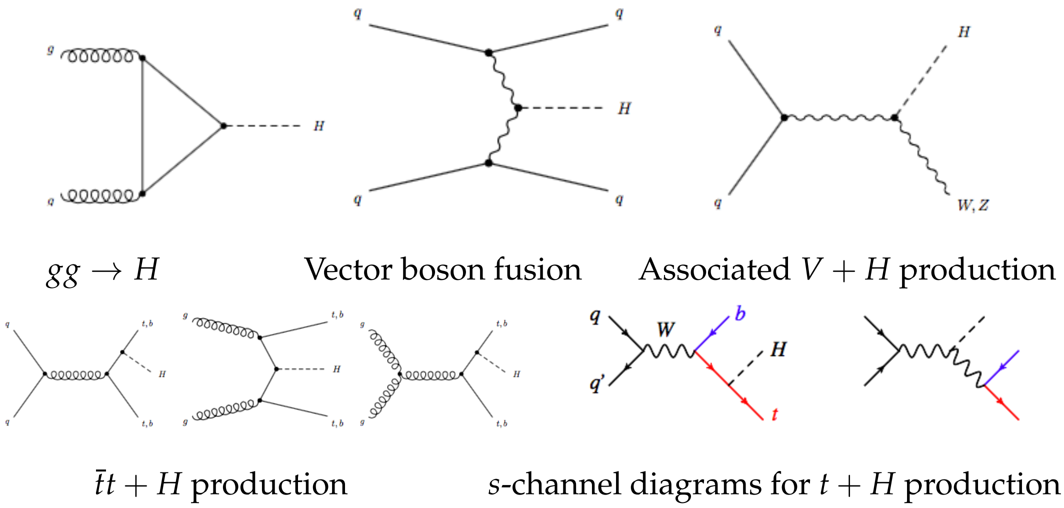


Figure 1.4: The leading order diagrams for Higgs production modes, taken from [46].

The gluon fusion production mode, $gg \to H$, where the Higgs boson is produced via a quark loop as shown in Fig. 1.4, has the highest cross-section among the production channels. Fig. 1.5 presents the calculated value of different Higgs production modes at the LHC.

The vector boson fusion channel (VBF) is produced via fusion of two W or Z bosons. The VBF is depicted by Feynman diagrams in Fig. 1.4. The VBF has the second-largest cross-section among the production modes.

The top associated production mode ($t\bar{t}$H), is one mode of the general quark associated production of Higgs. The $t\bar{t}$H has a cross-section of order 0.13 pb. The $t\bar{t}$H is unique as the top quark has the strongest coupling to the Higgs boson and is the heaviest elementary particle.

The Higgs associated production with a vector boson W/Z (VH), has a cross-section of the order 0.41-0.69 pb, this is the leading production channel for the Higgs in its decays to the b-quarks due to better triggering for the leptonic decays of the vector boson. The Higgs boson decays to the b-quarks is a probe of the Higgs Yukawa-couplings to the third generation fermions.

1.4.2 Higgs decay modes

In $gg \to H$ channel decays, $H \to ZZ \to 4\mu/2e2\mu$, has excellent mass resolution with good control over the background regarding the leptonic final state. The $H \to WW \to 2l + 2\nu$ channel is very similar to ZZ but does not have a clear mass-peak, backgrounds can be suppressed by requiring cuts on 2-leptons spatial separations, WW has the second-highest branching ratio for Higgs, see Fig. 1.5. The $H \to \gamma\gamma$ decay, although having a branching ratio of 0.23%, due to the good control over background processes and the design attributes of the LHC detectors provides excellent mass resolution and opportunities to study properties of the Higgs.

In the vector boson fusion channel, the WW and ZZ decaying to leptonic final states are very similar to the counterparts from gluon fusion. In studying the VBF decays to tau-leptons or hadronic decays finding techniques to remove Z+jets and QCD backgrounds is crucial.

In the top associated production of Higgs when Higgs decays to b-quarks (which has the highest branching ratio) properties of the top and Higgs system is probed. This decay channel has backgrounds from $t\bar{t}$+jets and W+jets and challenges in reconstructing the $t\bar{t}$ system and its combinatorics.

For the VH production mode with Higgs decaying to b-quarks in the final state, the $ZH \to llb\bar{b}$ (2-lepton), $ZH \to \nu\bar{\nu}b\bar{b}$ (0-lepton) and $WH \to l\nu b\bar{b}$ (1-lepton) are studied in this thesis. The main challenges are identifying b-jets and improving their mass resolution. Control over W+jets and Z+jets backgrounds are also vital. Studying these decays in larger transverse momenta regimes improves background rejection as decay products assume larger momenta.

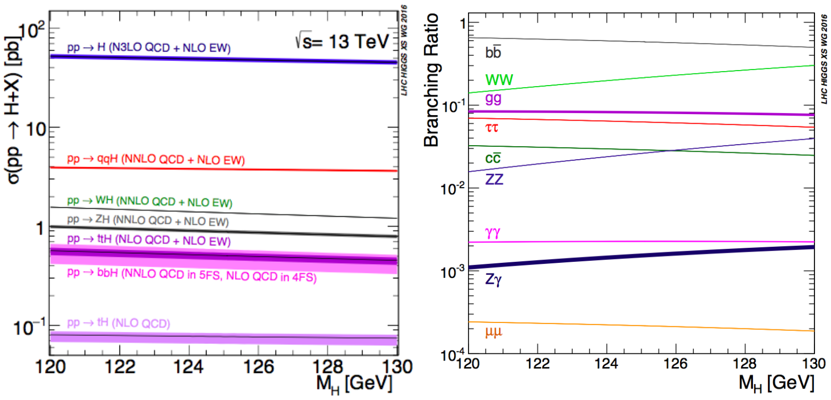


Figure 1.5: Higgs production cross-sections for dominant processes (left panel) branching ratios for Higgs decays (right panel) with $m_H \in [120, 130]$ GeV [51]. The uncertainties are represented with colored bands widths, taken from [45].

A summary of the status of the different production modes of Higgs at LHC and the reference to papers are given in Table 1.1. Figure 1.6 shows the Higgs couplings to fermions as presented in [97] from CMS collaboration, which reports the first evidence for the decay of the Higgs boson to muons with a signal strength relative to the SM prediction equal to $\mu = 1.19 \pm 0.40\,(\text{stat}) \pm 0.15\,(\text{syst.})$ from the combination of the Run 1 and 2 data taking periods. For an overview of Higgs physics at the LHC, see [63].

The latest results from the CMS collaboration for the VH production channel when the Higgs decays to b-quarks and the vector boson has leptonic decays,

Production mode	Status, observed (expected)-significance ATLAS	CMS
ggF	observed [4]	observed 6.6σ (7.4σ) [74]
VBF	observed 5.4σ (4.6σ) [3]	
VH	observed 5.3σ (4.8σ) [1]	evidence 4.9σ (4.8σ) [98]
$t\bar{t}$H	observed 5.8σ (4.9σ) [2]	evidence 5.2σ (4.2σ) [99]

Table 1.1: Status of the Higgs production modes at the LHC.

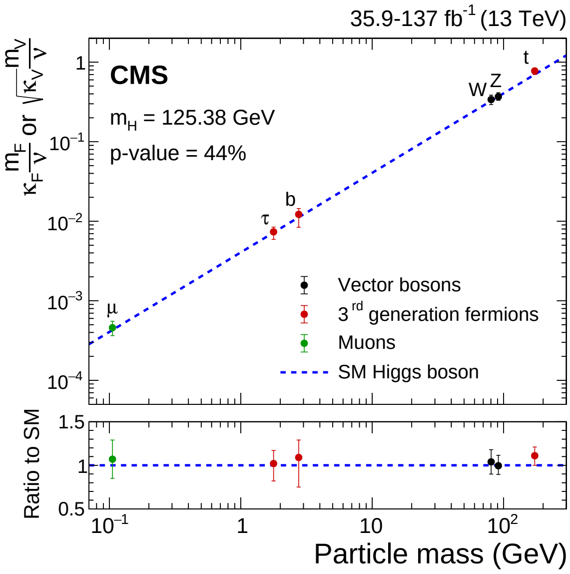


Figure 1.6: Higgs couplings to fermions and bosons [97]

previous to the work in this thesis, is shown in Fig. 1.7. The combination of VH($b\bar{b}$) 2016 and 2017 datasets [87] with the ones from Run 1 [27] of the CMS, resulted in a signal strength of $\mu = 1.01 \pm 0.18\,(\text{stat}) \pm 0.14\,(\text{syst.})$ and a discovery of the Higgs and b-quark Yukawa coupling with more than 5 σ significance was achieved. The dijet mass plot that shows the Higgs mass contribution is depicting only the 2016 and 2017 data taking periods.

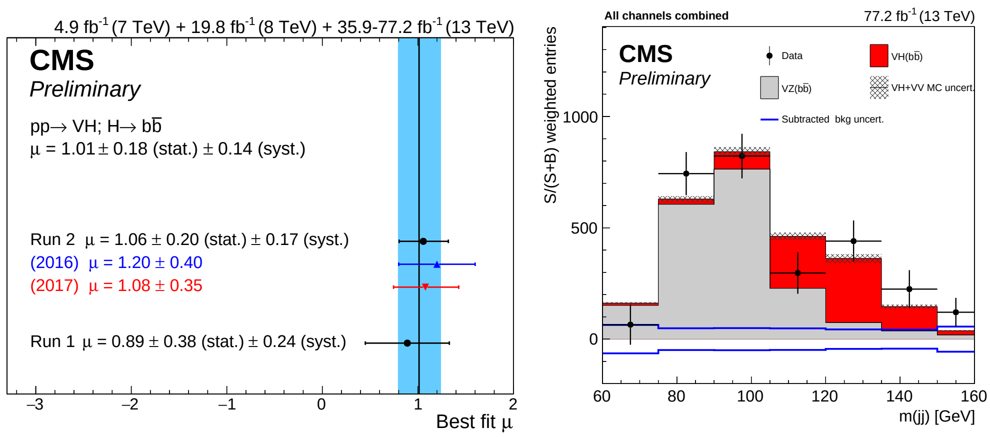


Figure 1.7: The signal strength best-fit value (left) shown as combination and individual results for 2016 and 2017 data taking periods with the uncertainty depicted as a blue band. The background subtracted distribution of the dijet mass for the VH (red) and the VZ (grey) with the uncertainties on the signal and background as blue and on the data points with error bars [86].

1.4.3 Simplified Template Cross Sections (STXS)

The main focus of Higgs measurements during the LHC Run 1 was the inclusive signal strengths, and multiplicative coupling modifiers. The collection of more data at the LHC during the Run 2 (2015-2018) and the needs from the theory community in performing combination studies and interpretations based on new physics, motivated the introduction of a universal framework for expressing the measurement results, called the simplified template cross section framework (STXS) [41].

The STXS for the measurement in this thesis is schematically depicted in Fig. 1.8. The cuts on vector boson transverse momenta and cuts on number of jets per process are meant to reduce the dependence to theoretical uncertainties and the underlying physics model. Often some bins are combined due to the limitations arising from the detector acceptance, the scheme used in this thesis is explained in detail in Section 4.5.1.

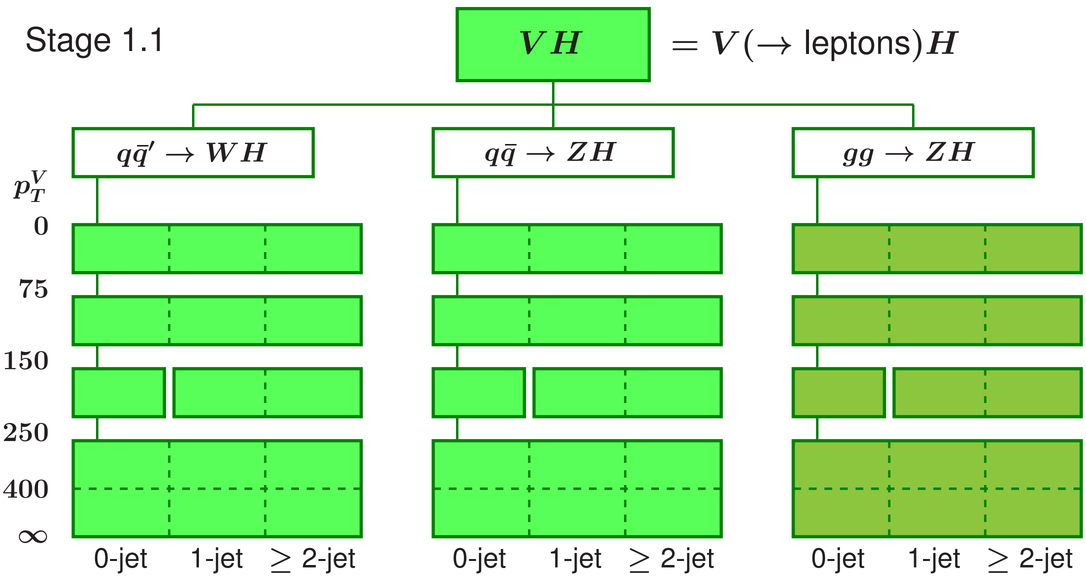


Figure 1.8: STXS stage 1.1 for the VH channel, taken from [41]

Chapter 2

The CMS experimental apparatus at the LHC

The Higgs processes mentioned in Section 1.4, are measured with the Compact Muon Solenoid (CMS) experimental detector apparatus at the Large Hadron Collider (LHC). This chapter summarizes the properties of the experimental setup.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [47], was designed to discover and study properties of the Higgs boson and search for possible new physics and deviations from the SM. It is a two-ring superconducting synchrotron accelerator complex (Fig. 2.1) that collides high-energy protons or Ions. It uses the 26.7 km tunnel originally created for the Large Electron Positron collider (LEP) stationed at the border of France and Switzerland. The collision center of mass energies were 13 TeV for the data used in this thesis.

The LHC needs numerous high intensity proton bunches, that have small transverse emittance. This is satisfied by the following accelerator chain and linear acceleration operations. The injector chain Linac2 supplies the LHC with protons. Then these protons enter the Booster (Proton Synchrotron Booster, PSB), where protons are further accelerated to 1.4 GeV, then protons enter the Proton Synchrotron (PS), where they reach an energy of 24 GeV. Finally, the Super Proton Synchrotron (SPS), feeds the LHC rings with 450 GeV protons.

To reach the 6.5 TeV beam energy, LHC relies on its superconducting magnets that operate at temperatures below 2 $K(-271.25\ ^{\circ}C)$ [47].

There are mainly four collision points corresponding to the four LHC experiments namely, A Toroidal LHC Apparatus (ATLAS) [13], the Compact Muon Solenoid (CMS) [104] both are multipurpose detectors that look for multiple scenarios, A Large Ion Collider Experiment (ALICE) [8] which studies lead-lead collisions, and LHC beauty (LHCb) [78] which focuses on studies in flavor physics.

The number of collision events per time interval for a certain process with

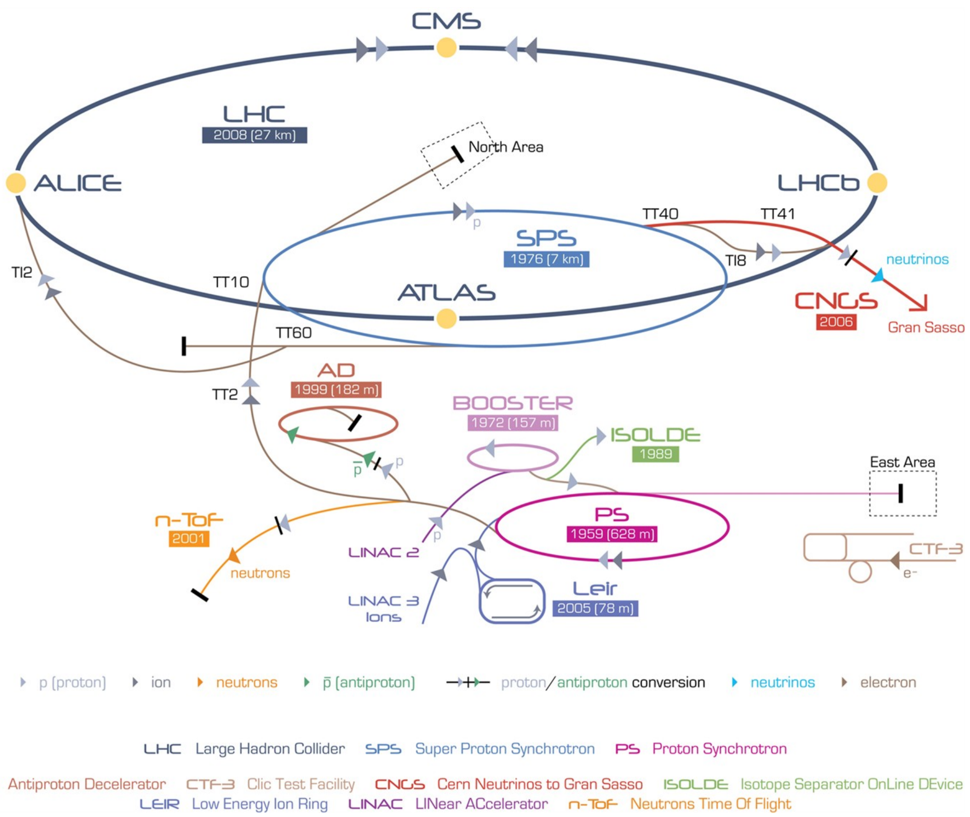


Figure 2.1: Schematic drawing of the LHC accelerator complex at CERN, figure taken from [30].

cross-section σ is related to the instantaneous luminosity $\mathcal{L}$, by $N_{events} = \sigma \times \mathcal{L}$. The instantaneous luminosity, defined by Eq. (2.1), is a measure of the performance of the collider;

$$\mathcal{L} = f \frac{N_b N_p^2}{4\pi\sigma_x\sigma_y}\mathcal{F}, \tag{2.1}$$

where f is the frequency of proton bunches, N_p the number of protons per bunch, N_b is the maximum number of proton bunches per beam, and $\sigma_{x,y}$ are the transverse beam sizes at the interaction point [47]. $\mathcal{F}$ is a geometric factor defined as $\mathcal{F} = \frac{1}{\sqrt{1+(\frac{\sigma_s}{\sigma_{xing}}\frac{\alpha}{2})^2}}$ where σ_s is the r.m.s bunch length and the σ_{xing} is the transverse beam size in the crossing plane, and α is the full crossing angle [70]. The cumulative progression of luminosity over time during the CMS Run 2 data taking period is shown in Fig. 2.2.

A summary of the main LHC machine beam parameters is presented in Table 2.1 and compared to the design beam configuration for the data taking years 2012, during Run 1, and 2016, 2017 and 2018 during Run 2.

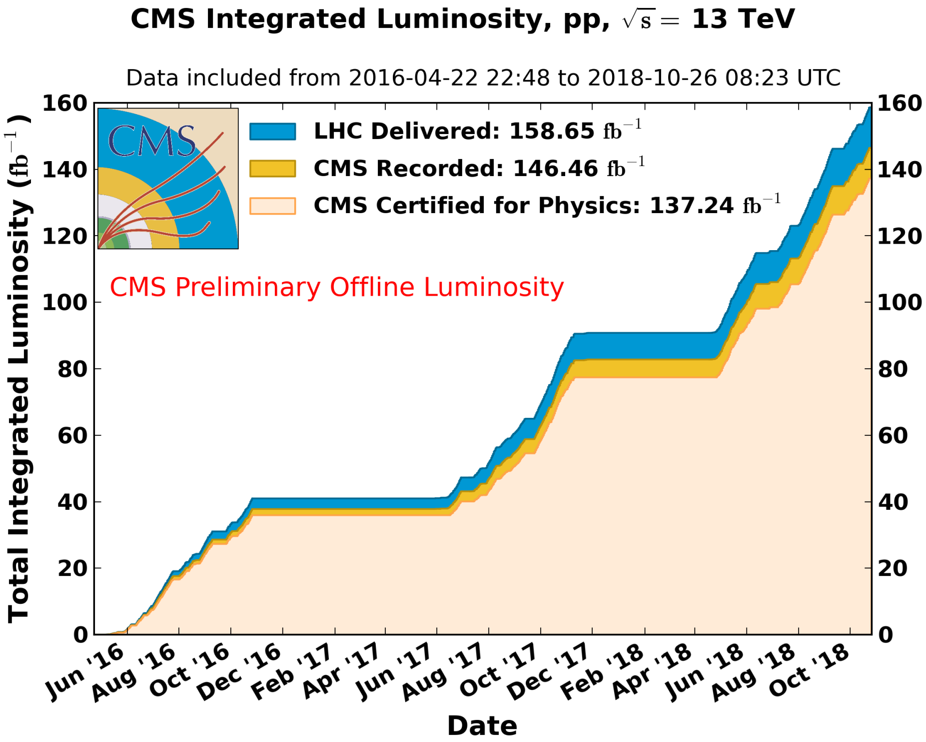


Figure 2.2: The plot shows the cumulative curves for the luminosity delivered by LHC (azure), recorded by CMS (orange) and certified as good for physics analysis during stable beams (light orange). The luminosity validated for physics analysis corresponds to data recorded with all detectors and reconstructed physics objects showing good performance. Taken from [81].

Beam parameter	Design	2012	2016	2017	2018
beam energy [TeV]	7	4	6.5	6.5	6.5
bunch spacing [ns]	25	50	25	25	25
N_p [10^{11} ppb]	1.15	1.65	1.1	1.15	1.15
N_b	2808	1374	2220	2556	2556
$\mathcal{L}_{peak}$ [10^{34} cm^{-2}s^{-1}]	1	0.75	1.4	2.06	2.01
$<\mu>$	20	21	27	38	37

Table 2.1: The main LHC machine parameters for the data production years 2012 in Run 1, and 2016, 2017, 2018 in Run 2 compared to the design beam configuration [110]. The mean number of interactions per bunch crossing (pile-up) denoted with $<\mu>$ is taken from [81]. To compare the pile-up conditions at different values of the center-of-mass-energy, the theoretical prediction of the pp inelastic cross-section derived with PYTHIA [102] is used, i.e. $\sigma_{in}^{pp}(8\text{ TeV}) = 73.0$ mb and $\sigma_{in}^{pp}(13\text{ TeV}) = 80.0$ mb.

2.2 The CMS experiment

The CMS detector was designed to cover a wide variety of physics searches and precision measurements from the Higgs boson to Dark matter, SUSY and others. The name of this detector stems from the compact multilayered design which is realized by placing the 12.5 m long and 6 m wide, 4 T superconducting solenoid magnet in a way that encapsulates most of the other detector parts. Only the muon chambers, return yoke and parts of the calorimeter system are hosted on its periphery. This design results in a magnetic field configuration that have a large bending power for charged particles and results in precise measurement of track momentum [30]. The total length of the CMS detector is around 21.6 meter while being 14.6 meter wide, and it weighs around 12,500 tons.

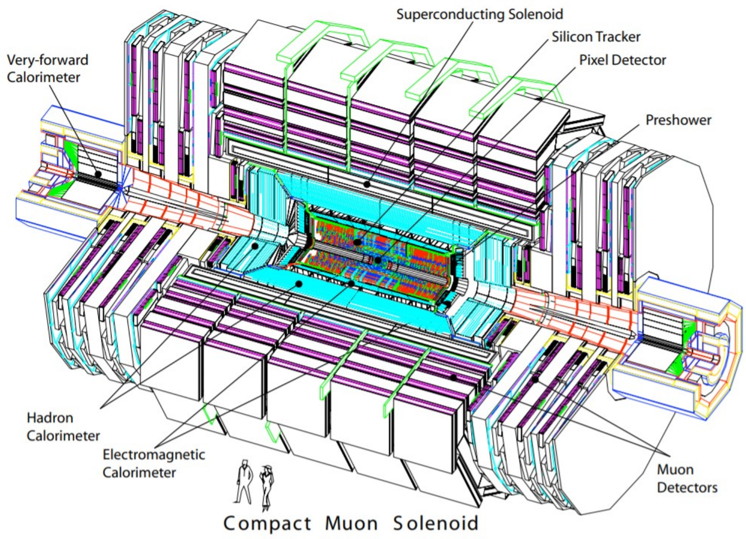


Figure 2.3: 3D schematics of the CMS detector, taken from [30]

The subdetectors of the CMS detector shown in Fig. 2.3, are designed to deliver the following; good muon identification and resolution in the Muon system, good tracking and reconstruction of collision vertex for trajectories of the collision products, in the tracker system; good resolution on electromagnetic energy deposites, which is crucial for diphoton final states, in the Electromagnetic Calorimeter (ECAL); good resolution of energy deposits from hadrons, in the Hadronic Calorimeter (HCAL).

2.2.1 The coordinate system

CMS uses a right-handed coordinate system with the origin at the nominal collision vertex. The z-axis coincides with the beam direction and the y-axis extends upwards, the x-axis on the other hand points in the direction of the LHC ring's center. The coordinate system is shown in Fig. 2.4 where the angles θ

and ϕ and their relation to the axes are depicted and have the following ranges, $0 < \theta < 2\pi$ and $-\pi < \phi < \pi$.

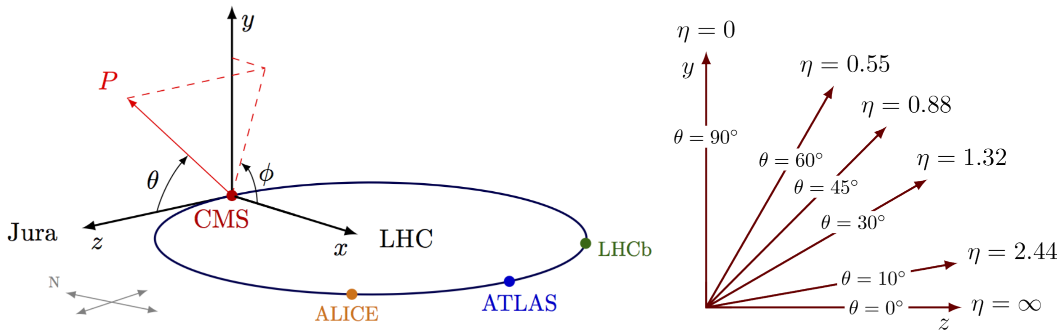


Figure 2.4: The z-axis coincides with the beam direction and the y-axis extends upwards, the x-axis points in the direction of the LHC ring's center, $0 < \theta < 2\pi$ and $-\pi < \phi < \pi$ (left panel), pseudorapidity (right panel), taken from [84].

Another helpful convention in particle physics is to describe kinematical properties in terms of pseudorapidity, which is used with the following definition,

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right). \tag{2.2}$$

The spatial distances in $\eta\phi$-plane is often expressed in terms of ΔR quantity, expressed as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{2.3}$$

The momenta of particles and their energies are expressed in terms of their projection on the transverse plane p_T and E_T. Finally, the imbalance in the measured momenta of all particles in the event is defined as the missing transverse energy (MET), denoted by $\not{E}_T$.

2.2.2 Tracker

The innermost part of the CMS detector hosts the Tracker system, it reconstructs the tracks of each charged particle traveling in the magnetic field in three-dimensions. The kinematical properties of each particle (e.g. p_T) and the primary and the secondary vertices of the event, which are the primary point of the proton-proton interaction and the secondary interaction point which corresponds to the decayed particle respectively, are deduced from these reconstructed tracks.

As seen in the Fig. 2.5, the tracker system surrounds the interaction point and extends from 4.4 cm to 120 cm along the denoted r-axis of the plot. The total length of this system reaches up to approximately 2.7 m in each direction along the z-axis, and with its two major subcomponents namely the pixel tracker and the strip tracker, the tracker system covers an area of 200 m^2 and a pseudorapidity range up to $|\eta| < 2.5$.

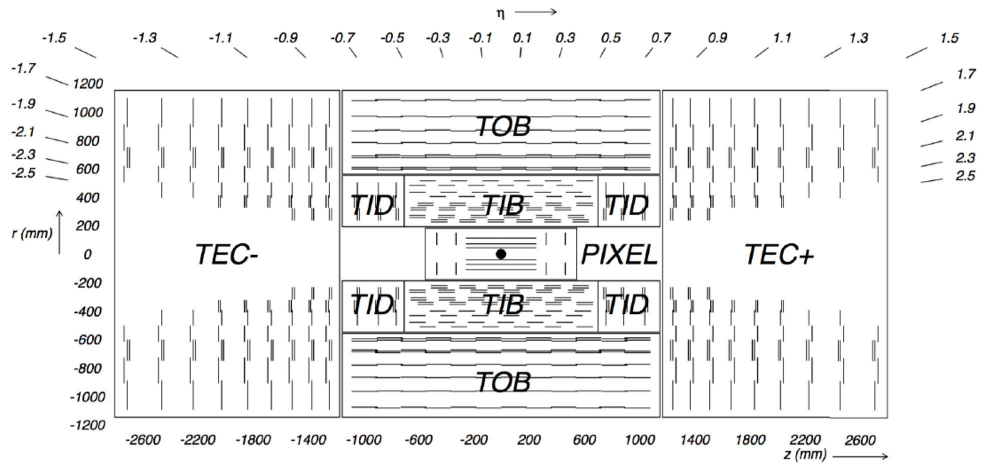


Figure 2.5: Longitudinal section view of the CMS tracker, demonstrating the positioning of the modules and components. Sections are denoted as, the tracker inner barrel (TIB), the tracker outer barrel (TOB), the tracker inner discs (TID) and the tracker endcaps (TEC), taken from [106].

The design based on semiconductor modules offers several benefits that are crucial in good tracking and vertexing. Low charge collection time which results in faster signal processing compared to gaseous solutions, high density which means large number of electron-hole pairs created when a minimum ionising particle traverse the modules ($\sim$80 electron-hole pairs are created per μm for Si [14]) are examples of the benefits from this design choice. Furthermore, the silicon material exhibits mechanical stability and is well-studied for radiation damage effects.

Pixel tracker

The pixel tracker is made of approximately 1,440 pixel modules (100 $\mu m \times$ 150 μm modules) placed in three cylindrical barrels (BPIX) and of two disks on each end of the barrel (FPIX).

In 2017, the pixel detector was upgraded and completely replaced, as the radiation damage compromised the performance. Additional pixel layers were added for barrel and endcaps, to improve the performance [92]. The added layers are shown and compared to the old system in Fig. 2.6, the top half of the figure shows the new arrangement which covers more area and increases the resolution of the tracks compared to the bottom half of the figure which depicts the old system.

Strip tracker

The strip tracker is the outer part of the tracker system engulfing the pixel tracker. It has 15148 silicon modules and roughly 9.3 million strips. Similar to the pixel tracker, it has sections which are organized in cylindrical barrels of different sizes (TIB and TOB) and disks of different sizes on each end of these

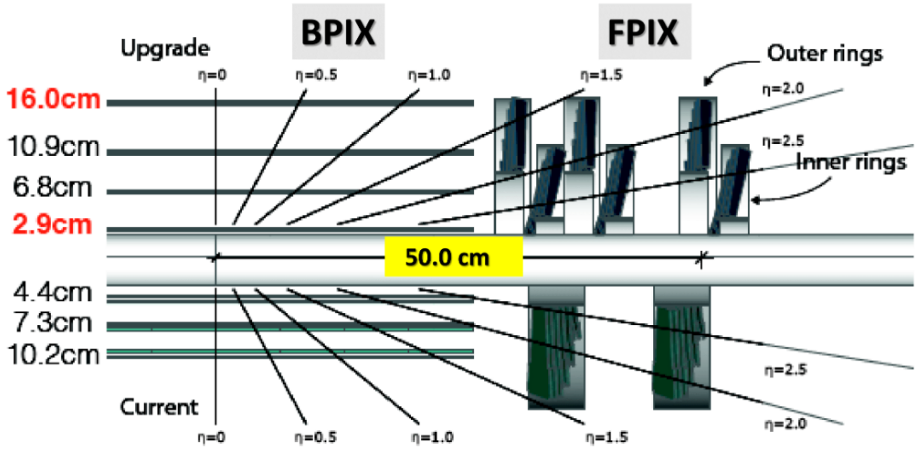


Figure 2.6: Geometrical scheme of the upgraded pixel detector (top) compared with the 2016 legacy system (bottom), taken from [92].

barrels (TID and TEC), refer to Fig. 2.5 for the geometrical placement.

2.2.3 Electromagnetic calorimeter

As previously stated, the energy of the collision products needs to be measured precisely. The electromagnetic calorimeter (ECAL) measures the energy of the incident electrons and photons, by stopping them through interaction with the dense lead-tungstate scintillating crystals.

In the ECAL, 61,200 lead-tungstate crystals are used which are kept at a fixed temperature using a water cooling system. Its properties such as short radiation length (0.89 cm) and small Moliere radius (2.2 cm) [105] allows ECAL to be highly granular and compact.

The ECAL submodules are as follows; a cylindrical arrangement of the crystals which covers the pseudorapidity range $|\eta| < 1.479$ called ECAL barrel (EB), in Fig. 2.7 one of the barrel submodules is depicted with yellow color, two disks on each end of the barrel structure called ECAL endcaps (EE) which can be seen in Fig. 2.7 in green, this module covers $1.479 < |\eta| < 3.0$ and two additional lead absorbers to reject neutral pion decay to two photons called the pre-shower subdetector (ES) depicted with pink color in Fig. 2.7.

The relative ECAL energy resolution σ_E/E from test beam measurement [6] is:

$$\frac{\sigma_E}{E} = \frac{2.8\%}{\sqrt{E(\text{GeV})}} \oplus \frac{12\%}{E(\text{GeV})} \oplus 0.3\%, \qquad (2.4)$$

where the $\oplus$ symbol means that the uncertainties have to be added in quadrature.

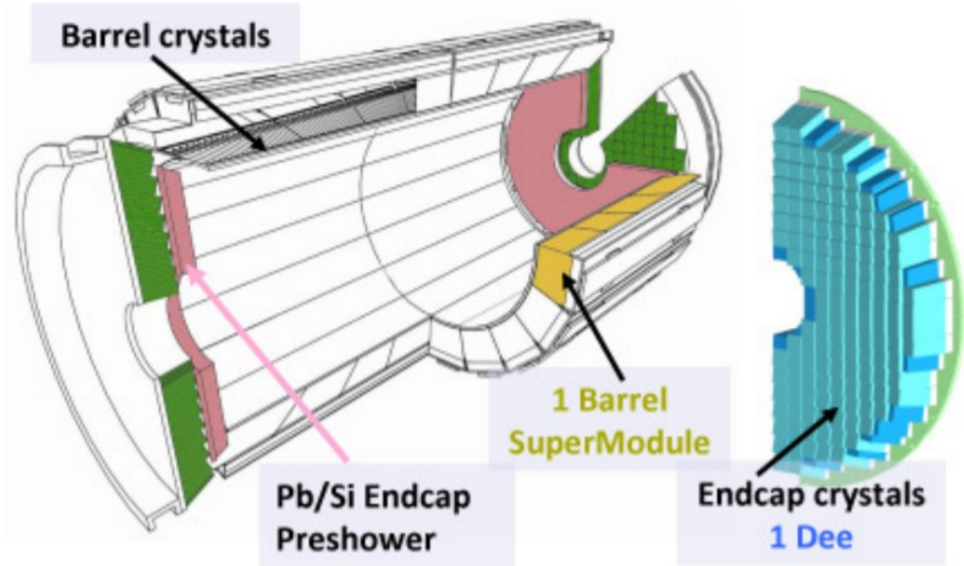


Figure 2.7: The general schematics of the electromagnetic calorimeter ECAL, taken from [17]

2.2.4 Hadronic calorimeter

Similar to the ECAL, the Hadronic Calorimeter (HCAL) is a calorimeter, i.e. it measures energies by stopping the particles through interactions. It is designed with scintillating (plastic tiles) and absorber (brass) material layers spacing one another, therefore it is a sampling calorimeter [32].

The hadronic calorimeter has extensions to additionally capture the energy leakage from the HCAL barrel (annotated with HB in Fig. 2.8) called the outer layer and annotated with HO in Fig. 2.8. Where the HO uses magnet and iron yoke as absorbers and plastic tiles as scintillators; the forward sections annotated with HF in Fig. 2.8 uses iron absorbers and quartz fibers parallel to the beam as scintillators; the HCAL endcap (annotated with HE in Fig. 2.8) similar to the HB uses brass as absorber and plastic tiles as scintillators; these pieces provide a $|\eta| = 5.2$ coverage.

The energy resolution of HCAL combined with ECAL for hadrons, from [25], is

$$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E(\mathrm{GeV})}} \oplus 5\%, \tag{2.5}$$

where the $\oplus$ symbol means that the uncertainties to be added in quadrature.

2.2.5 Muon system

The detection of muons with good resolution and accuracy is one of the design goals of CMS. To gain good resolution for muons with high momenta which traverse the calorimeters with small energy loss, the muon system is designed to cover large areas and long distances.

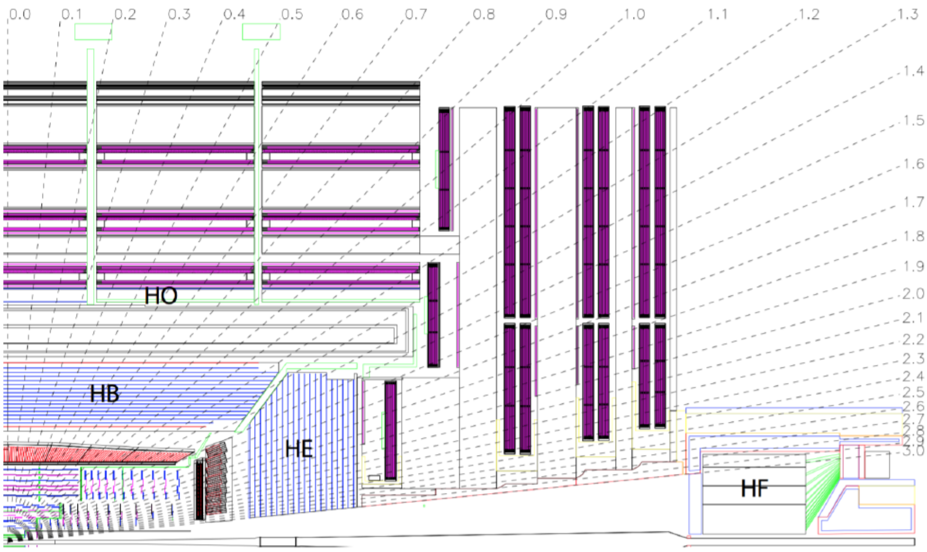


Figure 2.8: Longitudinal slice of a quarter of the CMS detector, showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters.

In CMS, to achieve this goal, three types of gaseous detector modules were designed and implemented. In gaseous detectors, the incident particle produces a signal by ionizing the gas inside chambers.

The three components are as follows, the Cathode Strip Chambers (CSC), which as the name suggests is made of cathode strips and anode wires, these strips and wires are arranged perpendicular to one another, the CSC layers are stationed in the endcaps, see Fig. 2.9. The Drift tubes (DT), stationed in the barrel, are arranged perpendicular to the possible muon trajectory. The Resistive plate chambers (RPC), are anode and cathode plates separated by a gas volume. They are present both in endcaps and barrel, see Fig. 2.9.

2.2.6 Trigger and data acquisition

In LHC, collisions occur at a very high rate and volume and the storage of all corresponding data is not feasible. The trigger system provides an efficient and reliable way to select online the relevant data and discard the rest.

To this end, in CMS, the trigger system is composed of two levels of filtering. The Level 1 (L1), which is hardware level filtering, uses the information from calorimeters and the muon system as depicted in Fig. 2.10 to bring down the rate of the data input stream from bunch collisions with a rate of 40 MHz to a few 100 kHz.

The second level, which is called the High Level Trigger (HLT), is implemented as a software system and by adding the information from different parts of the detector reduces the data rate further to few kHz and decides on the storage. The event reconstruction takes roughly 40 seconds in CMS where the

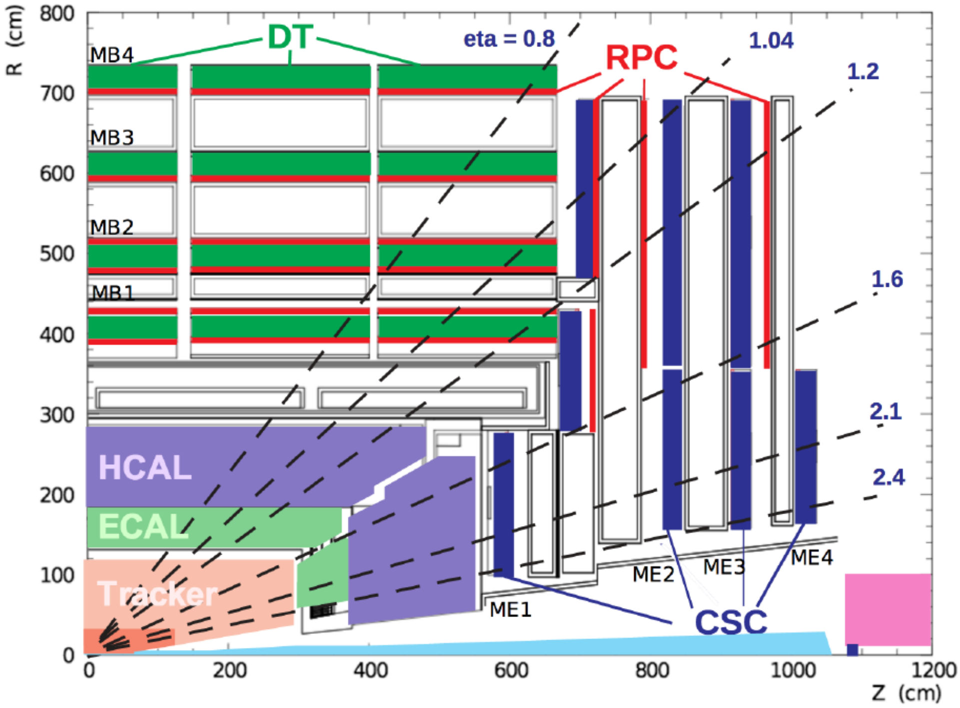


Figure 2.9: Layout of the CMS Muon system [30].

L1 step takes roughly 4 μs and HLT around 300 ms. These time scales rely on incorporation of Field Programmable Gate Arrays (FPGA), Application Specific Integrated Circuits (ASIC) and high performance computing solutions chosen by the CMS collaboration.

The Data stored in data centers are further controlled and certified by examining the status of the detector in the acquisition time, in the Data Quality Monitoring (DQM) system of CMS before it is used in physics analysis.

The collected raw data is further processed to reconstruct the physics objects from detector information, the next chapter discusses the physics objects.

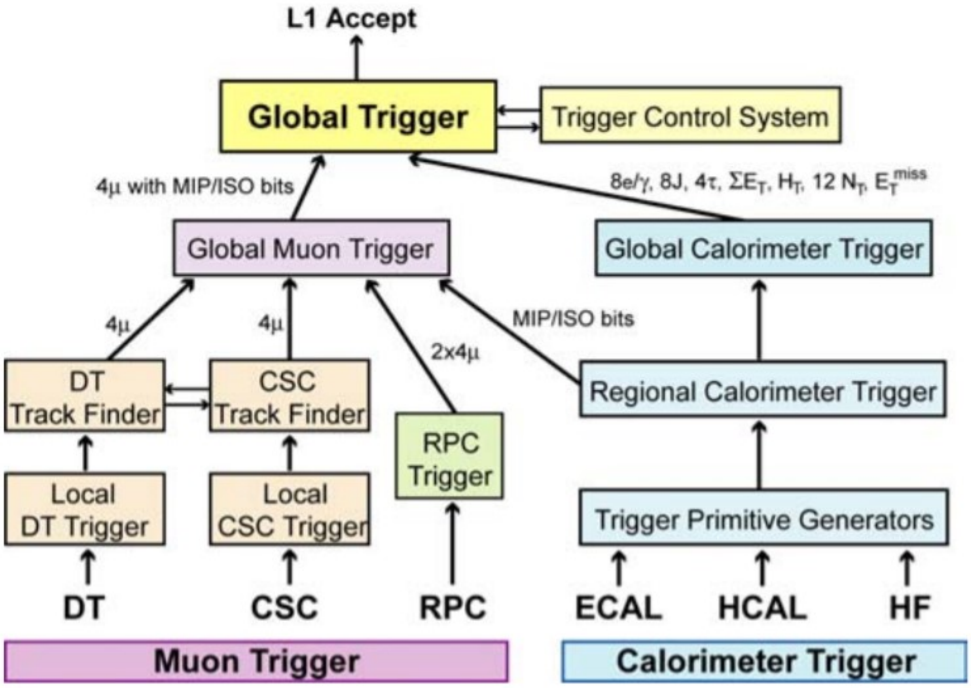


Figure 2.10: Overview of the L1 Trigger architecture, taken from [23].

Chapter 3

Event Simulation and Reconstruction

In this chapter, an overview of the physics objects and the simulation for a proton-proton collision event at the LHC is given.

In this thesis, the VH process as defined in Section 1.4.2 is studied in final states with two electrons or two muons as $ZH \rightarrow ll b\bar{b}$ denoted by two-lepton channel, or with one electron or one muon as $WH \rightarrow l v b\bar{b}$ denoted by one-lepton channel and with no muon or electrons present in the final state as $ZH \rightarrow v\bar{v}b\bar{b}$ denoted by zero-lepton channel. All channels have also two b-quarks in the final state. The physics object reconstruction relevant for the study of these channels is described in this section.

3.1 Simulation

To analyze the data generated by collisions in the CMS detector, there is the need of computer generated simulations of the relevant physics processes (generator-level) and the detector behavior in response to the particles created by these processes (reconstruction-level).

The proton-proton collision simulation needs to deal with multiple levels of complexity. Figure 3.1 demonstrates the approach used by general purpose Monte Carlo generators used in particle physics.

The red circle in the center of the graph denotes the hard process, in which calculations are performed in perturbative regime, it is the result of the collision of the hardest momentum partons (constituents of hadrons, here the incident proton) that carry a fraction of the momentum of the proton. The behavior of these fractions are determined experimentally, and they are formulated as Parton Distribution Functions (PDF) [103].

The red branched out structures, next to the red circle, are radiations and splittings, the blue lines represent the Initial and Final State radiations (ISR and FSR), all of which are simulated with parton showers [68].

The final state hadrons are demonstrated with green circles, the process of hadronization is modeled with phenomenological and effective models [108] as in the confinement regime perturbation methods cannot be applied. Additional Multi-Parton Interactions (MPI) often have small momentum transfers and are modeled with similar models, these are shown as purple circle and lines.

The details on the Monte Carlo samples and generators used in this thesis are given in Section 4.3.3.

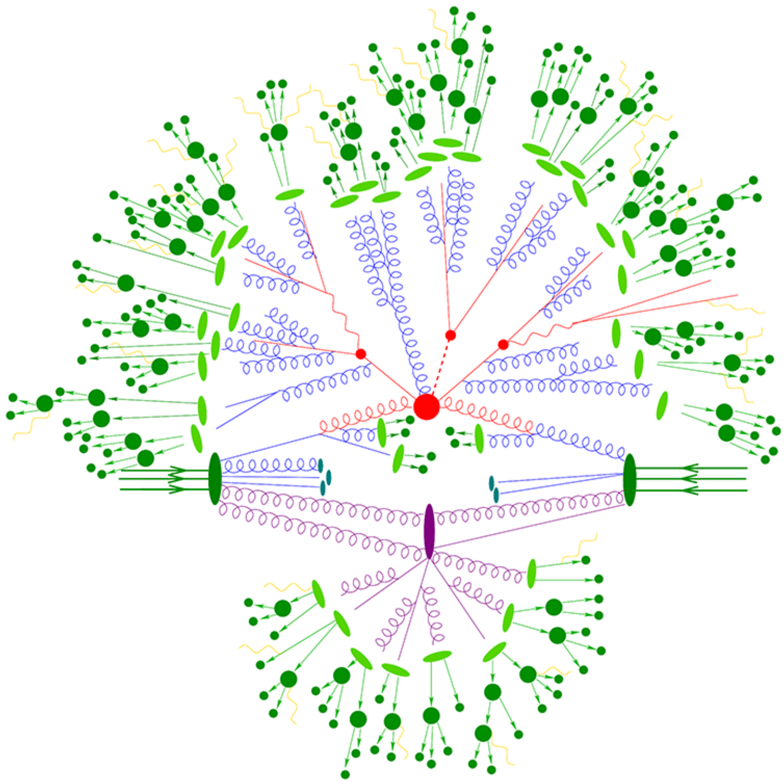


Figure 3.1: Schematic of a proton-proton collision Monte Carlo simulation, taken from [59].

3.2 Physics event and object reconstruction in CMS

The data collected from the CMS detector comprises the response generated from different detector components due to their interactions with the particles emerging from the hadronic collision at the collision point.

The detector is designed so that each particle has its own signature (Fig. 3.2), therefore the identification and measurement of its properties (reconstruction) is possible. In the following, the reconstruction process used in CMS and for objects that are relevant to the study mentioned in this thesis is briefly described.

3.2.1 Particle-flow algorithm

In order to reconstruct all the stable particles (photons, charged and neutral hadrons, muons and electrons) in an event, a global reconstruction algorithm

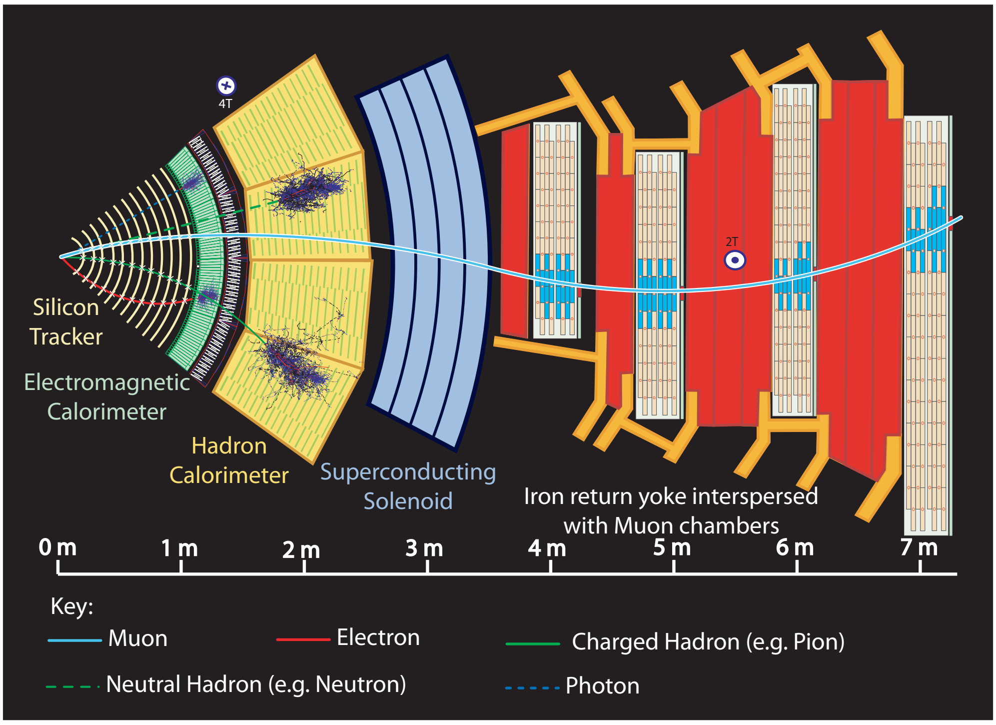


Figure 3.2: Signatures of different particles in the CMS detector, taken from [31].

called Particle-flow (PF) [44] algorithm is used in CMS. It is called a global algorithm as it uses and combines the information from all sub-detectors. It reconstructs the particles to the degree that the output can be compared and treated similar to the physics simulations generated by general purpose Monte Carlo generators [95].

An example of this algorithm is shown in Fig. 3.3, as the charged particle tracks from tracker are combined with calorimeter-clusters. Therefore, the charged hadron is identified when the tracks associated with it, have counterparts in one or more calorimeter clusters. For photons and neutral hadrons only the calorimeter clusters are used. In case of electrons, the cluster in electromagnetic calorimeter is considered with associated track and no leftover signal in the hadronic calorimeter. The muon-identification in the PF-algorithm uses the tracks from the tracker and the muon system.

3.2.2 Tracks

As tracks play a fundamental role on reconstruction of other physics objects, their precise reconstruction is crucial. To reconstruct them, the signal caused by passage of a charged-particle through the silicon modules in the tracker system, which is called a hit is used. A seed of 2-3 hits corresponding to a track (starting with tracks that have relatively large p_T and are close to the interaction region) is iteratively used and removed for the next iteration to reduce the combinatorial complexity in the Combinatorial Track Finder (CTF) [43] algorithm, which is adapted from the Kalman Filter (KF) [73]. The comparison between the CTF

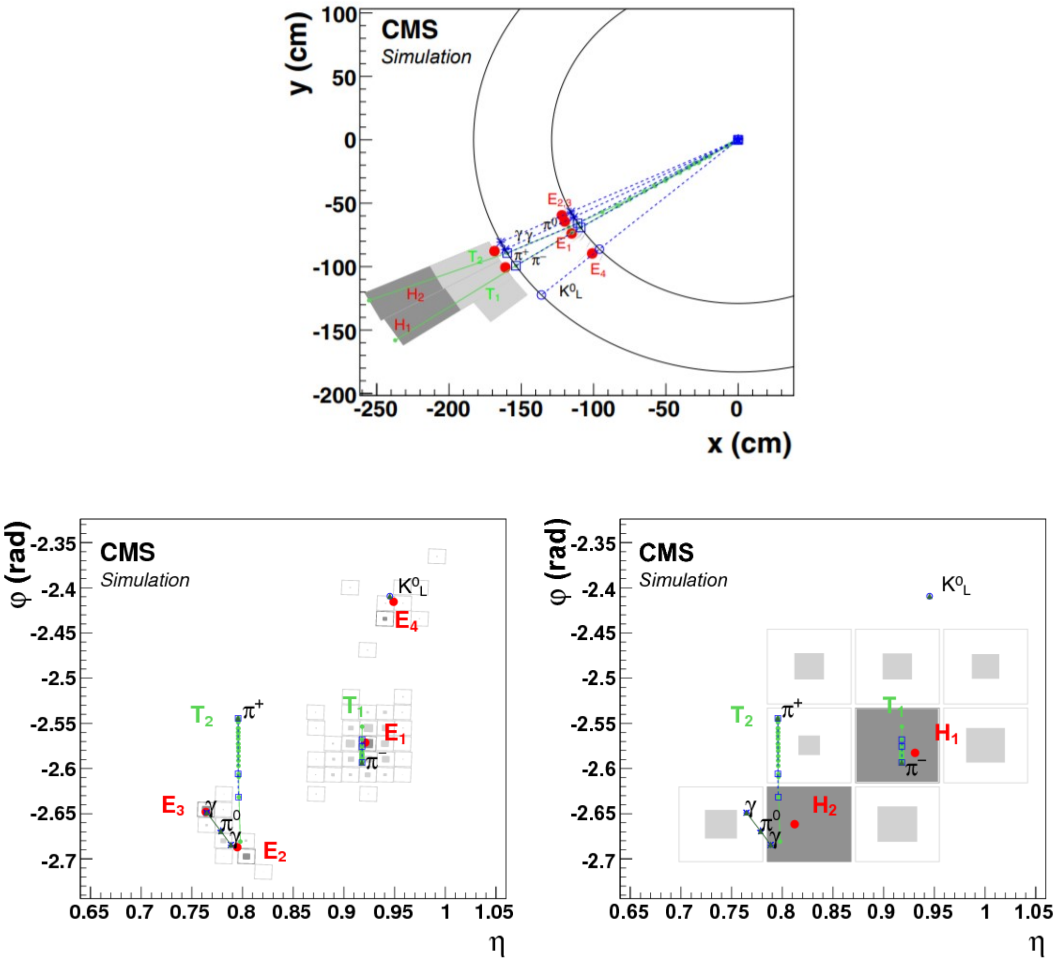


Figure 3.3: Event display, to illustrative particle-flow for a jet consisting of five particles. The associated tracks (top), the (η, ϕ)-view on the ECAL surface (left) and the HCAL surface (right) are shown. The ECAL and HCAL cells are demonstrated as squares. The cluster positions are represented by dots and dashed lines represent simulated particles. $E_{1,2,3,4}$ are the associated ECAL clusters, $T_{1,2}$ are charged particle tracks and $H_{1,2}$ denote the HCAL clusters, taken from [44, 96].

and the iterative tracking method is shown in Fig. 3.4

It is expected that about 1000 charged particles produced by on average more than twenty proton-proton interactions, traverse through the tracker at each bunch crossing [43]. These multiple interactions that may also have contributions from prior or later bunch crossings are known as pileup. The mean number of interactions per crossing (pile-up (PU)) is shown in Fig. 3.5 for multiple data taking eras.

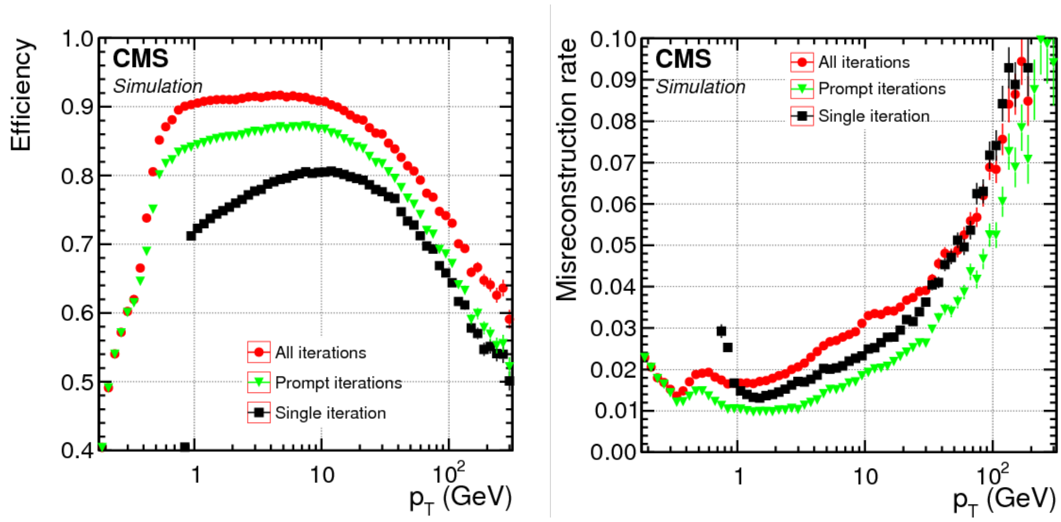


Figure 3.4: Efficiency (left) and mis-reconstruction rate (right) of the global combinatorial track finder (black squares); and of the iterative tracking method (green triangles: prompt iterations based on seeds with at least one hit in the pixel detector; red circles: all iterations, including those with displaced seeds), as a function of the track p_T, for charged hadrons in multi-jet events without pileup interactions. Only tracks with $|\eta| < 2.5$ are considered in the efficiency and mis-reconstruction rate determination. The efficiency is displayed for tracks originating from within 3.5 cm of the beam axis and $\pm$ 30 cm of the nominal center of CMS along the beam axis, taken from [44].

3.2.3 Vertices

Primary vertices (PV) are reconstructed to measure the position of all proton-proton collision in the event. To this end tracks are selected based on their compatibility with the beam spot, number of hits and fit quality. The selected tracks are then clustered into primary vertex candidates on the basis of their z-coordinates at their point of closest approach to the centre of the beam spot [43] and a vertex fit is performed using the deterministic annealing algorithm (DA) [91] to find the vertex candidates. Next, those candidates based on the DA clustering in z with at least two tracks are fitted using Adaptive Vertex Fitter (AVF)[55] to improve the estimate of the vertex parameters. In the AVF, each track in the vertex is assigned a weight from 0 to 1, which reflects the likelihood of belonging to the vertex.

3.2.4 Jets

Fluxes of stable particles coming from hadronization of quarks or gluons are normally considered as jets. To find the properties of the original parton, jets need to be reconstructed.

Jets are reconstructed using the anti–k_T clustering algorithm [21]. The anti–k_T is a sequential-recombination clustering method where a distance is defined between the constituents of the jet and is used as the metric to decide for in-

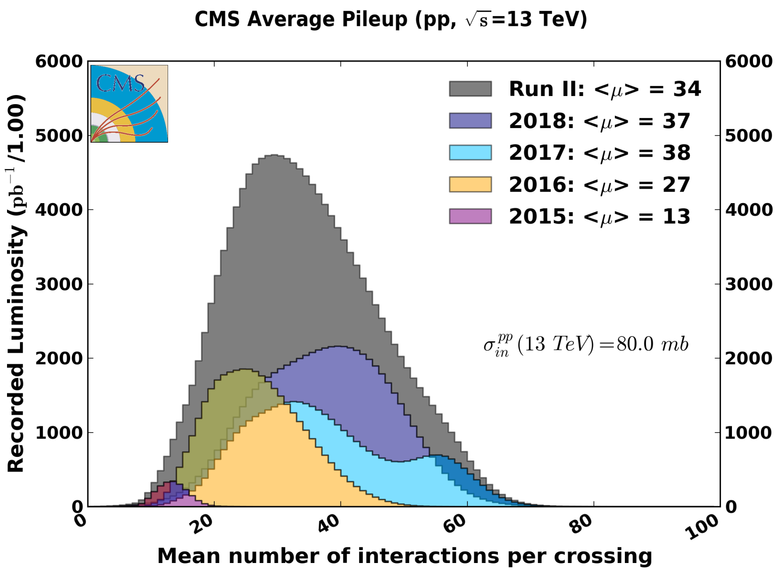


Figure 3.5: Mean number of interactions per bunch crossing (pile-up) for the Run II proton-proton collisions, taken from [81].

clusion or exclusion. The metric is calculated iteratively by addition of each subcomponent of the event to find the best candidate. For each subcomponent i with transverse momentum $k_{T,i}$, the distance from the beam axis is calculated using the following:

$$d_{iB} = k_{T,i}^{2p}, \tag{3.1}$$

and for each possible pairs of the subcomponents the d_{ij} distance is calculated as defined in the following:

$$d_{ij} = \min\left(k_{T,i}^{2p}, k_{T,j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2}, \qquad \text{where} \qquad \Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{3.2}$$

where $k_{T,i}$, ϕ_i, y_i representing respectively the transverse momentum, the azimuthal angle and the rapidity of the constituent i and $p = -1$. To cluster the jets, iteratively, the smallest distance is found using particles or groups of particles. If the smallest distance is d_{iB}, then the entity is assigned to a jet and if the smallest distance is d_{ij}, then i and j are merged together and considered again as a new entity. In Fig. 3.6, such clusters are schematically presented, the active area of the anti–k_T is referred to as the jet ΔR cone.

The anti–k_T jets in this thesis are classified into two general categories, jets detected with a spatial cone of $\Delta R = 0.4$ are called resolved and jets with $\Delta R = 0.8$ are called fat jets.

One of the issues that might arise in the jet reconstruction, is caused by overlap of multiple jets that may result in jets with high p_{T}. The probability of overlapping jets increases with the number of proton-proton interactions present in the event. The correct p_T of the jet is extracted by removing the overlapping jets with a multivariate method documented in [89].

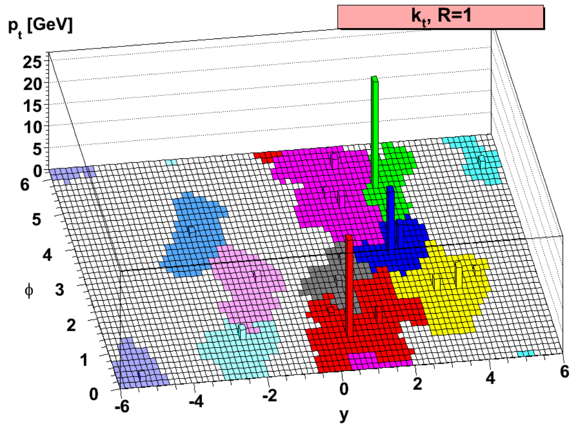


Figure 3.6: An example an event clustered with the anti–k_T jet algorithm, this illustrates how the active area of the clustered jets might look like, this active area is referred to as the jet cone. taken from [21].

To reconstruct the energy of the jet so that it corresponds to the energy of the underlying parton, the jet energy needs to be corrected. The mismatch in energies are coming from non-linear effects in calorimeter response, noise from electronic components and PU. The set of corrections, called Jet Energy Corrections (JEC), corrects the four-vector of momentum for jets. The corrections are applied in multiple stages, the first removes the effect of PU on energy, the second corrects for the non-linear response of the detector with the information from Monte Carlo simulation of the detector behavior. The next two steps correct the residual differences between the data and Monte Carlo simulations and finally the effect arising from differences between the flavor of underlying parton is corrected, see Figs. 3.7 and 3.8. On top of JEC, the jet p_T of the simulated jets are smeared to count for differences between the resolution of jets in data and Monte Carlo simulation. This correction is called the Jet Energy Resolution (JER) correction.

The jet energy composition for an anti–k_T jet clustered with a $\Delta R = 0.4$ for the data taking periods 2016-2018 is shown in Fig. 3.8 as a function of the jet p_T, the ratio plot shows how well each fraction is modeled.

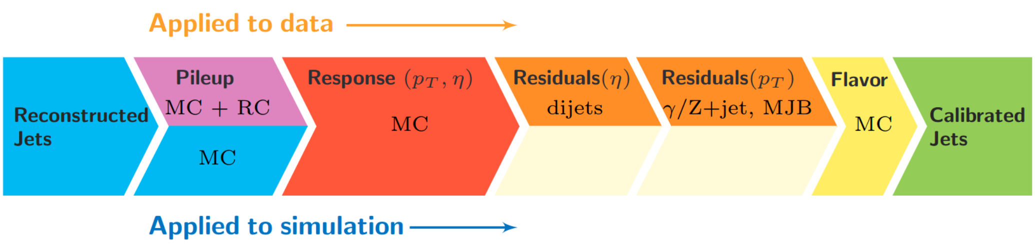


Figure 3.7: Consecutive stages of JEC, for data and MC simulation. All corrections marked with MC are derived from simulation studies, MJB refers to the analysis of multi-jet events, taken from [75].

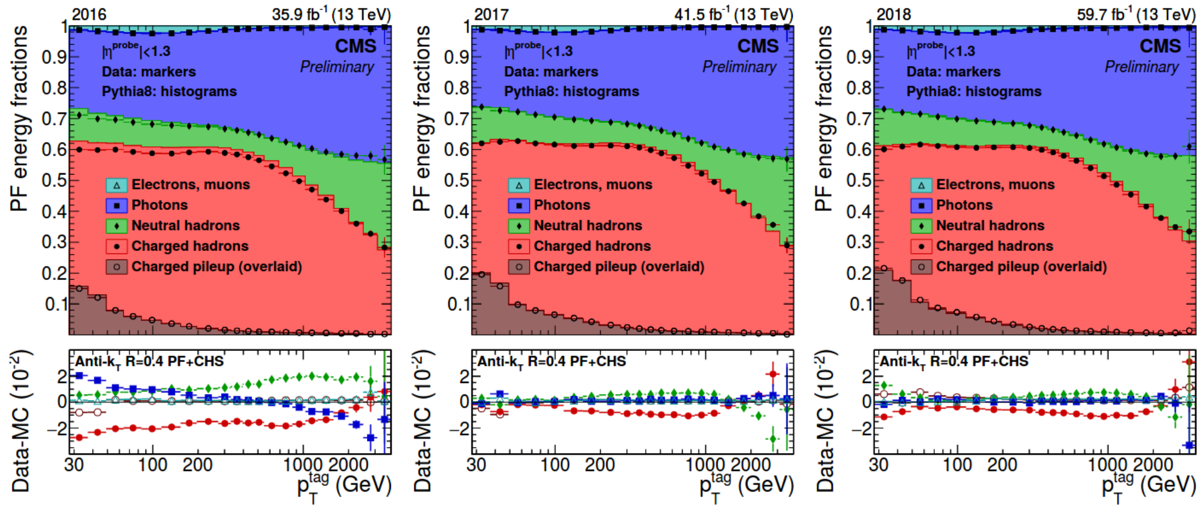


Figure 3.8: Jet energy composition in observed and simulated events as a function of p_{T}, taken from [44, 72].

Heavy flavor tagging

The jets that are originating from a heavy parton (b or c -quark), due to the high mass of the parton have decay products with longer lifetime and higher momenta compared to the jets from gluons or light-quarks (u, d, s). The differences in the properties is the motivation for algorithm and methods that identify the heavy flavor jets. In this thesis, since final states with b-quarks are considered, identifying the jets coming from b-quarks (b-tagging), known as b-jets is a fundamental part.

One of the main inputs to the flavor-tagging algorithms are the secondary vertices, which are the displaced vertices compared to the interaction point for the parton, this is the point where trajectories of its decay products meet. The Inclusive Vertex Finding (IVF) [94] algorithm is used to reconstruct secondary vertices.

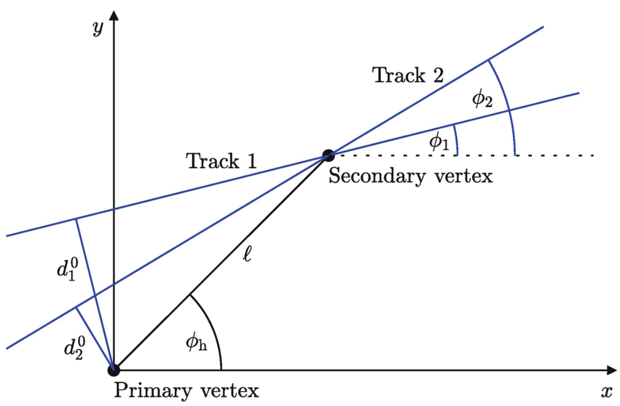


Figure 3.9: The secondary vertex and its relation to the PV, taken from [56].

In order to identify the jet flavors in this thesis, the DeepCSV and the DeepAK8 tagger has been used. These algorithms use secondary vertex and track-based lifetime information alongside other kinematic parameters. The DeepCSV algorithm uses a deep neural network multivariate method, with 5 hidden layers of a width of 100 nodes each. The DeepAK8 algorithm is a multi-class particle identification algorithm [82, 101] capable of identifying hadronic decays of highly boosted top quarks and W, Z, and Higgs bosons and classifying the different decay modes (e.g. $H \rightarrow b\bar{b}, \rightarrow c\bar{c}$ or $\rightarrow q\bar{q}$), based on the anti–k_T jets clustered with a $\Delta R = 0.8$. The DeepAK8 algorithm uses a deep convolutional neural network (CNN) to process the PF candidates and secondary vertices associated with the jet. The performance of each b-tagging algorithm is shown in Fig. 3.10.

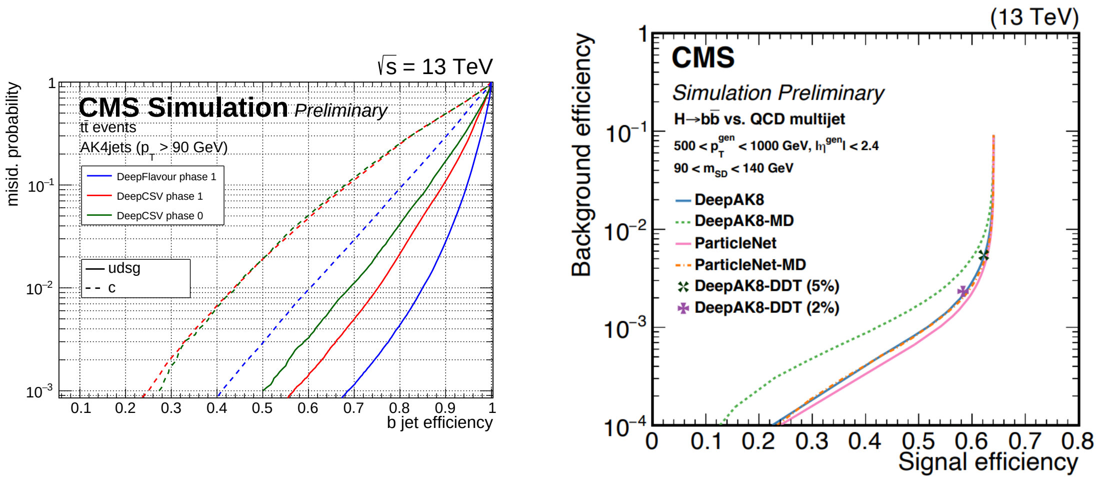


Figure 3.10: Performance of the DeepCSV and DeepJet (also known as DeepFlavour) b-jet identification algorithms demonstrating the probability for jets from partons of other flavor to be misidentified as b jet, as a function of the efficiency to correctly identify b jets (left) and the Performance of the DeepAK8 algorithm variations in comparison to the ParticleNet algorithm for the Higgs decaying to a pair of b-quarks versus QCD multi-jet (right), taken from [71, 88]

The DeepCSV algorithm efficiencies are calculated with different working points, these working points which are denoted by loose, medium and tight correspond to cuts on the DeepCSV discriminant that has probability of misidentifying a jet from partons of other flavor as a b jet of 10% for the loose, 1% for the medium and 0.1% for the tight working points.

Soft activity

Soft hadronic activity, which is a measure of the soft hadronic interactions underlying the primary hard interaction present in the event, can be used to reject backgrounds like $t\bar{t}$, where a large contribution of soft hadronic activity is expected.

To measure soft activity and reconstruct it, high purity tracks with $p_T > 300$ MeV are selected, when they are not associated with the W, Z boson or the selected b-jet in the event. These tracks then are fed to the anti–k_T algorithm with $\Delta R = 0.5$ to be clustered as soft-track jets [24]. A cut on the multiplicity of these soft activity jets with momentum larger than 5 GeV is used to reject backgrounds.

Muons

The muon reconstruction is achieved with information from the muon chambers and the tracker. If the muon is reconstructed by considering all silicon tracker tracks to be potential muon candidates and checking the hypothesis by looking for compatible signatures in the calorimeters and in the muon system, the reconstructed muon is called a tracker muon. On the other hand, if the reconstruction starts with candidates from the muon chambers which are then associated with a well reconstructed track, the muon is called a global muon. The muons selected in this thesis pass the following selection to make sure they are compatible with the channels studied. Some muons might appear as decay products of weak decays of hadrons inside jets and their contribution is controlled by the additional cuts on the relative isolation Eq. (3.3).

The muons are considered with looser selections for the 2-lepton channel and tighter selection for the 1-lepton channel. The loose selection is as follows; the muon candidates can be either a tracker or a global muon, the muon candidates are required to have $p_T > 5\,\mathrm{GeV}$, to be inside the $|\eta| < 2.4$ area of the tracker and are required to be compatible with the primary vertex, therefore satisfying $d_{xy} < 0.5\,\mathrm{cm}$ and $d_z < 1.0\,\mathrm{cm}$. The muons are required to pass the relative isolation cut of $I_{PF} < 0.4$.

The tight selection for muons is as follows; the muon candidate are required to be global muons, the muon candidates are required to have $p_T > 25\,\mathrm{GeV}$, and to have at least five hits in the inner tracker and at least one hit in the pixel detector. The global muon track is required to have a good fit with $\chi^2/ndof < 10$ and required to pass the relative isolation cut of $I_{PF} < 0.06$.

Electrons

The electrons are reconstructed with the Gaussian Sum Filter algorithm [5] (GSF Electrons). A transverse momentum cut of $p_T > 7\,\mathrm{GeV}$, cuts on distances with respect to the primary vertex $d_{xy} < 0.05$ cm, $d_z < 0.2$ cm, and relative isolation cut of less than 0.15 for the 2-electron channel and less than 0.06 for the 1-electron channel are applied for the selection of electron candidates. Then a multivariate approach [12] (rejecting fakes) is utilized with two working points loose and tight corresponding to 90% and 80% efficiencies for one and two electron channels respectively. The use of MVA approach requires the following additional selection:

- The electron candidate is required to have $p_T > 15\,\mathrm{GeV}$.
- The energy deposits in HCAL are required to be less than % of the ECAL energy deposit along the electron track.
- The track sum p_T is required to be less than 18% of the electron p_T.

- Cut of $1.4442 < |\eta| < 1.5660$ to count for the gap in ECAL geometry between endcap and barrel.
- For $|\eta| < 1.4442$ the electron candidate is in the ECAL barrel region and required to have the following addition cuts:
  - The shower shape must satisfy $\sigma_{i\eta i\eta} < 0.012$ which is a covariance variable in terms of the crystal spacing.
  - Isolation (Eq. (3.3)) required to be less than 0.4 for ECAL clusters, and less than 0.25 for HCAL clusters.
  - The location of the electron measured in the cluster and based on the track location is required to be small, within $\Delta|\eta| < 0.0095$ and $\Delta|\phi| < 0.065$.
- For electrons with $|\eta| > 1.5660$ in the ECAL endcap region:
  - The shower shape is required to have $\sigma_{i\eta i\eta} < 0.033$.
  - Isolation (Eq. (3.3)) in the ECAL cluster must be less than 0.45, and isolation in the HCAL must be less than 0.28.

Isolation

The relative isolation is defined as follows, it improves the accuracy of picking prompt (from the primary interaction) leptons:

$$I_{\mathrm{PF}} \equiv \frac{1}{p_{\mathrm{T}}^{\ell}} \Big(\sum p_{\mathrm{T}}^{\mathrm{charged}} + \max \Big[0, \sum p_{\mathrm{T}}^{\mathrm{neutral}} + \sum p_{\mathrm{T}}^{\gamma} - p_{\mathrm{T}}^{\mathrm{PU}}(\ell) \Big] \Big), \tag{3.3}$$

where the sum is over all particles inside cones of different sizes for electron and muons, in the (η, φ) plane around the lepton momentum.

3.2.5 Missing transverse energy

The Missing Transverse Energy (MET) is the undetected part of the energy, expected from the physics process (e.g. decay to neutrinos) in the detector, considering the conservation of energy and momentum.

$$\not{E}_T = \Big| - \sum_{measured} \vec{p_{\mathrm{T}}} \Big|, \tag{3.4}$$

In this work, two approaches in reconstruction were considered. The first one is the reconstruction based on fully reconstructed particle flow particles (PF MET) and the second one, called the tracker MET which considers the missing energy in calorimeters while removing the energy depositions of measured tracks. Furthermore, cuts on $sig(\not{E}_T) = \not{E}_T / \sqrt{\sum_i |\vec{p}_T(\mathrm{jet})|}$ which is known as the MET significance is used, to improve the selection criteria.

3.2.6 Missing hadronic transverse momentum

The Missing Hadronic Transverse (MHT) momentum, is defined as,

$$\not{H}_T = \Big| - \sum_{measured} \vec{p}_{\mathrm{T}_{jets}} \Big|, \tag{3.5}$$

where the sum is performed over all reconstructed jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 5.2$.

3.2.7 Muon efficiency and scale factors

Particles produced in the collision might escape detection by not reaching the detector subcomponent associated with them, or not be reconstructed by the reconstruction algorithm, therefore it is crucial to determine the detector's efficiency for each particle.This process is done centrally in the CMS detector for the general selection criteria.

Since in this thesis the full potential of the muon identification is exploited, the muons are subject to tighter selection criteria than the default approach for which external studies on the efficiency is performed centrally in the CMS collaboration, namely relative isolation cuts of 0.25 and 0.06 for two-lepton and one-lepton muon channels, respectively. This requires to calculate the detector efficiency again with the introduced cuts. An estimation of this efficiency using the simulations would not satisfy the requirement, as simulations themselves need to be calibrated with the collected data. Therefore, a data driven method called tag and probe has been used in this thesis to extract the efficiency and find the ratio between the simulation and data (calibrate the simulation with data), this ratio is called scale factor (SF).

The efficiency for a selection on the muon is considered as follows,

$$\epsilon_\mu = \frac{N_p}{N_p + N_f} \tag{3.6}$$

where N_p is the number of muons passing the selection criteria and N_f is the number of those failing the selection.

Since the selection on the muon isolation is imposed in addition to selection cuts corresponding to the muon identification and tracking, and the trigger selection added after the isolation cut, changing the isolation cuts requires calculating the two other efficiencies as well. Assuming no correlation between the different steps, the efficiency of the total selection can be expressed as,

$$\epsilon^\mu_{ID} \cdot \epsilon^\mu_{ISO|ID} \cdot \epsilon^\mu_{Trigger|ISO}, \tag{3.7}$$

where the relation between the efficiencies for identification (ID), isolation (ISO) and trigger (Trigger) selections are presented as fractions, meaning that each step is calculated after requiring the previous step. The identification and isolation is described in detail in Section 3.2.4. For the trigger, there are two separate p_T thresholds in the double muon trigger requirement, these two thresholds are 8 and 17 GeV, while for the single muon trigger the threshold is 24 GeV for 2016 and 2018 and 27 GeV for the 2017 data-taking. Each of these thresholds are considered separately and the efficiency has been measured for them.

The tag and probe method

The tag and probe method uses a resonance decaying to a pair of particles to find the efficiencies. In the case of muons and in this thesis, the Z is the resonant particle. First, events with exactly one pair of oppositely charged muons are selected, requiring one of the muons to be well identified by requiring stringent selection criteria. This muon is called a tag. The second muon is selected from the most relaxed possible selection for a muon and called a probe muon. The selection under study is applied to the probe muon and then two mass distributions are extracted for the dimuon system created by the probe and tag. One mass distribution is considering the probes passing the selection criteria and the second one considering the probes failing the selection, an example of these distributions can be seen in Fig. 3.11. To extract the efficiencies these distributions are fitted by considering a Voigtian, which is the convolution of a Breit-Wigner function with a Gaussian distribution for the Z mass peak. To model the background shape, an exponential distribution is used for the masses larger than the mass of the Z-boson and an error function is used for the region below or equal to the Z-boson mass. The analytical form is as follows,

$$f_{bkg} = \begin{cases} e^{-(x-Z_{mass})*\gamma} & x > Z_{mass} \\ e^{-(x-Z_{mass})*\gamma} \otimes \text{erf}((\alpha - x) * \beta) & x \leq Z_{mass} \end{cases} .$$

where x is the mass of the dimuon system and the free parameters are constrained by the fit. The fits are performed in five p_{T} and four η bins, see Fig. 3.13.

For the double muon trigger, the calculation of trigger efficiency requires considering permutations of the two muons. The efficiency, for the event passing the double muon trigger, is calculated as follows,

$$\epsilon_{event} = \frac{\epsilon^2_{8\text{GeV\_leg}}(1)\epsilon_{17\text{GeV\_leg}}(2) + \epsilon^2_{8\text{GeV\_leg}}(2)\epsilon_{17\text{GeV\_leg}}(1)}{\epsilon_{17\text{GeV\_leg}}(2) + \epsilon^2_{8\text{GeV\_leg}}(2)} \tag{3.8}$$

where (1) refers to the p_{T} and η of the first muon and similarly (2) refers to the second muon in the event ordered in p_{T} and ϵ denotes efficiency.

The distribution of the efficiency for the isolation cut for one-muon channel for the 2017 data taking period is shown in Fig. 3.12 and the trigger efficiency for each leg of the double-muon trigger for the 2018 data taking period is shown in Figs. 3.13 and 3.14.

The range of scale factors is shown in Table 3.1 which are close to one, the systematic uncertainties are estimated by manipulating the Z boson mass window for the Voigtian function and the number of bins in the fit (the total errors on scale factors are found to be 1-2%).

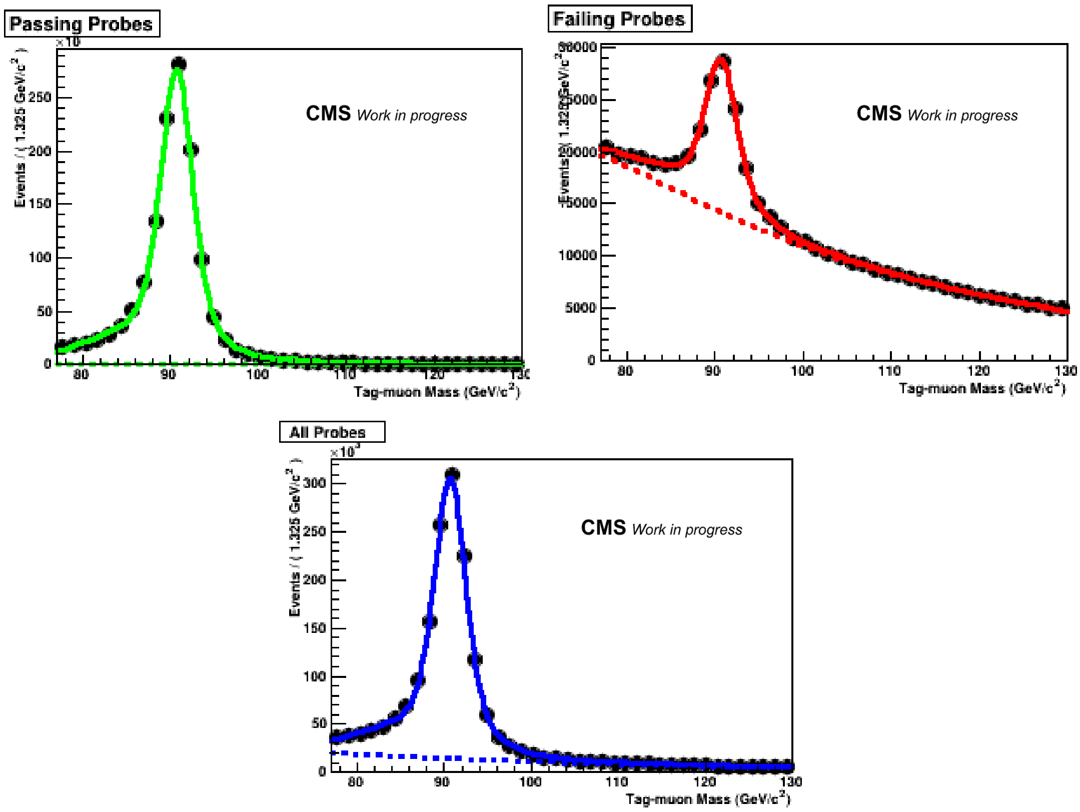


Figure 3.11: The fitted lines to mass distributions of passing probes (top left), failing probes (top right) and all probes (bottom).

Year	Ident. SF	Isolation SF	Single-muon trig. SF	Double-muon trig. SF
2018	0.97-1.0	0.96-1.01	0.98-1.08	0.96-1.03
2017	0.97-0.99	0.98-1.01	0.80-1.00	0.96-0.99
2016	0.96-0.99	0.90-0.98	0.82-1.01	0.88-1.02

Table 3.1: The range of scale factor values for each year and each category.

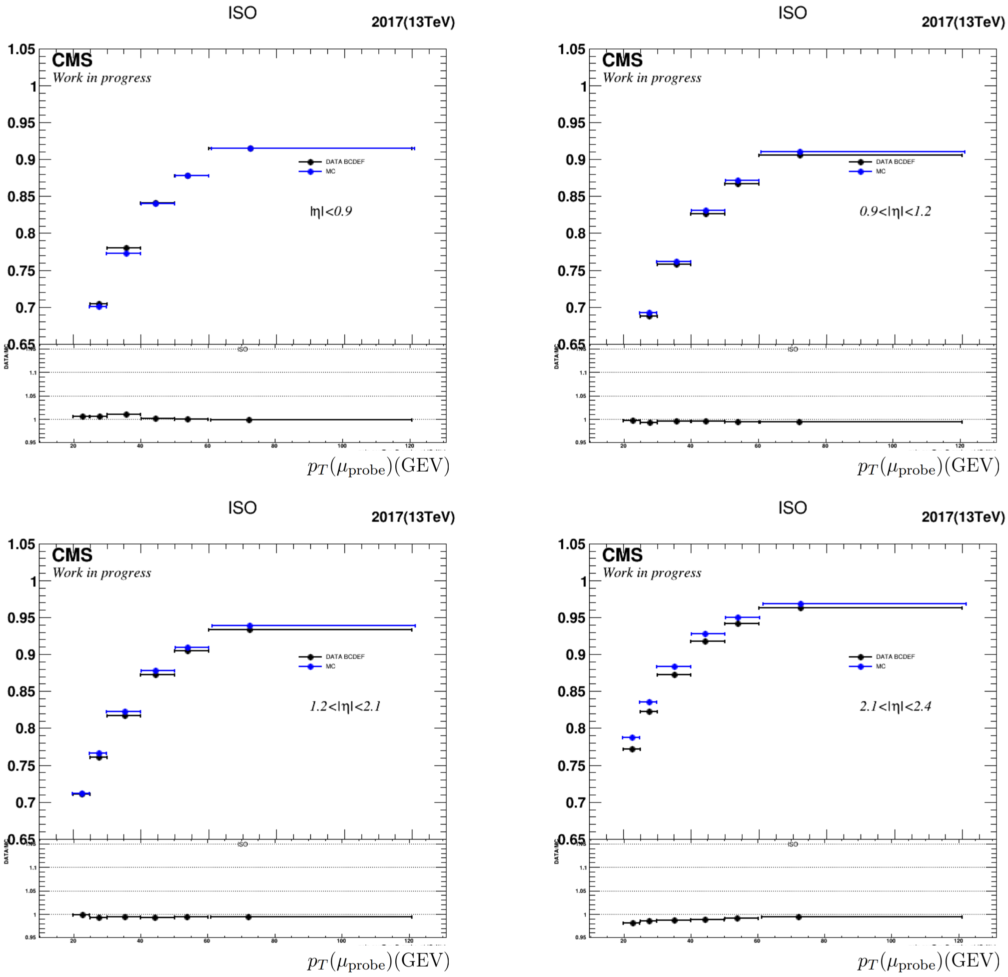


Figure 3.12: Efficiency ($\epsilon^{\mu}_{ISO|ID}$) for a muon to pass the relative isolation (in a cone of $\Delta R < 0.4$) < 0.06 cut after tight ID, as a function of the muon p_T for the 2017 data and MC. The efficiency have been computed in four $|\eta|$ bins. From top left to bottom right: $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$, $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$

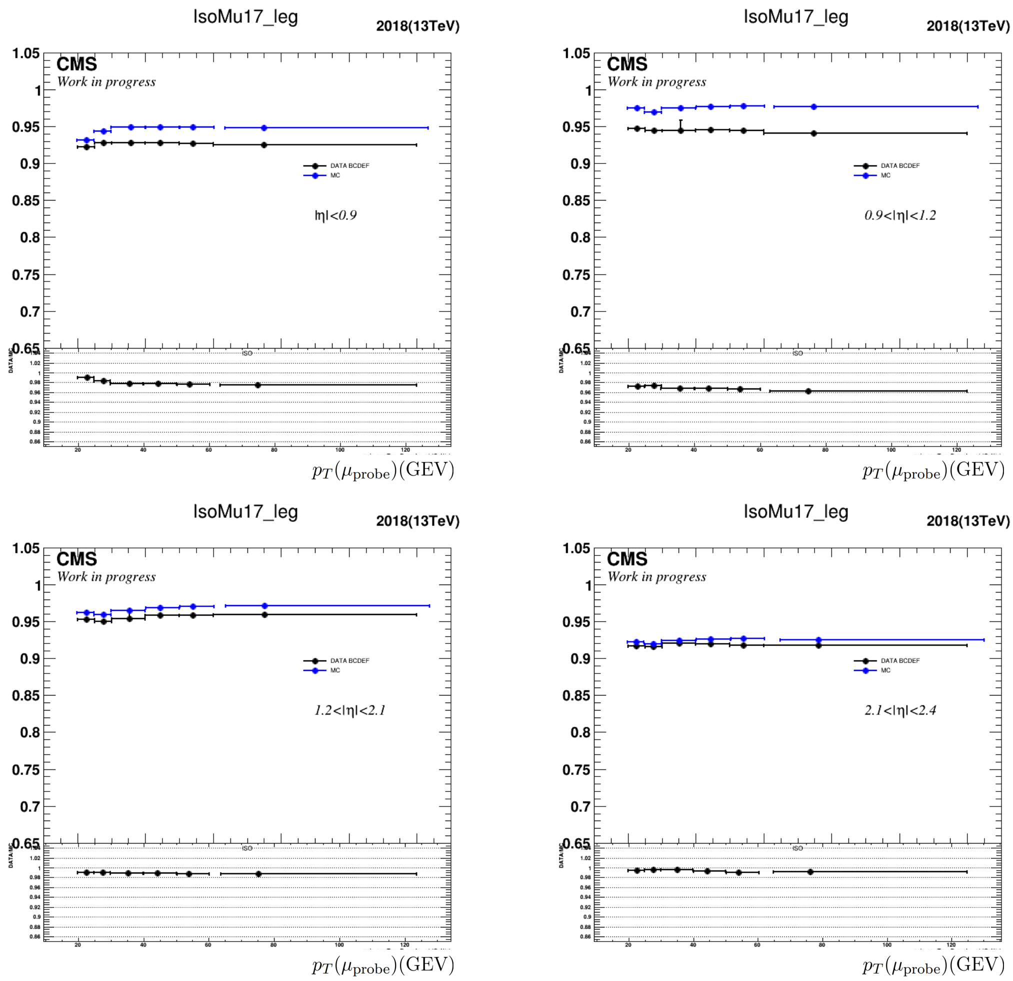


Figure 3.13: Efficiency ($\epsilon^{\mu}_{Trigger|ISO}$) for a muon to pass the 17 GeV threshold trigger, as a function of the muon p_T for the 2018 data and MC. The efficiency have been computed in four $|\eta|$ bins. From top left to bottom right: $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$, $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

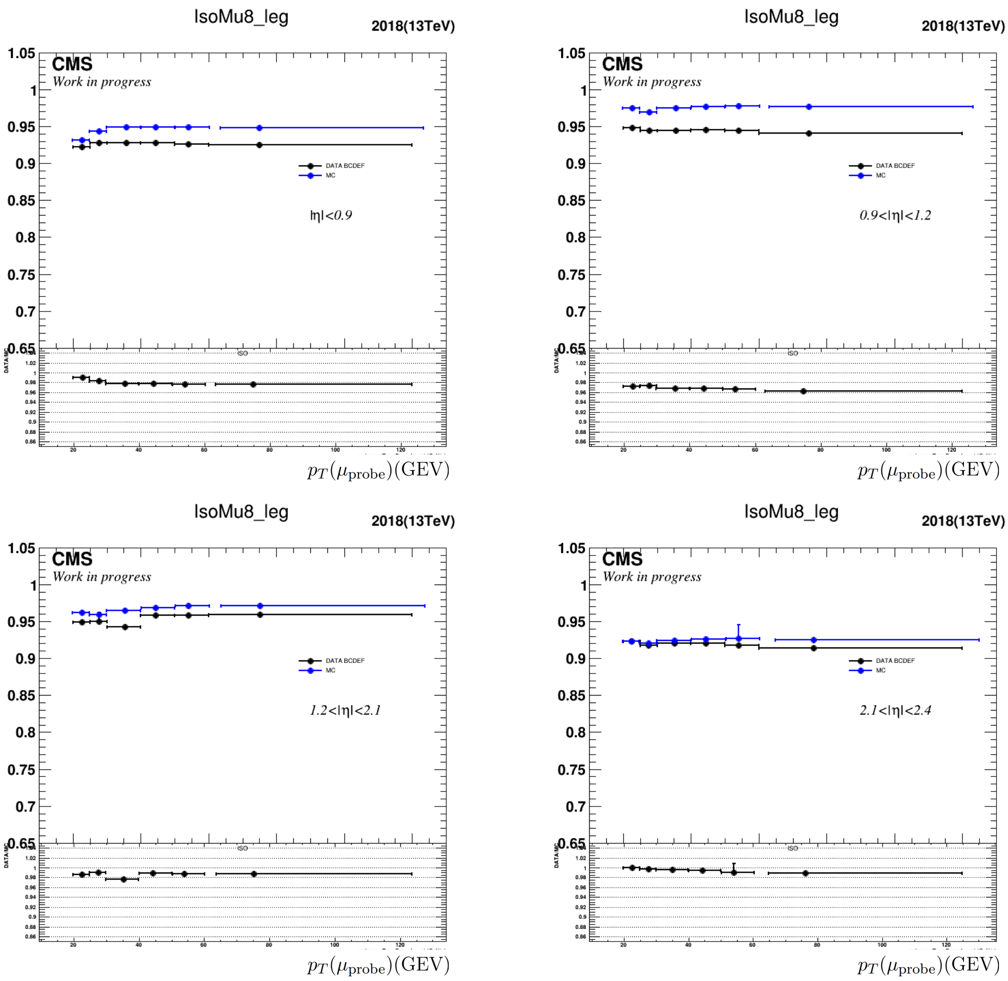


Figure 3.14: Efficiency ($\epsilon^{\mu}_{Trigger|ISO}$) for a muon to pass the 8 GeV threshold trigger, as a function of the muon p_T for the 2018 data and MC. The efficiency have been computed in four $|\eta|$ bins. From top left to bottom right: $|\eta| < 0.9$, $0.9 < |\eta| < 1.2$, $1.2 < |\eta| < 2.1$, $2.1 < |\eta| < 2.4$.

Chapter 4

Higgs associated production with a vector boson and Higgs decaying to a pair of b-quarks

In this chapter, the study of Higgs bosons produced in association with a vector boson (W/Z boson), the VH process, where the Higgs boson decays to a pair of b-quarks, using the full Run 2 data collected in CMS, is presented.

4.1 Introduction

The observation of Higgs boson decays to a pair of b-quarks is a direct test of the Yukawa coupling to down type quarks.

The decays to b-quarks has the highest branching ratio (58.1% [50]), despite this large branching ratio the successful study of the VH process depends on good heavy flavor tagging, improving the mass resolution and control over W+jets and Z+jets backgrounds. For a brief overview of the Higgs study channels, refer to Section 1.4.

The study of the VH production channel with leptonic decays of the vector boson and Higgs boson decaying to a pair of b-quarks is presented in three channels, zero-lepton ($ZH \to \nu\bar{\nu}b\bar{b}$, $\ell = e, \mu$), one-lepton ($WH \to l\nu b\bar{b}$, $\ell = e, \mu$) and two-leptons ($ZH \to llb\bar{b}$, $\ell = e, \mu$). In addition resolved and boosted topologies are studied, where resolved or fat b-jets are considered respectively (for resolved and fat jet, see Section 3.2.4). By considering boosted W or Z boson, QCD multi-jet background and $t\bar{t}$ contribution to the background processes can be reduced [20].

The first attempt to measure this channel in CMS used 1.1 fb^{-1} of data in 2011 [29, 107], later the complete 5 fb^{-1} 2011 dataset was used and the result was published in Ref. [27]. This analysis used five channels (Z($\ell\ell$)H, Z($\nu\nu$)H, W($\ell\nu$)H, with $\ell = e, \mu$ and no $e\mu$ channel) and considered a cut-based approach. The result set limits at four times the standard model expectation for this process. Later, full LHC Run 1 data (5 + 19 fb^{-1} at 7 and 8 TeV) was analyzed and the results were published [28], and a signal strength $\mu = 1.0 \pm 0.5$ relative to the SM was reported. Finally, combining 2016 and 2017 datasets [87] with those from Run 1 of CMS, and adding the other H $\to b\bar{b}$ production channels, an observed (expected) significance of 5.6σ (5.5σ) with a signal strength of $\mu = 1.04 \pm 0.20$ was reported by the CMS collaboration.

As stated in the last paragraph, the discovery of the H $\to b\bar{b}$ decay in CMS is well established. This fact together with the quantity of data collected in the CMS detector during Run 2 of data taking, justifies studying the processes with finer bins and more details. To this end, the STXS framework [16], described in Section 1.4.3 is used in this thesis to analyze the 138 fb^{-1} of data collected.

4.2 Analysis strategy

To clarify the strategy of the STXS measurement in this thesis, it is helpful to give a brief and more general view of the strategy employed. The physics process under study, here the VH($b\bar{b}$) , is referred to as the signal process and the other physics processes that have similar detector signatures and are to be suppressed or controlled are called the background processes.

The strategy is to determine the signal strength modifier μ_{signal} by simultaneously fitting the signal and background shape templates of an observable for each bin of the STXS framework. The signal region is a region of the phase space that is selected to be enriched in the VH($b\bar{b}$) signal, this is called the signal selection criteria. The control regions are defined as regions of phase space that have selections orthogonal to the signal selection criteria while ideally enriched in one single important background process.

The templates for signal and background processes are generated using the Monte Carlo generator simulations, including all the detector effects. All variations due to detector and calibration effects are included, either with detailed shape information or as multiplicative normalization factors. The strengths of each source of variation is steered by nuisance parameters in the likelihood function, see Eq. (4.4). The signal strength modifier is considered as a free parameter in the fitting procedure. Furthermore, for important background processes, normalization factors are considered as nuisance parameters in the fit and are set as free parameters so that the data points from collision reduce the impacts of modeling uncertainties in simulations.

To demonstrate this strategy, an example with two background processes and three regions of phase space is shown in Fig. 4.1, the collision data and simulations pass through a selection criterion, to create separate templates for the signal and control regions. Then these templates are simultaneously fitted to the data to extract the signal strength μ_{signal}, and the normalization factors for each background contributions. In this thesis, to improve the accuracy of the fit, a multivariate method is used to separate the signal and background in the signal region.

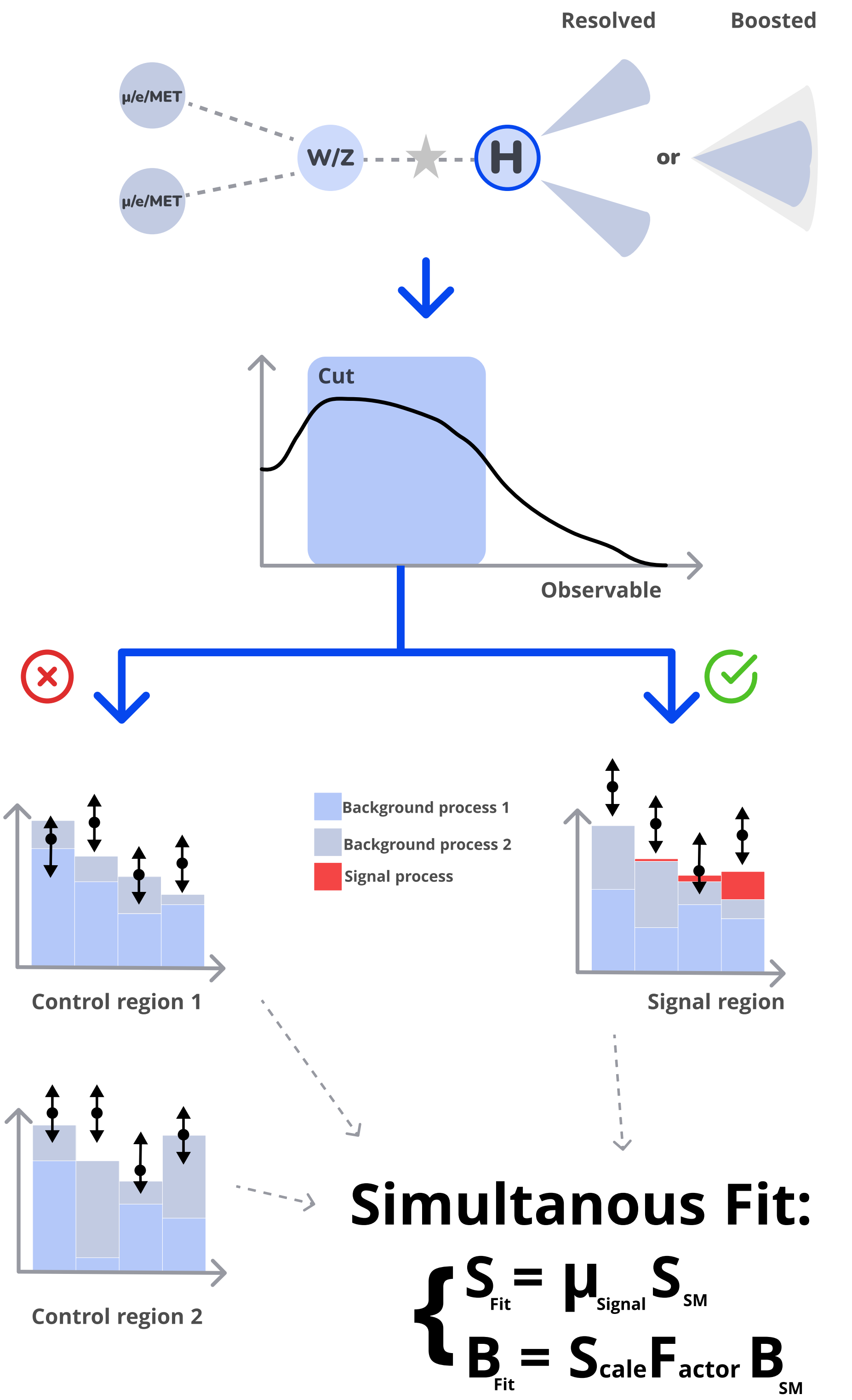


Figure 4.1: Simplified schematic of the strategy employed in this thesis. The signal and control regions are created after applying the selection criteria to the collision data and simulation. The resulting templates are fitted simultaneously to extract the signal strength and normalization factors for the background processes (scale factors).

4.3 Signal and backgrounds

In this section, the signal and background processes are introduced in more detail.

4.3.1 Signal processes

The possible Feynman diagrams for the signal processes in this thesis are shown in Fig. 4.2. The W($\ell\nu$)H (with $\ell = e, \mu$) channel is produced solely from the quarks, while for the Z($\ell\ell$)H, Z($\nu\nu$)H (with $\ell = e, \mu$) channels the contributions from the gluon induced production is also included. The cross-section of the gluon induced production modes is around 0.127 fb^{-1} which is small compared to the roughly 0.58 fb^{-1} cross-section of the quark mode. This fact makes the separation of these processes challenging, therefore in this thesis VH($b\bar{b}$) includes the gluon induced production as the signal process.

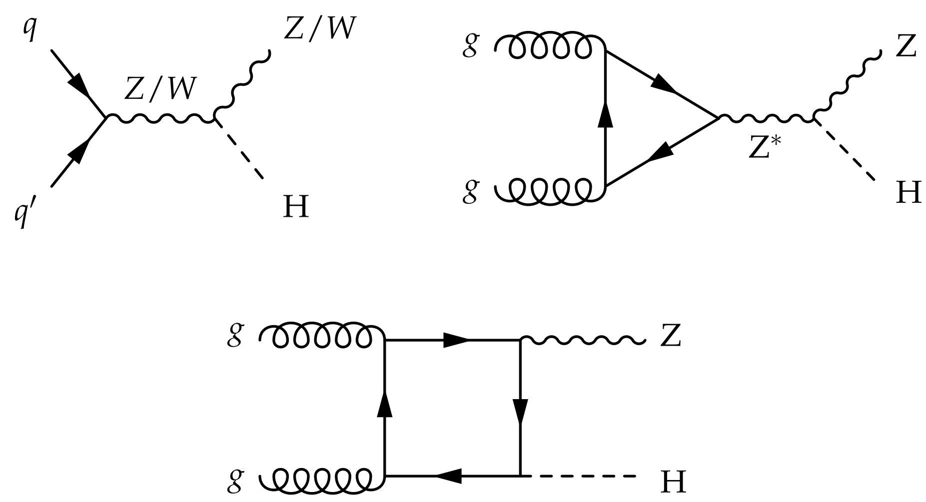


Figure 4.2: The leading order Feynman diagrams corresponding to the VH($b\bar{b}$) signal process. The gluon induced production mode contributes to the zero and two lepton channels (top right and bottom diagrams).

The Higgs boson in the VH($b\bar{b}$) signal can have different signatures in the detector, one of these that is instructive to consider is the decay of the b-quark dijet system back to back with the vector boson, roughly illustrated in Fig. 4.1 at the top. In case of the two-lepton channel, the vector boson can be reconstructed from the four momenta of the two-lepton, for the one-lepton channel from the combination of the lepton and MET and for the zero-lepton channel from the MET.

The final states with τ-leptons are not considered explicitly, but as τ decays to muons, 17.39 % of the times [63], it contributes to the leptonic channels for signal and background processes. The semi-hadronic decays of the τ will result in light flavor jets that contribute to the corresponding control regions and treated as background.

4.3.2 Background processes

In this section, the background processes are introduced and ways to control or reduce their impacts are discussed.

Vector boson + jets

The background from the production of vector bosons associated with jets are results of processes similar to the diagrams shown in Fig. 4.3, where a radiated gluon creates two b-quark jets. These b-jets can resemble the signal signature in the detector, therefore the contribution from this process is constrained by applying cuts. The differences compared to the signal signature are different dijet mass distributions due to the decay originating from a gluon instead of the Higgs resulting in lower p_T for the jets and the vector boson in the background compared to the signal and difference in the distribution of the multiplicity of additional jets present in the event.

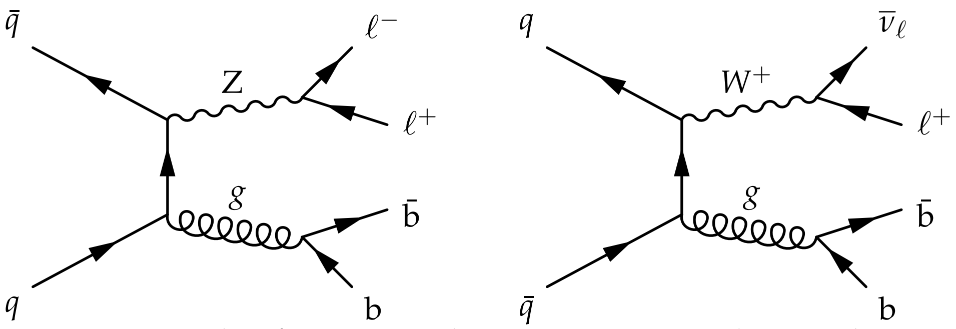


Figure 4.3: An example of Feynman diagrams corresponding to the Z + jets (left) and W + jets (right) background processes.

By applying cuts on the b-tagging working points, p_T and mass of the two b-jets, the contribution from this background is reduced, but since this process has a high cross-section, a large contribution remains in the signal regions (for two-lepton this is the largest background process). To further control the effects, the contributions are separated according to the parton flavor based on counting the number of matched generator-level B and D hadrons with $|\eta| < 2.6$: into V + light-jets, corresponding to zero D/B hadrons; V+ c-jets, corresponding to zero B but one or more D hadrons; and the V + b-jets, corresponding to one or more B hadrons. A free floating rate-parameter is assigned to each of the V + light-jets, V + c-jets and V + b-jets in the fit.

Top pair

The top pair process is a significant background due to its large cross-section, some modes of production are shown in Fig. 4.4. The top quark decays to a W boson and a b-quark, and further decays of the W boson to a lepton and neutrino creates signatures imitating the signal process.

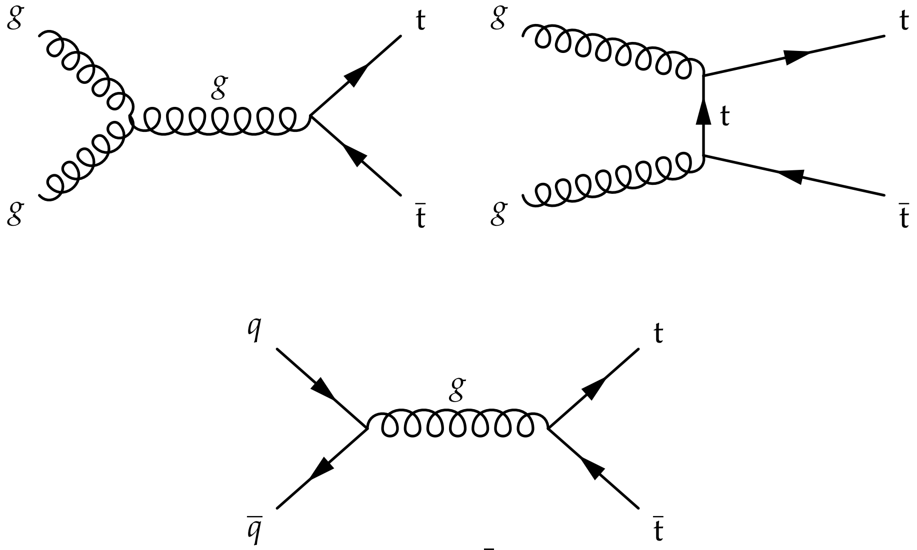


Figure 4.4: Leading order diagrams for $t\bar{t}$ production, the top quark decaying to a W boson and a b-quark, with the W decaying to a lepton and neutrino creates signatures imitating the signal process.

For a top pair background imitating the signal signature more neutrinos are expected in the decay products compared to the real signal, this results in larger MET contribution to be expected for the background compared to the signal in the events dominated by top pair background. Also the spatial distribution of the b-quark jets relative to the reconstructed vector boson is different from the signal event. Using cuts on MET, and the angle between the vector boson and the b-quark dijet system, and the multiplicity of the additional jets present in the event, this background can be suppressed.

Single Top

Single top production is similar to $t\bar{t}$ in ways of imitating the signal signature but has a smaller cross-section compared to $t\bar{t}$. Nevertheless, it has a non-negligible contribution due to the kinematical properties which fake the signal signature. Some single top channel production modes are shown in Fig. 4.5.

Diboson

The WW, WZ and ZZ processes can imitate the signal signature with their decays to two b-jets and leptons. The main handle to control the contributions from these processes is to constrain the b-quark dijet system mass. The leading order production modes are shown in Fig. 4.6.

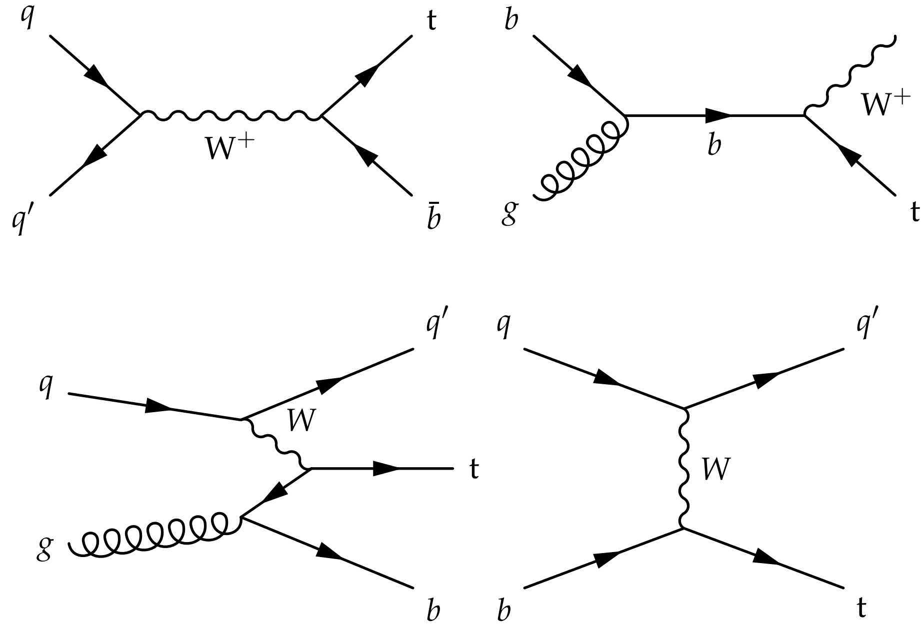


Figure 4.5: Production modes of the single top represented by Feynman diagrams.

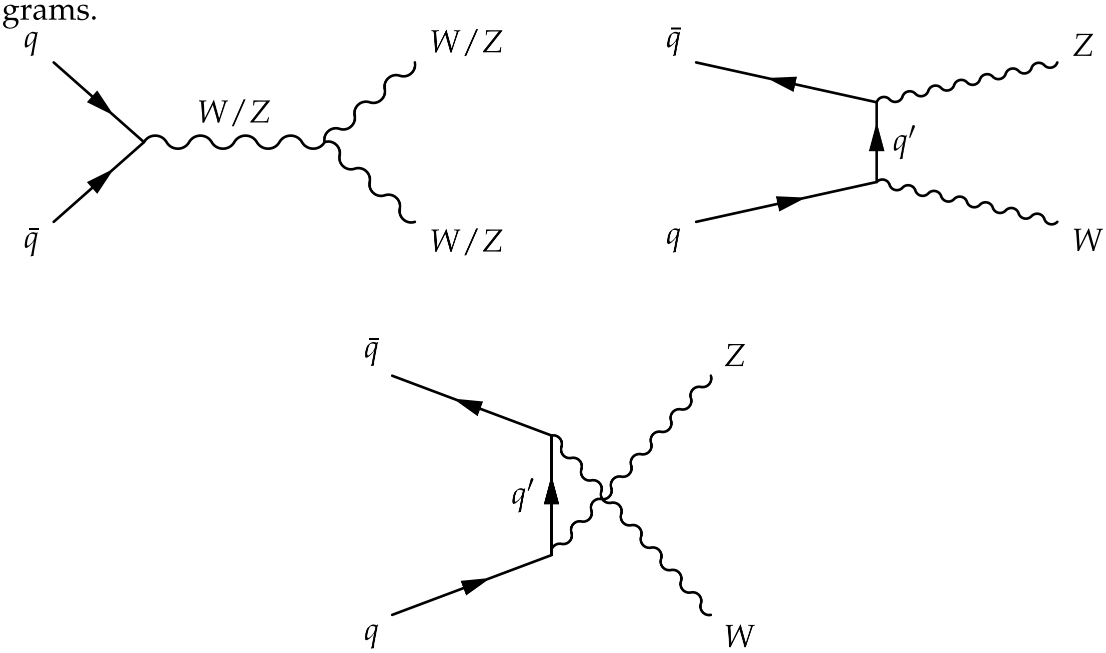


Figure 4.6: Leading order Feynman diagrams for the diboson contributions, t-channel at the top right, s-channel at top left and u-channel at the bottom.

QCD multi-jet

The QCD contributions can imitate the signal signature in multiple ways. For the zero-lepton channel, QCD events with two b-jets with MET due to detector effects (e.g. mis-measurement of a jet) imitate the signal signature. For the one and two-lepton channels, QCD events having two b-jets wrongly reconstructed as leptons and with misreconstructed MET can imitate signatures very close to the signal in these channels. This background is challenging to control with the Monte Carlo simulations, as it needs a very large MC statistics to create the phase space required for the extraction of corrections or constraints. In this thesis, the contribution from QCD is reduced with cuts on the phase space so that the contribution could be neglected in the signal region. The stringent cuts on the lepton isolation present in this thesis are among the main cuts introduced to this end; these cuts reduce the number of non-prompt leptons which are leptons originating from the decay of a hadron or misidentified leptons.

4.3.3 Simulation datasets

The event generation with Monte Carlo was briefly discussed in Section 3.1. In this section, the simulated Monte Carlo datasets (referred to as samples) used for signal and background processes with their cross-sections are listed, see Table 4.1 for signal processes and Table 4.2 for background process.

To reduce the statistical uncertainties from the generation process, events are produced with higher numbers than is expected to appear for that processes in the real collision event. To retrieve the correct number of events, a weight is assigned to each event,

$$w_{event} = \sigma \times \mathcal{L} \times \frac{w_{generator}}{\sum w_{generator}}, \tag{4.1}$$

where the $w_{generator}$ is assigned by the Monte Carlo generator for each generated event and are not always constant, in some Monte Carlo generators these weights include negative values when next-to-leading order (NLO) accuracies are included, e.g. in the case of MADGRAPH5_aMC@NLO [10].

The signal samples for the quark induced production of ZH and WH are generated by the POWHEG [9, 54, 83] v2 event generator extended with the MiNLO procedure [64, 80] at NLO accuracy. The gluon induced signal samples on the other hand have leading order accuracy and are produced with POWHEG [9, 54, 83] v2, see Table 4.1.

To produce the diboson background processes ZZ, WW and WZ the MADGRAPH5_aMC@NLO v2.3.3 at NLO using the FxFx merging scheme [52] was used. This generator was also used to generate QCD multijet and the V+jets processes at LO accuracy with the MLM matching scheme [11]. The V+jets samples are produced with two additional b-enrichment configurations, one enriches the samples with more b quarks by generating only the matrix-element of the orthogonal b-quark decays, while the other creates this enrichment at parton shower level with considering the showers that contribute to the b-quark final states. The $t\bar{t}$ and the single top sample in t-channel are generated with POWHEG v2. The single top quark samples in the tW and s-channel are produced with POWHEG v1. An overview of all background simulated MC datasets are given in Table 4.2.

The cross-sections for production of physics processes included in the signal, V+jets and the diboson samples are scaled to next-to-next-to-leading order (NNLO) QCD and NLO electroweak accuracy by using the calculations from the following generators, the VHNNLO [18, 48, 49], VH@NNLO [18, 65], and HAWK v2.0 [42]. These corrections are applied in bins of the vector boson p_{T}, a detailed description of the procedure is presented in [50]. For $t\bar{t}$, the cross-section is scaled with the results from Top++ v2.0 [39] to NNLO accuracy.

For NLO accurate samples, the NLO NNPDF3.0 set [15] parton distribution functions (PDFs) and for the LO accurate samples the LO NNPDF3.0 set is used.

The parton showering and hadronization procedures are also added to the generated events by interfacing the POWHEG and MADGRAPH5_aMC@NLO generators with showering and for hadronization the samples are interfaced with PYTHIA 8.212 [102]. The parameters for the effective models that describe the underlying events are based on the new set of PYTHIA 8 tunes, CP5 for 2017-2018 data-taking periods and the CUETP8M1 for the 2016 [100].

The detector response is simulated using the GEANT4 package [7] generator, the simulation uses a very detailed description of the CMS detector and its components.

Sample	σ (pb)	k-factor	Event Generator
$pp \to ggZH; H \to b\bar{b}, Z \to l^+l^-, M_H = 125GeV$	0.01437	1.0	POWHEG v2
$pp \to ggZH; H \to b\bar{b}, Z \to \nu\bar{\nu}, M_H = 125GeV$	0.01437	1.0	POWHEG v2
$pp \to ZH; H \to b\bar{b}, Z \to l^+l^-, M_H = 125GeV$	0.04718	1.0	POWHEG v2 + MiNLO
$pp \to ZH; H \to b\bar{b}, Z \to \nu\bar{\nu}, M_H = 125GeV$	0.09322	1.0	POWHEG v2 + MiNLO
$pp \to W^-H \to b\bar{b}, M_H = 125GeV$	0.10899	1.0	POWHEG v2 + MiNLO
$pp \to W^+H \to b\bar{b}, M_H = 125GeV$	0.17202	1.0	POWHEG v2 + MiNLO

Table 4.1: Summary of Monte Carlo datasets for signal processes (All hadronized by PYTHIA8), where k-factors are multiplicative factors calculated to correct the leading order (LO) cross-sections to next to leading order (NLO).

Merging the leading order V+jets MC datasets

The V+jets MC datasets used are produced in different bins of transverse hadronic momenta H_T, and in bins of the vector boson p_T for the b-enrichment configurations. The b-enriched MC datasets are employed to increase statistics in heavy flavor regions. To avoid double counting in the regions, similar generator-level phase-spaces are reweighted to match the expected SM cross-section, Fig. 4.7 shows how this plays out in practice by color coding each V+jets MC dataset. The H_T-binned MC datasets are generated in 8 bins (shown in Fig. 4.7 legend starting with "HT"), and the b-enriched MC datasets are generated in two $p_T(V)$ bins for each of the two configurations (shown in Fig. 4.7 legend starting with "BGen" referring to the b-enrichment using matrix-element level b-enrichment and "BJet" referring to the parton-shower level b-enrichment).

4.3.4 NLO V+jets MC datasets

The accurate modeling of the V+jets background process is important as it has large non-reducible contributions to the signal regions. To improve the modeling, NLO order MC datasets of Table 4.3 were used in two separate ways in this thesis.

Sample	σ (pb)	k-factor	Event Generator
$Z^0/\gamma^* \to l^+l^- + \mathrm{B-Jets},\ 100 < P_T(Z) < 200GeV$	3.206	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{B-Jets},\ P_T(Z) > 200GeV$	0.3304	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 100 < P_T(Z) < 200GeV$	2.662	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ P_T(Z) > 200GeV$	0.3949	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 100 < H_T < 200GeV$	160.8	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 1200 < H_T < 2500GeV$	0.1931	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 200 < H_T < 400GeV$	48.63	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ H_T > 2500GeV$	0.003513	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 400 < H_T < 600GeV$	6.982	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 600 < H_T < 800GeV$	1.756	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ 800 < H_T < 1200GeV$	0.8094	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \mathrm{Jets},\ M_{Z^0/\gamma^*} > 50GeV$	5343.0	1.23	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 1000 < H_T < 1500GeV$	1088.0	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 1500 < H_T < 2000GeV$	99.11	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ H_T > 2000GeV$	20.23	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 200 < H_T < 300GeV$	1547000.0	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 300 < H_T < 500GeV$	322600.0	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 500 < H_T < 700GeV$	29980.0	1.0	MADGRAPH5_aMC@NLO
$\mathrm{Multijet-QCD},\ 700 < H_T < 1000GeV$	6334.0	1.0	MADGRAPH5_aMC@NLO
Single Top production (s-channel)	3.74	1.0	POWHEG
Single anti-Top production (t-channel)	80.95	1.0	POWHEG
Single Top production (t-channel)	136.02	1.0	POWHEG
Single anti-Top production (tW-channel inclusive)	35.85	1.0	POWHEG
Single anti-Top production (tW-channel leptonic)	19.56	1.0	POWHEG
Single Top production (tW-channel inclusive)	35.85	1.0	POWHEG
Single Top production (tW-channel leptonic)	19.56	1.0	POWHEG
Hadronic $t\bar{t}$	377.96	1.0	POWHEG
$t\bar{t} \to l\nu$	88.29	1.0	POWHEG
Semi-leptonic $t\bar{t}$	365.34	1.0	POWHEG
$W \to l\nu + \mathrm{B-Jets},\ 100 < P_T(W) < 200GeV$	5.527	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{B-Jets},\ P_T(W) > 200GeV$	0.7996	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 100 < H_T < 200GeV$	1392.0	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 1200 < H_T < 2500GeV$	1.084	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 200 < H_T < 400GeV$	410.3	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ H_T > 2500GeV$	0.008067	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 400 < H_T < 600GeV$	57.85	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 600 < H_T < 800GeV$	12.95	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 800 < H_T < 1200GeV$	5.45	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 100 < P_T(W) < 200GeV$	20.49	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ P_T(W) > 200GeV$	2.935	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + \mathrm{Jets},\ 70 < H_T < 100GeV$	1353.0	1.21	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{B-Jets},\ 100 < P_T(Z) < 200GeV$	6.195	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{B-Jets},\ P_T(Z) > 200GeV$	0.6293	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 100 < P_T(Z) < 200GeV$	1.679	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ P_T(Z) > 200GeV$	0.2468	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 100 < H_T < 200GeV$	303.4	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 1200 < H_T < 2500GeV$	0.3425	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 200 < H_T < 400GeV$	91.71	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ H_T > 2500GeV$	0.005263	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 400 < H_T < 600GeV$	13.1	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 600 < H_T < 800GeV$	3.248	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + \mathrm{Jets},\ 800 < H_T < 1200GeV$	1.496	1.23	MADGRAPH5_aMC@NLO
WW	117.6	1.0	MADGRAPH5_aMC@NLO
WZ	48.1	1.0	MADGRAPH5_aMC@NLO
ZZ	17.2	1.0	MADGRAPH5_aMC@NLO

Table 4.2: Summary of Monte Carlo Samples for background processes (All hadronized by PYTHIA8), where k-factors are calculated multiplicative factors to correct the leading order (LO) cross-sections to next to leading order (NLO).

Sample	σ (pb)	k-factor	Event Generator
$Z^0/\gamma^* \to l^+l^- + 1 - \text{Jet},\ 50 < P_T(Z) < 150GeV$	316.6	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 1 - \text{Jet},\ 150 < P_T(Z) < 250GeV$	9.543	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 1 - \text{Jet},\ 250 < P_T(Z) < 400GeV$	1.098	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 1 - \text{Jet},\ P_T(Z) > 400GeV$	0.1193	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 2 - \text{Jets},\ 50 < P_T(Z) < 150GeV$	169.6	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 2 - \text{Jets},\ 150 < P_T(Z) < 250GeV$	15.65	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 2 - \text{Jets},\ 250 < P_T(Z) < 400GeV$	2.737	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 2 - \text{Jets},\ P_T(Z) > 400GeV$	0.4477	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 0 - \text{Jets, inclusive}$	5333.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 1 - \text{Jets, inclusive}$	965.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + 2 - \text{Jets, inclusive}$	362.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \text{Jets},\ 50 < P_T(Z) < 100GeV$	409.8	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \text{Jets},\ 100 < P_T(Z) < 250GeV$	97.26	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \text{Jets},\ 250 < P_T(Z) < 400GeV$	3.764	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \text{Jets},\ 400 < P_T(Z) < 650GeV$	0.5152	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma^* \to l^+l^- + \text{Jets},\ P_T(Z) > 650GeV$	0.0483	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 0 - \text{Jets, inclusive}$	54500.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 1 - \text{Jets, inclusive}$	8750.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 2 - \text{Jets, inclusive}$	3010.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + \text{Jets},\ 50 < P_T(W) < 100GeV$	3570.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + \text{Jets},\ 100 < P_T(W) < 250GeV$	770.8	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + \text{Jets},\ 250 < P_T(W) < 400GeV$	28.06	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + \text{Jets},\ 400 < P_T(W) < 650GeV$	3.591	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + \text{Jets},\ P_T(W) > 650GeV$	0.5495	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1 - \text{Jet},\ 50 < P_T(Z) < 150GeV$	596.3	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1 - \text{Jet},\ 150 < P_T(Z) < 250GeV$	17.98	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1 - \text{Jet},\ 250 < P_T(Z) < 400GeV$	2.045	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1 - \text{Jet},\ P_T(Z) > 400GeV$	0.2243	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2 - \text{Jets},\ 50 < P_T(Z) < 150GeV$	325.7	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2 - \text{Jets},\ 150 < P_T(Z) < 250GeV$	29.76	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2 - \text{Jets},\ 250 < P_T(Z) < 400GeV$	5.166	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2 - \text{Jets},\ P_T(Z) > 400GeV$	0.8457	1.0	MADGRAPH5_aMC@NLO

Table 4.3: Summary of the NLO V+jets Monte Carlo Samples.

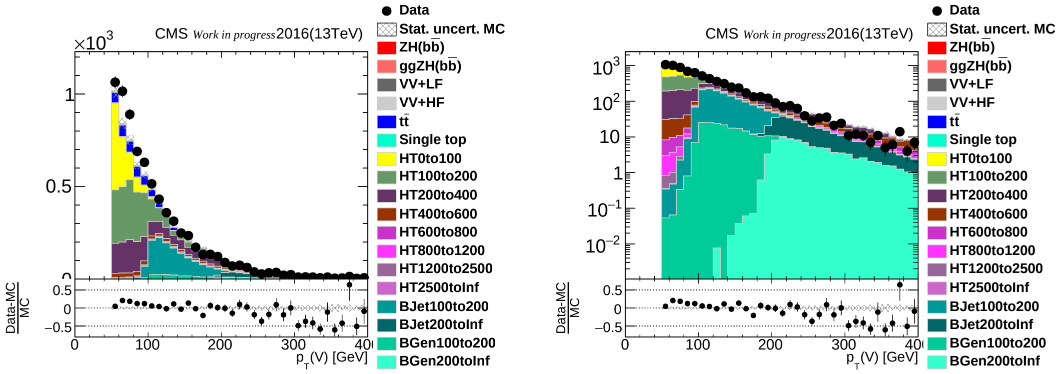


Figure 4.7: The linear (left) and logarithmic (right) histograms of the $p_{\mathrm{T}}(\mathrm{V})$ in 2-lepton heavy flavour control region for the 2016 data taking period, with V+jets samples shown with separate colors to show the stitching between different samples. The H_T-binned MC datasets are generated in 8 bins (shown in Fig. 4.7 legend starting with "HT"), and the b-enriched MC datasets are generated in two $p_{\mathrm{T}}(\mathrm{V})$ bins for each of the two configurations (shown in Fig. 4.7 legend starting with "BGen" or "BJet").

For the 2016 data taking era, the leading order MC datasets are reweighed to the NLO accuracy in bins of $p_{\mathrm{T}}(\mathrm{V})$ and number of b-hadrons (nB) to improve the data and MC agreement. The corrections are found by linear fits to the ratio (NLO/LO) of the $p_{\mathrm{T}}(\mathrm{V})$ for all V+jets samples and polynomial fits to the ratio of $\Delta\eta_{bb}$ (of the two leading b-jets) for the $Z \to \nu\bar{\nu} + \text{jets}$ samples. The values are presented in Tables 4.4 and 4.5. This choice is made to make use of the higher statistics of the leading order MC datasets compared to the available MC statistics for the NLO MC datasets for the 2016 data taking era.

channel	nB	NLO/LO
$Z \to \nu\bar{\nu} + \text{Jets}$	0	1.688±0.002 - (1.785±0.007)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$Z \to \nu\bar{\nu} + \text{Jets}$	1	1.575±0.007 - (1.754±0.027)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$Z \to \nu\bar{\nu} + \text{Jets}$	2	1.424±0.010 - (1.539±0.041)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$W \to l\nu + \text{Jets}$	0	1.628±0.005 - (1.339±0.020)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$W \to l\nu + \text{Jets}$	1	1.586±0.027 - (1.531±0.112)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$W \to l\nu + \text{Jets}$	2	1.440±0.048 - (0.925±0.203)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$Z \to \nu\bar{\nu} + \text{Jets}$	0	1.650±0.002 - (1.707±0.020)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$Z \to \nu\bar{\nu} + \text{Jets}$	1	1.534±0.010 - (1.485±0.080)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$
$Z \to \nu\bar{\nu} + \text{Jets}$	2	1.519±0.019 - (1.916±0.140)·10^{-3} $p_{\mathrm{T}}(\mathrm{V})$

Table 4.4: NLO/LO weight as function of flavor and $p_{\mathrm{T}}(\mathrm{V})$ for the three types of V+jets samples.

For 2017 and 2018 data taking eras, the NLO samples were used directly. There are regions in the NLO samples that are not well modeled, namely the region with $\Delta R(jj) < 1$ (of the two leading selected jets in p_{T}), for which a reweighting is calculated per lepton channel and $p_{\mathrm{T}}(\mathrm{V})$ bins in the light flavor control region and extrapolated to the signal and heavy flavor control regions

nB	p_T(V)	NLO/LO
0	150-170	$(8.902e-01)+(1.171e-01)x+(-4.028e-02)x^2+(5.413e-03)x^3$
0	170-200	$(8.902e-01)+(1.171e-01)x+(-4.028e-02)x^2+(5.413e-03)x^3$
0	200-250	$(9.055e-01)+(1.016e-01)x+(-4.262e-02)x^2+(6.732e-03)x^3$
0	250-300	$(9.258e-01)+(6.067e-02)x+(-2.653e-02)x^2+(4.067e-03)x^3$
0	300-400	$(9.446e-01)+(8.139e-02)x+(-4.740e-02)x^2+(8.163e-03)x^3$
1	150-170	$(9.509e-01)+(4.284e-02)x+(-2.355e-02)x^2+(6.062e-03)x^3$
1	170-200	$(9.476e-01)+(6.023e-02)x+(-4.144e-02)x^2+(9.121e-03)x^3$
1	200-250	$(9.424e-01)+(8.237e-02)x+(-4.904e-02)x^2+(9.142e-03)x^3$
1	250-300	$(9.838e-01)+(-5.520e-03)x+(-1.425e-02)x^2+(4.843e-03)x^3$
1	300-400	$(1.012e+00)+(5.707e-03)x+(-3.566e-02)x^2+(1.008e-02)x^3$
2	150-170	$(8.635e-01)+(4.481e-02)x+(8.009e-02)x^2+(-1.883e-02)x^3$
2	170-200	$(8.942e-01)+(6.551e-02)x+(2.721e-02)x^2+(-7.706e-03)x^3$
2	200-250	$(8.709e-01)+(8.808e-02)x+(1.561e-02)x^2+(-6.648e-03)x^3$
2	250-300	$(9.174e-01)+(2.229e-02)x+(4.194e-02)x^2+(-1.010e-02)x^3$
2	300-400	$(9.712e-01)+(-1.882e-02)x+(5.351e-02)x^2+(-1.159e-02)x^3$

Table 4.5: NLO/LO weight as function of x=$\Delta\eta_{bb}$ of the two hardest generator jets, for $Z \to \nu\bar{\nu}$ + Jets sample.

(see Section 4.5 for region definitions). The other region affected by modelling issues is the region where the DeepCSV score is lower than the loose working point value while $\Delta R(jj) < 1$, this is resolved by a two-dimensional reweighting where the DeepCSV score for the leading and sub-leading jets are the dimensions. This correction is derived in the $\Delta R(jj) > 1$ of the light flavor control region and extrapolated to the $\Delta R(jj) < 1$ light flavor control region (other regions are not affected due to the b-tagging requirement). Dedicated systematic uncertainties are added for each correction, the impact on the signal strength is observed to be negligible.

4.3.5 Trigger

The CMS trigger system was introduced in Section 2.2.6. In this thesis, multiple HLT triggers are used to record data which corresponds to the signature of the signal processes under study.

For the zero-lepton channel, the same thresholds for the MET and the MHT ($\not{H}_T$) were applied during the reconstruction at the HLT level to trigger the data acquisition. These thresholds are 110 GeV in 2016 and 120 GeV in the 2017 and 2018 data-taking periods.

For one-lepton channel, a p_T threshold of 24 GeV for 2016 and 2018 and a threshold of 27 GeV for the 2017 data-taking period has been applied. Similarly for electrons a threshold of 27 GeV for 2016 and of 32 GeV for 2017 and 2018 on p_T was used.

For the two-lepton channel, the muons have p_T thresholds of 17 and of 8 GeV and the electrons have 23 and 12 GeV thresholds.

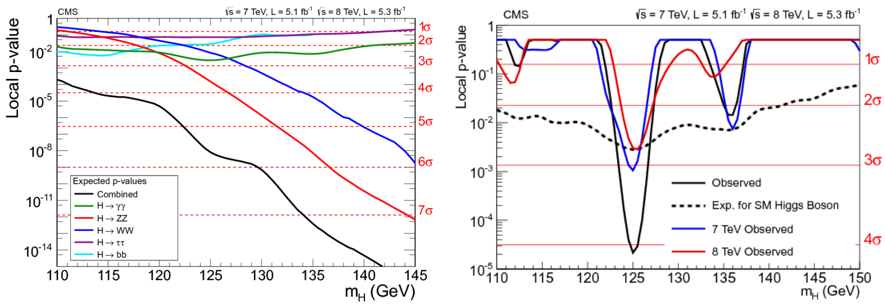


Figure 4.8: Some local p-values and the corresponding significance for the Higgs discovery in 2012, expected on the left and observed on the right. [26]

4.4 Statistical procedure

4.4.1 Statistics in HEP

The general approach towards a search for new phenomena using the frequentist statistical test, starts with defining the null hypothesis, H_0, representing only known processes (called backgrounds in the context of particle physics).

This hypothesis is going to be tested against the alternative hypothesis H_1 (containing both backgrounds and the new phenomenon we call signal in this context). Normally, one uses the p-value to quantify the compatibility of observed data with a given hypothesis H. It can be translated to number of standard deviations (significance Z) using the one-sided Gaussian (denoted N) tail convention (Eq 4.2).

$$p - value = \int_Z^\infty N(x;0,1)\,dx\,. \tag{4.2}$$

Figure 4.8 demonstrates some local p-values and the corresponding significance for the Higgs boson discovery [26].

In this thesis, we use likelihood ratios as test statistics (As the Neyman-Pearson lemma suggests that this gives the highest power test of H_1 agains H_0).

The parameter of interest is the signal strength (the ratio of the production cross-section of the physics phenomenon under study over the one for the standard model expectation). Here $\vec{\theta}$ denotes nuisance parameters,

$$\lambda(\mu) = \frac{L(\mu,\hat{\hat{\vec{\theta}}})}{L(\hat{\mu},\hat{\vec{\theta}})}\,. \tag{4.3}$$

$\hat{\hat{\vec{\theta}}}$ in the numerator corresponds to the value of $\vec{\theta}$ that maximizes L for a fixed μ (MLE for fixed μ). The denominator corresponds to MLE considering the nuisance parameters changing μ's profile.

The likelihood, given a set of measurements n_j with s_j signal, b_j backgrounds and Gaussian distributed nuisances $\vec{\theta}$ is as follows:

$$L(\mu,\vec{\theta}) = \prod_j \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_k e^{-\frac{1}{2}\theta_k^2}\,. \tag{4.4}$$

For the discovery of a positive signal, we need to reject the background-only hypothesis where $\mu = 0$. The test statistics can be expressed as:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0\,, \\ 0 & \hat{\mu} < 0\,, \end{cases} \tag{4.5}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in Eq. (4.3). Then the p-value is as follows

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0\,, \tag{4.6}$$

where f denotes the probability distribution function for the observed, and using 4.2 we can convert it to significance.

4.4.2 Impact of statistical fluctuations in background models

Here, a simple test is presented, which shows the difficulties that may arise when one deals with systematic variations and bin-by-bin uncertainties at the same time in a maximum likelihood fit similar to the one used in this thesis.

To take into account the effects of the MC simulation statistics per bin, an extra factor is added to the likelihood function to account for the uncertainties arising from the limited statistics from the simulations (in the fit model the Barlow-Beeston-lite [37] model has been incorporated). Furthermore, the shape systematic uncertainties arise as variations on the bin counts per each bin, these effects are added by assuming a morphing function as follows in the main fit model used in this thesis:

$$h(\theta) = \frac{1}{2}\left((\lambda_+ - \lambda_-)\theta + \frac{1}{8}(\lambda_+ + \lambda_-)(3\theta^6 - 10\theta^4 + 15\theta^2)\right), \tag{4.7}$$

where the shapes (fraction of events h in each bin) are interpolated using a spline for $\theta\epsilon[-1,1]$ and are interpolated linearly for $\theta\notin[-1,1]$ and the variations of the expected number of events due to some systematic error are represented as λ_+ or λ_-. To see the effect of the fluctuations here, we only assume the linear interpolations. Adding v_is as free parameters to control the bin-by-bin statistical uncertainties for a MC simulation with uncertainty σ the morphing is as follows:

$$h(\theta, \lambda_0, \lambda_+, \lambda_-) = \begin{cases} \lambda_0 + \theta * (\lambda_+ - \lambda_0) - v_i\sigma & \theta \geq 0 \\ \lambda_0 - \theta * (\lambda_0 - \lambda_-) - v_i\sigma & \theta \leq 0 \end{cases}$$

where λ_0 is the expected number of events in a bin.

Considering the above function, the likelihood can be written as:

$$L(\mu, \vec{\theta}, \lambda_0, \lambda_+, \lambda_-, \vec{v}) = \prod_j \frac{(\mu\, h)^{n_j}}{n_j!} e^{-(\mu\, h)} \prod e^{-\frac{1}{2}\vec{\theta}} \prod_j e^{-\frac{1}{2}v_j^2}\,. \tag{4.8}$$

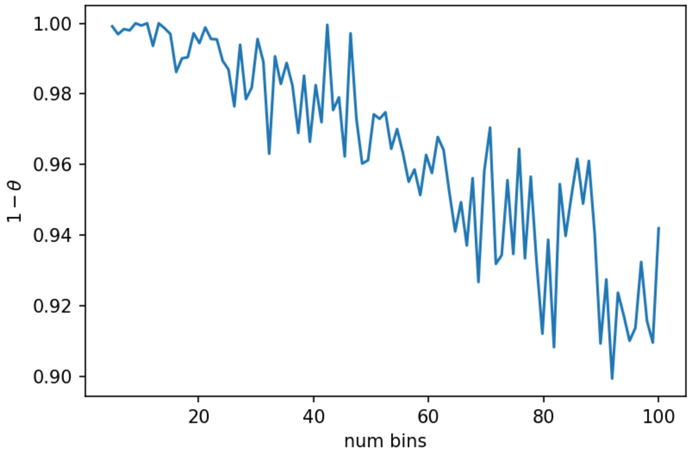


Figure 4.9: The post fit constraint on the nuisance parameter θ is dependent on the number of bins, as the number of bins increases the nuisance parameter is more constrained.

Considering Eq. (4.8), the maximum likelihood minimization will receive penalties from the $\theta^2/2$ and $v^2/2$ terms in the $-2\ln L(\mu, \vec{\theta}, \lambda_0, \lambda_+, \lambda_-, \vec{v})$. This dependence means that care should be taken in cases with large MC statistical errors in each bin. These cases arise when a very fine binning is selected in regions of phase-space where there are not enough statistics from the MC. This may result in the post fit constrains on the nuisance parameter θ that are not physical. This is demonstrated in Fig. 4.9, where the post fit constraint on theta is dependent on the number of bins. The study is performed by generating pseudo data with a Bradford distribution [77]. The distributions for 20 and 80 bins points of the Fig. 4.9 are shown in Fig. 4.10. To avoid similar behavior for the post fit constraints on the nuisance parameters in this thesis, the noise in templates have been smoothed out before the fit is performed.

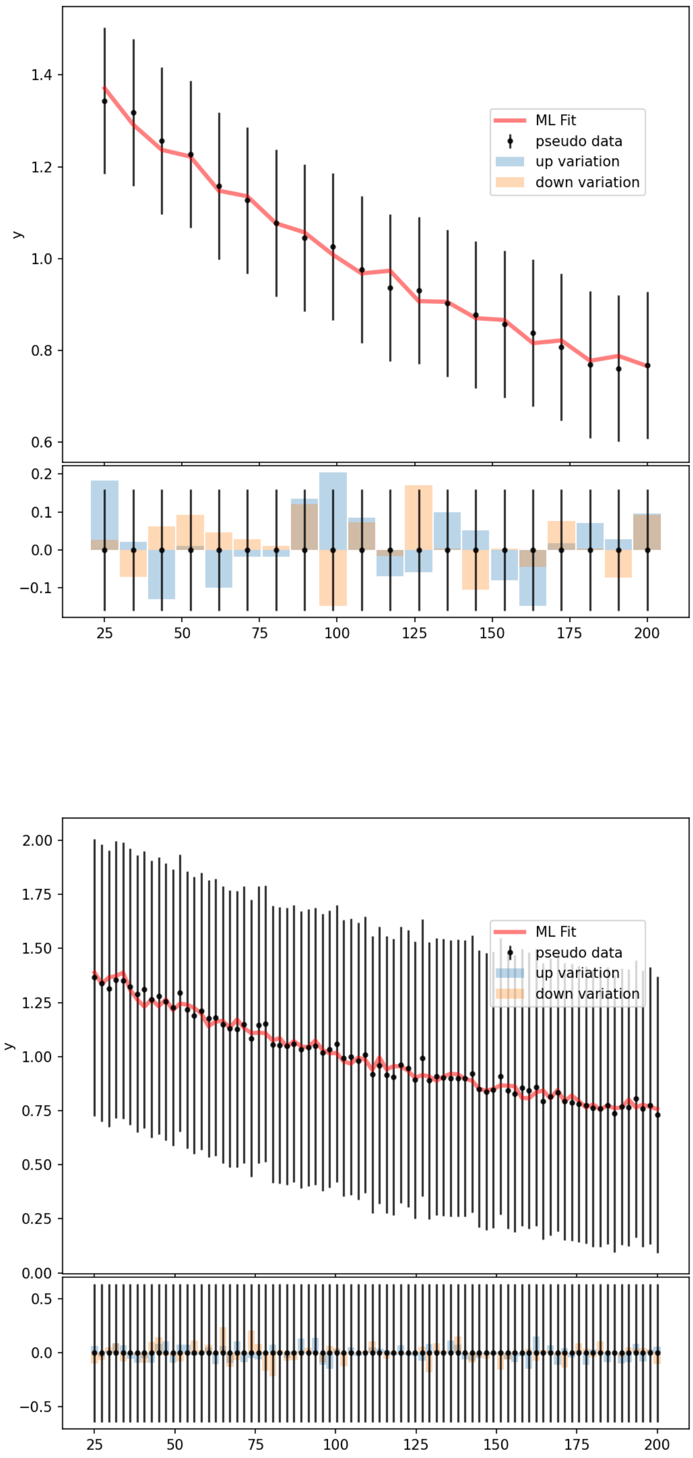


Figure 4.10: The post fit distributions and the maximum likelihood fit (in red) are shown for two number of bins, 20 (top) and 80 (bottom), the post fit constraint on the nuisance parameter is shown in Fig. 4.9.

4.5 Event selection

Following the strategy described in Section 4.2, the signal and control regions are defined as follows, the signal region (SR) enriched in VH($b\bar{b}$), the $t\bar{t}$ control region where $t\bar{t}$ has the highest contribution, the V+HF for the vector boson associated with heavy-flavour jets and the V+LF for the vector boson associated with light-flavour jets. These regions are partitioned according to the STXS framework. Furthermore, the analysis exploits the fact that events with boosted vector boson (high momentum transfer) have fewer QCD contributions in the signal region. Therefore, dedicated regions for the boosted topologies are considered with $p_{\mathrm{T}}(\mathrm{V}) > 250$ GeV.

The definition of the boosted region, which has a single fat b-jet instead of the expected two resolved b-jets, requires a decision for the events that are present in both topologies. In this thesis, the overlap events are considered in the resolved topology, as it improves the overall sensitivity of the analysis.

4.5.1 Simplified template cross-section bins

Due to the lack of sensitivity of the official STXS framework in some bins and the fact that some of these bins are out of the acceptance region of the selection criteria, some bins are merged. Also, the quark and gluon induced productions are combined for the same reason. The merging scheme is summarized in Fig. 4.11.

The final bins considered are:

- WH process:
  - $150 < p_{\mathrm{T}}(\mathrm{V}) < 250$ GeV (one-lepton channel)
  - $250 < p_{\mathrm{T}}(\mathrm{V}) < 400$ GeV (one-lepton channel, both resolved and boosted topologies contribute)
  - $p_{\mathrm{T}}(\mathrm{V}) \geq 400$ GeV (one-lepton channel, boosted topology contribution is dominant)
- Quark induced and gluon induced processes (qqZH and ggZH)
  - $75 < p_{\mathrm{T}}(\mathrm{V}) < 150$ GeV (two-lepton channel)
  - $150 < p_{\mathrm{T}}(\mathrm{V}) < 250$ GeV with zero additional jets (zero/two-lepton channel)
  - $150 < p_{\mathrm{T}}(\mathrm{V}) < 250$ GeV with one or more additional jets (zero/two-lepton channel)
  - $250 < p_{\mathrm{T}}(\mathrm{V}) < 400$ GeV (zero/two-lepton channel, both resolved and boosted topologies contribute)
  - $p_{\mathrm{T}}(\mathrm{V}) \geq 400$ GeV (zero/two-lepton channel, boosted topology contribution is dominant)

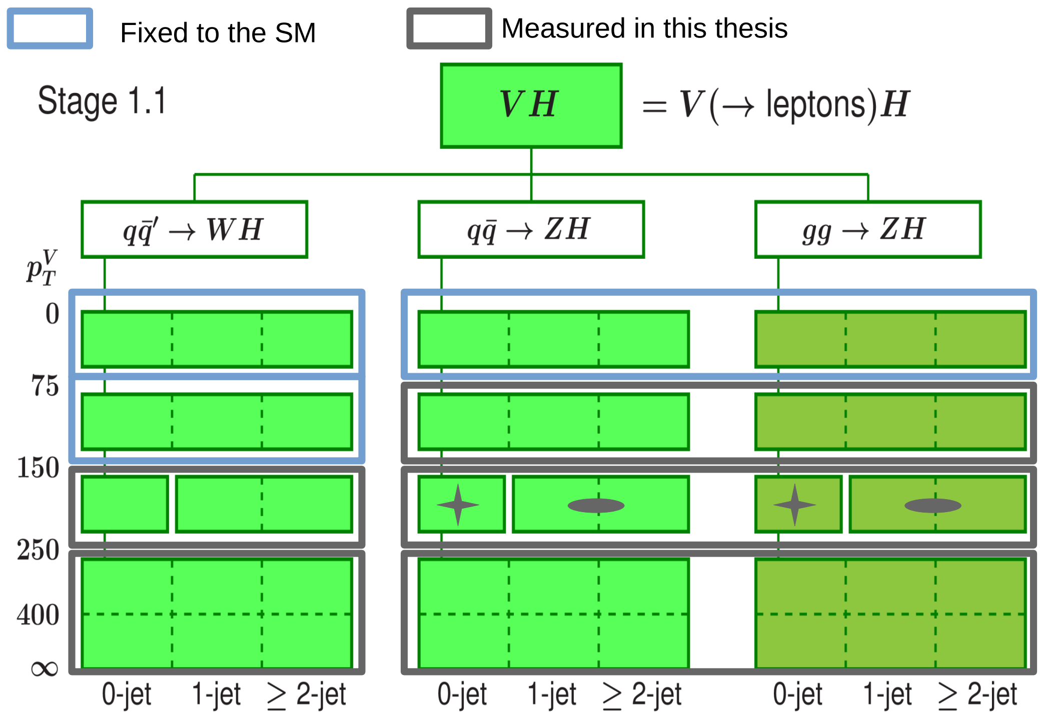


Figure 4.11: The bins annotated with blue boxes are not accessible, the bins in dark gray are the ones measured, the star and oval show the merging of jet multiplicity bins between qqZH and ggZH processes, annotations added to the plot taken from [41].

4.5.2 Pre-selection

All events considered in this thesis are required to pass a set of preliminary selections before being split to SR and CRs, these selections are listed in Table 4.6. This selection criteria ensure that the required objects for splitting events into the main categories are present.

The cuts on the leading and sub-leading jet p_T and the requirement on the MET are significantly different for the zero-lepton channel in comparison to the two other channels and has a high impact.

Anti QCD

In order to effectively reject the QCD background in final states with no muons or electrons and with large missing transverse momenta contribution, events with jets azimuthally within 0.5 radians of the MET and fulfilling the following criteria are rejected:

- p_T> 30 GeV
- tight jet ID (selecting jets with high purity)
- pile-up rejection for jets

	zero-lepton channel	one-lepton channel	two-lepton channel
$p_{T,V}$	-	> 150 GeV	> 75 GeV
MET	> 170 GeV	-	-
min(MET, $\not{H}_T$)	> 100 GeV	-	-
$\max(p_{T,j_1}, p_{T,j_2})$	> 60 GeV	> 25 GeV	> 20 GeV
$\min(p_{T,j_1}, p_{T,j_2})$	> 35 GeV	> 25 GeV	> 20 GeV
$p_{T,\mu}$	-	> 25 GeV	> 20 GeV
$p_{T,e}$	-	> 30 GeV	> 20 GeV
Isolation$_{rel,\mu}$	-	< 0.06	< 0.25
Isolation$_{rel,e}$	-	< 0.06	< 0.15
$\lvert\eta_\mu\rvert$	-	< 2.4	< 2.4
$\lvert\eta_e\rvert$	-	< 2.5	< 2.5

Table 4.6: Pre-selection cuts, these cuts are applied before the definition of the SR and CRs.

- $\Delta\phi$(jet, MET) < 0.5

These cuts are denoted as Anti QCD cuts in this thesis.

4.5.3 Resolved topology selection criteria for SR and CR

For all channels, in the signal region, the highest-p_T (leading) b-jet is required to pass the medium working point for the DeepCSV b-tagger and the next highest-p_T (sub-leading) b-jet is required to pass the loose working point selection (for the definition of working points see Section 3.2.4). All channels also pass the selection criteria on the mass of the two b-quark jets, namely to be centered on the expected Higgs mass and between 90 GeV and 150 GeV.

Signal and control regions selection for the zero-lepton

The zero-lepton channel signal signatures are a pair of b-quark jets produced from the decay of the H boson and a large contribution of MET which is distributed back to back with the four momenta of the dijet system coming from the decay of the Z boson to neutrinos. For the signal region, the selection requires the events to not have any additional high-p_T prompt lepton. To reject the QCD multi-jet background, the Anti-QCD cut described in Section 4.5.2, is applied. To be consistent with the requirements in the triggering phase, the same threshold of 100 GeV is applied on the MET and the MHT. The full list of the signal region cuts are summarized in Table 4.7.

The cuts for the orthogonal control-regions $t\bar{t}$, ZH+HF and ZH+LF are summarized in Tables 4.8 to 4.10 respectively. Examples of distributions from 0-lepton control-regions are shown in Fig. 4.12.

Signal and control regions selection for the one-lepton channel

A characteristic feature of the one-lepton channel signal signature is the recoil of the W boson which decays to a lepton and MET against the pair of b-quarks

	zero-lepton channel	one-lepton channel	two-lepton channel
b-tag max	$>$ medium	$>$ medium	$>$ medium
b-tag min	$>$ loose	$>$ loose	$>$ loose
m_{jj}	90 GeV $< m_{jj} <$ 150 GeV	90 GeV $< m_{jj} <$ 150 GeV	90 GeV $< m_{jj} <$ 150 GeV
$p_{T,jj}$	$>$ 120 GeV	$>$ 100 GeV	-
$p_{T,V}$	$>$ 170 GeV	$>$ 150 GeV	$>$ 75 GeV
m_{ll}	-	-	75 GeV $< m_{ll} <$ 105 GeV
$n_{add.\,\mathrm{lep}}$	$= 0$	$= 0$	-
$n_{add.\,\mathrm{jet}}$	-	≤ 1	-
$\Delta\phi(V, H)$	> 2.0	> 2.5	-
$\Delta\phi(\mathrm{MET}, Tk\mathrm{MET})$	< 0.5	-	-
$\Delta\phi(\mathrm{MET}, lep)$	-	< 2.0	-
Anti-QCD	True	-	-
$\min(\mathrm{MET}, \not{H}_T)$	$>$ 100 GeV	-	-

Table 4.7: Signal-region selection cuts for the resolved topology.

coming from the Higgs boson. No additional lepton and not more than one additional jets are required for the signal region. The cuts for the signal region are summarized in Table 4.7.

The cuts for the orthogonal control regions are found in Tables 4.8 to 4.10. Examples of distributions from 1-lepton control-regions are shown in Fig. 4.13.

	zero-lepton channel	one-lepton channel	two-lepton channel
b-tag max	$>$ medium	$>$ tight	$>$ tight
b-tag min	$>$ loose	-	$>$ loose
m_{jj}	50 GeV $< m_{jj} <$ 500 GeV	50 GeV $< m_{jj} <$ 250 GeV	$m_{jj} >$ 50 GeV
$p_{T,jj}$	$>$ 120 GeV	$>$ 100 GeV	-
m_{ll}	-	-	$m_{ll} \notin [0, 10], \notin [74, 120]$ GeV
$n_{add.\,jet}$	≥ 2	≥ 2	-
$\Delta\phi(V, H)$	> 2.0	-	-
$\min\Delta\phi(\mathrm{MET}, jet)$	< 1.57	-	-
Anti-QCD	True	-	-

Table 4.8: Selection cuts of the $t\bar{t}$ control-region.

Signal and control regions selection for the two-lepton channel

In the two-lepton channel the Z bosons decay to two electrons or muons and the Z boson decay system recoils against a pair of b-quark jets associated with the Higgs candidate. Since the Z boson decays to two leptons, the signature is expected to have low MET contribution. A kinematic fit is used, exploiting the kinematic information from the leptons and the Z boson to reconstruct the MET back to b-jets and improve the resolution of the Higgs mass, as the pair of electrons or muons have a good resolution and a clean signal (more details in Section 4.5.6). The mass of the dilepton system is required to be close to the one expected from a Z boson for the signal region. The full list of cuts for the signal region are found in Table 4.7.

The selection for the orthogonal control regions are found in Tables 4.8 to 4.10. Examples of distributions from 2-lepton control-regions are shown in Fig. 4.14.

	zero-lepton channel	one-lepton channel	two-lepton channel
b-tag max	$<$ medium	$<$ medium & $>$ loose	$<$ loose
b-tag min	$>$ loose	$>$ loose	$<$ loose
m_{jj}	50 GeV $< m_{jj} <$ 500 GeV	50 GeV $< m_{jj} <$ 250 GeV	$m_{jj} \in [90, 150]$ GeV
$p_{T,jj}$	$>$ 120 GeV	$>$ 100 GeV	-
m_{ll}	-	-	$m_{ll} \in [75, 105]$ GeV
$n_{add.\,jet}$	< 2	-	-
$\Delta\phi(V, H)$	> 2.0	-	> 2.5
$\Delta\phi$(MET, TkMET)	< 0.5	-	-
Anti-QCD	True	-	-

Table 4.9: V+LF (light-flavor) control-region selection cuts

	zero-lepton channel	one-lepton channel	two-lepton channel
b-tag max	$>$ medium	$>$ medium	$>$ medium
b-tag min	$>$ loose	$>$ loose	$>$ loose
m_{jj}	$m_{jj} \in [50, 90]$ GeV, or $\in [150, 500]$ GeV	$m_{jj} \in [50, 90]$ GeV, or $\in [150, 250]$ GeV	$m_{jj} \in [50, 90]$ GeV, or $\in [150, \infty]$ GeV
$p_{T,jj}$	$>$ 120 GeV	$>$ 100 GeV	-
m_{ll}	-	-	$m_{ll} \in [85, 97]$ GeV
$n_{add.\,jet}$	$= 0$	< 2	-
$\Delta\phi(V, H)$	> 2.0	-	> 2.5
$\Delta\phi$(MET, TkMET)	< 0.5	-	-
Anti-QCD	True	-	-

Table 4.10: V+HF (heavy-flavor) control-region selection cuts

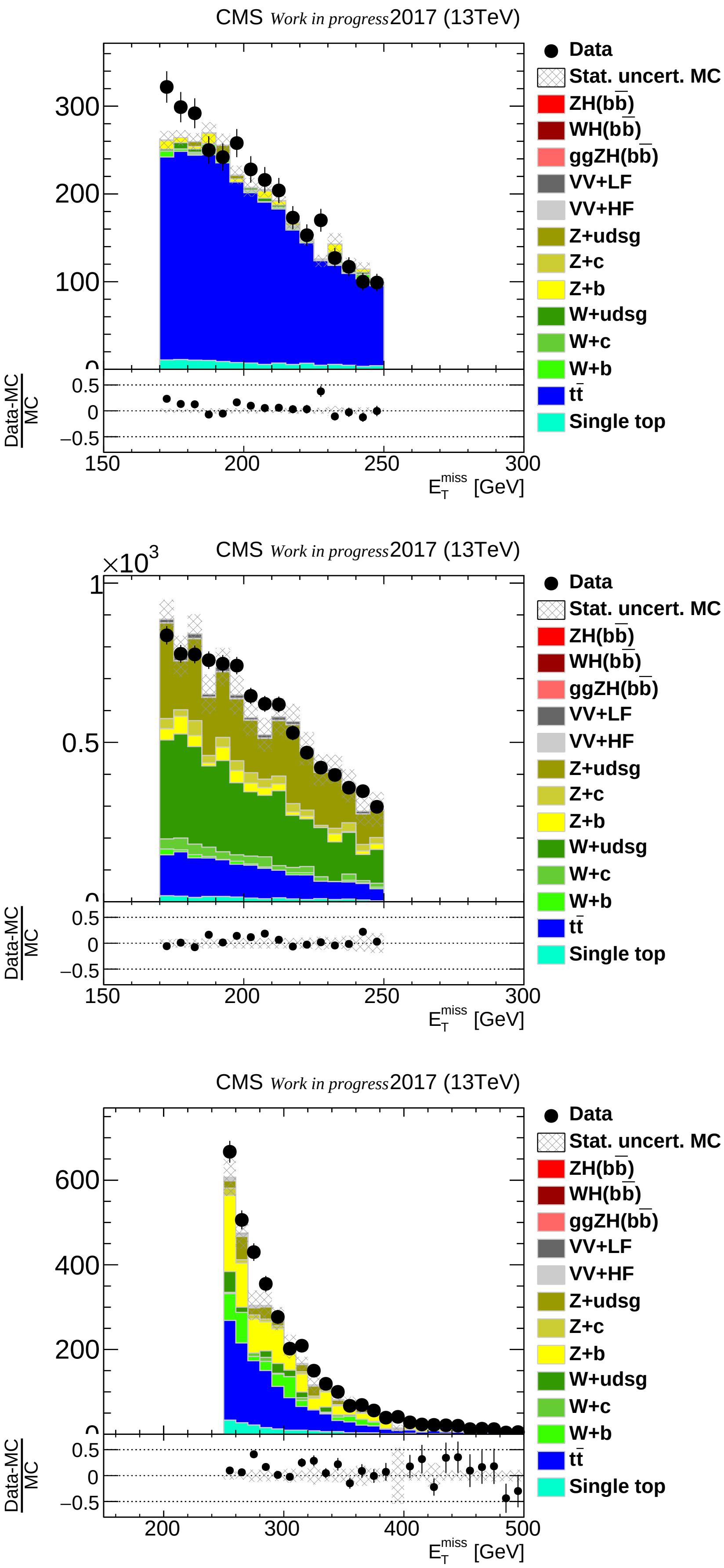


Figure 4.12: Distributions of the MET for zero-lepton channel in $t\bar{t}$ for STXS $150 < p_{\mathrm{T}}(\mathrm{V}) < 250$ GeV region, Z+LF for STXS $150 < p_{\mathrm{T}}(\mathrm{V}) < 250$ GeV region and Z+HF for STXS $250 < p_{\mathrm{T}}(\mathrm{V}) < 400$ GeV region, for the 2017 data-taking period.

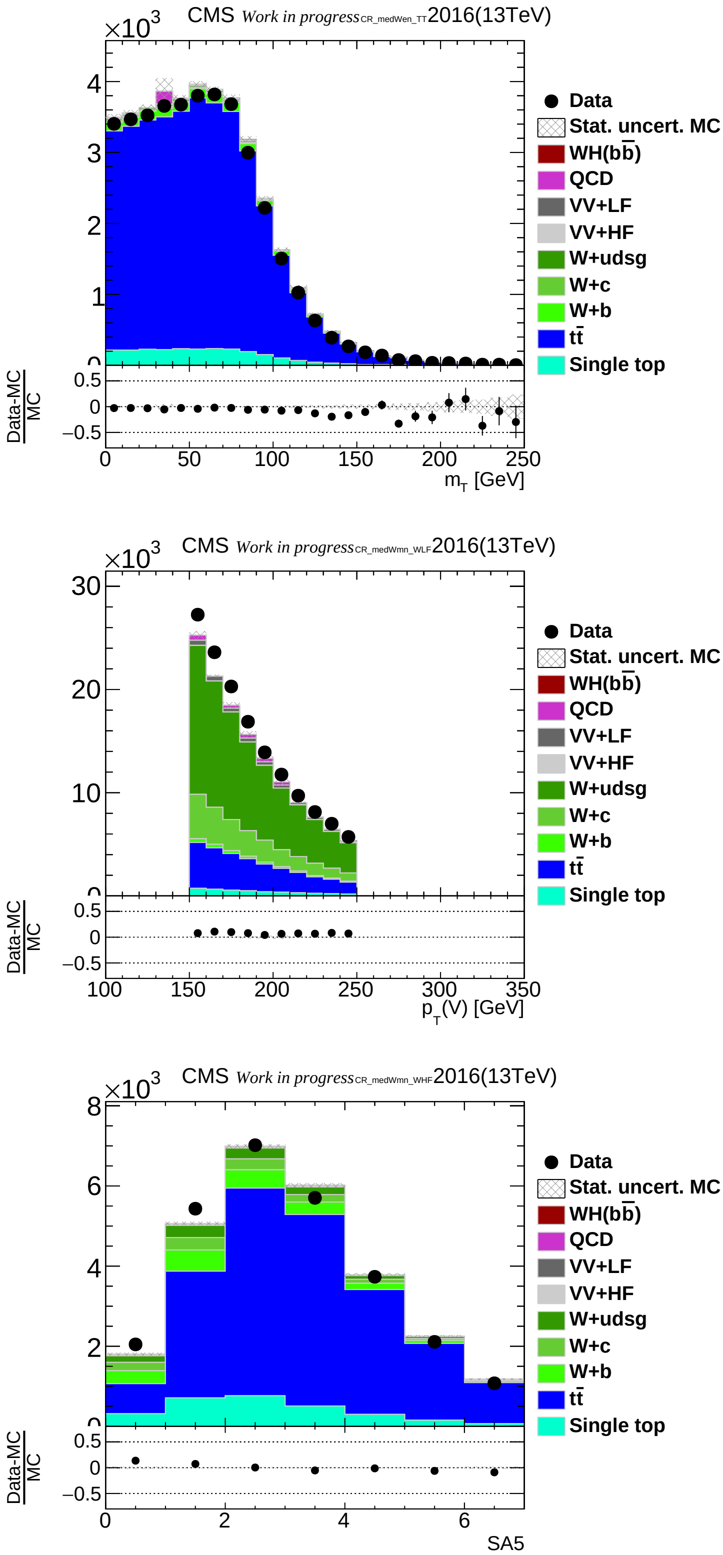


Figure 4.13: Different distributions for the one-lepton channel in $t\bar{t}$ for the $150 < p_{\mathrm{T}}(\mathrm{V}) < 250\,\mathrm{GeV}$ region, Z+LF for the $150 < p_{\mathrm{T}}(\mathrm{V}) < 250\,\mathrm{GeV}$ region and Z+HF for the $150 < p_{\mathrm{T}}(\mathrm{V}) < 250\,\mathrm{GeV}$ region, for the 2016 data-taking period.

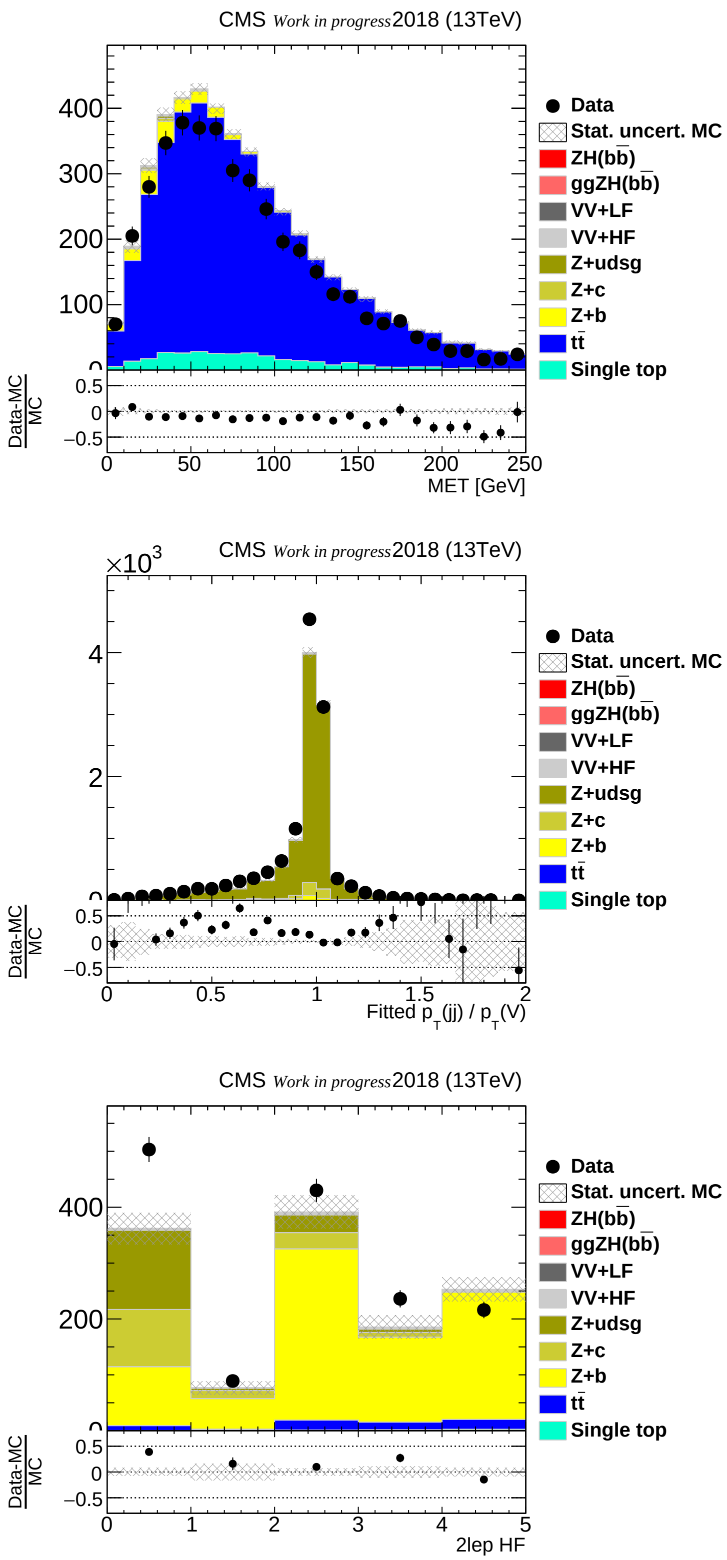


Figure 4.14: Various distributions for the two-lepton channel in $t\bar{t}$ for $150 < p_T(V) < 250\,\text{GeV}$ region, Z+LF for the $150 < p_T(V) < 250\,\text{GeV}$ region and Z+HF for the $150 < p_T(V) < 250\,\text{GeV}$ region, for the 2018 data-taking period.

4.5.4 Boosted topology selection criteria for SR and CR

As discussed previously, considering a boosted topology improves the results with reducing the multi-jet backgrounds. To be considered as the signal and control regions for the boosted topology, the STXS bins with $p_{\mathrm{T}}(\mathrm{V})$ higher than 250 GeV must pass selection on the DeepAK8 Double b-tagger discriminant. For the signal region, the soft-drop mass [76] of the fat jet must be within the Higgs boson mass window. Similarly to the resolved topology, orthogonal control regions are defined. The full list of cuts are available in Table 4.11. Some of the control region distributions are shown in Fig. 4.15.

The DeepAK8 algorithm is calibrated by considering the efficiency of the data against the Monte Carlo simulations in phase spaces enriched in boosted *b*-jets coming from gluon splitting events (g$\rightarrow$bb). As light, *c* and *b* boosted jets in top-quark decays are present in the V+LF, V+HF and $t\bar{t}$ control regions and there are no dedicated studies on the efficiencies of the DeepAK8 algorithm for these regions, free floating rate-parameters are assigned to the V+LF, V+HF and $t\bar{t}$ regions to account for the efficiency of this algorithm in these regions. These rate parameters are called "in-situ" scale factors in the context of this thesis, and are constrained in the simultaneous fit.

SR Variable	zero-lepton channel	one-lepton channel	two-lepton channel
Double b-tagger	> 0.8	> 0.8	> 0.8
m_{jj}	$\in [90, 150]$ GeV	$\in [90, 150]$ GeV	$\in$[90,150] GeV
$n_{add.\,lep}$	$= 0$	$= 0$	-
$n_{add.\,jet}$	$= 0$	$= 0$	-
V_{mass}	-	-	$\in [75, 105]$ GeV
Anti-QCD	True	-	-

V + HF Variable	zero-lepton channel	one-lepton channel	two-lepton channel
Double b-tagger	> 0.8	> 0.8	> 0.8
m_{jj}	$\notin [90, 150]$ GeV	$\notin [90, 150]$ GeV	$\notin [90, 150]$ GeV
$n_{add.\,lep}$	$= 0$	$= 0$	-
$n_{add.\,jet}$	$= 0$	$= 0$	-
V_{mass}	-	-	$\in [75, 105]$ GeV
Anti-QCD	True	-	-

V + LF Variable	zero-lepton channel	one-lepton channel	two-lepton channel
Double b-tagger	< 0.8	< 0.8	< 0.8
m_{jj}	> 50 GeV	> 50 GeV	> 50 GeV
$n_{add.\,lep}$	$= 0$	$= 0$	-
$n_{add.\,jet}$	$= 0$	$= 0$	-
V_{mass}	-	-	$\in [75, 105]$ GeV
Anti-QCD	True	-	-

$t\bar{t}$ Variable	zero-lepton channel	one-lepton channel	two-lepton channel
Double b-tagger	> 0.8	> 0.8	> 0.8
m_{jj}	> 50 GeV	> 50 GeV	> 50 GeV
$n_{add.\,lep}$	> 0	> 0	-
$n_{add.\,jet}$	> 1	> 1	-
V_{mass}	-	-	$\notin [90, 150]$ GeV
Anti-QCD	-	-	-

Table 4.11: Summary of all the cuts used to define signal and control regions for the boosted toplogies.

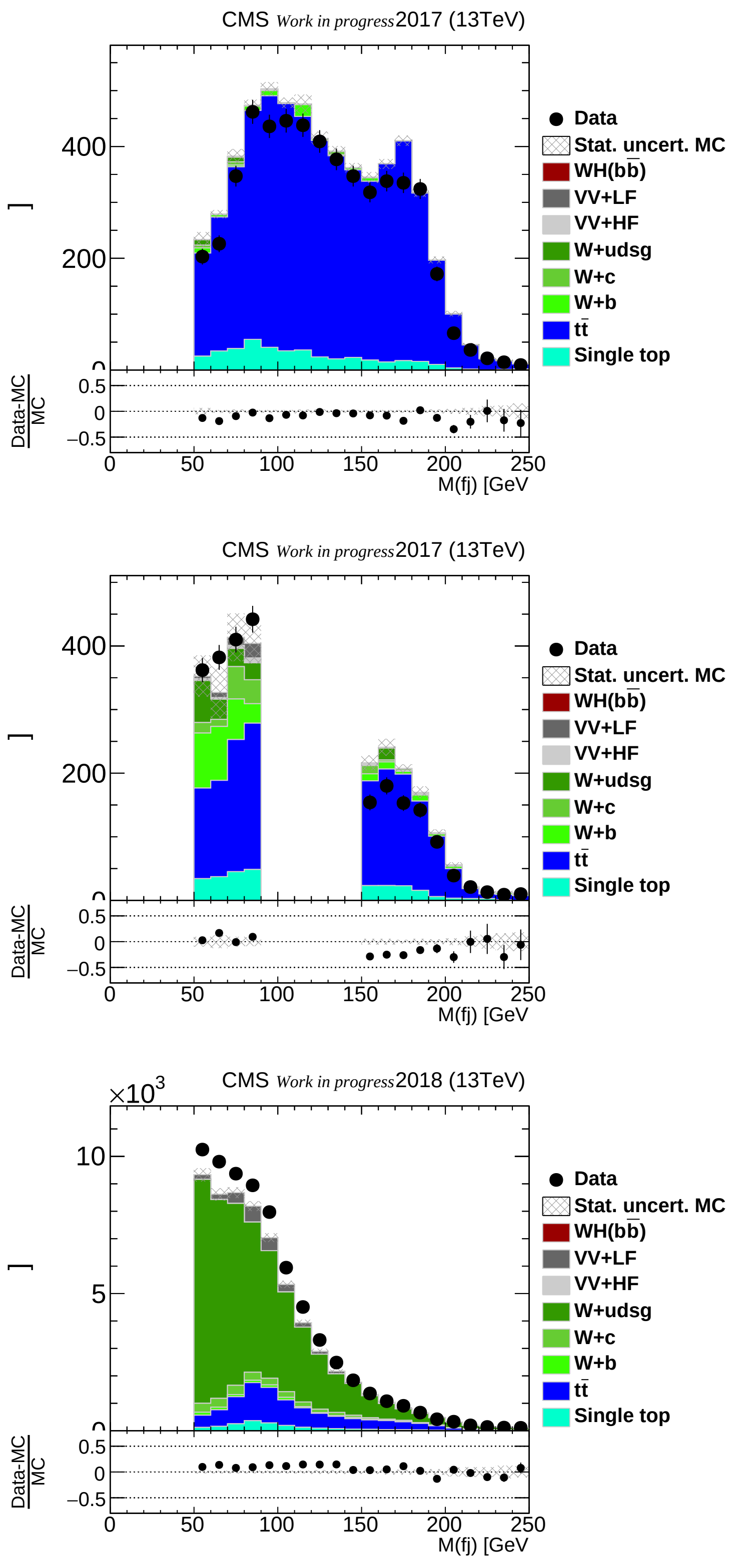


Figure 4.15: The soft drop mass distributions for the 1-lepton channel of the boosted topology in $t\bar{t}$(top left) and W+LF (top right) for the 2017 data taking period and W+HF (bottom) for the 2018 data-taking period.

4.5.5 Top quark reconstruction

In the events with one lepton and MET in the final states, it is helpful to reconstruct the top quark mass by assuming that the lepton and MET stem from the W boson decay and combine their four-momenta with that of the b-jet spatially close to this system to reconstruct the four-momentum of the top quark. Then the top quark mass estimate is used as an input to the multivariate method described in Section 4.6. An example of the reconstructed top quark mass distribution is shown in Fig. 4.16.

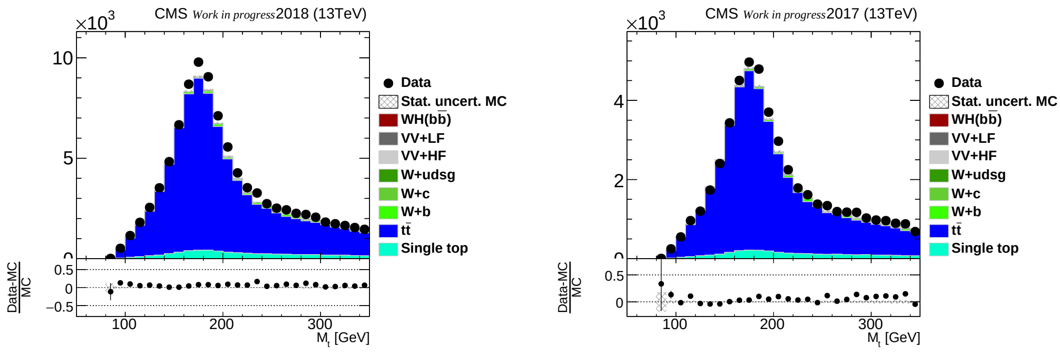


Figure 4.16: The distribution of the top quark mass in the $t\bar{t}$ enriched region, for the single muon channel with 2018 data (left) and the single electron channel with 2017 data (right).

4.5.6 Higgs boson reconstruction

The Higgs boson candidate is reconstructed from the four-vectors of the two highest p_T b-jets (the b-jets selection in the signal regions is described in Section 4.5.3) that pass all the selections. In order to improve the accuracy of this reconstruction, jets with p_T>30 GeV that are within a ΔR cone of less than 0.8 around any of the two selected b-jets are attributed to FSR jets. The FSR jets four-momenta are added to the four-momenta of the selected b-jets. Furthermore, the b-jet regression and finally a kinematic fit for the two-lepton channel only, which corrects the MET and jet p_T using the information from the Z boson, complete the Higgs candidate reconstruction (Fig. 4.18). In Fig. 4.18, the distributions are fitted with the double shouldered crystal ball function [40] to find the width and mean.

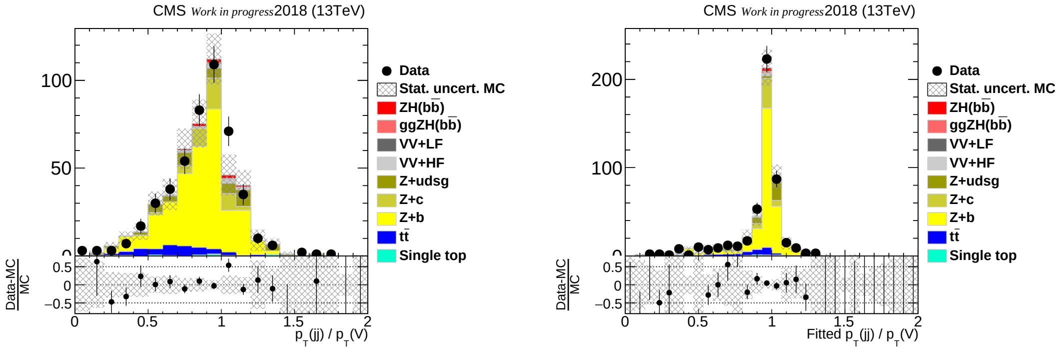


Figure 4.17: Improvement in resolution, the dijet p_{T} divided by the vector boson p_{T}, before (left) and after (right) kinematic fit, shown for the 2018 data taking period and the medium $p_{\mathrm{T}}(\mathrm{V})$ with no additional jets STXS signal region.

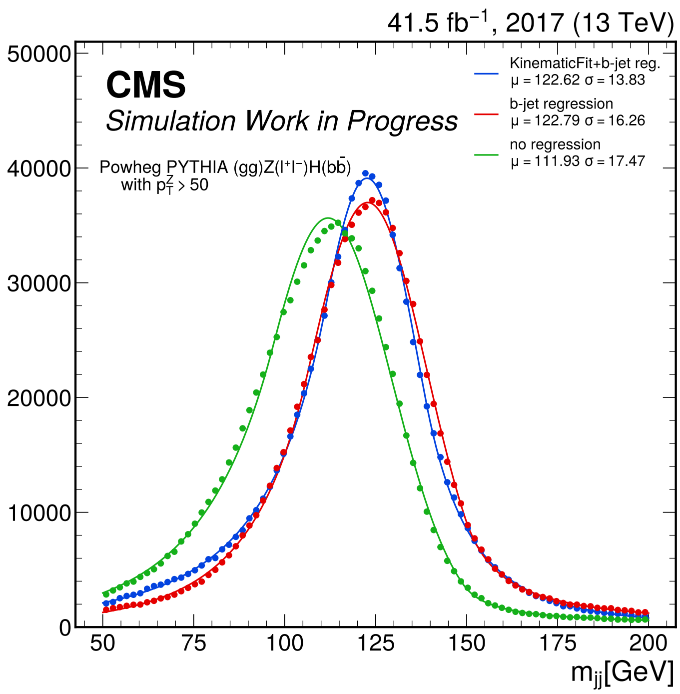


Figure 4.18: Comparison of Higgs candidate dijet system mass before and after applying corrections. Comparing the mean values of the fit, the b-jet regression pushes the Higgs candidate mass closer to the expected mass for Higgs. The kinematic fit significantly improves the resolution as one can see from the σ values of the fitted distributions.

4.6 Multivariate analysis

To separate the signal and background events, deep learning methods (DNN) were employed in the resolved topologies and boosted decision trees (BDT) in the boosted topologies. Furthermore, a multi-class deep learning method (HFDNN) is used in the V+HF control region to help alleviate possible mismodelings in flavor compositions.

4.6.1 Deep learning

Using an architecture with five hidden layers with sizes of 512, 256, 128, 64, 64 and 64 nodes with Leaky ReLU [112] activation for each layer, a deep multilayer classifier is built which is used both for the multi-class and the signal-background classification. The structure of this DNN is shown schematically in Fig. 4.19. The softmax ($softmax(x_i) = \frac{\exp x_i}{\sum \exp x_i}$) activation function is used in the final layer for assigning the probabilities per class. The Adam optimization algorithm is incorporated to minimize the cross entropy loss function for the training and optimization steps. Furthermore, skip connections [79] are added to every other layer to help the minimization. To train the classifier algorithm, the inputs summarized in Table 4.14 for each channel are used.

In case of the signal-background classification, all the signal process described in Section 4.3.1 are considered as one class (signal class) and background processes described in Section 4.3.2 as another (background class). The distribution of the signal-background classifier DNN is shown in Fig. 4.20.

The multi-classifier DNN is only used in the zero and one-lepton channels to add additional control over the single-top and $t\bar{t}$ backgrounds while creating templates to improve modelling of the b and c-quark contributions, five classes are defined as summarized in Table 4.12. Category classes 0-2 are defined to denote the V + light-jets, V + c-jets and V + b-jets as described in the vector boson + jets section in the background processes description (Section 4.3.2). The category classes 3 and 4 are the classes assigned to the single-top and the $t\bar{t}$ which correspond to the same named background processes and are also defined in Section 4.3.2. Confusion matrices for zero and one-lepton channel are reported in Figs. 4.22 and 4.23. The confusion matrices show the performance of categorization for each pair of category classes defined in Table 4.12.

In the two-lepton channel V+HF control region, instead of using the multi-classifier DNN, the DeepCSV b-tagger working points are used to define a template to help the fit in constraining b and c-quark contributions. The labeling is described in Table 4.13 where "T" refers to the tight working point selection, "M" to the medium working point selection and "L" to the loose working point selection.

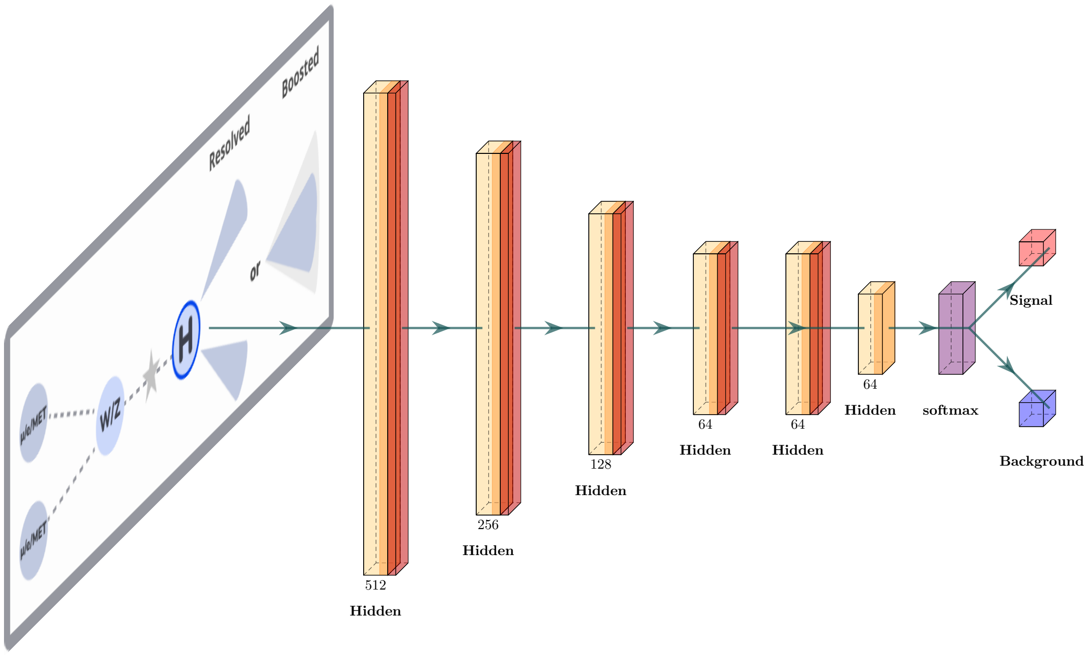


Figure 4.19: The architecture of the DNN, after each hidden layer a Leaky ReLU activation and on the last layer a softmax activation is used.

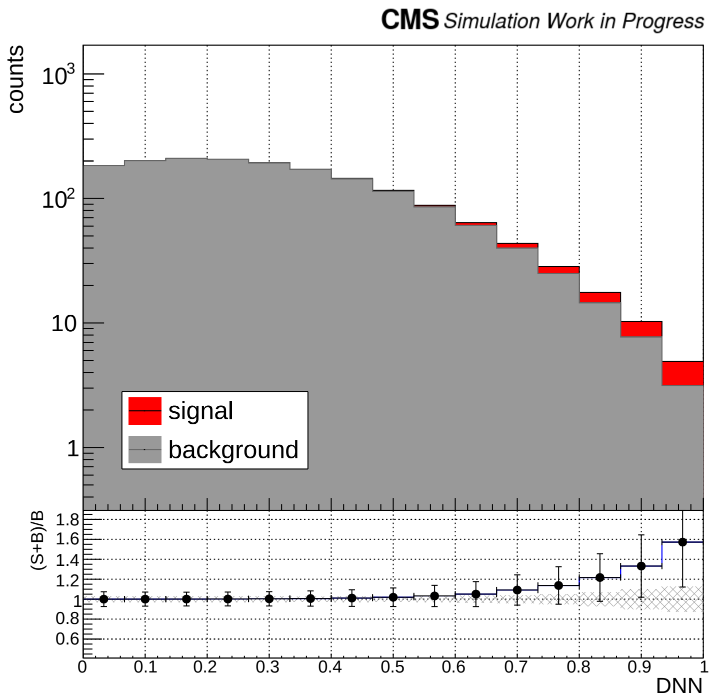


Figure 4.20: An example of the DNN output for the signal-background classification for the 2-lepton high $p_{\mathrm{T}}(\mathrm{V})$ region.

0	V+udsg (VL)
1	V+c (VC)
2	V+b (VB)
3	single-top (ST)
4	$t\bar{t}$ (TT)

Table 4.12: Classes used for the zero/one-lepton multi-background classifier.

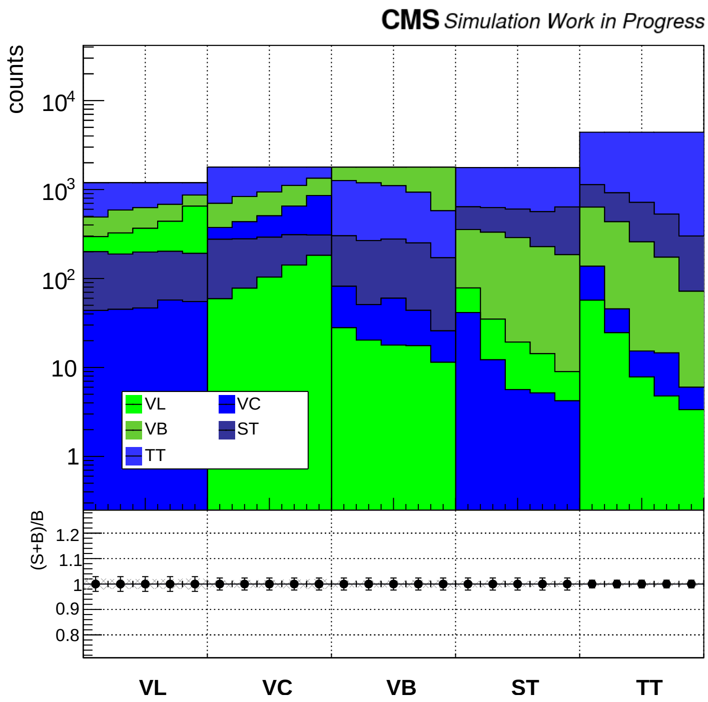


Figure 4.21: An example of the multi-class DNN output for the 1-lepton channel, where the VL, VC, VB, ST and TT correspond to the classes 0-4 in Table 4.12.

value	DeepCSV max	DeepCSV min
0	$<$ T	$<$ M
1	$<$ T	$>$ M
2	$>$ T	$<$ M
3	$>$ T	$>$ M, $<$ T
4	$>$ T	$>$ T

Table 4.13: Variable used for template fit in 2-lepton HF control region.

Variable	Description	0-lepton	1-lepton	2-lepton	
					kinematic fit applied
m(jj)	dijet invariant mass	✓	✓	✓	✓
pT(jj)	dijet transverse momentum	✓	✓	✓	✓
pT(MET)	transverse momentum of MET	✓	✓	✓	✓
V(mt)	transverse mass of vector boson		✓		
V(pt)	transverse momentum of vector boson		✓	✓	✓
pT(jj)/pT(V)	ratio of transverse momentum of vector boson and higgs boson		✓	✓	✓
$\Delta\phi(V,H)$	azimuthal angle between vector boson and dijet directions	✓	✓	✓	✓
$\mathrm{btag}_{\mathrm{max}}$ WP	1,2,3 if b-tagging discriminant (DeepCSV) score of leading jet is above T, M, L WP resp.	✓	✓	✓	✓
$\mathrm{btag}_{\mathrm{min}}$ WP	1,2,3 if b-tagging discriminant (DeepCSV) score of sub-leading jet is above T, M, L WP resp.	✓	✓	✓	✓
$\Delta\eta(jj)$	pseudorapidity difference between leading and sub-leading jet	✓	✓	✓	
$\Delta\phi(jj)$	azimuthal angle between leading and sub-leading jet	✓	✓		
$pT_{max}(j_1,j_2)$	maximum transverse momentum of jet between leading and sub-leading jet	✓	✓		
pT(j_2)	maximum transverse momentum of jet between leading and sub-leading jet	✓	✓		
SA5	number of soft-track jets with pT > 5GeV	✓	✓	✓	
N_{aj}	number of additional jets	✓	✓		
$\mathrm{btag}_{\mathrm{max}}(add)$	maximum btagging discriminant score among additional jets	✓			
$\mathrm{pT}_{\mathrm{max}}$(add)	maximum transverse momentum among additional jets	✓			
$\Delta\phi(jet,MET)$	azimuthal angle between additional jet and MET	✓			
$\Delta\phi(lep,MET)$	azimuthal angle between lepton and MET		✓		
M_t	Reconstructed top quark mass		✓		
$pT(j_1)$	transverse momentum of leading jet			✓	✓
M_t	transverse momentum of sub-leading jet			✓	✓
$m(V)$	Reconstructed vector boson mass			✓	
$\Delta R(V,H)$	angular separation between vector boson and Higgs boson			✓	✓
$\Delta R(V,H)$	angular separation between leading and sub-leading jets			✓	
$\sigma(m(jj))$	resolution of dijet invariant mass				✓
N_{rec}	number of recoil jets				✓

Table 4.14: List of input variables to the DNN and HFDNN used in the measurement analysis.

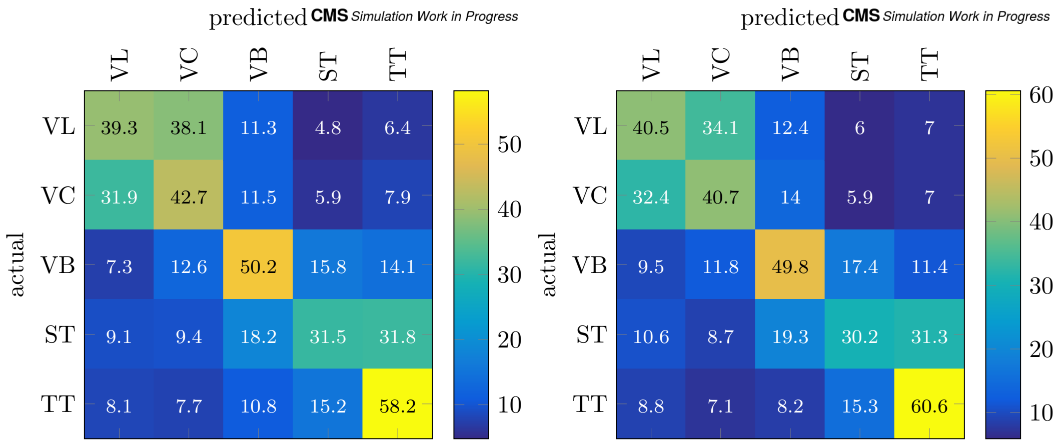


Figure 4.22: Confusion matrix for 0-lepton HF CR 5-process multi-background classifier with new flavor scheme for V+jets: udsg/c/b in medium (left) and high (right) $p_{\mathrm{T}}(\mathrm{V})$regions

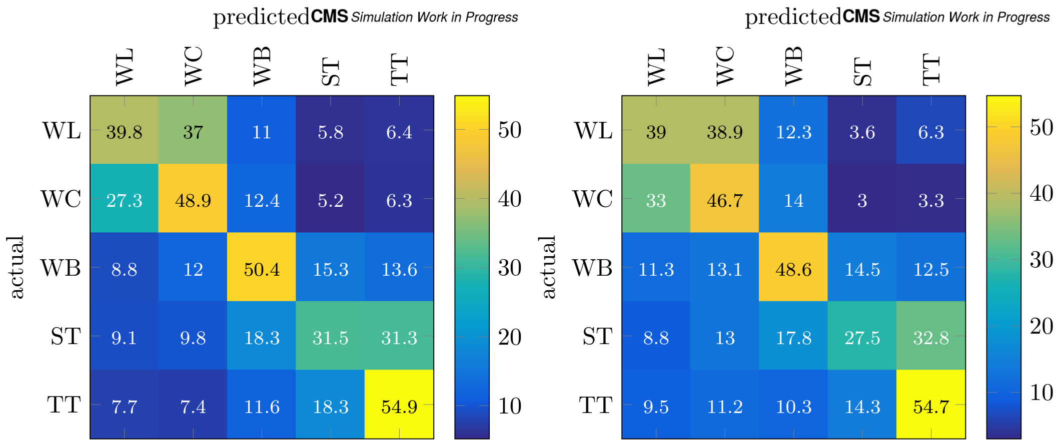


Figure 4.23: Confusion matrix for 1-lepton HF CR 5-process multi-background classifier with new flavor scheme for V+jets: udsg/c/b in medium (left) and high (right) $p_{\mathrm{T}}(\mathrm{V})$regions

4.6.2 Boosted decision tree

A boosted decision tree classifier is used to separate the signal and background in the boosted signal region, where all the signal processes described in Section 4.3.1 are considered as one class (signal class) and background processes described in Section 4.3.2 as another (background class).

In order to have a smooth transition for the overlap events between the boosted and resolved topologies all the inputs in Table 4.14 are included in the training of the BDT in addition to the following variables, where for the overlap events the values of those inputs are nullified:

- Soft-drop mass of the FatJet candidate
- Transverse momentum of the FatJet candidate
- Transverse momentum of the vector boson
- Double b-tagger output node trained on boosted topology (DeepAK8 algorithm)

The distributions of the BDT outputs are shown for all channels and data taking years in Fig. 4.24.

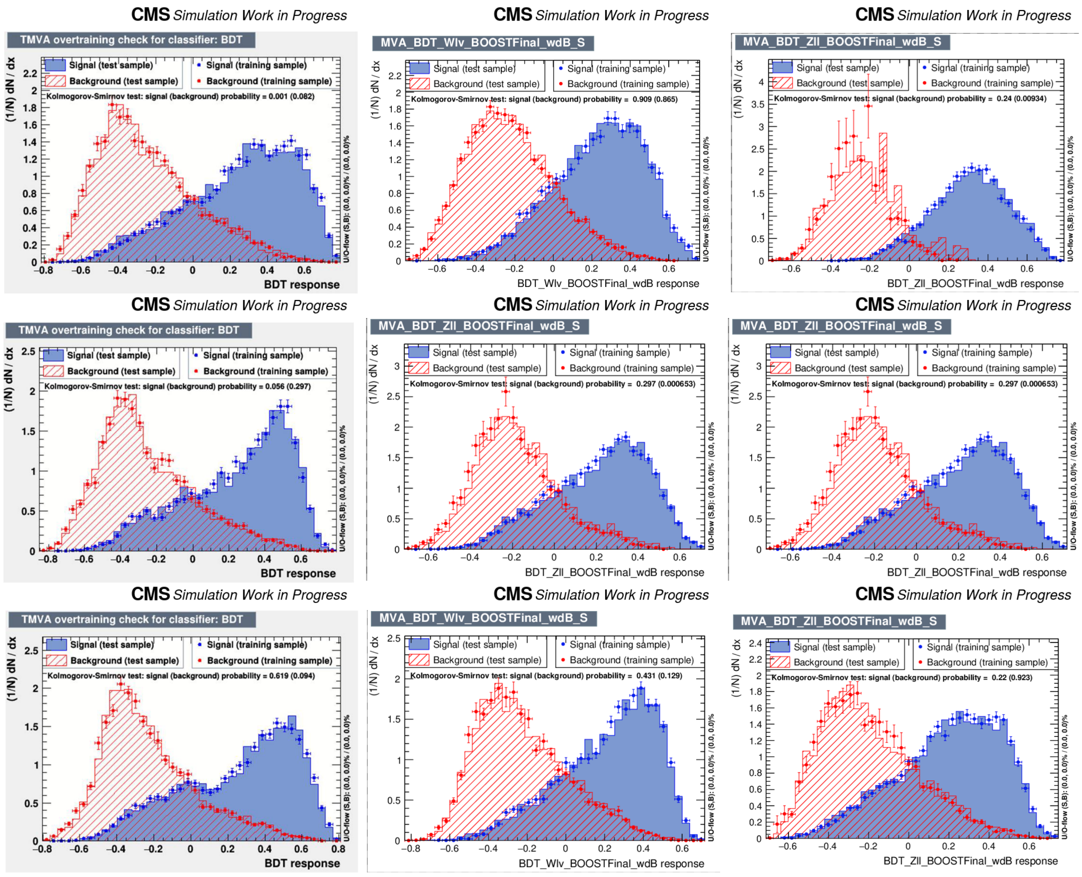


Figure 4.24: Overtraining tests for the boosted topology BDT for the 2016, 2017 and 2018 and for the zero-lepton, one-lepton and two-lepton channels. All the signal processes described in Section 4.3.1 are considered as one class (signal class) and background processes described in Section 4.3.2 as another (background class).

4.7 Systematic uncertainties

In this analysis many sources of systematic uncertainties need to be considered. Several have effects on the normalization of signal or background processes, many others on the shape of the key observables. In the following, a summary of these sources is given.

4.7.1 Uncertainties affecting only the normalization

Luminosity

The uncertainty in the integrated luminosity measurement is 2.5% in 2016 and 2018, and 2.3% in 2017 [34–36]. These uncertainties are partially correlated between the different data-taking periods.

H $\rightarrow b\bar{b}$ branching ratio

The uncertainty in the H $\rightarrow b\bar{b}$ branching ratio is 0.5% [50].

Theoretical (scale) uncertainties in the signal cross section

Theoretical uncertainties from inclusive QCD scale variations for the STXS bins are estimated and included in the analysis. Plots of the inclusive variations for each signal sample and in STXS bins are shown in Fig. 4.25.

Uncertainties in the pdf+α_s for the signal

The errors ΔPDF and $\Delta\alpha_s$ induced by uncertainties in the PDFs and α_s, respectively, are combined as $\Delta\text{PDF} \oplus \Delta\alpha_s$, which is calculated from the 68% CL interval using the PDF4LHC15_nnlo_mc PDF set [19, 22]. These uncertainties amount to 1.6% (ZH) and 1.9% (WH) [51].

Uncertainties in the theoretical $p_{\mathbf{T}}(V)$ spectrum for the signal

As the analysis is performed at high vector boson p_{T}, corrections are applied to correct the p_{T} spectra for the signal MC datasets. This considers both an NLO electroweak and an NNLO QCD correction; the associated uncertainties are 2% and 5%, respectively [51].

Uncertainties in the background normalization

For the backgrounds not measured in control regions, which are single top and diboson, a 15% uncertainty is assigned. This corresponds to the uncertainty of the measured cross sections.

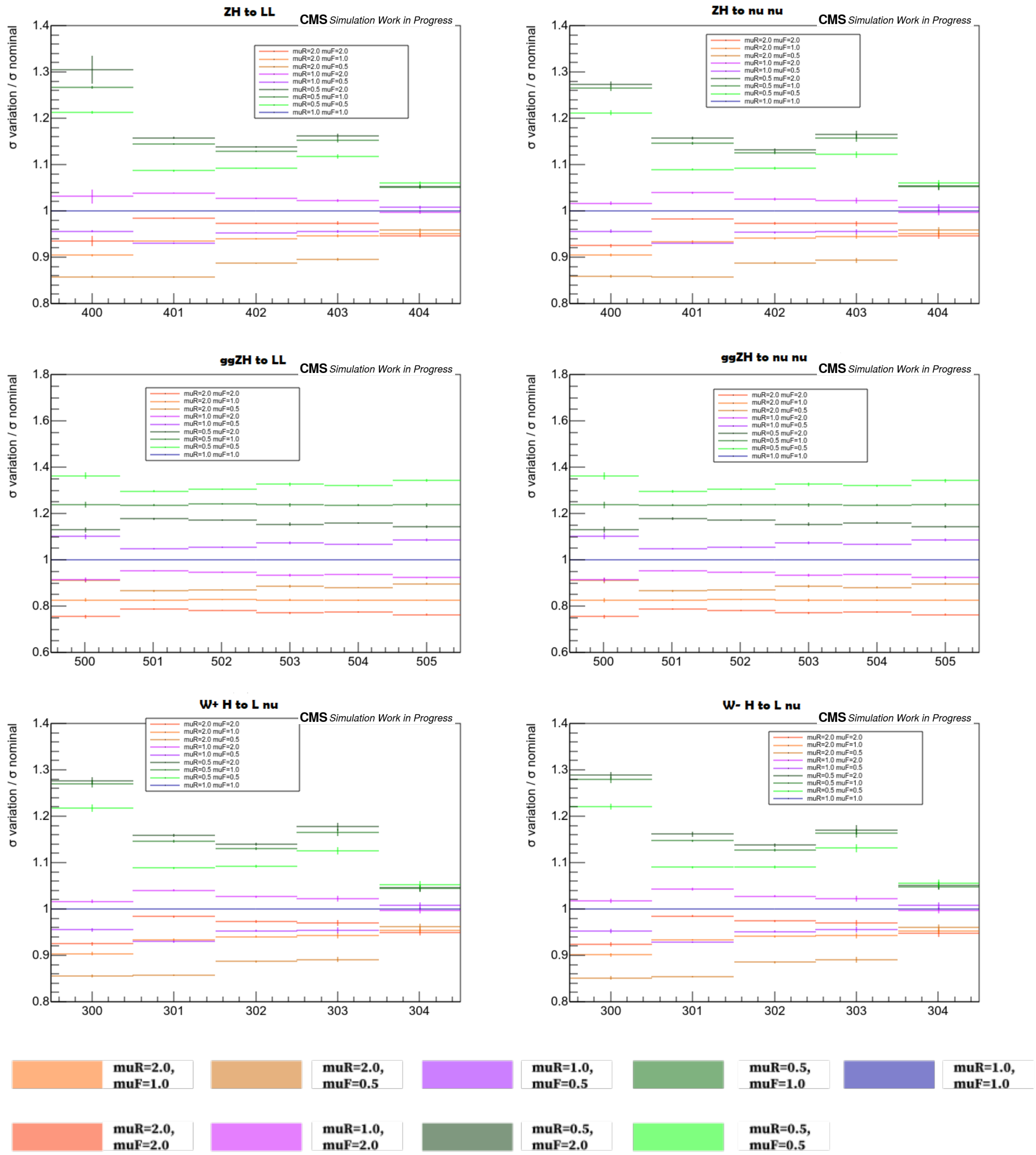


Figure 4.25: Variations of the μ_R and μ_F for generator level STXS categories. The bins denoted as 300, 400 and 500 correspond to the full width. The bins 301 and 401 correspond to the 0 GeV $< p_T^V <$ 150 GeV STXS category. The bins 302, 402 to the 150 GeV $< p_T^V <$ 250 GeV STXS category with zero additional jets and 303, 403 to 150 GeV $< p_T^V <$ 250 GeV STXS category with at least one additional jets. The bins 304 and 404 correspond to $p_T^V >$ 250 GeV STXS category. The bin 501 correspond to 0 GeV $< p_T^V <$ 75 GeV and 502 to 75 $< p_T^V <$ 150 STXS categories. The bin 503 correspond to 150 $< p_T^V <$ 250 STXS category with zero additional jets and 504 to 150 GeV $< p_T^V <$ 250 GeV STXS category with at least one additional jets. 3XX denotes the quark induced WH, 4XX the quark induced ZH and 5XX is the gluon induced ZH signal samples.

Lepton efficiencies

Uncertainties in the electron and muon ID, isolation, and trigger efficiencies amount to 2%, this is estimated by changing the parameters used to perform the efficiency measurement and define the range of the scale factors, then the effects of the variations are estimated in the analysis selection regions.

MET trigger efficiencies

Uncertainties in the MET trigger efficiency measurement amount to 1%, the number is estimated similar to the one for leptons.

4.7.2 Uncertainties with both normalization and shape effects

These uncertainties are considered as variations on the observable shape template and are constrained by the fit.

Jet energy scale (JES)

The jet energy scale uncertainties [75] are categorized as seen in Table 4.15. The uncertainty denoted as "Absolute" measures the absolute jet response relative to the p_{T} scale of the precisely measured object (Z or photon) in $Z(\rightarrow e^+e^-)$+jets, $Z(\rightarrow \mu^+\mu^-)$+jets, γ+jets processes while considering the out of cone showering, underlying event, ISR and FSR effects. The jet response is studied using p_{T}-balance and MPF (missing transverse momentum projection fraction) defined as follows:

$$\begin{aligned} R_{\mathrm{jet},p_T} &= \frac{p_{T,\mathrm{jet}}}{p_{T,\mathrm{ref}}} \\ R_{\mathrm{jet,MPF}} &= 1 + \frac{\vec{p}_T^{\,\mathrm{miss}} \cdot \vec{p}_{T,\mathrm{ref}}}{(p_{T,\mathrm{ref}})^2}. \end{aligned} \tag{4.9}$$

The uncertainty denoted as "Fragmentation" in Table 4.15 is the uncertainty of estimating the effect of the fragmentation and underlying event modeling on the absolute scale of the jet response. The Single Pion response uncertainties account for the detector calibration effect on the absolute scale of the jet response. The Flavor uncertainty is derived by checking the differences between Pythia and Herwig generators in modeling jet flavors.

The relative corrections are measured by considering the response of all jets relative to the response for central jets ($|\eta| < 1.3$). The uncertainty denoted as "RelativeJER" in Table 4.15 is estimated by varying the data/MC scales when applying jet p_{T}-resolution smearing to simulation for different regions of the detector. The uncertainty denoted as "RelativeBal" in Table 4.15 estimates the difference from estimating the JES with p_{T}-balance or MPF. The uncertainty denoted as "RelativeSample" in Table 4.15 estimates the differences among the dijet, Z+jets, γ+jets samples. The uncertainty denoted as "RelativeFSR" in Table 4.15 is estimating the ISR and FSR effects using the Herwig generator as data and comparing the effect of each method on JES. The uncertainty denoted as "RelativeStat" in Table 4.15 estimates the uncertainties due to the number of data events in each region of the detector.

The additional contribution to the jet energy and momentum from the pile-up is estimated and called the pile-up offset, the uncertainties denoted as "PileUpDataMC" and "PileUpPt" in Table 4.15 account for the effects coming from simulation and the p_T-balance in estimating the offset.

The size and shape of JES uncertainties are depicted in Figs. 4.26 and 4.27. The size of the uncertainties for the 2017 data taking era is larger in comparison to the ones from 2016 and 2018, mainly due to detector effects.

Absolute	Uncertainties on the absolute scale
Fragmentation	Difference of Fragmentation and UE between Pythia/Herwig
SinglePionECAL	Single-pion response in ECAL, 3%
SinglePionHCAL	Single-pion response in HCAL, 3%
Flavor	Variation of possible color mixtures
RelativeJER	Jet p_T resolution
RelativeBal	MPF (Missing Transverse Energy Projection Fraction) vs. p_T-balance
RelativeSample	Difference among dijet, Z+jets, γ+jets
RelativeFSR	ISR+FSR correction
RelativeStat	Statistical uncertainty
PileUpDataMC	Data vs. MC simulation offset
PileUpPt	Jet p_T-dependent offset

Table 4.15: The jet energy scale uncertainties groups used, each of which can be subdivided into several uncertainties, each of which covers distinct methodologies, samples, or detector locations, this grouping is based on [75].

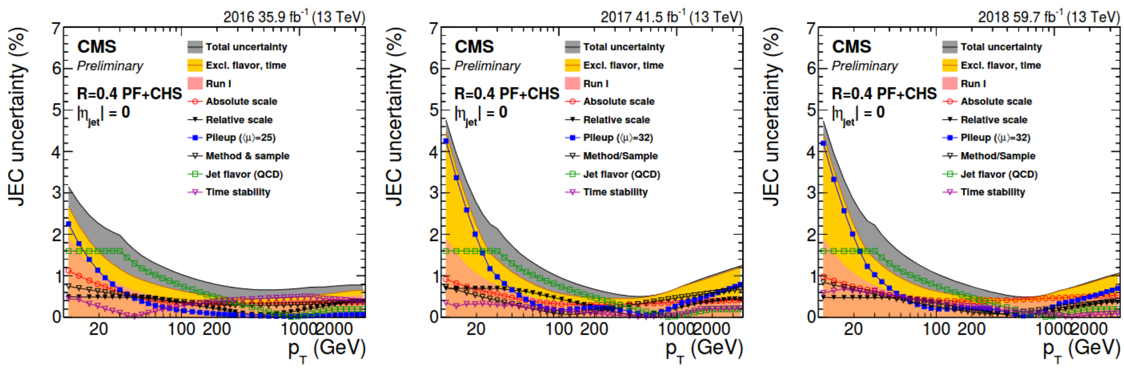


Figure 4.26: JES uncertainty sources and total uncertainty (quadratic sum of individual uncertainties) as a function of p_T^{Jet}, taken from [72]

Jet energy resolution

The resolution for jet energies in recorded data does not match exactly those from the MC simulations, since the jets in recorded data have lower energy resolution. Therefore, the resolution for the jets in the MC simulations are smeared

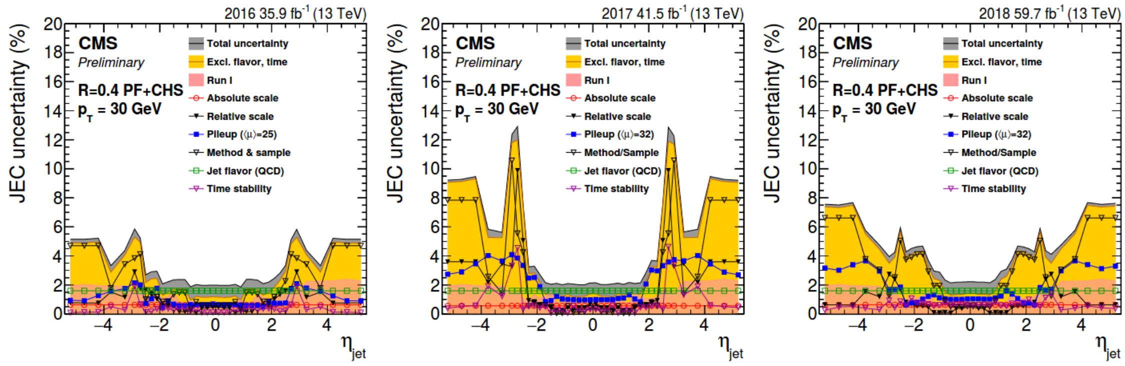


Figure 4.27: JES uncertainty sources and total uncertainty (quadratic sum of individual uncertainties) as a function of η^{Jet}, taken from [72]

to match the ones from data. The smearing is applied to the MC sample jets as follows,

$$p_{T,smeared} = s_{\mathrm{JER}} \times (p_{T,gen} + (p_{T,reco} - p_{T,gen})(1 + s_{\mathrm{diff}})), \tag{4.10}$$

where the $p_{T,gen}$ are the jet p_{T} without the simulated detector effects and the $p_{T,reco}$ after considering the detector effects. The s_{JER} and s_{diff} are listed in Table 4.16.

Year	Scaling (s_{JER})	Resolution Difference (s_{diff})
2016	0.998 ± 0.019	0.017 ± 0.060
2017	1.020 ± 0.023	0.088 ± 0.071
2018	0.985 ± 0.019	0.080 ± 0.073

Table 4.16: The smearing corrections for each data taking era as a percent of the jet's p_T.

B-tagging

Uncertainties in the DeepCSV b-tagger's calibration are represented as shape changes for nine uncorrelated sources, which can be categorized into three groups:

- Uncertainties stemming from the jet energy scale
- Flavor contamination, light+charm in heavy-flavor control region and b+charm in light-flavor
- Statistical uncertainties corresponding to each flavor of the tagged jets

Each of these sources are implemented in bins of p_{T} and η.

For the boosted topology, DeepAK8 Double b-tagger is used. The "in-situ" scale factors are used to account for the efficiency and variations (not constrained prior to the final fit) for the background. Shape uncertainties with variations up to 10 % are used for signal processes in bins of the Double b-tagger score and p_{T}^{Jet} (scores from 0.8-0.97 and from 0.97-1.0).

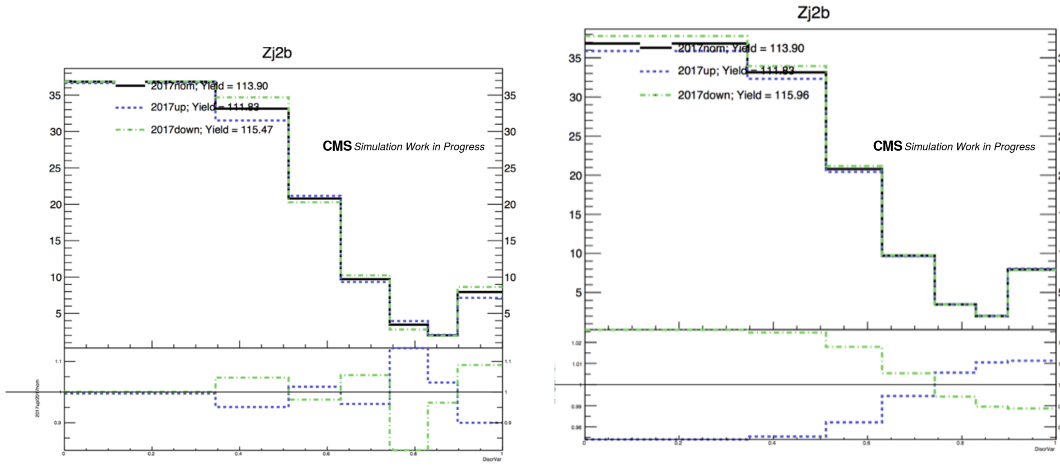


Figure 4.28: Impact of the template smoothing on the pile-up uncertainty template associated to the Z+bb process of the 2017 data taking era, shown as an example for the $150 < p_{\mathrm{T}}(\mathrm{V}) < 250\,\mathrm{GeV}$ two-lepton signal region. Before smoothing the up and down variations are fluctuating (right) and after smoothing the shape has a smooth variation which matches the expected physical behavior (left).

Uncertainties due to the limited number of simulated events

These are taken into account using the Barlow-Beeston-lite [37] approach.

To avoid nonphysical constraints as described in Section 4.4.2, the observable shape templates are smoothed using the method described in [53]. The effect of smoothing is shown in Fig. 4.28 for one of the signal region templates for the 2017 data taking era.

Chapter 5

Invariant mass cross-check analysis

The main approach to extract the signal strength for the resolved topologies in this analysis is to fit the distribution of the discriminant of the deep learning multivariate classifier described in Section 4.6. This is generally referred to as the DNN-based approach in this thesis. To validate this approach, a cross-check analysis is performed, using the distribution of the mass of the leading two b-jets in the signal region as the observable, where the jets have passed FSR recovery and b-jet regression as described in Section 4.5.6. This mass is denoted as m_{jj} in this chapter. This approach results in direct physical interpretation based on the mass distributions associated with the $Z \to bb$ and $H \to bb$ processes.

5.1 Mass extracted DNN (MEDNN)

One of the main challenges for the mass cross-check analysis is that any cuts on the DNN discriminant change the shape of the m_{jj} distribution as each event is effectively transformed based on similarity to the events that are signal Higgs events. To restore the physical m_{jj} distribution, one can remove the dependence of the DNN on m_{jj} and parameters that are highly correlated to it.

To find the correlated parameters and to rank them, the relative importance is defined by means of a regressor over all input parameters to the DNN to predict m_{jj}, where the regressor is trained again and again removing the features one by one resulting on training on a subset of input features. Then using the mean value of Asimov Median Significance (AMS) [38] as the metric, the relative importance of a feature j is evaluated as follows:

$$I_{rel}(j) = \frac{\langle AMS \rangle|_{j \epsilon S_n} - \langle AMS \rangle|_{j \epsilon S_{n-1}}}{\langle AMS \rangle|_{j \epsilon S_{n-1}}} \tag{5.1}$$

where S_{n-1} denotes the subset of features including j and removing another input feature. This way of defining the importance ensures that the correlations between features are considered. The feature ranking for each channel with the value of the relative importance is shown Figs. 5.1 to 5.3.

The highest ranked features are fixed in the evaluation of the DNN to the mean value of the background, this way the dependence to these parameters are removed while using exactly the same DNN as used in the DNN-based analysis, resulting in restoring the unweighted m_{jj} distributions as shown in Figs. 5.4 to 5.5 for each channel, where the distribution before (left) and after (right) the mass extraction is shown for each channel and for roughly the same events. The new DNN which has little correlation with the m_{jj} is referred to as the Mass Extracted DNN (MEDNN).

rank	feature	rel. importance [%]
1	$\Delta\phi$(jj)	6.92±1.35
2	$\Delta\eta$(jj)	5.23±1.48
3	max jet p_T	4.86±1.50
4	min jet p_T	2.50±1.37
5	H p_T	1.72±1.15
6	$\Delta\phi$(H,MET)	-0.52±1.20
7	MET_Pt	-1.65±0.87
8	n add jet	-1.80±0.77
9	add jet d phi	-1.92±0.82
10	max btag WP	-2.20±0.84
11	SA5	-2.38±0.72
12	add jet pt	-2.39±0.69
13	const. zero	-2.68±0.68
14	min btag WP	-2.85±0.82
15	add jet btag	-2.92±0.67

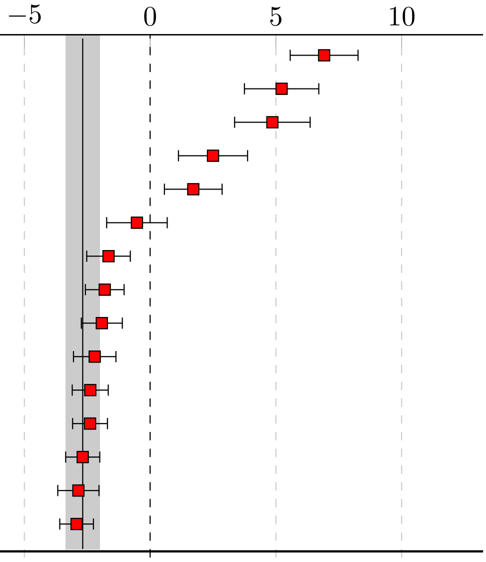


Figure 5.1: Impact of each input parameter using the relative importance (See Eq. (5.1)) measure, for zero-lepton channel.

5.2 Selection

The selection for the DNN-based approach is explained in detail in Section 4.5. In the signal region definition for the DNN-based analysis, m_{jj} variable is limited to 90-150 GeV range and a complementary control region is defined with considering m_{jj} variable outside this limit, namely the V+HF control region. For the mass cross-check analysis, it is desired to have a wider range for the m_{jj} to probe the behavior of the $Z \to bb$ and $H \to bb$ processes alongside one another. Therefore the mass based cross-check analysis follows the DNN-based selection, but with removal of the V+HF control region. These changes to the selection are shown schematically in Fig. 5.7.

To improve the performance of the cross-check analysis, the signal region is divided into three regions for the zero and one-lepton channel and five regions for the two-lepton channels by applying cuts on the MEDNN discriminant distribution. These cuts are shown in Tables 5.1 to 5.3 where for the two-lepton there are two separate MEDNNs with $p_{\mathrm{T}}(\mathrm{V}) \leq 150$ GeV (high) and

rank	feature	rel. importance [%]
1	$\Delta\eta$(jj)	2.39±0.43
2	max jet p_T	2.23±0.44
3	p_T(H)	1.54±0.43
4	min jet p_T	1.16±0.44
5	m_{top}	-0.17±0.24
6	p_T(V)/p_T(H)	-0.30±0.31
7	p_T(V)	-0.53±0.23
8	$\Delta\phi$(V,H)	-0.56±0.32
9	MET	-0.59±0.23
10	min btag WP	-0.62±0.21
11	m_T(V)	-0.67±0.22
12	dPhiLepMet	-0.70±0.22
13	const. zero	-0.82±0.21
14	max btag WP	-0.85±0.21
15	SA5	-0.90±0.21
16	n add jet	-0.94±0.20

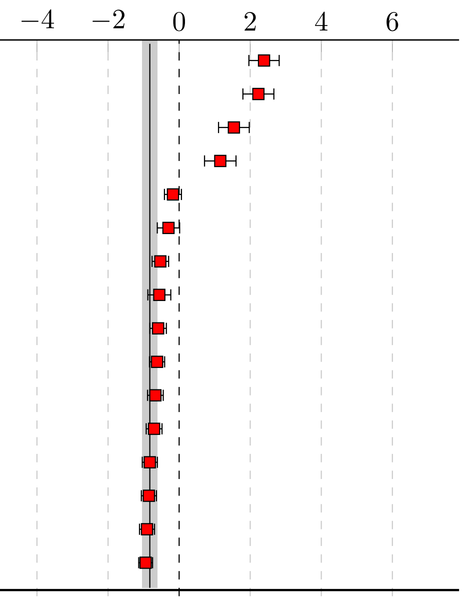


Figure 5.2: Impact of each input parameter using the relative importance (See Eq. (5.1)) measure, for one-lepton channel.

$p_{\mathrm{T}}(\mathrm{V}) < 150$ GeV (low) corresponding to the two separate DNNs trained for these regions. The value of each cut for these boundaries are determined by optimizing the AMS value for each choice. As an example the optimization for the one-lepton channel is shown in Fig. 5.8 where the value of AMS for the corresponding cuts are shown for the sequence of optimization iterations on the right and as a 2D representation on the left.

Channel	Boundaries
zero-lepton	[0.0 ; 0.474 ; 0.744 ; 0.744 ; 1.0]
one-lepton	[0.0 ; 0.400 ; 0.689 ; 0.689 ; 1.0]
two-lepton low	[0.0 ; 0.401 ; 0.611 ; 0.611 ; 1.0]
two-lepton high	[0.0 ; 0.574 ; 0.574 ; 1.0]

Table 5.1: 2018 MEDNN category boundaries

Channel	Boundaries
zero-lepton	[0.0 ; 0.427 ; 0.737 ; 0.737 ; 1.0]
one-lepton	[0.0 ; 0.315 ; 0.437 ; 0.437 ; 1.0]
two-lepton low	[0.0 ; 0.451 ; 0.669 ; 0.669 ; 1.0]
two-lepton high	[0.0 ; 0.560 ; 0.560 ; 1.0]

Table 5.2: 2017 MEDNN category boundaries

rank	feature	rel. importance [%]
1	kinFit $\sigma(m_{jj})$	13.75±0.78
2	ΔR(jj)	10.65±0.89
3	kinFit min jet p_T	8.76±1.02
4	kinFit p_T(H)	5.35±1.15
5	kinFit max jet p_T	2.77±0.97
6	p_T(H)	1.64±0.79
7	$\Delta\eta$(jj)	1.32±0.84
8	min jet p_T	1.12±0.87
9	kinFit p_T(H)/p_T(V)	0.89±0.80
10	p_T(H)/p_T(V)	0.68±0.74
11	max jet p_T	0.67±0.83
12	p_T(V)	-1.98±0.84
13	kinFit n recoil jets	-2.03±0.71
14	kinFit p_T(V)	-2.09±0.72
15	m(V)	-2.20±0.78
16	kinFit ΔR(H,V)	-2.28±0.79
17	MET	-2.43±0.68
18	min btag	-2.69±0.69
19	n add jet	-2.83±0.72
20	$\Delta\phi$(H,V)	-2.95±0.71
21	const. zero	-3.03±0.73
22	kinFit $\Delta\phi$(H,V)	-3.25±0.64
23	max btag	-3.55±0.67
24	ΔR(H,V)	-3.64±0.65
25	SA5	-3.70±0.64
26	kinFit m(V)	-3.81±0.63

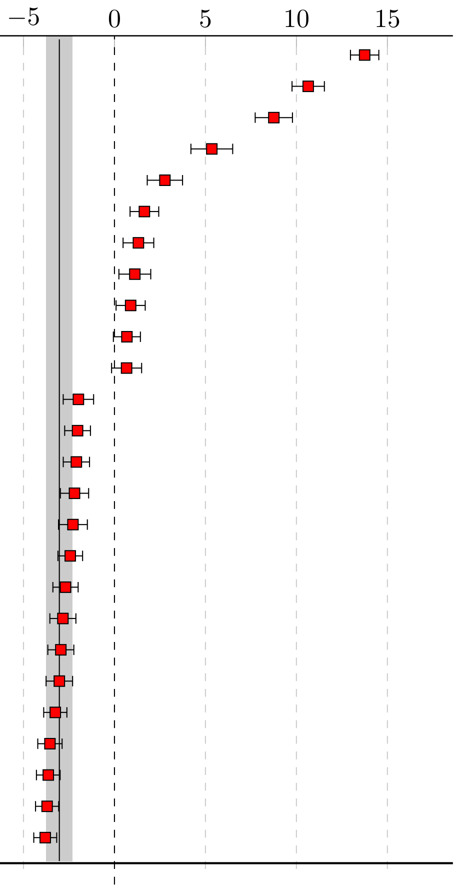


Figure 5.3: Impact of each input parameter using the relative importance (See Eq. (5.1)) measure, for two-lepton channel.

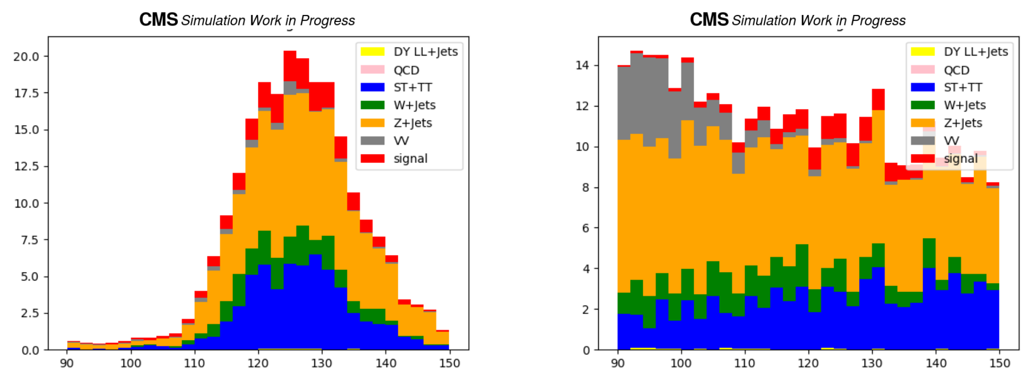


Figure 5.4: Demonstration of the effect of removing the bias from m_{jj} distribution. m_{jj} distribution for zero-lepton channel after applying a similar cut on the DNN discriminant (left) and MEDNN discriminant (right); the m_{jj} distribution on the right is less biased towards the Higgs mass.

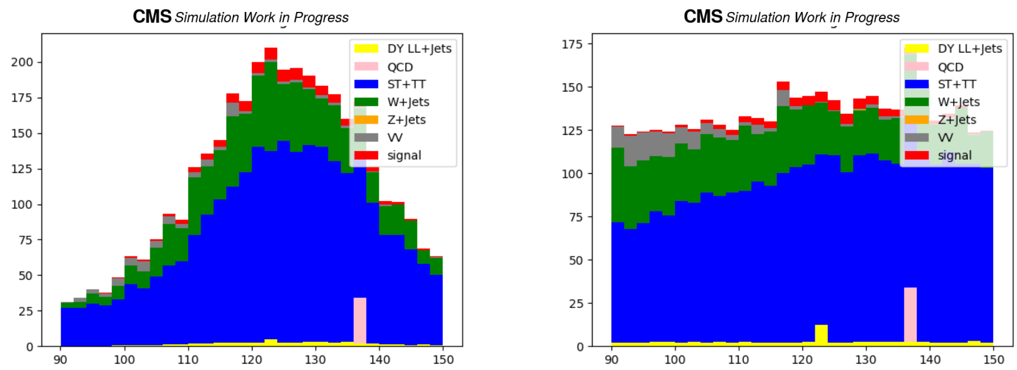


Figure 5.5: Demonstration of the effect of removing the bias from m_{jj} distribution. m_{jj} distribution for one-lepton channel after applying a similar cut on the DNN discriminant (left) and MEDNN discriminant (right); the m_{jj} distribution on the right is less biased towards the Higgs mass.

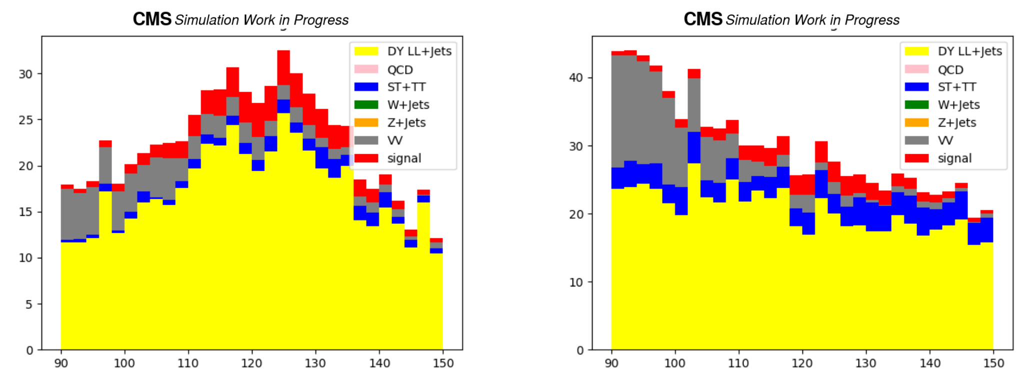


Figure 5.6: Demonstration of the effect of removing the bias from m_{jj} distribution. m_{jj} distribution for two-lepton channel after applying a similar cut on the DNN discriminant (left) and MEDNN discriminant (right); the m_{jj} distribution on the right is less biased towards the Higgs mass.

Channel	Boundaries
zero-lepton	[0.0 ; 0.520 ; 0.810 ; 0.810 ; 1.0]
one-lepton	[0.0 ; 0.330 ; 0.470 ; 0.470 ; 1.0]
two-lepton low	[0.0 ; 0.400 ; 0.620 ; 0.620 ; 1.0]
two-lepton high	[0.0 ; 0.610 ; 0.610 ; 1.0]

Table 5.3: 2016 MEDNN category boundaries

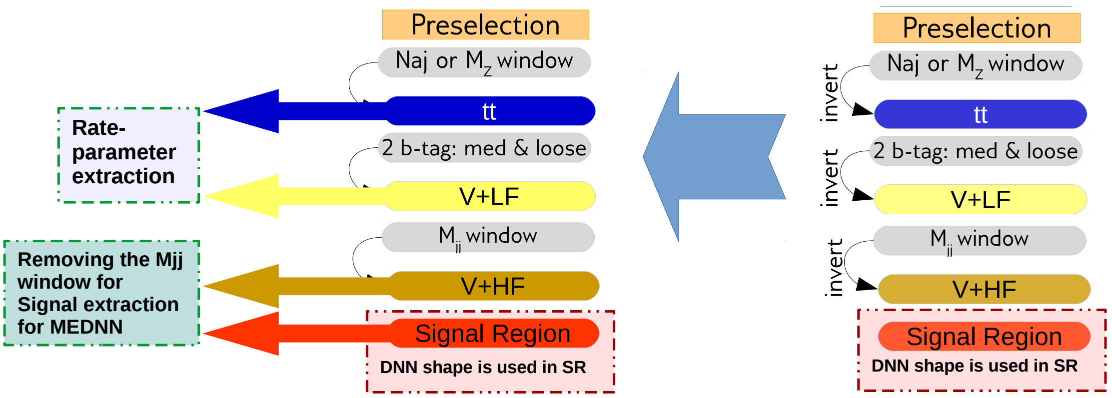


Figure 5.7: Selection for the Invariant mass cross-check analysis, the main change in the new selection stems from removing the Higgs mass window, therefore merging the V+HF control region with the signal region.

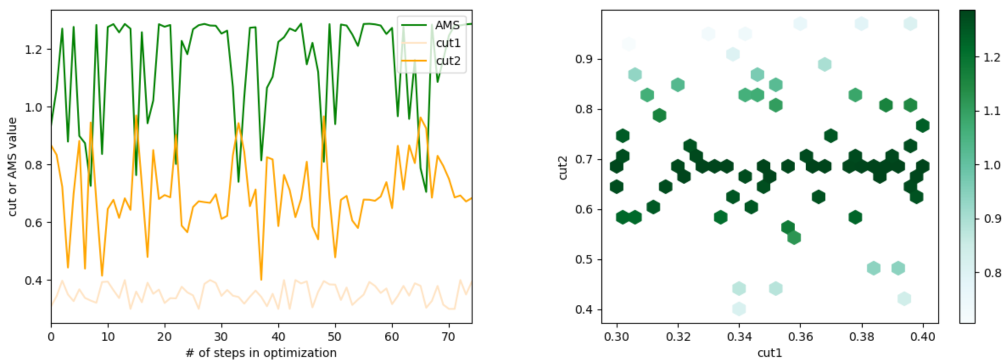


Figure 5.8: The value of AMS for the corresponding cuts on the MEDNN are shown for the sequence of optimization iterations on the right and as a 2D representation on the left for the one-lepton channel.

Chapter 6

Results

In this chapter, the results of the analysis work described in this thesis is presented.

6.1 Signal strengths modifiers for the STXS categories and the inclusive VH$(b\bar{b})$

To extract the result, a combined simultaneous signal and background likelihood fit in control and signal regions is performed as described in Section 4.2. The templates used in each region of the fit are summarized in Table 6.1.

	SR	$t\bar{t}$ CR	V+LF CR	V+HF CR
zero-lepton, resolved	DNN	$p_{\mathrm{T}}(\mathrm{V})$	$p_{\mathrm{T}}(\mathrm{V})$	HFDNN
one-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger
one-lepton, resolved	DNN	$p_{\mathrm{T}}(\mathrm{V})$	$p_{\mathrm{T}}(\mathrm{V})$	HFDNN
one-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger
two-lepton, resolved	DNN	$p_{\mathrm{T}}(\mathrm{V})$	$p_{\mathrm{T}}(\mathrm{V})$	DeepCSV scores
two-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger

Table 6.1: The variables for the distributions used for each signal and control region in the fit. In the signal regions, the DNN and BDT distributions are used. For the V+LF resolved control region, the $p_{\mathrm{T}}(\mathrm{V})$ is used. In the V+LF and V+HF boosted control regions and V+HF two-lepton resolved control region, the b-tagging discriminant distribution is used and for the rest of the resolved topology the HFDNN is used.

In the resolved signal regions, the DNN as described in Section 4.6.1 is used while for the boosted signal regions the BDT (Section 4.6.2) distributions are used. For the V+LF resolved control region, the $p_{\mathrm{T}}(\mathrm{V})$ variable is used. In the V+LF and V+HF boosted control regions the DeepAK8 discriminant for the light-flavour discrimination is used and for the V+HF two-lepton resolved control region the DeepCSV b-tagging discriminant distribution (binned in the working point cuts introduced in Table 4.13) is used. For the rest of the resolved control regions, the HFDNN of Section 4.6.1 is used. The relevant selections are described in detail in Section 4.5.

Furthermore, as described in Section 4.2 for each template in the fit, the overall normalization and shapes are varied within the range of statistical and systematic uncertainties discussed in Section 4.7. The treatment of the nuisance parameters in the fit model is described in Section 4.4.1. The free parameters associated with the normalization of important background processes, namely the $t\bar{t}$, V+$udsg$, V+c and V+b processes in the fit (also called scale factors) are constrained by the control region templates, then extrapolated to the signal regions. These scale factors are inclusive in $p_{\mathrm{T}}(\mathrm{V})$. For the boosted topologies in the high $p_{\mathrm{T}}(\mathrm{V})$ regions, the dedicated in-situ scale factors described in Section 4.5.4 are used.

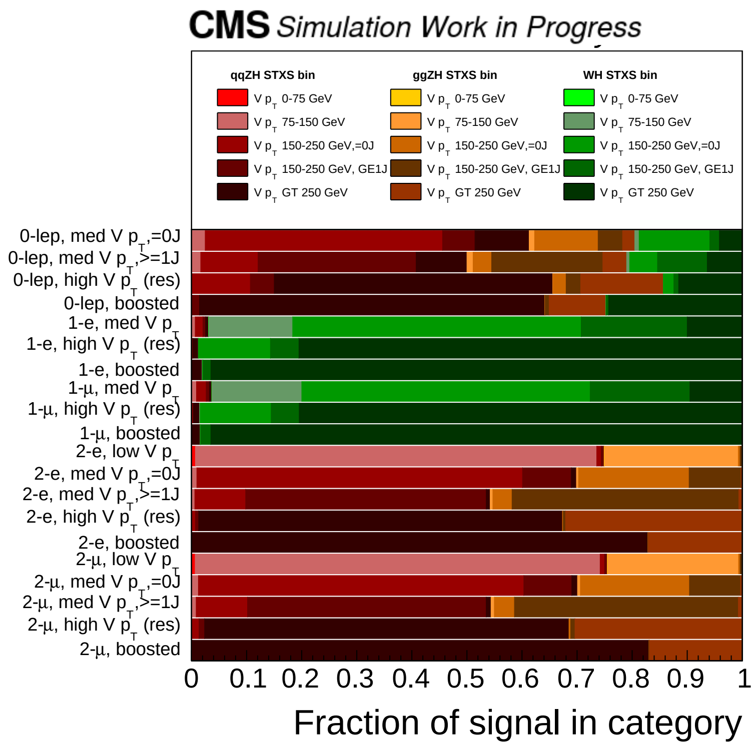


Figure 6.1: The composition of the signal processes in bins of the STXS are shown as fractions for the 2017 signal regions.

The composition of the signal processes for the reconstructed-level STXS categories are shown as fractions for the 2017 signal regions in Fig. 6.1, the 2016 and 2018 are expected to follow the same pattern. This plot shows the contamination from other signal processes and migrations between the generator-level STXS categories. The Fig. 6.1 shows that the reconstructed categories measured in this analysis are compatible with the STXS categorization.

6.1.1 Combination of 2016, 2017 and 2018 data taking periods

The combined result of all the data taking periods is presented with different configurations of the combination in this section, the inclusive signal strength extracted from the fit is $\mu = 0.71 \pm 0.22$, this result is compatible with the standard model (an expected signal strength of 1) within the 1.5 σ. The p-value for the compatibility to the SM for this result is 13%. The observed (expected) significance of the VH($b\bar{b}$) signal is 3.6 σ (4.7 σ). This result is further split between the W and Z boson and presented in Fig. 6.3.

The result of the per channel combinations is presented in Fig. 6.2 (right) on top of the inclusive result.

The combined results for the STXS bins are shown in Fig. 6.4, the post-fit plots of a selected DNN and BDT templates for the signal regions are shown in Fig. 6.5. The values of the scale-factors after the fit, are reported in Tables 6.2 to 6.4. The in-situ scale factors for the DeepAK8 c-components often deviate significantly from one, this is partially due to the low statistics for this contribution in control regions to have a good post fit constraint and also partially due to the DeepAK8 architecture as similar behavior has been seen in other external studies using this tagger. The correlations of the signal strength with the c-jet process and in-situ scale factors have been checked and found to be lower than 3%. The impacts of systematic uncertainties are grouped and are presented in Table 6.5.

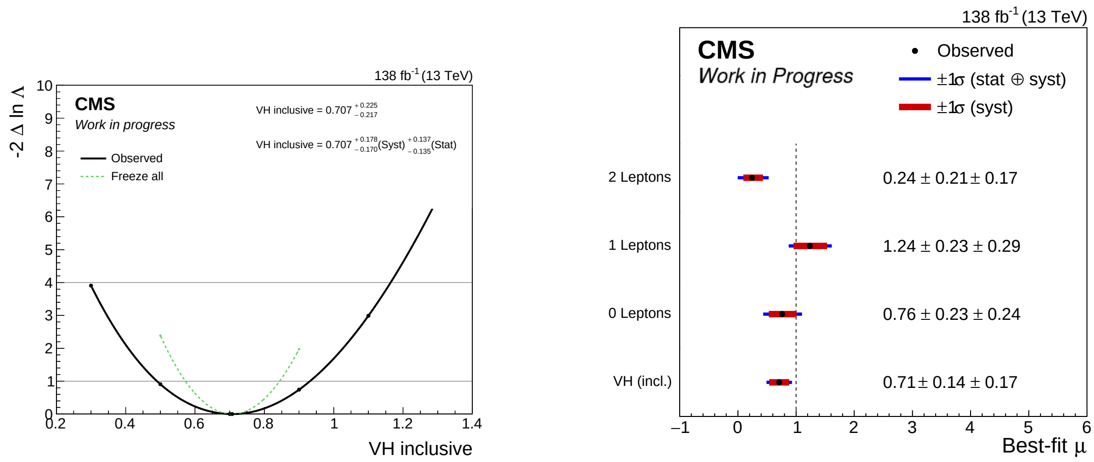


Figure 6.2: Observed results of the full Run 2 combination for the per-channel signal extractions and the inclusive signal strength.The compatibility of the fit with respect to the SM is 13% for the inclusive fit.

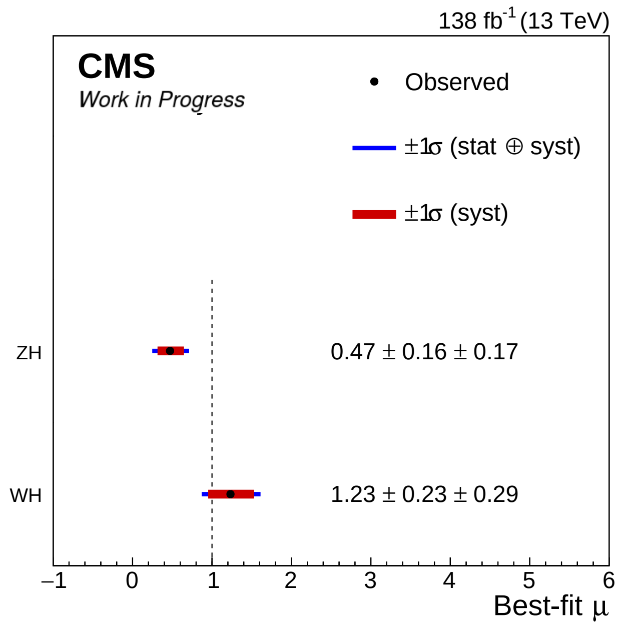


Figure 6.3: Observed results of the full Run 2 combination for the production mode split (ZH/WH) signal extractions.

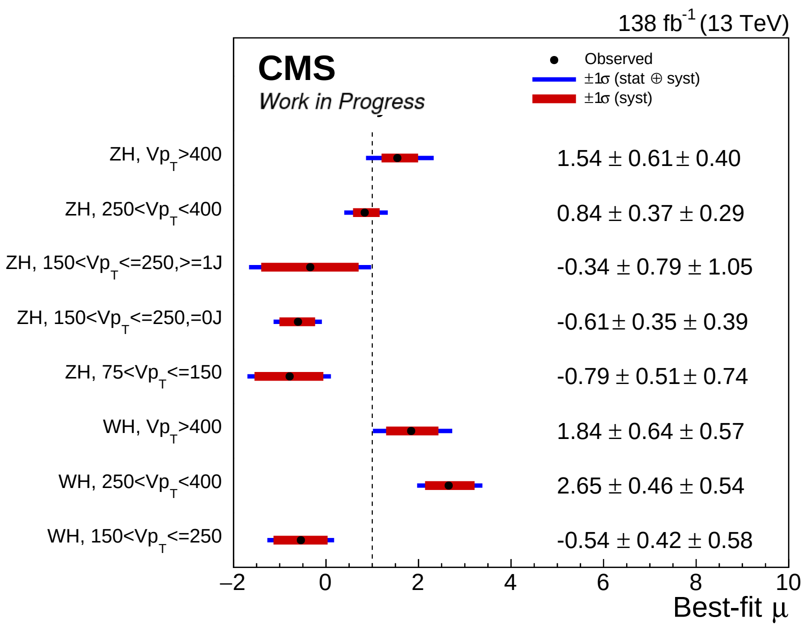


Figure 6.4: Observed results of the full Run 2 combination for the STXS-based signal extractions.

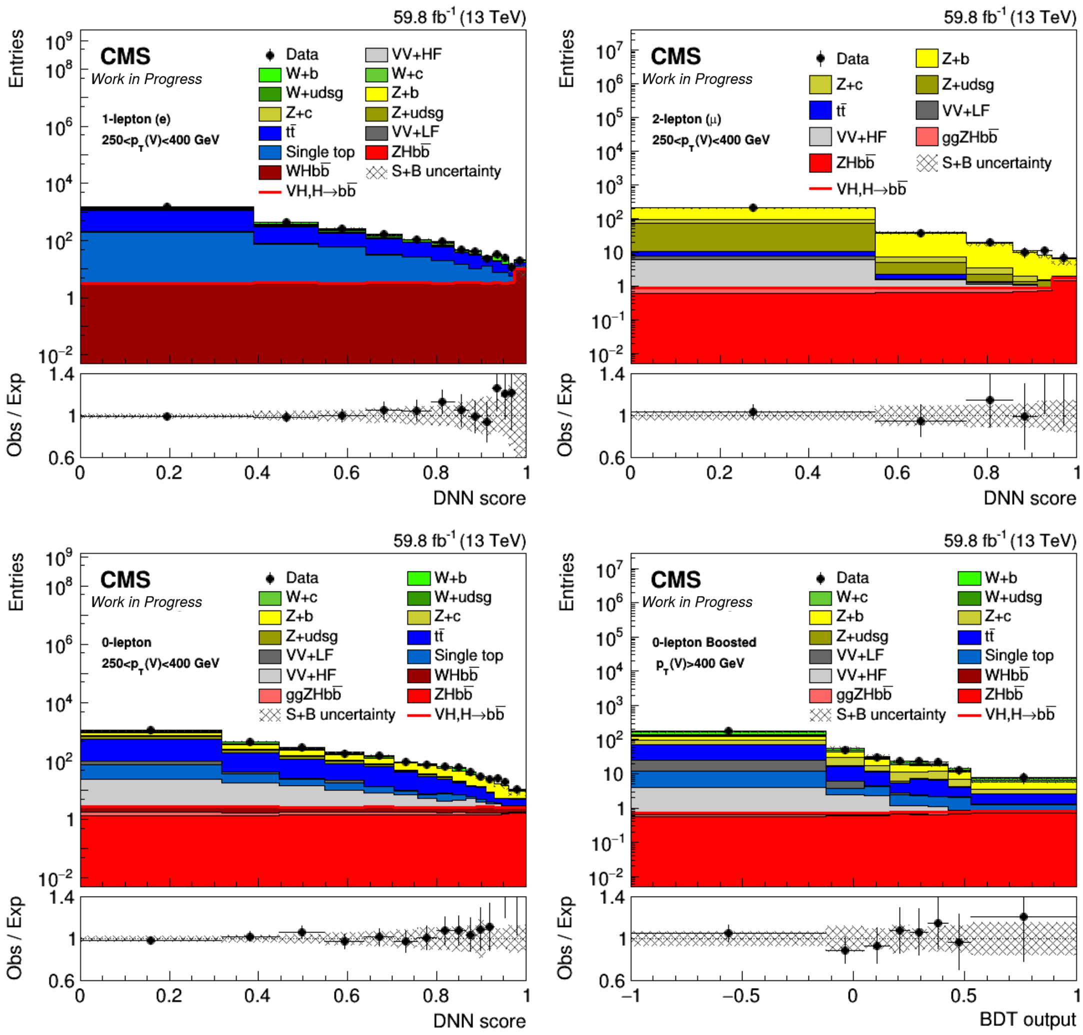


Figure 6.5: Examples of postfit templates for DNN and BDT for resolved and boosted signal regions, respectively, for the 2018 data taking era.

scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
$t\bar{t}$	0.97±0.08	0.67±0.05	0.82±0.07	0.73±0.11	0.90±0.12
V+b,bb	1.79±0.14	0.87±0.11	1.21±0.12	1.16±0.13	1.17±0.12
V+udsg	1.11±0.21	0.61±0.09	0.70±0.10	1.09±0.12	1.09±0.12
V+c	2.68±0.58	1.17±0.27	1.68±0.34	1.45±0.43	2.74±1.13

scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
in-situ $t\bar{t}$	0.87±0.02	0.87±0.02	0.87±0.02	0.87±0.02	0.87±0.02
in-situ heavy flavor	0.62±0.09	0.48±0.13	0.54±0.14	0.48±0.13	0.54±0.14
in-situ light flavor	1.75±0.29	1.13±0.08	1.10±0.07	1.13±0.08	1.10±0.07
in-situ c (pass b-tag)	0.74±0.11	0.74±0.11	0.74±0.11	0.74±0.11	0.74±0.11
in-situ c (fail b-tag)	0.36±0.06	0.36±0.06	0.36±0.06	0.36±0.06	0.36±0.06

Table 6.2: Normalization scale factors for the combined fit in the zero-lepton, one-lepton, two-lepton channels (top) and boosted in-situ (bottom) for the 2016 data taking period. The errors include both statistical and systematic uncertainties.

scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
$t\bar{t}$	0.96±0.08	0.87±0.07	0.90±0.07	1.03±0.13	1.05±0.17
V+b,bb	1.02±0.07	1.13±0.16	1.40±0.18	1.06±0.09	1.04±0.08
V+udsg	1.38±0.14	0.89±0.07	0.85±0.07	1.12±0.20	1.02±0.08
V+c	0.40±0.13	0.99±0.18	1.00±0.19	0.84±0.11	1.01±0.23
scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
in-situ $t\bar{t}$	0.93±0.02	0.93±0.02	0.93±0.02	0.93±0.02	0.93±0.02
in-situ heavy flavor	0.84±0.12	0.76±0.13	0.74±0.11	0.76±0.13	0.74±0.11
in-situ light flavor	0.91±0.05	0.91±0.04	0.95±0.04	0.91±0.04	0.95±0.04
in-situ c (pass b-tag)	4.10±0.91	4.10±0.91	4.10±0.91	4.10±0.91	4.10±0.91
in-situ c (fail b-tag)	2.24±0.40	2.24±0.40	2.24±0.40	2.24±0.40	2.24±0.40

Table 6.3: Normalization scale factors for the combined fit in the zero-lepton, one-lepton, two-lepton channels (top) and boosted in-situ (bottom) for the 2017 data taking period. The errors include both statistical and systematic uncertainties.

scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
$t\bar{t}$	1.10±0.09	0.89±0.07	0.96±0.08	1.27±0.16	1.01±0.18
V+b,bb	1.15±0.09	0.92±0.14	1.35±0.16	1.02±0.07	1.10±0.16
V+udsg	0.91±0.15	1.02±0.08	1.03±0.08	0.45±0.16	1.00±0.07
V+c	0.61±0.28	0.40±0.18	0.38±0.17	0.86±0.11	0.53±0.15
scale-factor	zero-lepton	one-lepton e	one-lepton μ	two-lepton e	two-lepton μ
in-situ $t\bar{t}$	0.87±0.02	0.87±0.02	0.87±0.02	0.87±0.02	0.87±0.02
in-situ heavy flavor	0.94±0.11	0.71±0.12	0.78±0.11	0.71±0.12	0.78±0.11
in-situ light flavor	1.05±0.07	0.90±0.04	0.92±0.04	0.90±0.04	0.92±0.04
in-situ c (pass b-tag)	4.09±1.51	4.09±1.51	4.09±1.51	4.09±1.51	4.09±1.51
in-situ c (fail b-tag)	3.84±1.37	3.84±1.37	3.84±1.37	3.84±1.37	3.84±1.37

Table 6.4: Normalization scale factors for the combined fit in the zero-lepton, one-lepton, two-lepton channels (top) and boosted in-situ (bottom) for the 2018 data taking period. The errors include both statistical and systematic uncertainties.

VH($b\bar{b}$)	Relative uncertainty on μ
Background (theory)	8%
Signal (theory)	11%
MC stats.	16.5%
Sim. modelling	10%
b-tagging	7%
Jet energy resolution	6%
Luminosity	3%
Jet energy scale	4%
LeptonID	1%
Trigger(MET)	0.03%

Table 6.5: Impacts of different nuisance groups in terms of relative uncertainties on the full Run 2 VH($b\bar{b}$) signal strength.

6.1.2 2016 data taking period results

The unblinded results for the 2016 data taking period is shown in Fig. 6.6, the V+jets modeling in this region is done using samples with the leading order accuracy in bins of the H_T and enriched in b-jets, as described in Section 4.3.3. The higher ranked nuisance parameters are shown with their impacts in Fig. 6.7.

This channel contributes significantly to the sensitivity of the analysis, as the lower statistical errors due to the higher statistics of the leading order V+jets Monte Carlo samples results in lower Monte Carlo statistical errors for this data taking period compared to the other two where the NLO V+jets has been used.

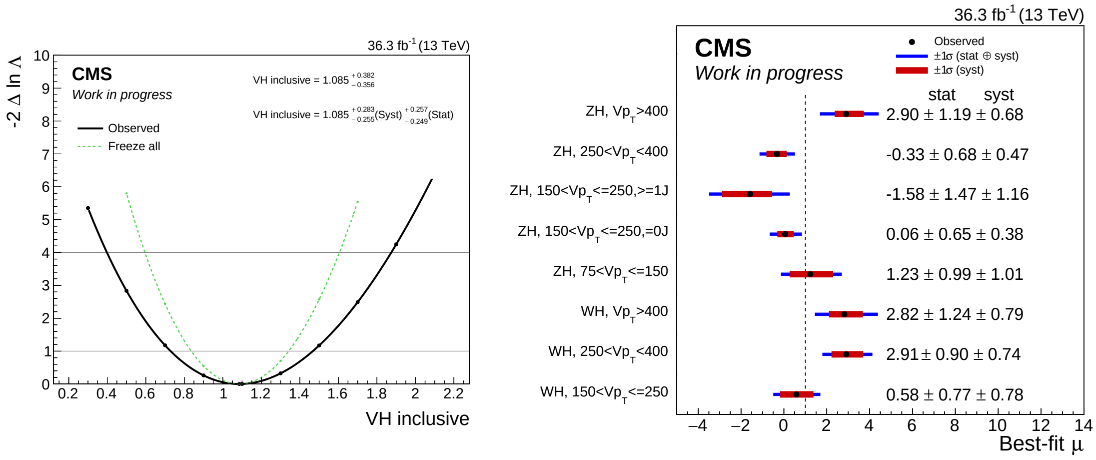


Figure 6.6: Observed results on 2016 VHbb analysis using inclusive and STXS-based signal extraction.

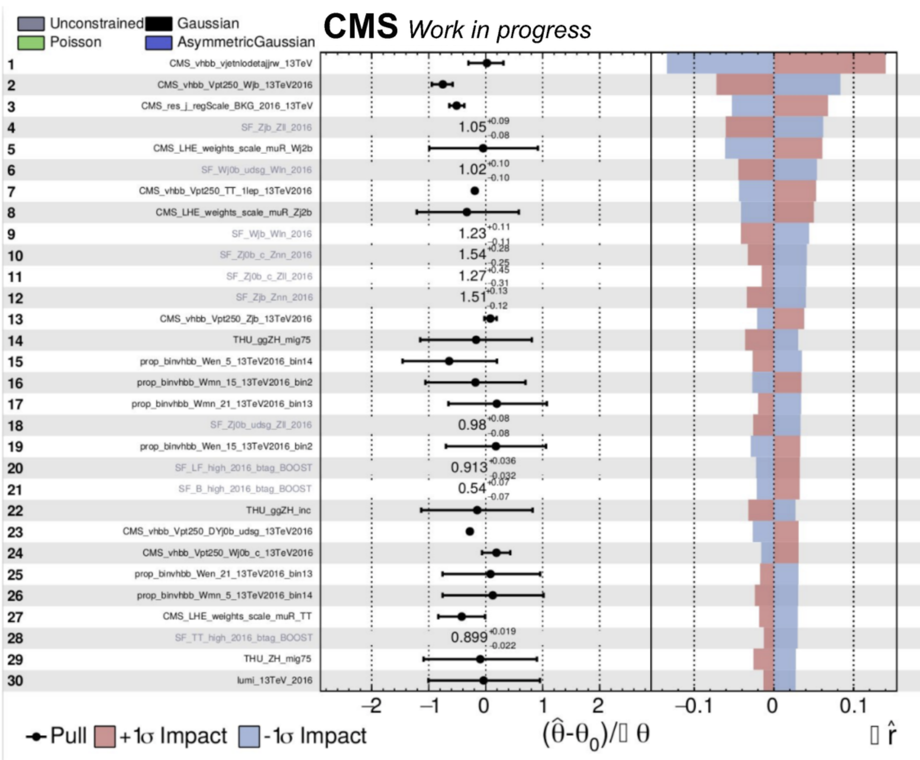


Figure 6.7: Impact on signal strenghts, pulls and constraints for the 2016 analysis post-fit nuisance parameters.

6.1.3 2017 data taking period results

The result for the 2017 data taking period is reported in Fig. 6.8 with impact of nuisance parameters in Fig. 6.9. This result for the 2017 data taking period is different with the inclusive result published with the same data in [86] by 2-2.5 σ.

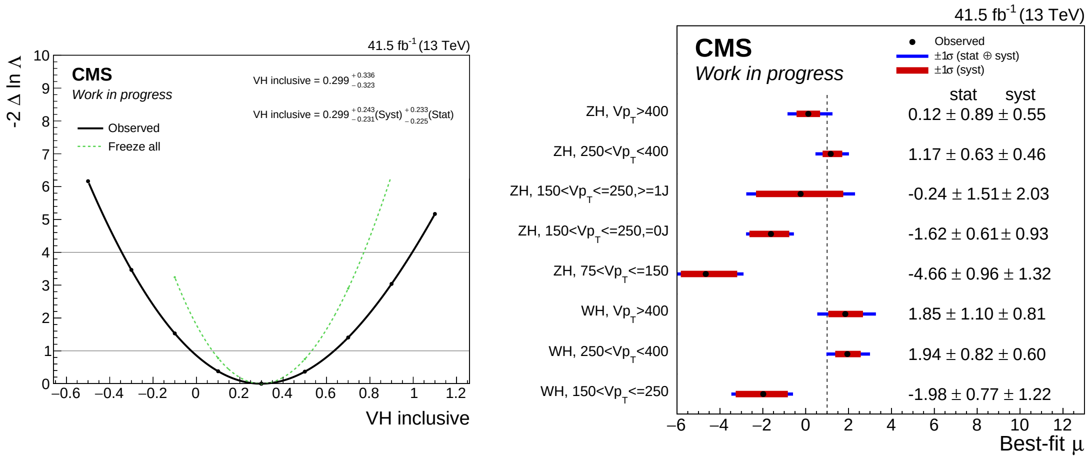


Figure 6.8: Observed results on 2017 VHbb analysis using inclusive and STXS-based signal extraction.

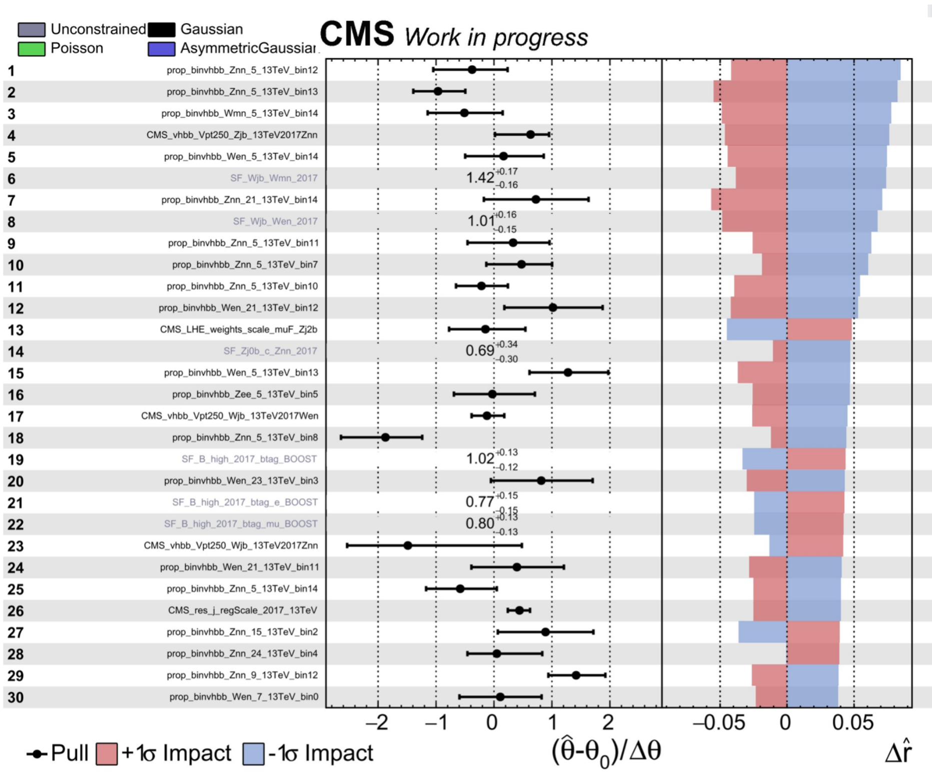


Figure 6.9: Impact on signal strenghts, pulls and constraints for the 2017 analysis post-fit nuisance parameters.

6.1.4 2018 data taking period results

The result for the 2018 data taking period is reported in Fig. 6.10 with impact of nuisance parameters in Fig. 6.11.

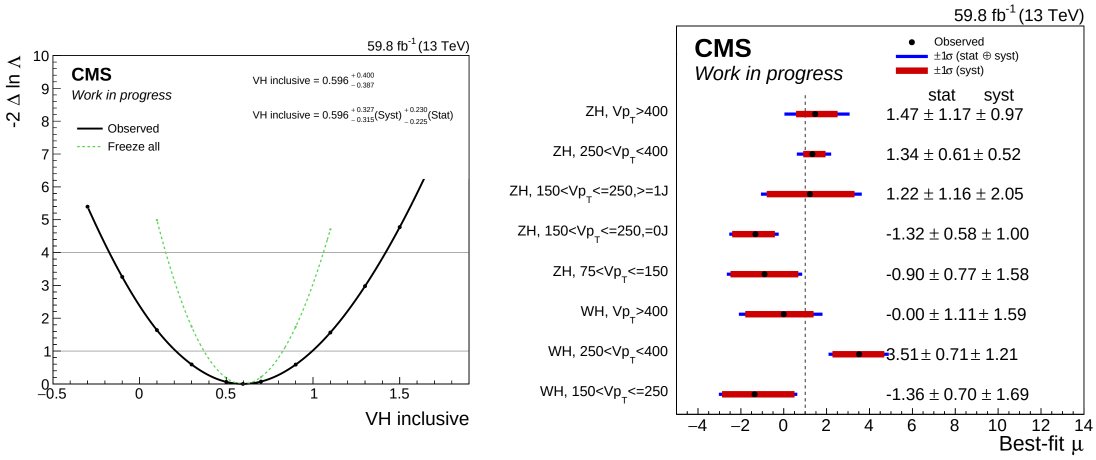


Figure 6.10: Observed results on 2018 VHbb analysis using inclusive and STXS-based signal extraction.

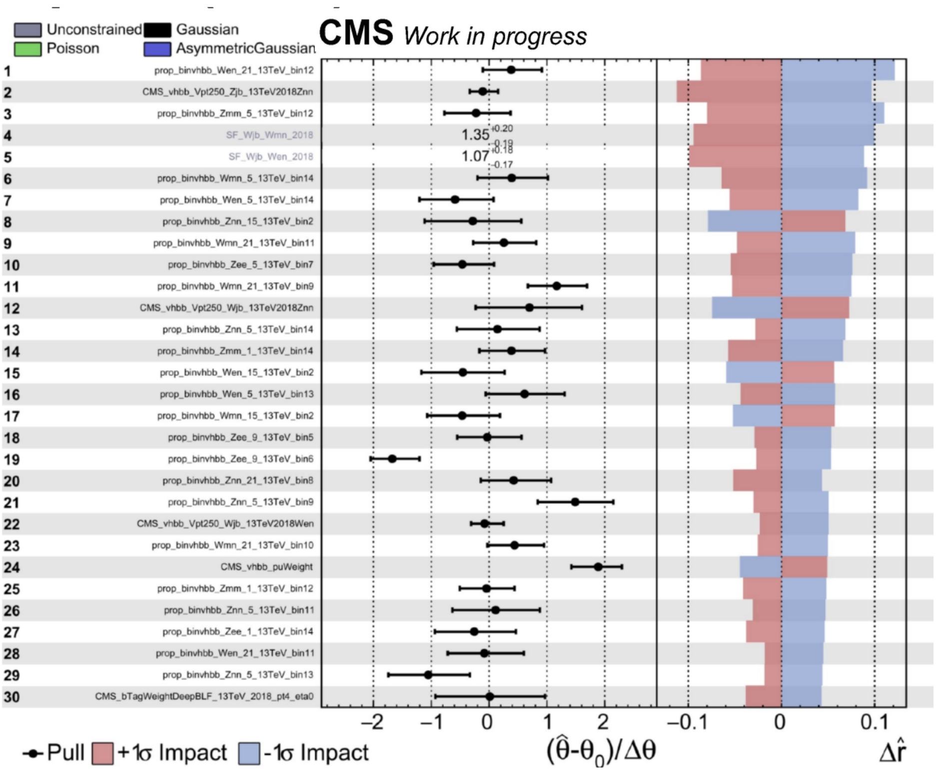


Figure 6.11: Impact on signal strenghts, pulls and constraints for the 2018 analysis post-fit nuisance parameters.

6.2 Mass cross-check analysis results

To extract the signal strength for the mass cross-check analysis the signal and control regions as defined in Section 5.2 are simultaneously fitted. The signal strengths for each channel is reported in Table 6.10 with the corresponding significance in Table 6.9. The signal regions are then combined and the resulting m_{jj} distributions for each year are shown in Figs. 6.12 to 6.14, where the distributions are weighted by the ratio of the number of events which stem from signal processes in each signal region divided by the sum of the number of events which stem from signal processes and the number of events which stem from background processes in the same region. In Figs. 6.12 to 6.14, the figure on the right is the distribution of m_{jj} while subtracting contribution of processes other than $Z \to bb$ and $H \to bb$.

The process scale-factors are also extracted from the fits and reported for each year in Tables 6.6 to 6.8. In comparison to the DNN-based analysis process scale factors reported in Tables 6.2 to 6.4, no lepton flavor separation and separate V+c and V+udsg process scale factors were assumed in the fit model. This choice for the process scale factors is to compensate for the reduced number of control regions in the cross-check analysis in comparison to the DNN-based analysis, which results in less constraining power over these processes.

The results are compatible with the expectations from the SM. In Fig. 6.12, few bins of the m_{jj} distributions for the 2016 data taking period show deviations, these bins are found not to be correlated with the signal strengths. The modeling for the V+jets process in the side bands of the m_{jj} is worse due to the removal of the V+HF control regions. This mismodelling is less present with 2017 and 2018 data taking eras that use the NLO V+jets samples.

The combination of all data taking periods is presented in Section 6.2.1.

Process	zero-lepton	one-lepton	two-lepton e	two-lepton μ
$t\bar{t}$	0.983±0.095	0.920±0.090	0.914±0.098	0.958±0.103
V+b,bb	1.325±0.085	1.196±0.111	1.103±0.072	1.199±0.076
V+udcsg	1.317±0.099	0.997±0.059	1.290±0.092	1.303±0.093

Table 6.6: Data-simulation scale factors for the MEDNN analysis in the zero-lepton, one-lepton, two-lepton channels from SR+CR m_{jj} fits for the 2016 data taking period. The quoted errors include both statistical and systematic uncertainties. The values are found to be compatible with the DNN-based 2016 analysis.

6.2.1 Combined result for the mass cross-check analysis

The dijet invariant mass distribution for the combination of all the three data taking eras, for the VH, $H \to b\bar{b}$ and the VZ, $Z \to b\bar{b}$ is shown in Fig. 6.15, with all background processes present on the left and background processes subtracted on the right. This distribution uses the DNN while removing any bias on the

Process	zero-lepton	one-lepton	two-lepton e	two-lepton μ
$t\bar{t}$	1.025±0.073	0.909±0.063	0.855±0.070	0.962±0.078
V+b,bb	1.016±0.090	1.904±0.290	1.006±0.118	1.044±0.088
V+udcsg	0.909±0.091	1.066±0.088	1.113±0.077	1.154±0.080

Table 6.7: Data-simulation scale factors for the MEDNN analysis in the zero-lepton, one-lepton, two-lepton channels from SR+CR m_{jj} fits for the 2017 data taking period. The quoted errors include both statistical and systematic uncertainties. The values are found to be compatible with the DNN-based 2017 analysis.

Process	zero-lepton	one-lepton	two-lepton e	two-lepton μ
$t\bar{t}$	0.982±0.070	0.930±0.064	0.961±0.077	0.935±0.075
V+b,bb	0.916±0.107	2.374±0.393	1.071±0.092	1.002±0.083
V+udcsg	0.989±0.079	0.948±0.051	1.091±0.074	0.978±0.067

Table 6.8: Data-simulation scale factors for the MEDNN analysis in the zero-lepton, one-lepton, two-lepton channels from SR+CR m_{jj} fits for the 2018 data taking period. The quoted errors include both statistical and systematic uncertainties. The values are found to be compatible with the DNN-based 2018 analysis.

	2016	2017	2018	combination
Expected Significance	1.73	1.32	1.33	2.57
Observed Significance	1.46	0.64	1.02	2.03

Table 6.9: Significance for the VH production cross-section times $H \to bb$ branching ratio.

	2016	2017	2018	combination
Expected	$1.000^{+0.581}_{-0.575}$	$1.000^{+0.749}_{-0.752}$	$1.000^{+0.741}_{-0.744}$	$1.000^{+0.384}_{-0.382}$
Observed	$0.856^{+0.593}_{-0.586}$	$0.690^{+0.809}_{-0.800}$	$0.778^{+0.745}_{-0.757}$	$0.964^{+0.381}_{-0.381}$

Table 6.10: Signal strength modifier for the VH production cross section times $H \to bb$ branching ratio.

dijet mass shape (MEDNN) as described in detail in Section 5.1. The events are weighted with S/(S+B) to emphasize the signal contribution in the distribution, where S corresponds to the signal and B to the background contributions. The data is consistent with the m_H=125 GeV Higgs boson and the signal strength is μ=0.964 $\pm$ 0.381 corresponding to a signal significance of 2.03 σ. The sensitivity is lower compared to the VH inclusive as the dijet mass is an important feature for the DNN and since this method removes the dependence of the DNN on

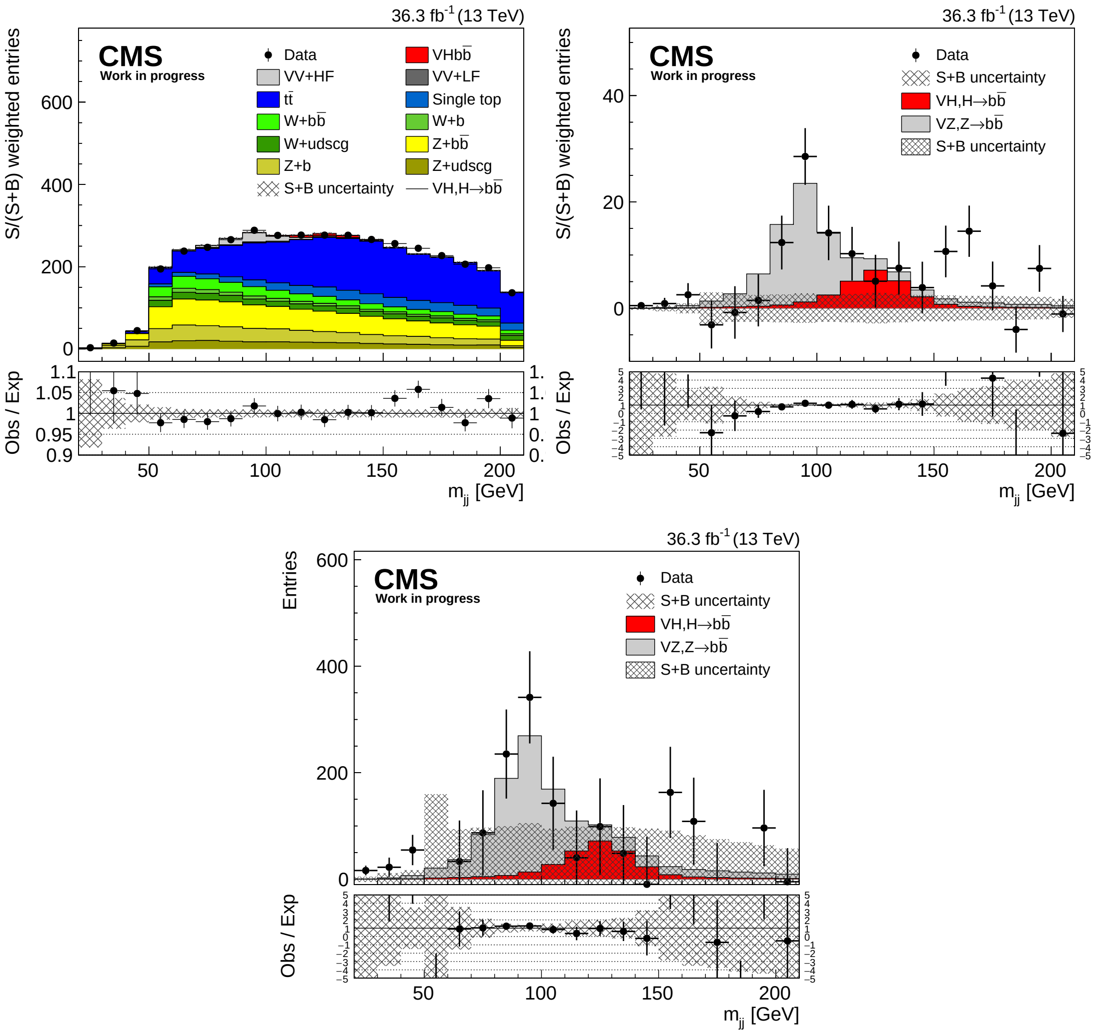


Figure 6.12: Invariant mass distributions for 2016 MEDNN analysis, without (left) and with background subtraction (right) and unweighted entries with background subtraction on the bottom.

the dijet mass the sensitivity of the DNN drops. The likelihood scan is shown in Fig. 6.16 and the impact of the nuisance parameters in Fig. 6.17.

m_{jj}	Relative uncertainty on μ
Background (theory)	14%
Signal (theory)	1%
MC stats.	30%
Sim. modelling	3%
b-tagging	9%
Jet energy resolution	4%
Luminosity	1%
Jet energy scale	3%
LeptonID	1%
Trigger(MET)	0.05%

Table 6.11: Impacts of different nuisance groups in terms of relative uncertainties on the full Run 2 MEDNN analysis signal strength.

m_{jj}	$\Delta\mu$
Background (theory)	+0.122 -0.121
Signal (theory)	+0.008 -0.002
MC stats.	+0.260 -0.261
Sim. modelling	+0.032 -0.027
b-tagging	+0.081 -0.080
Jet energy resolution	+0.037 -0.035
Luminosity	+0.015 -0.010
Jet energy scale	+0.021 -0.023
LeptonID	+0.007 -0.004
Trigger(MET)	+0.004 -0.005

Table 6.12: Impacts of different nuisance groups in terms of relative uncertainties on the full Run 2 MEDNN analysis signal strength.

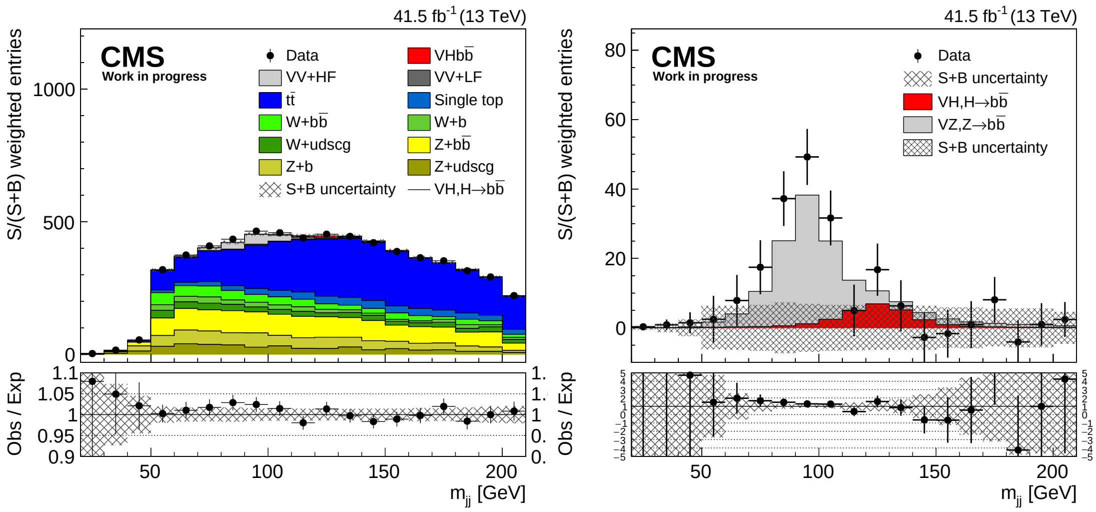


Figure 6.13: Invariant mass distributions for 2017 MEDNN analysis, without (left) and with background subtraction (right).

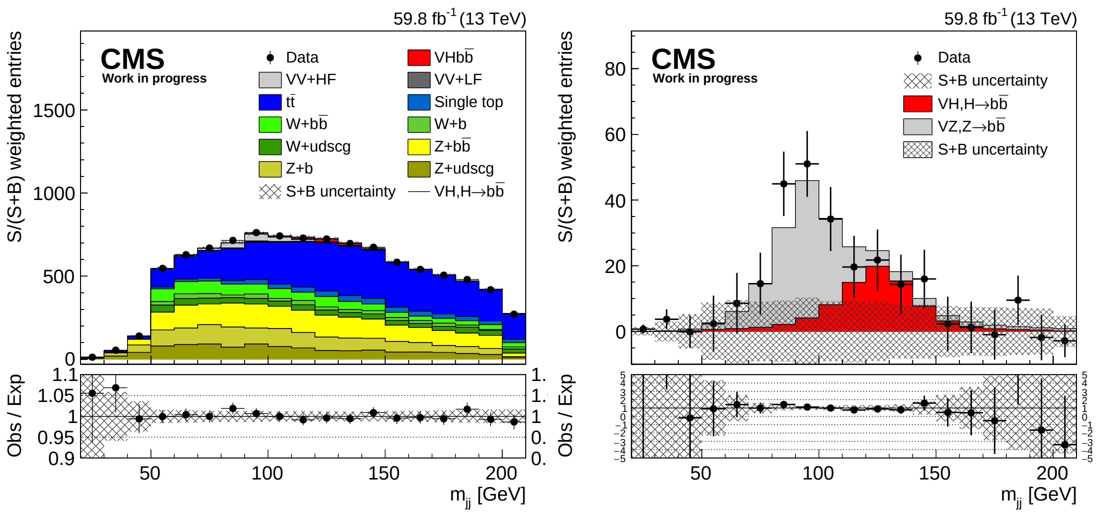


Figure 6.14: Invariant mass distributions for 2018 MEDNN analysis, without (left) and with background subtraction (right).

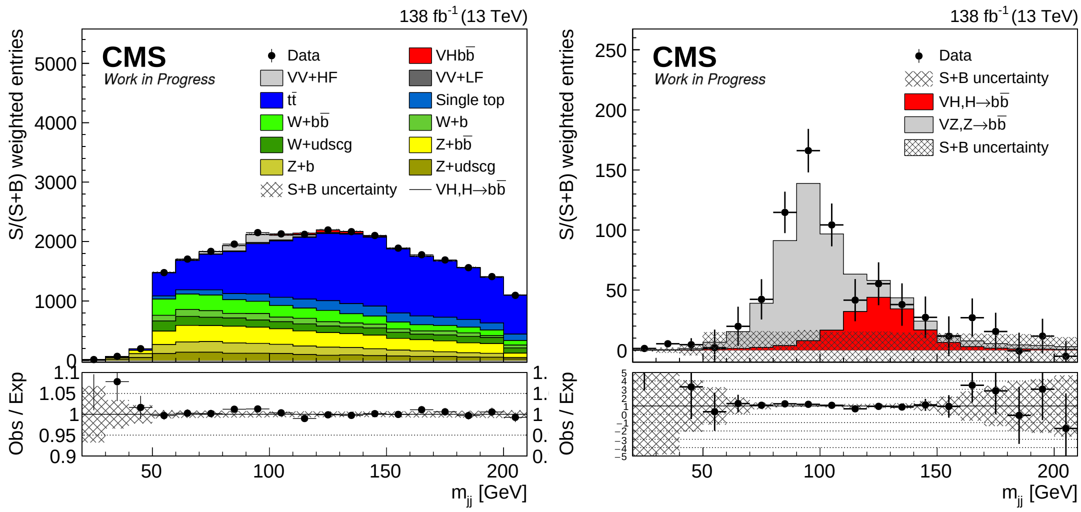


Figure 6.15: Invariant mass distributions for 2016+2017+2018 MEDNN analysis, without (left) and with background subtraction (right).

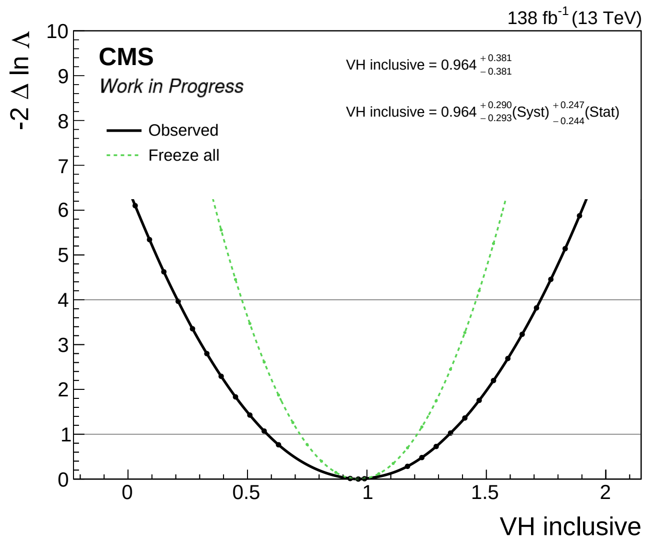


Figure 6.16: Signal strength scan for 2016-2018 data taking periods for the dijet mass cross check analysis.

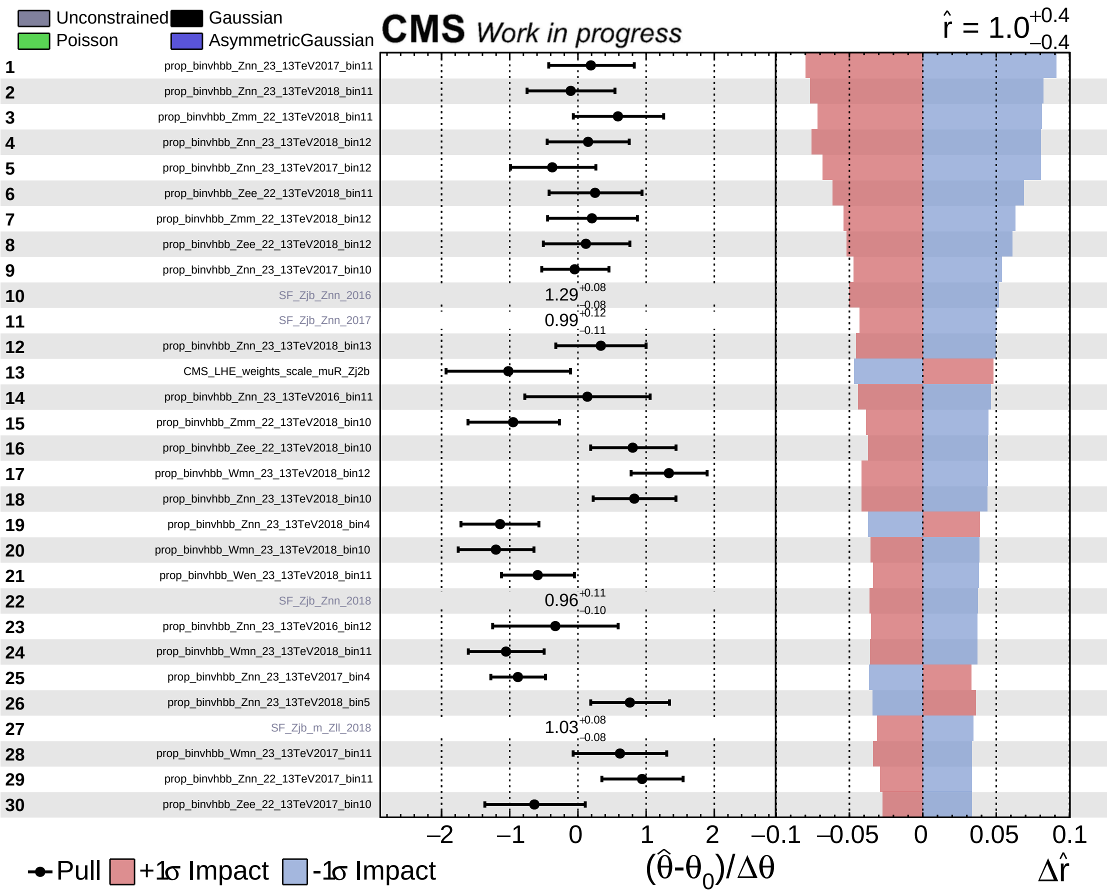


Figure 6.17: Main nuisance parameter impacts for 2016-2018 data taking periods, for the dijet mass cross check analysis.

Chapter 7

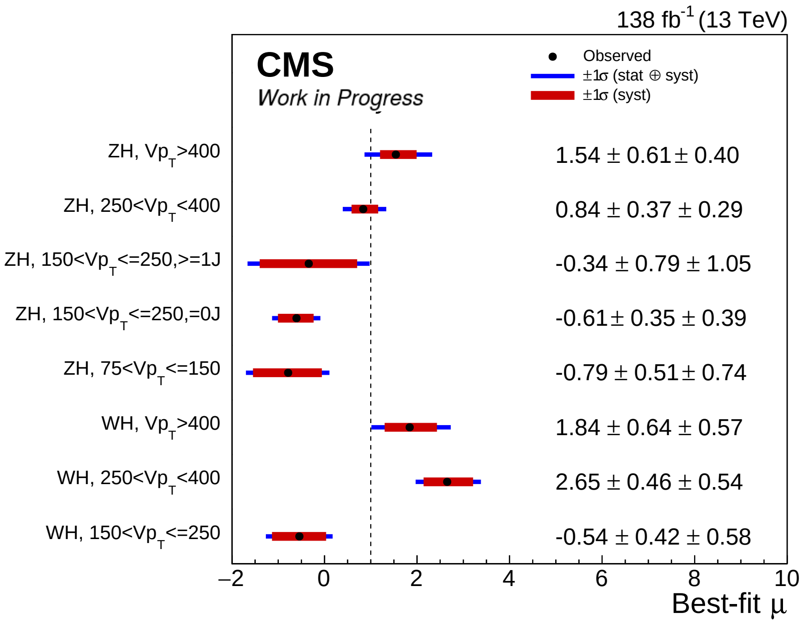
Summary

The full Run 2 (corresponding to a luminosity of 138 fb^{-1} at $\sqrt{s} = 13$TeV) analysis of the VH($b\bar{b}$) with the STXS categorization is reported, where the Higgs boson is produced in association with a vector boson and decays to a pair of b-quarks. The signal strength modifier for the inclusive analysis and combination of all the data taking periods is found to be $\mu = 0.71 \pm 0.22$. The analysis is performed in three channels with leptonic decays of the vector boson, zero-lepton ($ZH \to \nu\bar{\nu}b\bar{b}$, $\ell = e, \mu$), one-lepton ($WH \to l\nu b\bar{b}$, $\ell = e, \mu$) and two-leptons ($ZH \to llb\bar{b}$, $\ell = e, \mu$). For the Higgs decays to two b-quarks for each channel, two modes of resolved and boosted topologies for the b-quark jets are considered. The signal strength modifiers for the STXS categories and the instructive mass plot from the mass based cross-check analysis are shown in Fig. 7.1. The signal strength modifier for the mass based cross-check analysis is found to be μ=0.964 $\pm$ 0.381.

The main sources of uncertainty in this analysis are the MC statistics and the modeling of the V+jets background.

This analysis is can be used as a basis for an EFT interpretation [67] analysis or a differential cross section analysis for the VH, H$\to$b$\bar{\text{b}}$ process.

Additional improvements may be achieved by employing an additional multivariate method to explicitly remove the $t\bar{t}$ and single-top backgrounds contributions from the signal region and by introducing means to understand the effect of decorrelating the theory uncertainties such as two-point (fragmentation modeling) and continuous (higher-order corrections) uncertainties as mentioned in [57].



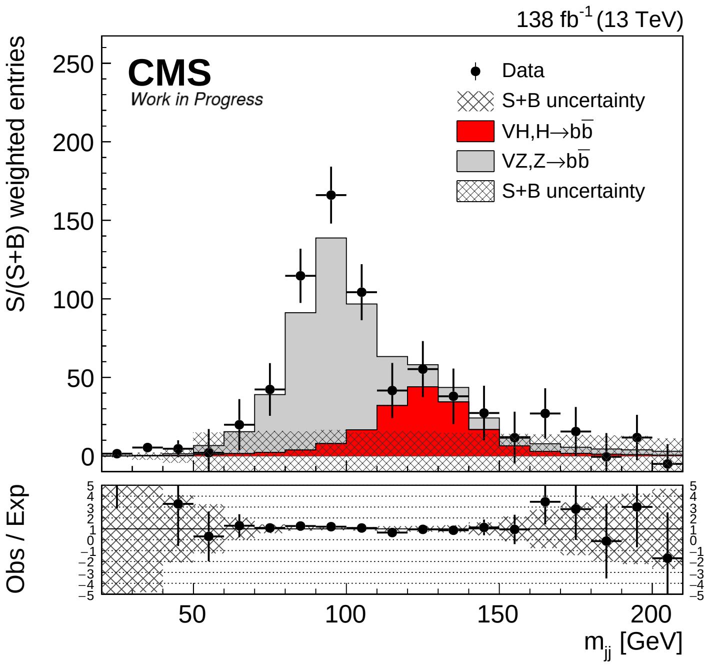


Figure 7.1: Full Run 2 combination for the STXS categories (top). The dijet invariant mass distribution for the combination of all the three data taking eras, for the VH, H→b$\bar{b}$ and the VZ, Z→b$\bar{b}$ with background subtraction and S/(S+B)-weighted (bottom).

Appendices

Appendix A

A Simple Bayesian model

To achieve a Bayesian inference, one can update the Eq.4.3 by multiplying the numerator and denominator with a prior $\pi(\mu)$ on the μ, but here we are going to take a more general approach using a linear line approximation based on [69].

Using a linear generative model (which is a quantitative parameterization of the statistical procedure capable of reasonably generating the data set under study). We assume $y = m\,f(signal)\,x + b$, where m (similar to slope of linear line) is the signal strength for a process and b is the intercept. Furthermore, we assume that deviations from this form can be generated by a Gaussian distribution centered around each point y with variance σ_y^2 (in high energy physics a Poisson distribution is more common for y). Given each bin center x_i and yields as y_i the frequency distribution is as follows:

$$p(y_i|x_i, \sigma_{yi}, m, b) = \frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,f(sig, x_i)\,x_i - m_{bkg}\,f(bkg, x_i)\,x_i - b]^2}{2\,\sigma_{yi}^2}\right), \tag{A.1}$$

where $f(signal, x_i)$ is the yield from the Monte Carlo for the bin x_i.

To incorporate any unmodeled but rare sources of noise and the possibility to reject outlier data points, we add handles on the information we get from the result of the fit using the approach from the previous section. We add to the generative model a set of N binary integers q_i, one per data point, each of which is unity if the ith data point is good, and zero if the ith data point is outlier. In addition, P_b (the probability that any individual data point is outlier), and parameters $(Y_\mathrm{b}, V_\mathrm{b})$ the mean and variance of the distribution of outlier points (in y). Considering $\boldsymbol{\theta} \equiv (m, b, \{q_i\}_{i=1}^N, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b})$, the posterior probability is

$$p(\boldsymbol{\theta}, I) = \frac{p(\{y_i\}_{i=1}^N|\boldsymbol{\theta}, I)}{p(\{y_i\}_{i=1}^N|I)}\,p(\boldsymbol{\theta}|I) \quad , \tag{A.2}$$

which we need to marginalize to get the desired posterior based on (m, b),

$$p(m, b|\{y_i\}_{i=1}^N, I) = \int \mathrm{d}\{q_i\}_{i=1}^N\,\mathrm{d}P_\mathrm{b}\,\mathrm{d}Y_\mathrm{b}\,\mathrm{d}V_\mathrm{b}\,p(\boldsymbol{\theta}, I) \quad , \tag{A.3}$$

Finally, we can write the likelihood

$$\begin{aligned}
\mathscr{L} &\equiv p(\{y_i\}_{i=1}^{N}|m, b, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b}, I) \\
\mathscr{L} &\equiv \prod_{i=1}^{N}\Big[(1-P_\mathrm{b})\, p(\{y_i\}_{i=1}^{N}|m, b, I)) + P_\mathrm{b}\, p(\{y_i\}_{i=1}^{N}|Y_\mathrm{b}, V_\mathrm{b}, I)\Big] \\
\mathscr{L} &\propto \prod_{i=1}^{N}\left[\frac{1-P_\mathrm{b}}{\sqrt{2\,\pi\,\sigma_{yi}^2}}\exp\left(-\frac{[y_i - m\,f(sig, x_i)\,x_i - m_{bkg}\,f(bkg, x_i)\,x_i - b]^2}{2\,\sigma_{yi}^2}\right)\right. \\
&\qquad \left. + \frac{P_\mathrm{b}}{\sqrt{2\,\pi\,[V_\mathrm{b}+\sigma_{yi}^2]}}\exp\left(-\frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b}+\sigma_{yi}^2]}\right)\right] \quad . \qquad \text{(A.4)}
\end{aligned}$$

Marginalization with Metropolis-Hastings method

The marginalization step can be performed in multiple ways, direct integration or using Monte Carlo methods.

Metropolis-Hastings method [66] is the choice as it provides the integral and gives a Monte-Carlo Markov chain sampling for free. Steps for this method can be summarized as *(a)* Randomly sample the parameter space. *(b)* Compute posterior using new points in parameter space. *(c)* Draw number R randomly while $0 < R < 1$. *(d)* Look at new over old posterior probability ratio, if R is less than the ratio, accept the step, append the new parameters to the chain; if greater, reject the step, re-append the old parameters instead.

A.1 Application to $ZH \to b\bar{b}$

As an example, the method is applied to the associated production of the Higgs in 2-lepton channel. Using Gaussian priors for each of the input parameters, we can gain distribution for signal strength instead of a single fit-value represented in Fig. A.2. The power of incorporating a technique like the one shown here is that one can preserve all the information about the uncertainties in the results.

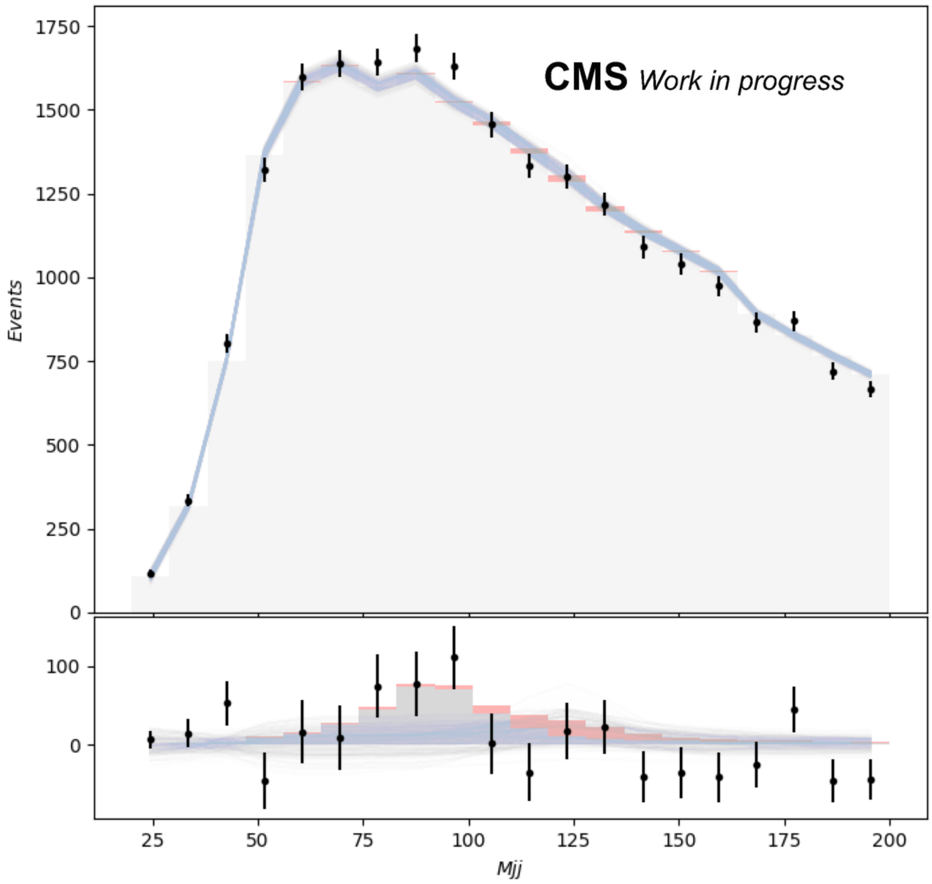


Figure A.1: This plot shows 1000 of the accepted samples in gray from the random walk sampling.

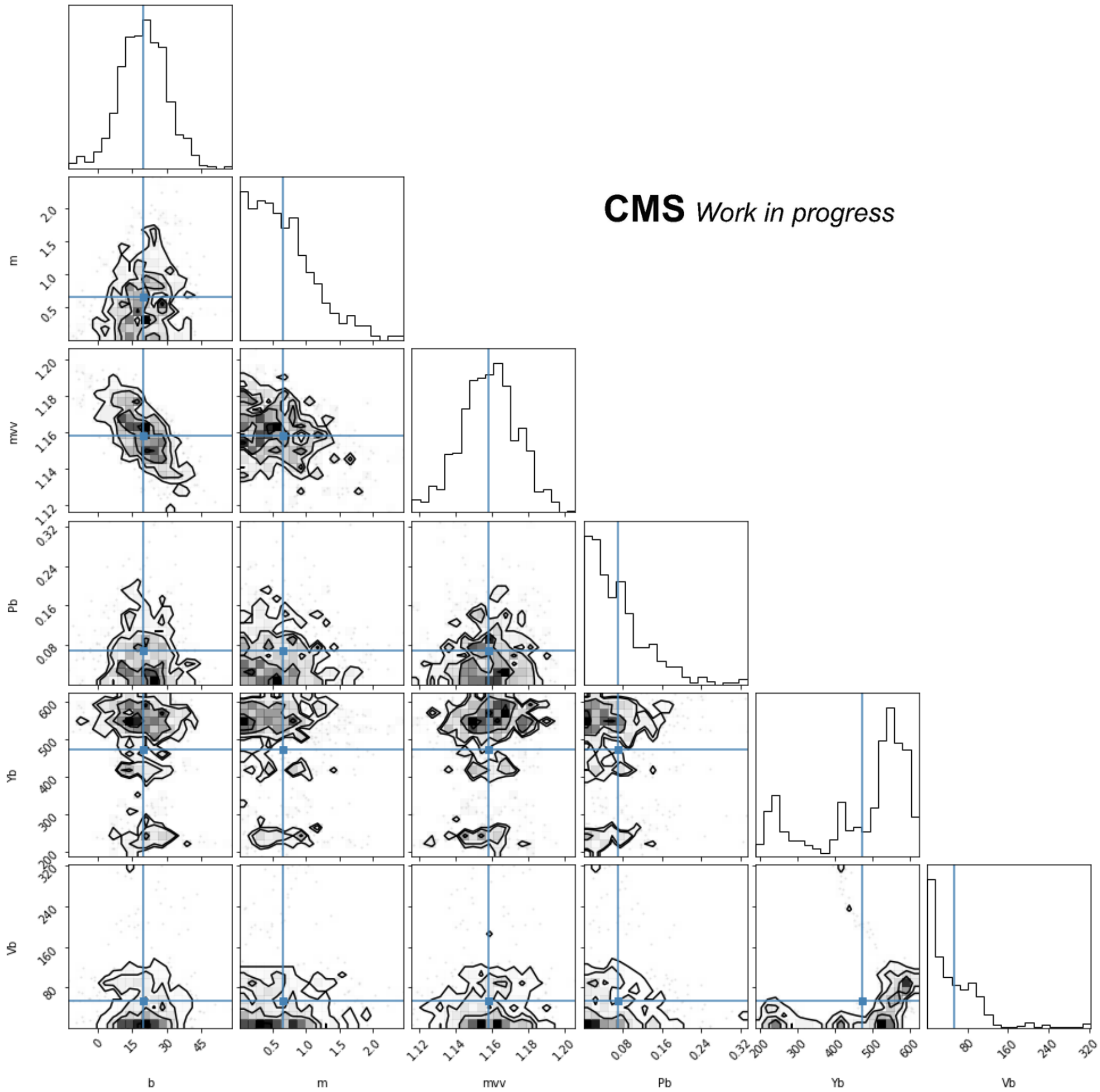


Figure A.2: The corner plot showing the distribution for the different free parameters of the fit along with the correlations between them.

Appendix B

M_{jj} results with leading order V+jets samples

The dijet invariant mass for VH, H→b$\bar{\text{b}}$ and the VZ, Z→b$\bar{\text{b}}$ is shown in Fig. B.1, with all background processes present on the left and background processes subtracted on the right. This distribution uses the DNN while removing any bias on the dijet mass shape as described in detail in Section 5.1. The events are weighted with S/(S+B) to emphasize the signal contribution in the distribution, where corresponds to the signal and B to the background contributions. The Data is consistent with the m_H=125 GeV Higgs boson and the signal strength is μ=1.04 $\pm$ 0.240 (stat) $\pm$0.163 (sys) corresponding to a signal significance of 3.63σ. The sensitivity is lower compared to the VH inclusive as the dijet mass is an important feature for the DNN and since this method removes the dependence of the DNN to the dijet mass the sensitivity of the DNN drops. The likelihood scan is shown in Fig. B.2 and the impact of the nuisance parameters in Fig. B.3.

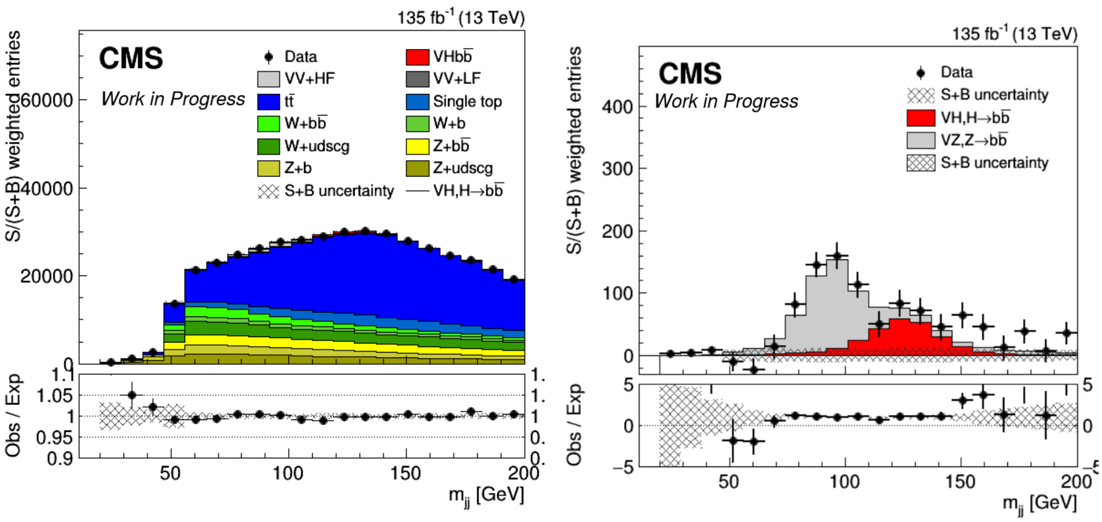


Figure B.1: Invariant mass distributions for 2016+2017+2018 MEDNN analysis, without (left) and with background subtraction (right).

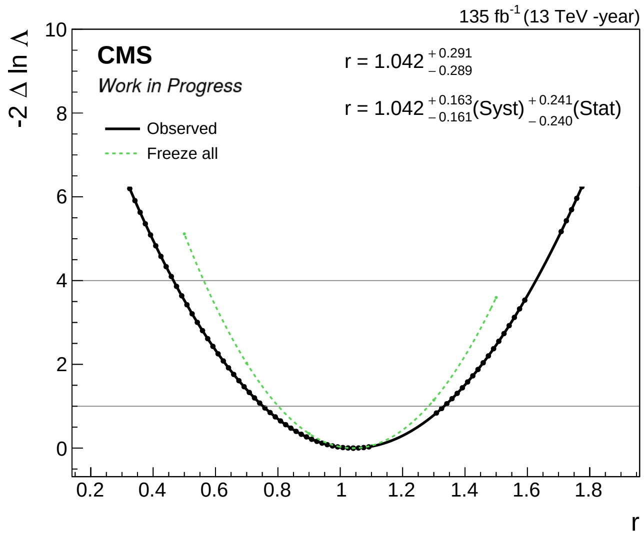


Figure B.2: Signal strength scan for 2016-2018 data taking periods for the dijet mass cross check analysis.

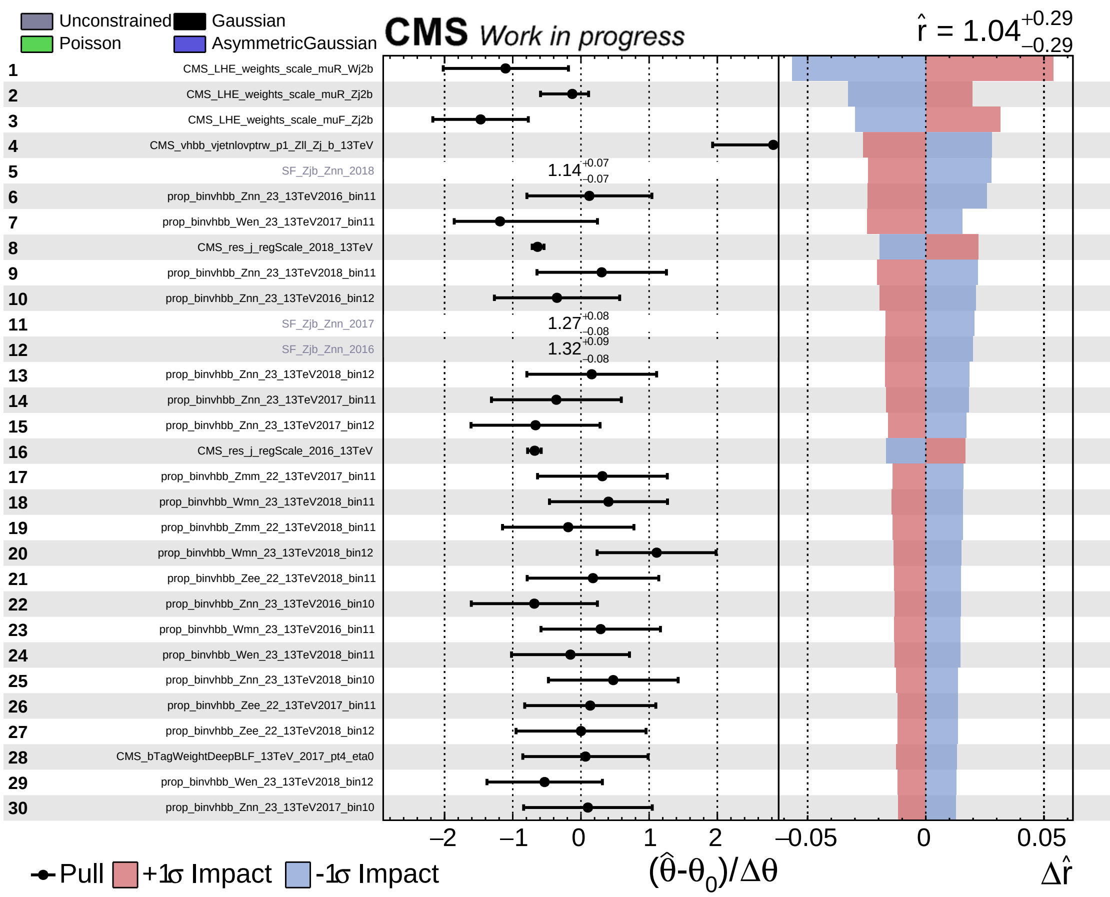


Figure B.3: Main nuisance parameter impacts for 2016-2018 data taking periods, for the dijet mass cross check analysis.

Appendix C

Additional control plots

0-lepton SR and CR plots

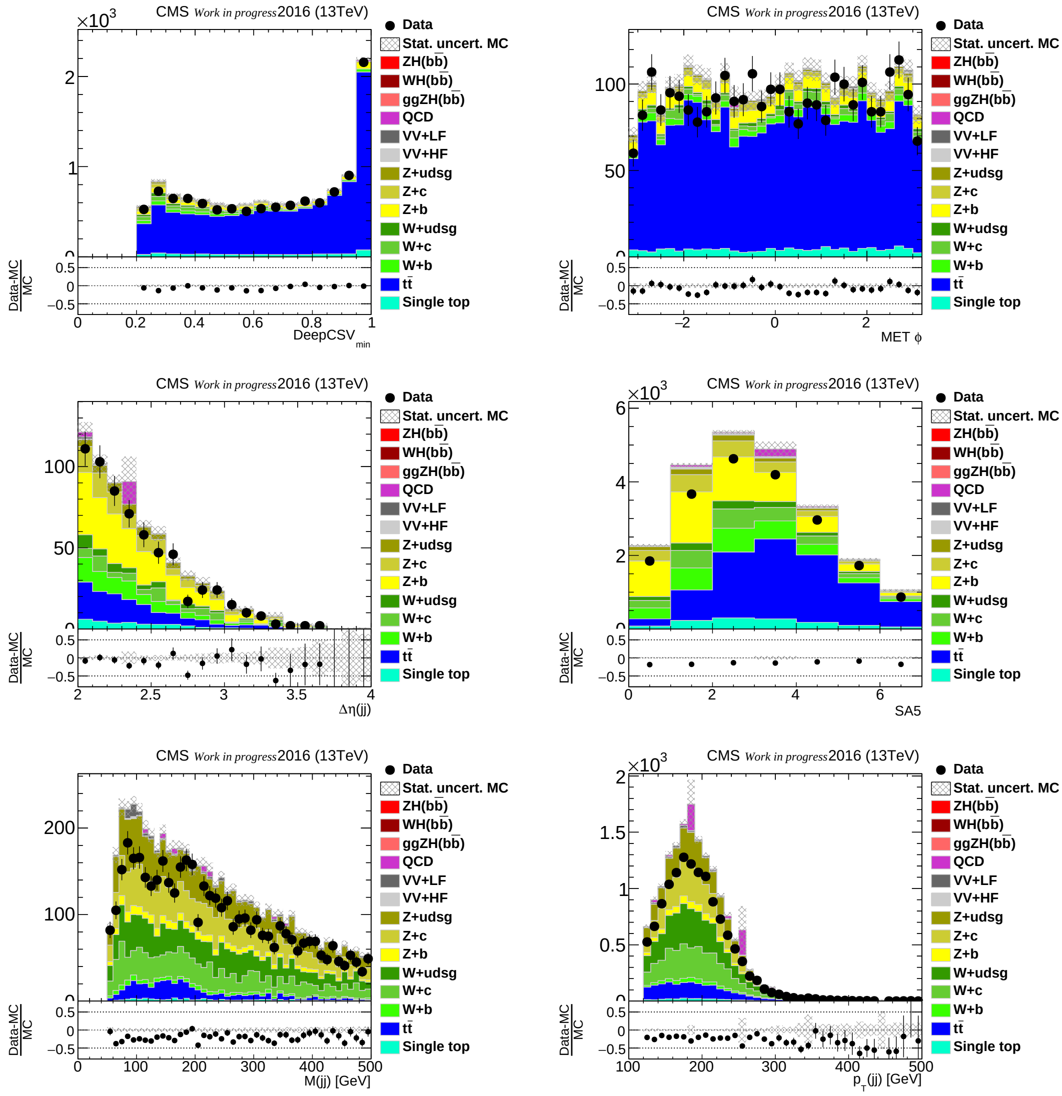


Figure C.1: From the top to bottom for the 0-lepton channel control regions and 2016 data taking era, distribution of the leading jet DeepCSV b-tagging score for the inclusive $t\bar{t}$ region (left), the ϕ for the MET in high $p_{\mathrm{T}}(\mathrm{V})$ $t\bar{t}$ region (right); The $\Delta\eta(jj)$ in high $p_{\mathrm{T}}(\mathrm{V})$ Z+HF region (left) and soft activity for medium $p_{\mathrm{T}}(\mathrm{V})$ Z+HF region (right); M(jj) for high $p_{\mathrm{T}}(\mathrm{V})$ Z+LF region (left) and $P_T(jj)$ for medium $p_{\mathrm{T}}(\mathrm{V})$ Z+LF region (left).

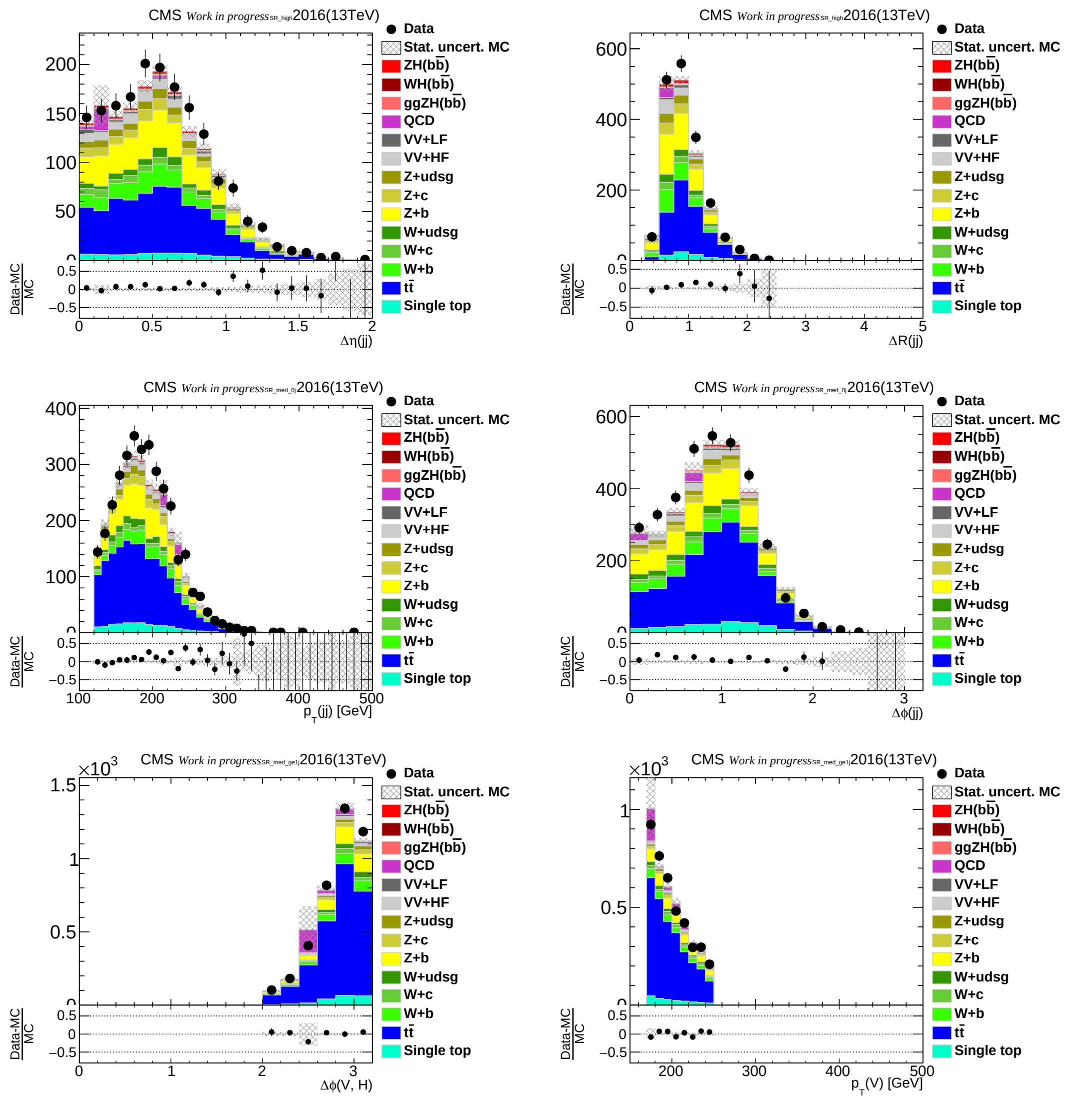


Figure C.2: From the top to bottom for the 0-lepton channel signal region and 2016 data taking era, distribution of the $\Delta\eta(jj)$ in the high $p_{\mathrm{T}}(\mathrm{V})$ STXS region (left), the $\Delta R(jj)$ in the high $p_{\mathrm{T}}(\mathrm{V})$ STXS region (right); The $P_T(jj)$ in the medium $p_{\mathrm{T}}(\mathrm{V})$ with zero additional jets STXS region (left), the $\Delta\phi(jj)$ in the medium $p_{\mathrm{T}}(\mathrm{V})$ with zero additional jets STXS region (right); The $\Delta\phi(V,H)$ in the medium $p_{\mathrm{T}}(\mathrm{V})$ with one additional jets STXS region (left), the $p_{\mathrm{T}}(\mathrm{V})$ in the medium $p_{\mathrm{T}}(\mathrm{V})$ with one additional jets STXS region (right).

1-lepton SR and CR plots

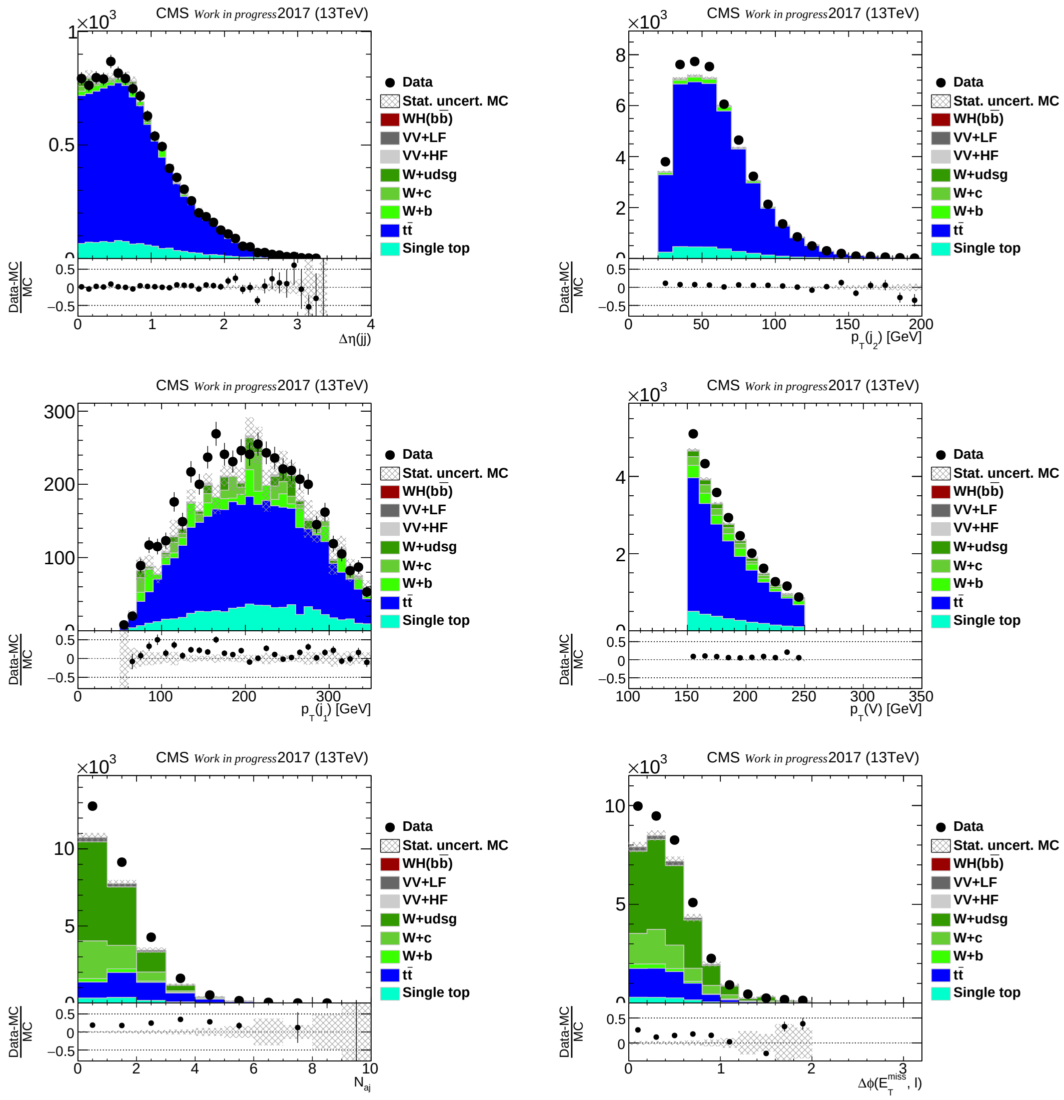


Figure C.3: From the top to bottom for the 1-lepton channel control regions and 2017 data taking era, distribution of the $\Delta\eta(jj)$ in high $p_{\mathrm{T}}(\mathrm{V})$ $t\bar{t}$ region (left), the P_T of the sub-leading jet in medium $p_{\mathrm{T}}(\mathrm{V})$ $t\bar{t}$ region (right); The P_T of the leading jet in high $p_{\mathrm{T}}(\mathrm{V})$ W+HF region (left) and the $p_{\mathrm{T}}(\mathrm{V})$ in medium $p_{\mathrm{T}}(\mathrm{V})$ W+HF region (right);The number of additional jets in high $p_{\mathrm{T}}(\mathrm{V})$ W+LF region (left) and $\Delta\phi(MET, l)$ in high $p_{\mathrm{T}}(\mathrm{V})$ W+LF region (left).

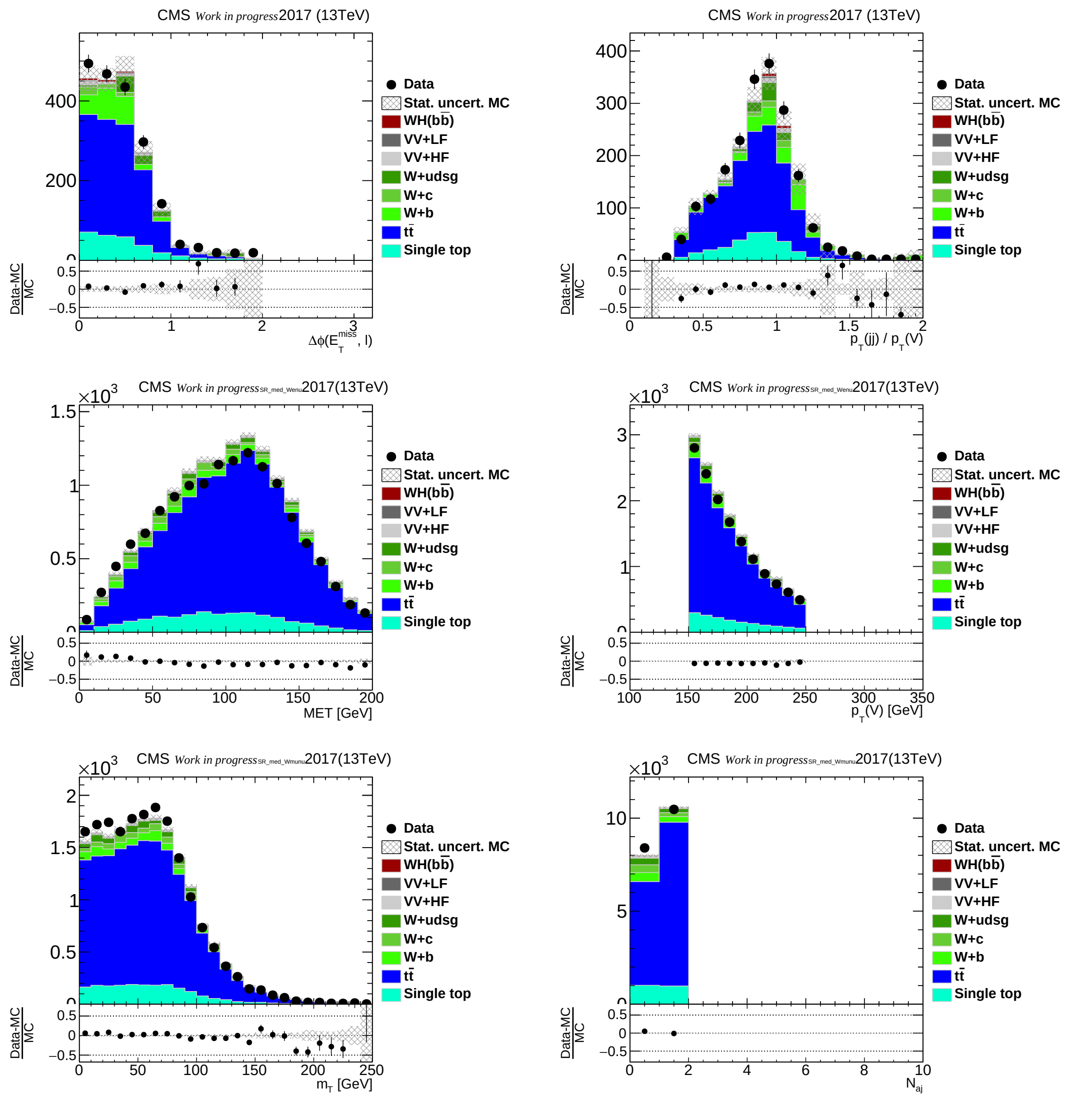


Figure C.4: From the top to bottom for the 1-lepton channel signal regions and 2017 data taking era, distribution of the $\Delta\phi(MET, l)$ in high $p_{\mathrm{T}}(\mathrm{V})$ STXS region (left), the ratio of the $p_T(jj)$ over $p_{\mathrm{T}}(\mathrm{V})$ in high $p_{\mathrm{T}}(\mathrm{V})$ STXS region (right); The MET medium $p_{\mathrm{T}}(\mathrm{V})$ STXS region (left) and the $p_{\mathrm{T}}(\mathrm{V})$ in medium $p_{\mathrm{T}}(\mathrm{V})$ STXS region (right);The transverse mass of the Vector boson in medium $p_{\mathrm{T}}(\mathrm{V})$ STXS region (left) and the number of additional jets in medium $p_{\mathrm{T}}(\mathrm{V})$ STXS region (right).

2-lepton SR and CR plots

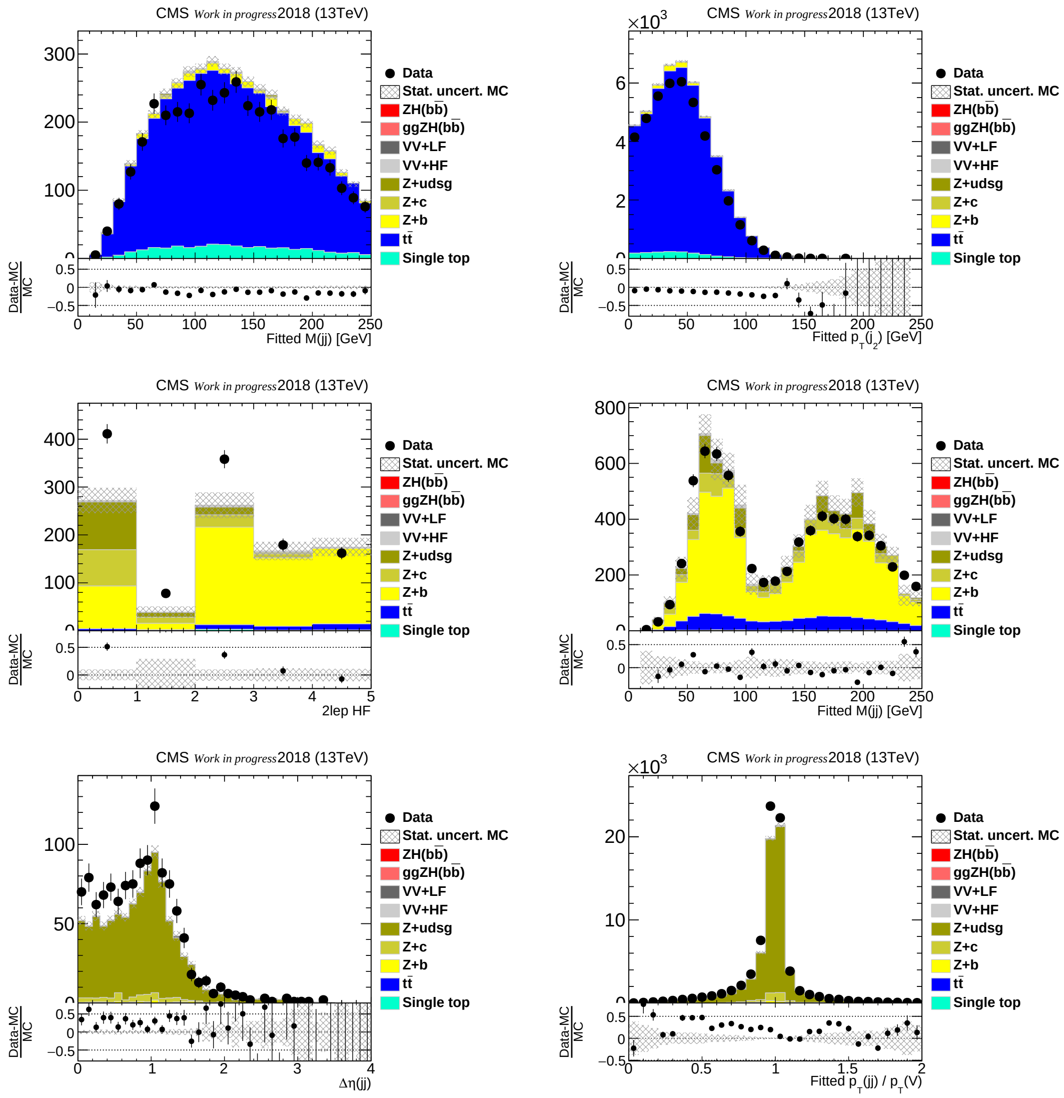


Figure C.5: From the top to bottom for the 2-lepton channel control regions and 2018 data taking era, distribution of the $M(jj)$ in medium $p_{\mathrm{T}}(\mathrm{V})$ $t\bar{t}$ $(\mu\mu)$ region (left), the P_T of the sub-leading jet in low $p_{\mathrm{T}}(\mathrm{V})$ $t\bar{t}$ $(\mu\mu)$ region (right); The distribution of the b-tagging working points in medium $p_{\mathrm{T}}(\mathrm{V})$ Z(ee)+HF region (left) and the $M(jj)$ in low $p_{\mathrm{T}}(\mathrm{V})$ Z($\mu\mu$)+HF region (right);The $\Delta\eta(jj)$ in high $p_{\mathrm{T}}(\mathrm{V})$ Z(ee)+LF region (left) and the $p_T(jj)$ over $p_{\mathrm{T}}(\mathrm{V})$ in low $p_{\mathrm{T}}(\mathrm{V})$ Z($\mu\mu$)+LF region (left).

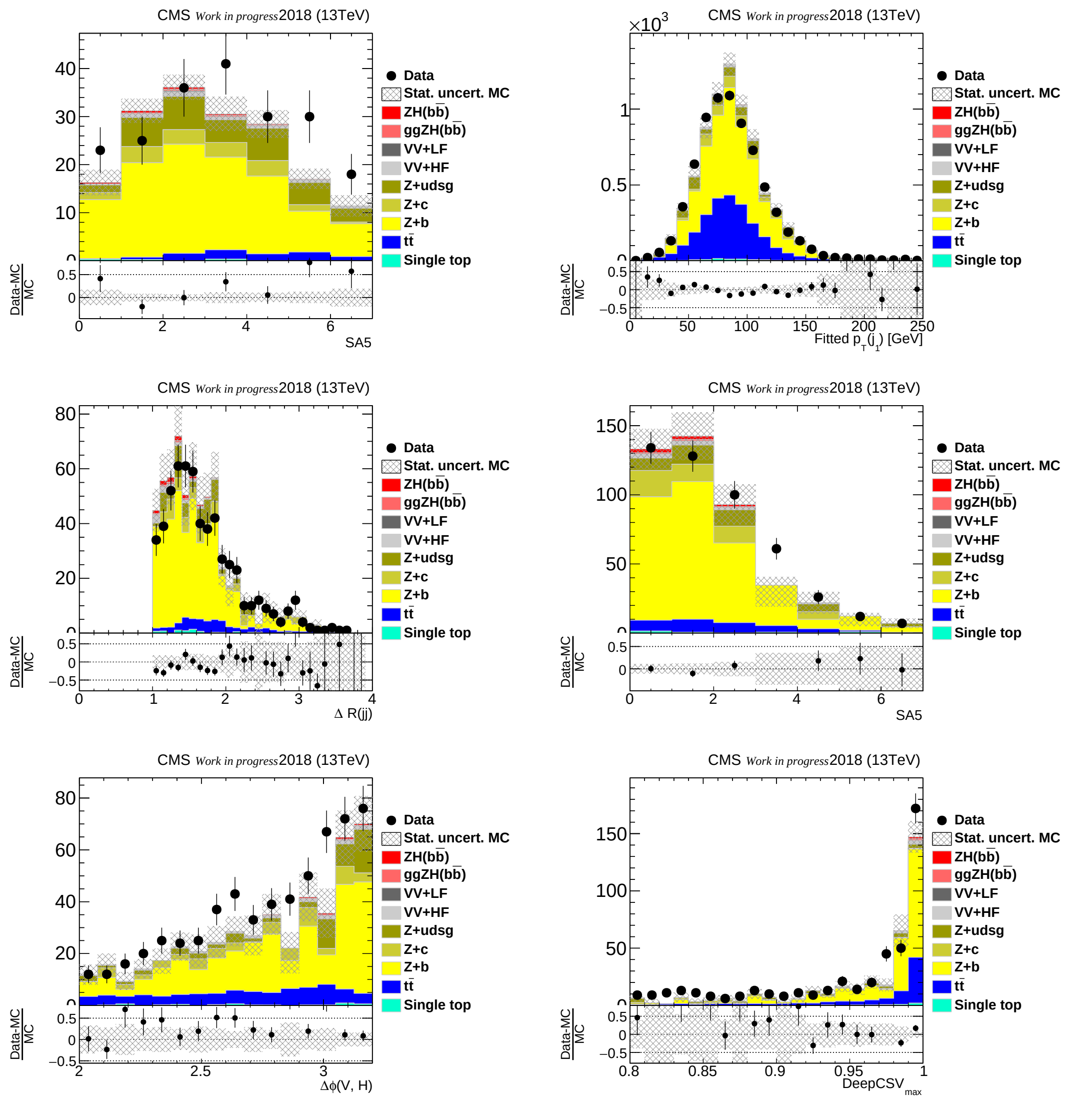


Figure C.6: From the top to bottom for the 2-lepton channel signal regions and 2018 data taking era, distribution of the soft activity in high $p_{\mathrm{T}}(\mathrm{V})$ $(\mu\mu)$ STXS region (left), the p_T of the leading jet in low $p_{\mathrm{T}}(\mathrm{V})$ (ee) STXS region (right); The $\Delta\phi(jj)$ in medium $p_{\mathrm{T}}(\mathrm{V})$ with zero additional jets $(\mu\mu)$ STXS region (left) and the soft activity in medium $p_{\mathrm{T}}(\mathrm{V})$ with zero additional jets (ee) STXS region (right);The $\Delta R(V,H)$ in medium $p_{\mathrm{T}}(\mathrm{V})$ with one additional jets $(\mu\mu)$ STXS region (left) and the DeepCSV distribution for the leading jet in medium $p_{\mathrm{T}}(\mathrm{V})$ with one additional jets (ee) STXS region (right).

Boosted SR and CR plots

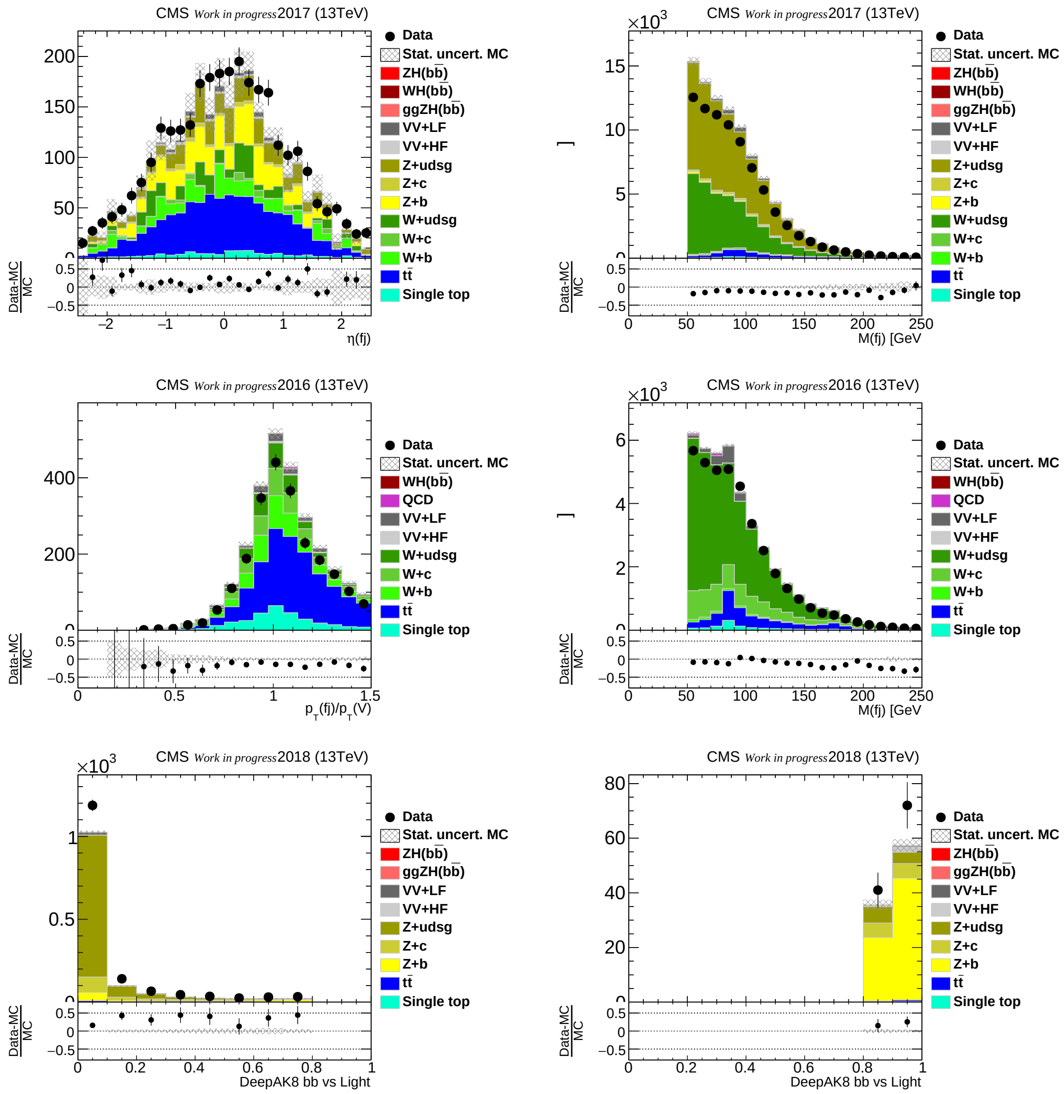


Figure C.7: From the top to bottom for the control regions of the boosted analysis, distribution of the η of the fat jet in 0-lepton boosted Z+HF region (left), the mass of the fat jet in 0-lepton boosted Z+LF region (right); The p_{T} of the fat jet over the $p_{\mathrm{T}}(\mathrm{V})$ in 1-lepton boosted W(μ)+HF region (left), the soft drop mass of the fat jet in 1-lepton boosted W(e)+LF region (right); The DeepAK8 output node distribution for the Z($\mu\mu$)+LF (left) and for the Z(ee)+HF (right).

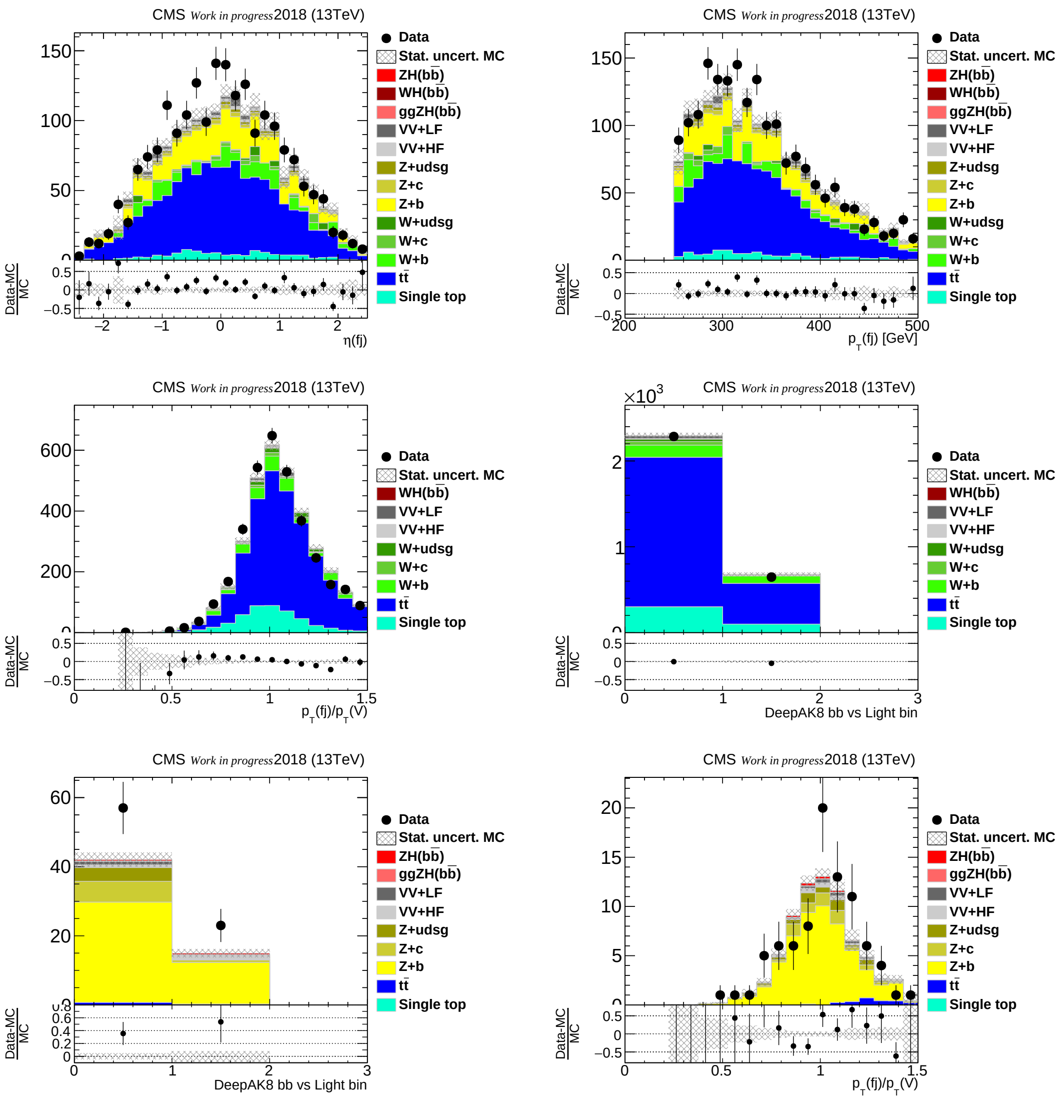


Figure C.8: From the top to bottom for the boosted signal regions for 2018 data taking era, distribution of the η of the fat jet in 0-lepton signal region (left), and the p_T of the fat jet in 0-lepton signal region (right); The p_T of the fat jet over the $p_T(V)$ in 1-muon signal region (left) and the DeepAK8 output node in 1-electron signal region (right); The DeepAK8 output node distribution for the Z(ee) signal region (left) and for the $Z(\mu\mu)$ signal region (right).

Bibliography

[1] M. Aaboud et al. "Observation of H→bb¯ decays and VH production with the ATLAS detector". In: *Physics Letters B* 786 (Nov. 2018), pp. 59–86. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.09.013. URL: http://dx.doi.org/10.1016/j.physletb.2018.09.013.

[2] M. Aaboud et al. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". In: *Physics Letters B* 784 (Sept. 2018), pp. 173–191. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.07.035. URL: http://dx.doi.org/10.1016/j.physletb.2018.07.035.

[3] G. Aad et al. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7 and 8 TeV$". In: *Journal of High Energy Physics* 2016.8 (Aug. 2016). ISSN: 1029-8479. DOI: 10.1007/jhep08(2016)045. URL: http://dx.doi.org/10.1007/JHEP08(2016)045.

[4] G. Aad et al. "Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7 and 8 TeV$ in the ATLAS experiment". In: *The European Physical Journal C* 76.1 (Jan. 2016). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-015-3769-y. URL: http://dx.doi.org/10.1140/epjc/s10052-015-3769-y.

[5] Wolfgang Adam et al. "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC". In: (Jan. 2005).

[6] P. Adzic et al. "Energy resolution of the barrel of the CMS electromagnetic calorimeter". In: *J. Instrum.* 2 (2007), P04004. DOI: 10.1088/1748-0221/2/04/P04004.

[7] S. Agostinelli et al. "GEANT4–a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[8] *ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC*. LHC technical proposal. Geneva: CERN, 1995. URL: http://cds.cern.ch/record/293391.

[9] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[10] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[11] Johan Alwall et al. "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions". In: *Eur. Phys. J. C* 53 (2008), pp. 473–500. DOI: 10.1140/epjc/s10052-007-0490-5. arXiv: 0706.2569 [hep-ph].

[12] Jonas Rembser and. "CMS Electron and Photon Performance at 13 TeV". In: *Journal of Physics: Conference Series* 1162 (Jan. 2019), p. 012008. DOI: 10.1088/1742-6596/1162/1/012008. URL: https://doi.org/10.1088%5C%2F1742-6596%5C%2F1162%5C%2F1%5C%2F012008.

[13] *ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN*. LHC technical proposal. Geneva: CERN, 1994. URL: http://cds.cern.ch/record/290968.

[14] T. Aziz et al. "Design, fabrication and characterization of the first AC-coupled silicon microstrip sensors in India". In: *JINST* 9 (2014), P06008. DOI: 10.1088/1748-0221/9/06/P06008. arXiv: 1402.2406 [physics.ins-det].

[15] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[16] et. al. Berger Nicolas. "Simplified Template Cross Sections - Stage 1.1". In: (2019). arXiv: 1906.02754 [hep-ph].

[17] Cristina Biino. "The CMS Electromagnetic Calorimeter: overview, lessons learned during Run 1 and future projections". In: *J. Phys. Conf. Ser.* 587.1 (2015), p. 012001. DOI: 10.1088/1742-6596/587/1/012001.

[18] Oliver Brein, Robert V. Harlander, and Tom J. E. Zirke. "vh@nnlo - Higgs Strahlung at hadron colliders". In: *Comput. Phys. Commun.* 184 (2013), pp. 998–1003. DOI: 10.1016/j.cpc.2012.11.002. arXiv: 1210.5347 [hep-ph].

[19] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph].

[20] Jonathan M. Butterworth et al. "Jet substructure as a new Higgs search channel at the LHC". In: *Phys. Rev. Lett.* 100 (2008), p. 242001. DOI: 10.1103/PhysRevLett.100.242001. arXiv: 0802.2470 [hep-ph].

[21] Matteo Cacciari, G. P. Salam, and Gregory Soyez. "The Anti-k(t) jet clustering algorithm". In: *J. High Energy Phys.* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[22] Stefano Carrazza et al. "A compression algorithm for the combination of PDF sets". In: *Eur. Phys. J. C* 75 (2015), p. 474. DOI: 10.1140/epjc/s10052-015-3703-3. arXiv: 1504.06469 [hep-ph].

[23] S. Chatrchyan et al. "The CMS Experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[24] Serguei Chatrchyan et al. "Measurement of the Hadronic Activity in Events with a Z and Two Jets and Extraction of the Cross Section for the Electroweak Production of a Z with Two Jets in pp Collisions at $\sqrt{s}$ = 7 TeV". In: *JHEP* 10 (2013), p. 062. DOI: 10.1007/JHEP10(2013)062. arXiv: 1305.7389 [hep-ex].

[25] Serguei Chatrchyan et al. "Missing transverse energy performance of the CMS detector". In: *JINST* 6 (2011), P09001. DOI: 10.1088/1748-0221/6/09/P09001. arXiv: 1106.5048 [physics.ins-det].

[26] Serguei Chatrchyan et al. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[27] Serguei Chatrchyan et al. "Search for the standard model Higgs boson decaying to bottom quarks in pp collisions at $\sqrt{s} = 7$ TeV". In: *Phys. Lett.* B710 (2012), pp. 284–306. DOI: 10.1016/j.physletb.2012.02.085. arXiv: 1202.4195 [hep-ex].

[28] Serguei Chatrchyan et al. "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks". In: *Phys. Rev.* D89.1 (2014), p. 012003. DOI: 10.1103/PhysRevD.89.012003. arXiv: 1310.3687 [hep-ex].

[29] CMS Collaboration. "Search for the Standard Model Higgs Boson Decaying to Bottom Quarks and Produced in Association with a W or a Z Boson". In: *CMS Physics Analysis Summary* CMS-PAS-HIG-11-012 (2011). URL: http://cdsweb.cern.ch/record/1376636.

[30] "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software". In: *Technical Design Report CMS* (2006). URL: https://cds.cern.ch/record/922757.

[31] "CMS slice raw illustrator files". In: (Aug. 2016). URL: https://cds.cern.ch/record/2204899.

[32] "CMS: The hadron calorimeter technical design report". In: (June 1997).

[33] CMS Collaboration. "A measurement of the Higgs boson mass in the diphoton decay channel". In: (2019).

[34] CMS Collaboration. *CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV*. CMS Physics Analysis Summary CMS-PAS-LUM-17-004. 2018. URL: https://cds.cern.ch/record/2621960.

[35] CMS Collaboration. *CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV*. CMS Physics Analysis Summary CMS-PAS-LUM-18-002. 2019. URL: https://cds.cern.ch/record/2676164.

[36] CMS Collaboration. *CMS luminosity measurements for the 2016 data taking period*. CMS Physics Analysis Summary CMS-PAS-LUM-17-001. 2017. URL: https://cds.cern.ch/record/2257069.

[37] J. S. Conway. *Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra*. 2011. arXiv: 1103.0354 [physics.data-an].

[38] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011). [Erratum: Eur.Phys.J.C 73, 2501 (2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an].

[39] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph].

[40] Souvik Das. "A simple alternative to the Crystal Ball function". In: (Mar. 2016). arXiv: 1603.08591 [hep-ex].

[41] Marco Delmastro et al. *Simplified Template Cross Sections – Stage 1.1*. Tech. rep. 14 pages, 3 figures. Geneva: CERN, Apr. 2019. arXiv: 1906.02754. URL: https://cds.cern.ch/record/2669925.

[42] Ansgar Denner et al. "HAWK 2.0: A Monte Carlo program for Higgs production in vector-boson fusion and Higgs strahlung at hadron colliders". In: *Comput. Phys. Commun.* 195 (2015), pp. 161–171. DOI: 10.1016/j.cpc.2015.04.021. arXiv: 1412.5390 [hep-ph].

[43] "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *J. Instrum.* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[44] Milos Dordevic. "The CMS Particle Flow Algorithm". In: *EPJ Web Conf.* 191 (2018), 02016. 7 p. DOI: 10.1051/epjconf/201819102016. URL: http://cds.cern.ch/record/2678077.

[45] John Ellis. "Higgs Physics". In: (Dec. 2013). 52 pages, 45 figures, Lectures presented at the ESHEP 2013 School of High-Energy Physics, to appear as part of the proceedings in a CERN Yellow Report, 117–168. 52 p. DOI: 10.5170/CERN-2015-004.117. arXiv: 1312.5672. URL: https://cds.cern.ch/record/1638469.

[46] John Ellis. *Topics in Higgs Physics*. 2017. arXiv: 1702.05436 [hep-ph].

[47] Lyndon R Evans and Philip Bryant. "LHC Machine". In: *JINST* 3 (2008). This report is an abridged version of the LHC Design Report (CERN-2004-003), S08001. 164 p. DOI: 10.1088/1748-0221/3/08/S08001. URL: https://cds.cern.ch/record/1129806.

[48] Giancarlo Ferrera, Massimiliano Grazzini, and Francesco Tramontano. "Associated ZH production at hadron colliders: the fully differential NNLO QCD calculation". In: *Phys. Lett. B* 740 (2015), pp. 51–55. DOI: 10.1016/j.physletb.2014.11.040. arXiv: 1407.4747 [hep-ph].

[49] Giancarlo Ferrera, Massimiliano Grazzini, and Francesco Tramontano. "Higher-order QCD effects for associated WH production and decay at the LHC". In: *JHEP* 04 (2014), p. 039. DOI: 10.1007/JHEP04(2014)039. arXiv: 1312.1669 [hep-ph].

[50] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: (2016). DOI: 10.2172/1345634,10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[51] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: 2/2017 (Oct. 2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[52] Rikkert Frederix and Stefano Frixione. "Merging meets matching in MC@NLO". In: *JHEP* 12 (2012), p. 061. DOI: 10.1007/JHEP12(2012)061. arXiv: 1209.6215 [hep-ph].

[53] J H Friedman. "Data analysis techniques for high energy particle physics". In: (Oct. 1974), 96 p. DOI: 10.5170/CERN-1974-023.271. URL: http://cds.cern.ch/record/695770.

[54] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[55] R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. *Adaptive Vertex Fitting*. Tech. rep. Geneva: CERN, Mar. 2007. URL: http://cds.cern.ch/record/1027031.

[56] Rudolf Frühwirth and Are Strandlie. "Secondary Vertex Reconstruction". In: *Pattern Recognition, Tracking and Vertex Reconstruction in Particle Detectors*. Cham: Springer International Publishing, 2021, pp. 159–165. ISBN: 978-3-030-65771-0. DOI: 10.1007/978-3-030-65771-0_9. URL: https://doi.org/10.1007/978-3-030-65771-0_9.

[57] Aishik Ghosh and Benjamin Nachman. "A cautionary tale of decorrelating theory uncertainties". In: *Eur. Phys. J. C* 82.1 (2022), p. 46. DOI: 10.1140/epjc/s10052-022-10012-w. arXiv: 2109.08159 [hep-ph].

[58] S. L. Glashow. "Partial Symmetries of Weak Interactions". In: *Nucl. Phys.* 22 (1961), pp. 579–588. DOI: 10.1016/0029-5582(61)90469-2.

[59] T. Gleisberg et al. "Event generation with SHERPA 1.1". In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

[60] Judith R Goodstein. "Ricci and Levi-Civita's tensor analysis paper: Lie Groups: History, Frontiers, and Applications, Volume 2. Translated and edited by Robert Hermann. Brookline, Massachusetts (Math. Sci. Press). 1975. 261 p. paper". In: *Historia Mathematica* 4.2 (1977), p. 228. ISSN: 0315-0860. DOI: https://doi.org/10.1016/0315-0860(77)90126-4. URL: https://www.sciencedirect.com/science/article/pii/0315086077901264.

[61] O. W. Greenberg. "Spin and Unitary-Spin Independence in a Paraquark Model of Baryons and Mesons". In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 598–602. DOI: 10.1103/PhysRevLett.13.598. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.598.

[62] David J. Gross and Frank Wilczek. "Ultraviolet Behavior of Non-Abelian Gauge Theories". In: *Phys. Rev. Lett.* 30 (26 June 1973), pp. 1343–1346. DOI: 10.1103/PhysRevLett.30.1343. URL: https://link.aps.org/doi/10.1103/PhysRevLett.30.1343.

[63] Particle Data Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104. eprint: https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf. URL: https://doi.org/10.1093/ptep/ptaa104.

[64] Keith Hamilton, Paolo Nason, and Giulia Zanderighi. "MINLO: Multi-Scale Improved NLO". In: *JHEP* 10 (2012), p. 155. DOI: 10.1007/JHEP10(2012)155. arXiv: 1206.3572 [hep-ph].

[65] Robert V. Harlander, Stefan Liebler, and Tom Zirke. "Higgs Strahlung at the Large Hadron Collider in the 2-Higgs-Doublet Model". In: *JHEP* 02 (2014), p. 023. DOI: 10.1007/JHEP02(2014)023. arXiv: 1307.8122 [hep-ph].

[66] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. ISSN: 0006-3444. DOI: 10.1093/biomet/57.1.97. eprint: https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf. URL: https://doi.org/10.1093/biomet/57.1.97.

[67] Chris Hays, Veronica Sanz Gonzalez, and Gabija Zemaityte. "Constraining EFT parameters using simplified template cross sections". In: (Oct. 2017). URL: https://cds.cern.ch/record/2290628.

[68] Stefan Höche. "Introduction to parton-shower event generators". In: *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders*. 2015, pp. 235–295. DOI: 10.1142/9789814678766_0005. arXiv: 1411.4085 [hep-ph].

[69] David W. Hogg, Jo Bovy, and Dustin Lang. *Data analysis recipes: Fitting a model to data*. 2010. arXiv: 1008.4686 [astro-ph.IM].

[70] Michael Hostettler et al. "Impact of the Crossing Angle on Luminosity Asymmetries at the LHC in 2016 Proton Physics Operation". In: (2017), TUPVA005. 4 p. DOI: 10.18429/JACoW-IPAC2017-TUPVA005. URL: https://cds.cern.ch/record/2289120.

[71] "Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques". In: (Jan. 2020). URL: https://cds.cern.ch/record/2707946.

[72] "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018". In: (Apr. 2020). URL: https://cds.cern.ch/record/2715872.

[73] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. ISSN: 0021-9223. DOI: 10.1115/1.3662552. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. URL: https://doi.org/10.1115/1.3662552.

[74] V. Khachatryan et al. "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV". In: *The European Physical Journal C* 75.5 (May 2015). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-015-3351-7. URL: http://dx.doi.org/10.1140/epjc/s10052-015-3351-7.

[75] Vardan Khachatryan et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[76] Andrew J. Larkoski et al. "Soft Drop". In: *JHEP* 05 (2014), p. 146. DOI: 10.1007/JHEP05(2014)146. arXiv: 1402.2657 [hep-ph].

[77] Ferdinand F. Leimkuhler. "THE BRADFORD DISTRIBUTION". In: *Journal of Documentation* 23 (1967), pp. 197–207.

[78] *LHCb : Technical Proposal*. Geneva: CERN, 1998. URL: http://cds.cern.ch/record/622031.

[79] Hao Li et al. "Visualizing the Loss Landscape of Neural Nets". In: *CoRR* abs/1712.09913 (2017). arXiv: 1712.09913. URL: http://arxiv.org/abs/1712.09913.

[80] Gionata Luisoni et al. "$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". In: *JHEP* 10 (2013), p. 083. DOI: 10.1007/JHEP10(2013)083. arXiv: 1306.2542 [hep-ph].

[81] *Luminosity: RunII proton-proton 13 TeV collisions*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults. Accessed: 2021-03-30.

[82] *Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment*. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2683870.

[83] Paolo Nason. "A New method for combining NLO QCD with shower Monte Carlo algorithms". In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146.

[84] Izaak Neutelings. *CMS coordinate diagrams*. https://tikz.net/axis3d_cms/. Accessed: 2022-01-30.

[85] E. Noether. "Invariante Variationsprobleme". ger. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918 (1918), pp. 235–257. URL: http://eudml.org/doc/59024.

[86] *Observation of Higgs boson decay to bottom quarks*. Tech. rep. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2633415.

[87] "Observation of Higgs boson decay to bottom quarks". In: *Phys. Rev. Lett.* 121.12 (2018), p. 121801. DOI: 10.1103/PhysRevLett.121.121801. arXiv: 1808.08242 [hep-ex].

[88] "Performance of b tagging algorithms in proton-proton collisions at 13 TeV with Phase 1 CMS detector". In: (June 2018). URL: https://cds.cern.ch/record/2627468.

[89] *Pileup Jet Identification*. Tech. rep. Geneva: CERN, 2013. URL: http://cds.cern.ch/record/1581583.

[90] H. David Politzer. "Reliable Perturbative Results for Strong Interactions?" In: *Phys. Rev. Lett.* 30 (26 June 1973), pp. 1346–1349. DOI: 10.1103/PhysRevLett.30.1346. URL: https://link.aps.org/doi/10.1103/PhysRevLett.30.1346.

[91] Kenneth Rose. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239.

[92] Anirban Saha. "Phase 1 upgrade of the CMS pixel detector". In: *Journal of Instrumentation* 12.02 (Feb. 2017), pp. C02033–C02033. DOI: 10.1088/1748-0221/12/02/c02033. URL: https://doi.org/10.1088%2F1748-0221%2F12%2F02%2Fc02033.

[93] Abdus Salam and John Clive Ward. "On a Gauge Theory of Elementary Interactions". In: *Nuovo Cim.* 19 (1961), pp. 165–170. DOI: 10.1007/BF02812723.

[94] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[95] A. M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[96] A. M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[97] A. M. Sirunyan et al. “Evidence for Higgs boson decay to a pair of muons”. In: *Journal of High Energy Physics* 2021.1 (Jan. 2021). ISSN: 1029-8479. DOI: 10.1007/jhep01(2021)148. URL: http://dx.doi.org/10.1007/JHEP01(2021)148.

[98] A. M. Sirunyan et al. “Observation of Higgs Boson Decay to Bottom Quarks”. In: *Physical Review Letters* 121.12 (Sept. 2018). ISSN: 1079-7114. DOI: 10.1103/physrevlett.121.121801. URL: http://dx.doi.org/10.1103/PhysRevLett.121.121801.

[99] A. M. Sirunyan et al. “Observation of tt¯H Production”. In: *Physical Review Letters* 120.23 (June 2018). ISSN: 1079-7114. DOI: 10.1103/physrevlett.120.231801. URL: http://dx.doi.org/10.1103/PhysRevLett.120.231801.

[100] Albert M Sirunyan et al. “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”. In: *Eur. Phys. J. C* 80.1 (2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179 [hep-ex].

[101] Albert M Sirunyan et al. “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques”. In: *JINST* 15.06 (2020), P06005. DOI: 10.1088/1748-0221/15/06/P06005. arXiv: 2004.08262 [hep-ex].

[102] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. “A Brief Introduction to PYTHIA 8.1”. In: *Comput. Phys. Commun.* 178 (2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[103] Davison E. Soper. “Parton distribution functions”. In: *Nucl. Phys. B Proc. Suppl.* 53 (1997). Ed. by C. Bernard et al., pp. 69–80. DOI: 10.1016/S0920-5632(96)00600-7. arXiv: hep-lat/9609018.

[104] *Technical proposal*. LHC technical proposal. Cover title : CMS, the Compact Muon Solenoid : technical proposal. Geneva: CERN, 1994. URL: http://cds.cern.ch/record/290969.

[105] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/349375.

[106] “The CMS Experiment at the CERN LHC”. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[107] VHbb Team. “Search for the Standard Model Higgs Boson Produced in Association with a W or Z and Decaying to Bottom Quarks”. In: *CMS Note* 2011/240 (2011). URL: %7Bhttp://cms.cern.ch/iCMS/jsp/openfile.jsp?tp=draft%5C&files=AN2011%5C_240%5C_v10.pdf%7D.

[108] Bryan Webber. *Hadronization*. 1994. arXiv: hep-ph/9411384 [hep-ph].

[109] Steven Weinberg. “A Model of Leptons”. In: *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[110] Jorg Wenninger. "LHC status and performance". In: *PoS* CHARGED2018 (2019), p. 001. DOI: 10.22323/1.339.0001.

[111] Kenneth G. Wilson. "Confinement of quarks". In: *Phys. Rev. D* 10 (8 Oct. 1974), pp. 2445–2459. DOI: 10.1103/PhysRevD.10.2445. URL: https://link.aps.org/doi/10.1103/PhysRevD.10.2445.

[112] Bing Xu et al. "Empirical Evaluation of Rectified Activations in Convolutional Network". In: *CoRR* abs/1505.00853 (2015). arXiv: 1505.00853. URL: http://arxiv.org/abs/1505.00853.