UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS FÍSICAS

TESIS DOCTORAL

Measurement of Higgs boson production via gluon fusion in the H WW decay channel with the CMS experiment

Medida de la producción del bosón de Higgs mediante fusión de gluones en el canal de desintegración H WW con el experimento CMS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Adrián Álvarez Fernández

DIRECTORA

Isabel Josa Mutuberría

Madrid

© Adrián Álvarez Fernández, 2021

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS

TESIS DOCTORAL

MEASUREMENT OF HIGGS BOSON PRODUCTION VIA GLUON FUSION
IN THE HWW DECAY CHANNEL WITH THE CMS EXPERIMENT

MEDIDA DE LA PRODUCCIÓN DEL BÓSON DE HIGGS MEDIANTE FUSIÓN DE GLUONES EN EL CANAL DE DESINTEGRACIÓN HWW CON EL EXPERIMENTO CMS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

ADRIÁN ÁLVAREZ FERNÁNDEZ

DIRECTOR

ISABEL JOSA MUTUBERRÍA

Measurement of Higgs boson production via gluon fusion in the $H \to WW$ decay channel with the CMS experiment

by

Adrián Álvarez Fernández

Doctoral Dissertation in Physics
submitted to

Universidad Complutense de Madrid
Facultad de Ciencias Físicas

UNIVERSIDAD
COMPLUTENSE
MADRID

Supervised by: **Dr. Isabel Josa Mutuberría**

Madrid, October 2020

Medida de la producción del bosón de Higgs mediante fusión de gluones en el canal de desintegración $H \to WW$ con el experimento CMS

por

Adrián Álvarez Fernández

Memoria de la tesis
presentada en la

Universidad Complutense de Madrid
Facultad de Ciencias Físicas

UNIVERSIDAD
COMPLUTENSE
MADRID

Supervisada por: **Dra. Isabel Josa Mutuberría**

Madrid, octubre de 2020

Abstract

This thesis reports measurements of the Higgs boson production in proton-proton collisions at a centre-of-mass energy of 13 TeV with data collected by the CMS experiment during the Run 2 operation period (2016-2018) of the Large Hadron Collider at CERN. The measurement is performed using a sample of event candidates to originate from the decay of a Higgs boson into a pair of W bosons, which then decay into a charged lepton and a neutrino resulting in a final state with two charged leptons of different flavour. The thesis focuses on the study of the Higgs boson production through the gluon fusion process, which is the most favoured one at a centre-of-mass energy of 13 TeV.

The optimisation of the physics object selection in the CMS configuration during the Run 2 is described first. Then, the evaluation of the main backgrounds are carefully scrutinised and the signal selection is decided. The main physics result is the Higgs boson production cross section, quantified in terms of a signal strength modifier, which is defined as the ratio of the observed cross section and the prediction from the Standard Model. The analysis is carried out in the inclusive phase space and in fine-grained exclusive phase space bins in the so-called Simplified Template Cross-Section (STXS) framework. The STXS framework has been developed with the aim to reduce the dependence on the underlying physics model and its theoretical uncertainties, while keeping a high sensitivity of the measurements.

The signal strength modifiers in both the inclusive and STXS analysis are obtained from a maximum likelihood fit, using as discriminant variables the dilepton invariant mass and the transverse mass defined with the transverse momentum of the dilepton system and the missing transverse momentum. The value of the systematic uncertainties in the analysis is also adjusted in the fit, and dedicated control regions are used to constrain the contribution from some background processes.

A multiple event categorisation is performed in the signal region for the inclusive analysis depending on the number of reconstructed jets with high transverse momentum, the flavour of the leptons, their charge and the transverse momentum of the subleading lepton, resulting in an optimised overall sensitivity to the Higgs boson signal. Signal strength modifiers are obtained for the 0- and 1-jet event categories, independently and combined, for the data collected in the three years. All signal strength modifiers are found to be consistent with 1, indicating that the Standard Model properly reproduces the experimental data. The combined result for the global signal strength using the full dataset is $\mu_{global} = 0.90^{+0.10}_{-0.09}$. The precision of the measurement is dominated by systematic effects coming from the theoretical model and instrumental sources.

Signal strength modifiers are determined for one STXS bin with 0-jets, 3 bins with 1-jet of different regions in the Higgs boson transverse momentum ($p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV, $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV and $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV) and a high Higgs boson transverse momentum bin ($p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV), which may be especially sensitive to new physics effects. These bins are specific to the gluon fusion production mode and are defined in a fully orthogonal way with respect to STXS bins characteristic of other production modes. The measured signal strength modifiers for the studied STXS bins are in all cases consistent with the Standard Model within their uncertainties.

Finally, production cross sections in the STXS bins are derived from the measured signal strength modifiers. The production cross section of the most sensitive 0-jet bin is measured to be $\sigma_{ggH0J} = 606 \pm 76$ fb, and the combined result of the STXS bins for the $\mathrm{gg} \rightarrow \mathrm{H} \times \mathcal{B}(\mathrm{H} \rightarrow \mathrm{WW})$ cross section is $\sigma_{ggH}^{STXS} = 1.08^{+0.12}_{-0.11}$ pb. These measurements are some of the main results of the CMS analysis of the Higgs boson production studies exploiting the decay channel into a pair of W bosons, with a leptonic decay of the W bosons.

Resumen

Esta tesis presenta el análisis de la medida de la producción del bosón de Higgs en colisiones protón-protón a una energía del centro de masas de 13 TeV. Los datos de dichas colisiones fueron tomadas por el experimento CMS durante el periodo de operación del Run 2 (2016-2018) del Gran Colisionador de Hadrones (LHC) en el CERN. Las medidas se han realizado utilizando una muestra de sucesos cuyas características serían compatibles con la desintegración de un bosón de Higgs a un par de bosones W, que a su vez se desintegrarían en un par de leptones cargados de distinto sabor y neutrinos. La tesis se centra en el estudio de la producción de bosones de Higgs con el proceso de fusión de dos gluones, que es el dominante a una energía de centro de masas de 13 TeV.

Se describe en primer lugar la selección de los elementos del análisis en la configuración usada por CMS. Después, los fondos principales se evalúan detalladamente y se decide cómo realizar una selección enriquecida en sucesos de señal. El principal resultado del estudio es la sección eficaz de producción del bosón de Higgs, cuantificada mediante un modificador a la intensidad de la señal, definido como el cociente entre la sección eficaz observada y la predicción del Modelo Estándar. El análisis se lleva a cabo en el espacio de fases inclusivo y en determinadas regiones del espacio de fases aisladas y con una granularidad fina utilizando un modelo de secciones eficaces simplificadas (Simplified Template Cross-Sections o STXS). Este entorno de trabajo ha sido desarrollado con el objetivo de reducir la dependencia en el modelo teórico subyacente y sus incertidumbres, manteniendo al mismo tiempo la máxima sensibilidad posible en las medidas.

Los modificadores a la intensidad de la señal tanto en el análisis inclusivo como para las STXS se obtienen de un ajuste de máxima verosimilitud, usando como variables discriminantes la masa invariante de los dos leptones y la masa transversa definida utilizando el momento del sistema dileptónico y el momento faltante en el plano transverso. El valor de las incertidumbres sistemáticas en el análisis también se adapta durante el ajuste, y se usan regiones de control para constreñir la contribución de algunos procesos de fondo.

Los eventos de la región de señal se dividen en múltiples categorías, dependiendo del número de jets reconstruidos con alto momento, el sabor de los leptones, su carga y el momento transverso del leptón de momento más pequeño, optimizando de esta manera la sensibilidad del análisis a la señal del bosón de Higgs. La intensidad de la señal se obtiene para las categorías de 0 y 1 jets, independientemente y de forma conjunta, para los datos recogidos durante cada uno de los tres años del Run 2 y su combinación. Todas las intensidades de señal medidas son compatibles con 1, indicando que el Modelo Estándar reproduce de manera correcta los datos experimentales. El resultado combinado para la intensidad global de la señal usando el conjunto de datos completo es $\mu_{global} = 0.90^{+0.10}_{-0.09}$. La precisión de la medida está dominada por efectos sistemáticos que provienen del modelo teórico y de fuentes instrumentales.

Los modificadores a la intensidad de la señal se determinan para la región de las STXS correspondiente a 0 jets, las 3 divisiones de la región de 1 jet dependiendo del momento transverso del bosón de Higgs ($p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV, $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV y $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV) y para la región de alto momento transverso del bosón de Higgs ($p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV), que puede ser especialmente sensible a efectos de nueva física. Estas regiones son específicas del modo de producción de fusión de gluones y se definen de manera completamente independiente con respecto a las regiones características de otros modos de producción. Los modificadores a la intensidad de señal medidos en cada una de las regiones son consistentes con la predicción del Modelo Estándar dentro de sus incertidumbres.

Finalmente, las secciones eficaces de producción se derivan de los modificadores a la intensidad de la señal en cada región de las STXS. La medida más precisa, correspondiente a la región de 0 jets es de $\sigma_{ggH0J} = 606 \pm 76$ fb, y el resultado combinado de todas las regiones STXS para la sección eficaz de gg$\rightarrow$ H $\times \mathcal{B}$(H $\rightarrow$ WW) es de $\sigma_{ggH} = 1.08^{+0.12}_{-0.11}$ pb. Estas medidas forman parte del conjunto principal de resultados de análisis de CMS que estudia la producción del bosón de Higgs en su canal de desintegración en un par de bosones W que se desintegran leptónicamente.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Standard Model of particle physics

The Standard Model (SM) is a gauge theory based on the symmetry group $\mathrm{SU(3)_C \times SU(2)_L \times U(1)_Y}$ that describes the fermionic constituents of matter and their interactions via boson exchange. The SM is the theory explaining most experimental phenomena as of today, providing very accurate predictions for a large variety of physical processes. It does not however explain everything, missing key items such as a description of the gravitational force consistent with general relativity or an explanation for dark matter, among other open questions in physics.

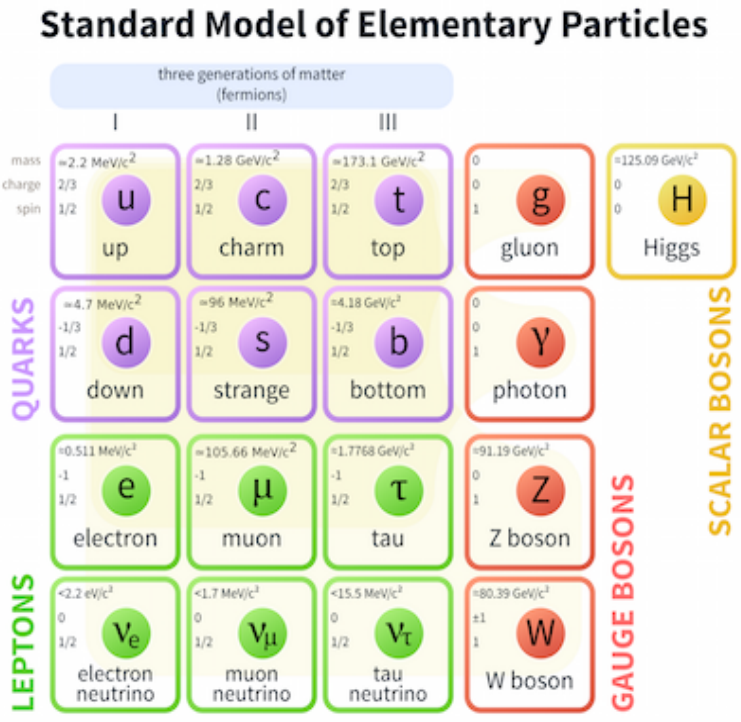


Figure 1.1: The Standard Model of particle physics. Picture taken from [1].

Figure 1.1 shows the elementary particles of the SM, including their masses and electric charges. Matter is divided into three families of particles with the same properties but different masses. The quarks are six strong-interacting particles with fractionary electric charge +2/3 (up, charm and top) or -1/3 (down, strange, bottom). Leptons are transparent to the strong nuclear force and exist in three families of negatively charged particles (electrons, muons and tau leptons) and three flavours of neutral particles, their neutrino counterparts. The fermionic content of the SM framework is completed when adding the antiparticles. They are associated to each fermion previously mentioned and possess the same mass, opposite electric charge and opposite flavour quantum numbers as their counterparts.

Electromagnetism, the strong force and the weak interaction are described by spin-1 gauge fields. Massless gluons and photons are the carriers of the strong and electromagnetic forces, respectively, whereas the weak interaction is mediated by the massive $W^{\pm}$ and Z bosons. The masses of the W and Z bosons are the result of spontaneous symmetry breaking. This mechanism also gives rise to a physical scalar particle that completes the SM, the Higgs boson.

1.1.1 The electroweak interaction

Electromagnetism and the weak interaction were unified by S. L. Glashow, A. Salam and S. Weinberg [2, 3] as a $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ local gauge theory. Fermions in the fundamental representation of this group appear as left-handed doublets and right-handed singlets and they are classified by the quantum numbers of isospin charge I, associated with the SU(2) symmetry group and the left-handed doublets, and the hypercharge Y, associated with the right-handed singlets and the U(1) symmetry group. The third component of the isospin charge I_3 and the weak hypercharge Y are related to the electric charge Q in this way:

$$Y = 2(Q - I_3) \ . \tag{1.1}$$

Therefore, in a field theory for electromagnetic and weak interactions, the Lagrangian must be invariant under gauge transformations of the type SU(2)×U(1). Additionally, this symmetry is broken by the Higgs mechanism down to the electromagnetic gauge symmetry in order for the $\mathrm{W}^{\pm}$ and Z bosons to acquire mass.

The full Lagrangian thus has a gauge field term $\mathcal{L}_{\mathrm{G}}$, a Higgs field term describing its interaction with the gauge bosons $\mathcal{L}_{\mathrm{H}}$, a fermion-gauge boson interaction term $\mathcal{L}_{\mathrm{F}}$ and the fermion mass term $\mathcal{L}_{\mathrm{Y}}$:

$$\mathcal{L}_{EWK} = \mathcal{L}_G + \mathcal{L}_H + \mathcal{L}_F + \mathcal{L}_Y \tag{1.2}$$

For the gauge part, the isospin operators I_1, I_2, I_3 and the hypercharge Y are each associated to a vector field. There is a triplet of vector fields W^i_μ for I_i (i =1,2,3) and a singlet field B^μ for Y. These define the field strength tensors

$$\begin{aligned} W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_2 \epsilon_{abc} W^b_\mu W^c_\nu \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu \ , \end{aligned} \tag{1.3}$$

where g_2 is the non-Abelian SU(2) gauge coupling constant. Consequently, the gauge field Lagrangian can be written as

$$\mathcal{L}_G = -\frac{1}{4} W^a_{\mu\nu} W a, \mu\nu - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \ . \tag{1.4}$$

The Higgs field is represented by a single complex scalar doublet field with hypercharge Y=1. It couples with the other gauge fields and with itself through

$$\mathcal{L}_H = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi), \tag{1.5}$$

where the gauge invariant covariant derivative and the Higgs self-interaction are defined like

$$\begin{aligned} D_\mu &= \partial_\mu + ig_2 W_\mu^a \frac{\tau^a}{2} + ig_1 \frac{Y}{2} B_\mu \\ V(\Phi) &= -\mu^2 \Phi^\dagger \Phi + \frac{\lambda}{4}(\Phi^\dagger \Phi)^2, \end{aligned} \tag{1.6}$$

with g_1 being the Abelian U(1) coupling and where λ and μ are constant parameters. τ^a are the Pauli matrices, which constitute a basis for the Lie algebra of SU(2). This potential has a vacuum expectation value of

$$< \Phi >= \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \text{ with } v = \frac{2\mu}{\sqrt{\lambda}}. \tag{1.7}$$

As a consequence, the bosonic fields in Eq.1.5 and the Higgs field acquire the following masses:

$$m_{\mathrm{W}} = \frac{v g_2}{2} \ , m_{\mathrm{Z}} = \frac{v}{2}\sqrt{g_1^2 + g_2^2} \text{ and } m_{\mathrm{H}} = v\sqrt{2\lambda} \tag{1.8}$$

The W and Z boson masses are related by the coupling constant and the ratio between them can be associated to the weak mixing angle or Weinberg angle. This angle is the angle by which spontaneous symmetry breaking rotates the original W^3 and B vector boson plane, producing as a result the Z boson and the photon.

$$\frac{m_W}{m_Z} = \frac{g_2}{\sqrt{g_1^2 + g_2^2}} = \cos\theta_W \quad , \quad \begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{bmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{bmatrix} \begin{pmatrix} B \\ W^3 \end{pmatrix} \tag{1.9}$$

The fermionic part of the Lagrangian has one term describing the interactions with the boson fields and one term where the Yukawa couplings between fermions and the Higgs boson give rise to the fermion masses. Given a family of fermions, e.g. the lighter family of quarks, we use this notation for the fermionic field:

$$\psi_1 = \begin{pmatrix} u \\ d \end{pmatrix}_L , \ \ \psi_2 = u_R, \ \ \psi_3 = d_R \tag{1.10}$$

The same may be defined for leptons using ν_e and e instead of u and d, and for the other quark and lepton families. With this definition, the interaction between gauge boson fields and fermions can be written using the covariant derivative of Eq. 1.6 as:

$$\mathcal{L}_F = \sum_j i\bar{\psi}_j(x)\gamma^\mu D_\mu \psi_j(x) \ . \tag{1.11}$$

For the Yukawa Lagrangian, we can choose the following gauge to simplify the term:

$$\Phi = \begin{pmatrix} \Phi^+(x) \\ \Phi^0(x) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \ , \tag{1.12}$$

which results in the simple expression for the fermion mass terms

$$\mathcal{L}_Y = -\sum_f m_f \bar{\psi}_f \psi_f - \sum_f \frac{m_f}{v} \bar{\psi}_f \psi_f H \ , \tag{1.13}$$

where $m_f = g_f \frac{v}{\sqrt{2}}$ is the mass of the fermions, which depends on the Yukawa coupling constants g_f.

With this Lagrangian the mass eigenstates for fermions are defined, but for quarks they do not coincide with the weak eigenstates. There is a mixing between flavours when there are charged currents involved, which is represented by a unitary matrix, called CKM (Cabibbo-Kobayashi-Maskawa) [4, 5]. The CKM matrix is defined by three angles and one CP-violating phase, four free parameters of the SM that are experimentally measured.

1.1.2 Quantum chromodynamics

Quantum chromodynamics (QCD) is the part of the SM that describes the strong nuclear interaction. It is a non-Abelian gauge theory represented by the $SU(3)_C$ group, where quarks constitute a colour triplet. Because of the three colours of the quarks and three anticolours of the antiquarks, there are eight possible independent gluon states that mediate the interaction, forming a colour octet.

Since gluons themselves carry colour charge, there are gluon-gluon interactions in QCD. This causes an antiscreening effect due to gluon pairs in the vacuum, which results in the running coupling constant of QCD, α_s, becoming stronger at higher distances. When separating colour charged particles, the energy needed to do so becomes increasingly larger, to the point where creation of quark-antiquark pairs from the vacuum is energetically favourable. This leads to the so-called *colour confinement*, an observed behaviour of QCD without yet an analytic proof, that implies the impossibility of isolating quarks or gluons. Coloured-charged particles are therefore always bound to others, resulting in the formation of hadrons.

On the other hand, the lower value of α_s at high energies implies that particles interacting at a short distance or with a high energy transfer behave almost like free particles, in what is called *asymptotic freedom*. In this energy regime, in which many of the interesting physics processes at the LHC happen, perturbation expansion is applicable.

The SM Lagrangian has two terms related to QCD; the gauge field propagation and self-interaction, and the quark-gluon interaction. A basis for the Lie algebra of SU(3) is formed by the eight Gell-Mann matrices λ^a, which have the structure constants f^{abc} defined by this commutator[6] :

$$[\lambda_a, \lambda_b] = if^{abc}\lambda_c \ . \tag{1.14}$$

The gauge field strength for the gluon is

$$G_a^{\mu\nu} = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu - g_s f^{abc} G_b^\mu G_c^\nu \ , \tag{1.15}$$

where the last term represents the gluon-gluon interaction, which is proportional to the strong coupling constant g_s. The QCD Lagrangian can thus be written as:

$$\mathcal{L}_{QCD} = -\frac{1}{4} G^{a,\mu\nu} G^a_{\mu\nu} + \mathcal{L}_F \ , \tag{1.16}$$

where the fermionic Lagrangian $\mathcal{L}_F$ of Eq. 1.11 needs an additional part to preserve the QCD gauge invariance, which is introduced in the covariant derivative:

$$D'_\mu = D_\mu - g_s \sum_f \bar{\psi}_f \gamma_\mu G_a^\mu \frac{\lambda^a}{2} \psi_f \tag{1.17}$$

1.2 Higgs boson production at the LHC

The SM predicts four main processes of Higgs boson production mechanisms that originate from proton-proton collisions at the LHC. Interaction between gluons and quarks forming the protons gives rise to a Higgs boson in isolation or in addition to other objects like vector bosons and quarks. The four dominant production mechanisms are shown in Feynman diagrams in Fig. 1.2 and the production cross section through those different processes is presented in Fig. 1.3.

- **Gluon fusion (ggH)**: the main production mode at the LHC at the center-of-mass energies of 7, 8 or 13 TeV and the one that mostly determines our analysis sensitivity. The gluons are massless and do not couple with the Higgs boson directly, but they do so through a heavy quark loop, as depicted in Fig. 1.2. As shown in the left plot of Fig. 1.3, the cross section for gluon fusion (top blue line) is the dominant process for a Higgs boson particle with a mass

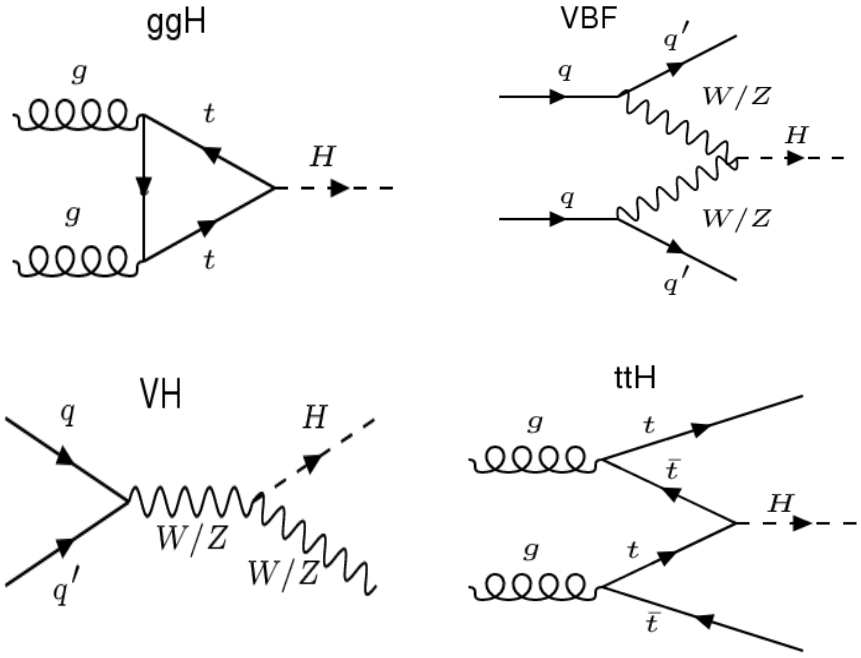


Figure 1.2: Feynman diagrams for the main Higgs boson production mechanisms in proton-proton interactions. These four processes are gluon fusion (ggH), vector boson fusion (VBF), vector boson associated production (VH) and top quark associated production (ttH).

less than 1 TeV. No other object appears in the final state at tree level, unlike the other production modes, that is why it is simply written as pp → H in the figure. The bottom plot in Fig. 1.3 shows the cross section for the mass range in which it was experimentally discovered in 2012.

- **Vector boson fusion (VBF)**: the process with the second highest cross section for a SM Higgs boson of 125 GeV and the highest cross section for a SM-like Higgs boson with a mass in the TeV range. In this case two quarks from the protons interact via electroweak W or Z bosons that merge to produce a Higgs boson. The topology of the final objects allows the identification of this process, since the outgoing quarks scatter forward after emitting the bosons and are therefore separated from each other and from the Higgs boson produced between them.

- **Vector boson associated production (VH)**: this process is the result of a vector boson emitting a Higgs boson and recoiling against it. Hence, it is also known as *Higgsstrahlung*. Figure 1.3 (bottom) shows that at m_H =125 GeV, WH and ZH cross sections are one order of magnitude smaller than ggH and of the same order the VBF cross section, but at higher masses (top plot) VH cross section falls rapidly, becoming several orders of magnitude smaller than ggH and VBF.

- **Top and bottom quark associated production ($t\bar{t}H$ and $b\bar{b}H$)**: these processes involve two vertices where $t\bar{t}$ or $b\bar{b}$ quark pairs are produced, and a third vertex in which a $t\bar{t}/b\bar{b}$ pair merges to produce a Higgs boson. The cross section for these processes at m_H =125 GeV is one order of magnitude smaller than VH and VBF and two orders of magnitude smaller than ggH. At very high mass values $t\bar{t}H$ separates from $b\bar{b}H$, surpassing VH at about 400 GeV but remaining several orders below ggH and VBF.

- There is also a single top associated production in Fig. 1.3, with a much lower cross section than the others at m_H =125 GeV. This process becomes relevant after 400 GeV, as it is the third with the largest cross section at very high masses.

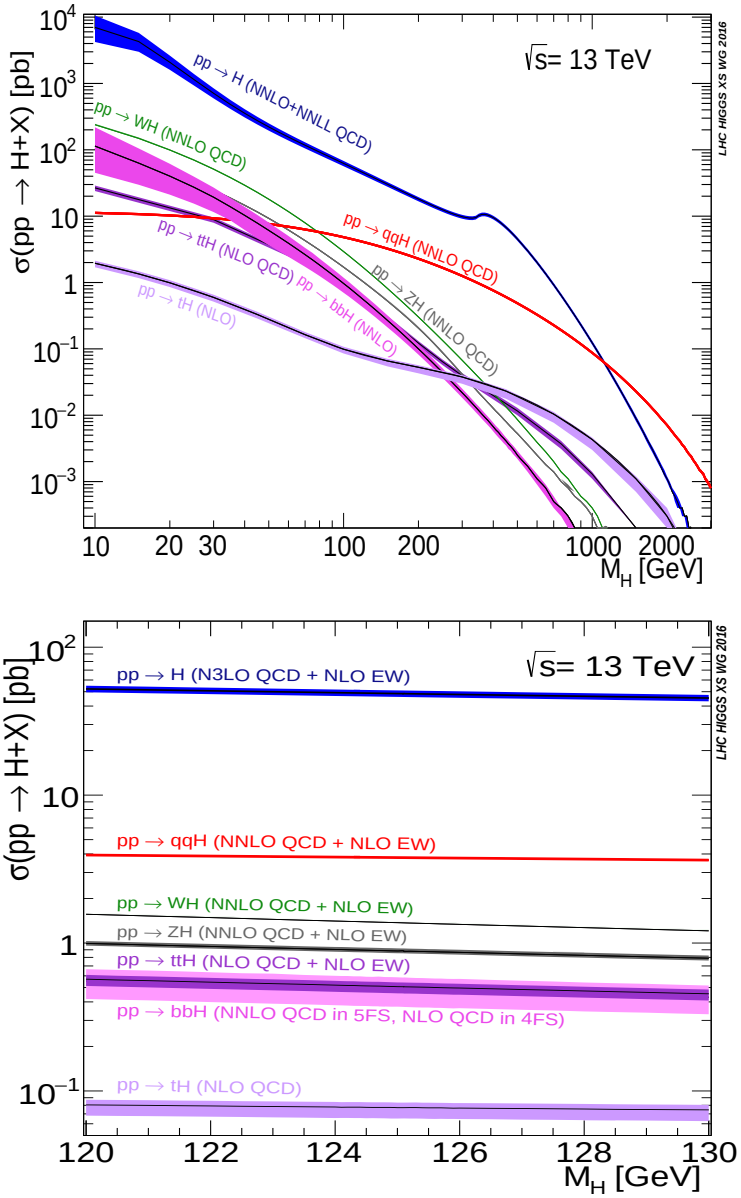


Figure 1.3: Higgs production cross sections for proton-proton collisions at a center-of-mass energy of 13 TeV, plotted as a function of the Higgs boson mass. Above: cross sections for Higgs boson masses up to 2 TeV. Below: cross sections for the mass region around 125 GeV. Picture taken from [7].

1.3 Higgs boson decay modes

The lifetime of the Higgs boson is of the order of 10^{-22}s. Therefore, it is from measurements of the kinematic variables of their decay products that it is inferred. In Fig. 1.4 the branching ratios for the SM predictions of the Higgs boson decays are presented. The decay into a pair of bottom quarks has the highest branching ratio for masses under approximately 135 GeV, then it falls rapidly, when the W and Z boson pair decays start to dominate. H → WW decay increases with the mass and is at higher Higgs boson mass values the main decay mode, having a maximum value for Higgs boson masses around 160 GeV or twice the W boson mass, since that would greatly enhance H → WW decay. ZZ decay also increases with respect to the Higgs boson mass,

then decreases temporarily around the WW mass peak and at higher masses it has the second highest branching ratio. Other Higgs decay products like charm quark pairs, tau lepton pairs or gluon pairs (produced through quark loops) are more relevant at lower masses and become less important where the boson pair production is dominant.

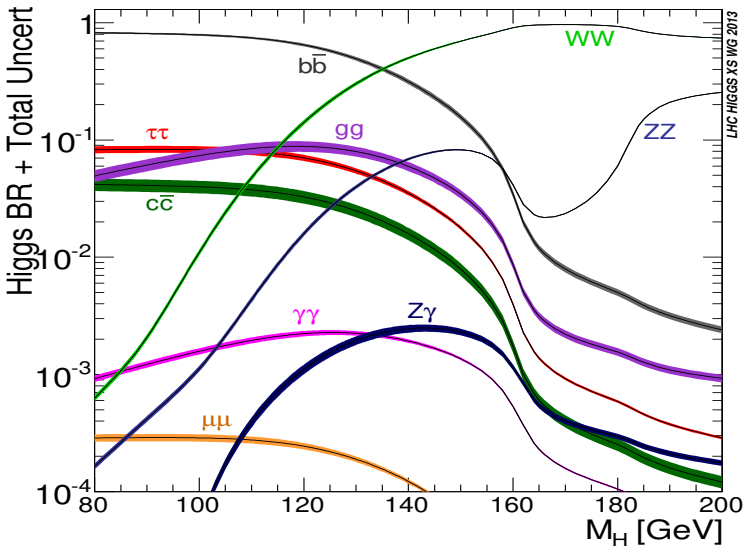


Figure 1.4: Higgs boson branching ratios with uncertainty widths. The decay of the Higgs boson to a pair of W bosons has the second highest branching ratio at a pole mass of 125 GeV and the highest for masses above 140 GeV. Picture taken from [7].

Focusing on the SM Higgs-like signal corresponding to a mass around 125 GeV, bottom quark pair is the main decay mode at that mass range. Then W boson pair has the second highest BR, followed by gluon pair, tau lepton pair and charm quark pair and Z boson pair production, all of which have branching ratios one order of magnitude below the first two processes. The branching ratios of other Higgs boson decay channels like $\mathcal{B}(\mathrm{H} \to \gamma\gamma)$, $\mathcal{B}(\mathrm{H} \to \mathrm{Z}\gamma)$ or $\mathcal{B}(\mathrm{H} \to \mu\mu)$ are at least two orders of magnitude smaller than the main decay modes.

Even though the bottom quark pair is the dominant decay mode at m_H= 125 GeV, it is however more difficult to experimentally observe than others, since there is a large background associated to this process. During Run 2 CMS observed this decay [8], but the channels in which the resonance was originally observed by both ATLAS and CMS collaborations during Run 1 were $\mathrm{H} \to \mathrm{ZZ} \to \ell^+\ell^-\ell^+\ell^-$ and $\mathrm{H} \to \gamma\gamma$. [9, 10, 11] These channels have a lower cross section, but a final state with no neutrinos or strong-interacting particles. This leads to a clear signal from isolated leptons and photons with a good signal-to-background ratio and a reconstructed invariant mass that can be precisely measured using the kinematic variables of the final state objects. The mass value for that resonance was found to be $m_H = 125.09 \pm 0.21$ (stat.) $\pm$ 0.11 (syst.) GeV. During Run 2 both ZZ [12, 13] and $\gamma\gamma$ [14, 15] channels provided new results.

Recently, other results have been made public with the Run 2 data for other Higgs decay modes. These studies include the tau lepton pair decay channel, the first observation of the Higgs boson coupling to leptons, which has a larger branching ratio than most other modes at low mass values [16, 17]. There was also a first result published by CMS in the $c\bar{c}$ quark pair decay channel [18], which follows a similar strategy to $b\bar{b}$, but since their smaller masses have an associated weaker coupling to the Higgs boson, they have a smaller associated branching ratio. The decay of a Higgs boson into $Z\gamma$ [19] has also been analyzed, even though it has a very small associated branching ratio. Very recently, evidence was found of the decay of a Higgs boson into a muon pair [20, 21], the first evidence for the decay of the Higgs boson to second generation fermions.

Finally we have the H → WW decay, which is an interesting channel to study the SM Higgs boson due to its large branching ratio, along with a clean signal from the isolated leptons which allows the study the Higgs boson in a variety of production modes. During Run 1 both ATLAS [22] and CMS [23] made an observation in this channel and during Run 2 CMS continued the work publishing results with 2016 data [24] and with the differential cross-section measurement using data of all three years [25]. The publication of the full run analysis containing the results of the different production modes is in progress and the work described in this thesis constitutes a relevant part of that analysis.

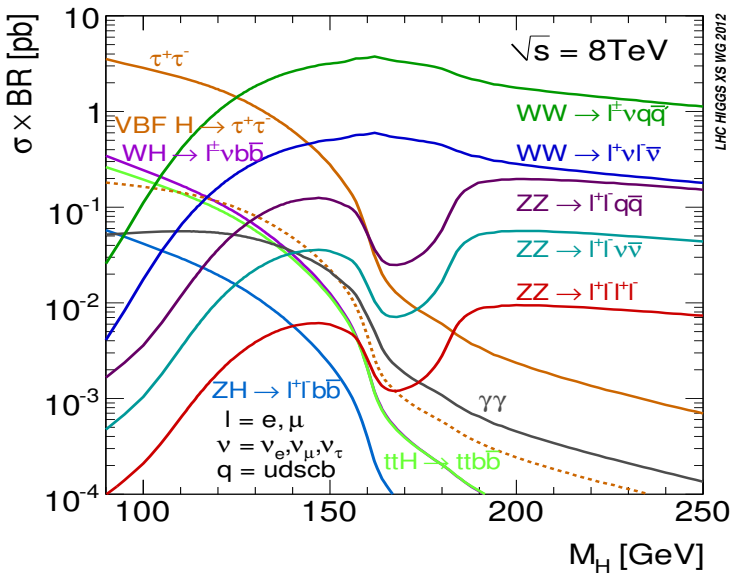


Figure 1.5: Higgs boson branching ratio multiplied by the decay cross sections at 8 TeV. The dark blue line represents the H → WW fully leptonic decay that is the focus of this work. Picture taken from [7].

Figure 1.5 shows the cross section times branching fraction for the different Higgs decay channels. b quark pair production is shown only for VH production, since it is the most accessible channel to observe experimentally. At a mass of 125 GeV, both H → WW channels are, after the gluon fusion produced H → $\tau\tau$, the decay processes with the highest cross sections. A W boson decays 67.41% ± 0.27 of the time into quarks and 10.86% ± 0.09 into each of the three lepton families [26]. The leptonic decay leads to more isolated final state particles, less background from QCD processes and clearer signals in our detector and therefore it is the one we chose for our work.

One last thing to note is that a Higgs boson with a mass close to 125 GeV has a mass less than twice the mass of a W boson (about 160 GeV) and therefore the decay into two real W bosons is not kinematically allowed. This means that at least one of them is produced *off-shell* and that technically the correct way to refer to the process is H → WW*.

1.4 A proton-proton collision: hard interaction, parton shower and hadronisation

There are several processes that occur between the beginning of a proton-proton collision and the measurement of particles in the final state. The first part of the collision depends on the *parton density functions* or PDF, which are the probability density functions that describe the amount of the proton momentum that a certain parton (i.e. quark or gluon) inside it may carry. If the incoming protons interact between them and there is a large momentum exchange between their partons there is a hard scattering. In that case, the energy of the process is high enough to be able to use perturbative QCD calculations, for what is then called a *hard interaction*. During the hard interaction between the quarks and gluons, resonances like the Higgs boson can be produced and decay. The matrix element that describes the parton-parton hard interaction can be obtained with Feynman diagrams, computing the contributions at different orders. Leading order (LO) and next-to-leading order (NLO) are the most used approximations, since even though higher orders are more accurate, the associated calculations makes them harder to compute.

Due to the strength of the strong interaction, partons can emit gluons at any stage in the process, either as *initial state radiation* (ISR) or as *final state radiation* (FSR). The emissions of radiation result in *parton showers*, where quarks emit gluons and gluons split into quark-antiquark or gluon-gluon pairs. With each emission or pair production the energy of the produced particles is lowered until the products reach an energy of around 1 GeV, where the production of more fundamental particles is no longer energetically favourable, non-perturbative QCD plays an important role and hadron formation becomes the dominant process.

In the final state the quarks cannot be found isolated, because of the colour confinement of QCD. Therefore, there is a *hadronisation* process in which quarks and gluons from the parton showers form colourless hadrons and mesons. This process has to be simulated using empirical models, since perturbative QCD is no longer valid at that scale. Most of these hadrons are unstable and they also decay afterwards.

Since protons in the beams travel in bunches containing a number of protons of the order 10^{11}, multiple proton interactions are very likely to happen. We refer to the additional interactions as *pile-up* interactions or simply 'pileup'. The consequences of the pileup are larger occupancies in the tracking and calorimeter subdetectors and added difficulties in reconstruction. Specially designed algorithms have been put in place in order to suppress the contribution from pileup, as it will be later explained in chapter 3. The number of interactions per bunch crossing is different for each data-taking period and has to be estimated separately. In Fig. 1.6 the pileup profile for each of the years is drawn and the average number of pileup interactions for each one is given. The average pileup was 23 for 2016 and 32 for 2017 and 2018.

Aside from the partons of the hard interaction other partons may also interact, with typically smaller transfers of momentum between them. These are *multiple parton interactions* (MPI), which together with the ISR, FSR, pileup events, beam remnants and detector noise constitute the *underlying event*, which represents everything that is not part of the hard interaction.

The matrix element calculations of the hard interaction, the parton shower and the hadronisation are the steps that are considered separately when simulating events of physical processes. Which generators are used for these simulations will be detailed in section 4.1.2.

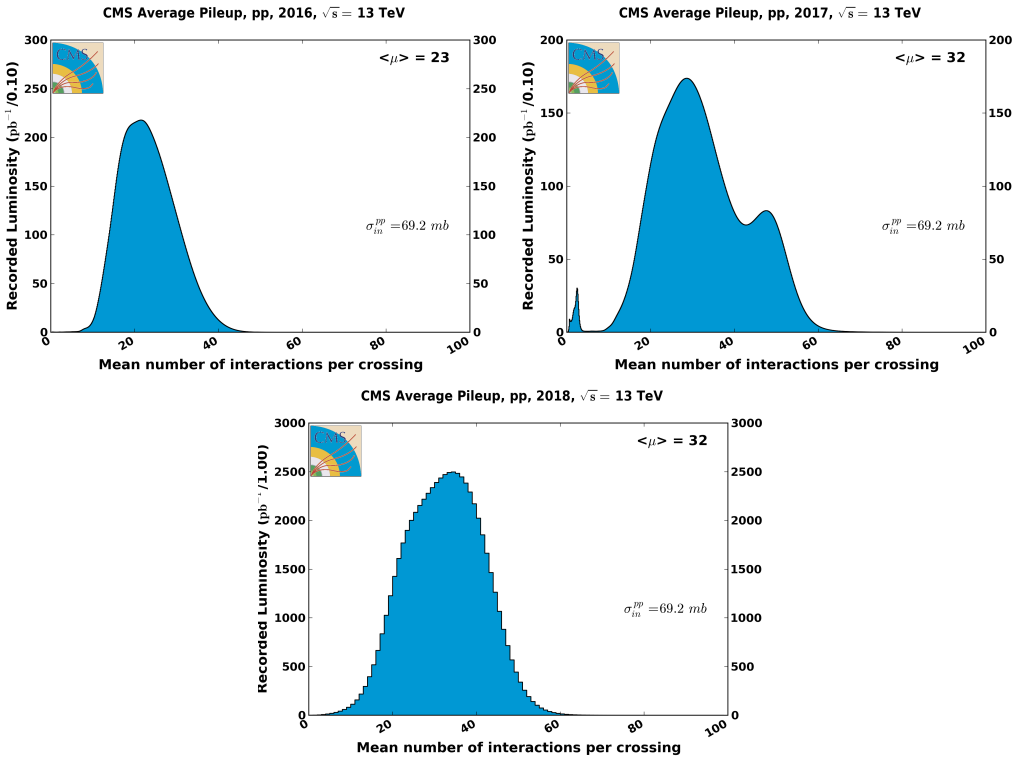


Figure 1.6: The pileup profiles during 2016, 2017 and 2018 data taking. Obtained from [27].

Chapter 2

The CMS detector and the LHC

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the highest energy particle collider in operation. Located at CERN (European Centre for Nuclear Research) near Geneva, Switzerland, it is a 27 km circular accelerator where two beams of protons or heavy ions circulating in opposite directions inside parallel vacuum pipes collide in four interaction points. It entered into operation in 2009, and since then it has provided large data samples of proton-proton collisions at a centre-of-mass energy of 2.36, 7, 8 and 13 TeV, as well as smaller datasets of heavy ion collisions at a centre-of-mass energies of 2.76 and 5.0 TeV.

The magnet system of the LHC includes more than 50 types of magnets. The main ones used to bend the particle beams are the dipoles, which generate a 8.3 T magnetic field. There are 1232 dipoles, each 15 m long and made from a niobium-titanium alloy that becomes superconducting at a temperature of 1.9 K, which is reached at the LHC by using superfluid helium. To keep the beam from spreading, 392 quadrupole magnets squeeze the beam horizontally and vertically. Hundreds of other higher order magnets (sextupoles, octupoles, decapoles and dodecapoles) further tune the magnetic field to keep the proton or ion beams in very precise trajectories [28].

The LHC is the last of a series of steps in the accelerating chain through which the particles acquire the operating energies. The CERN accelerator complex is shown in Fig. 2.1. Ionised hydrogen atoms are produced and injected to the LINAC2 where they are accelerated with electric fields to an energy of 50 MeV and then passed through a series of circular synchrotrons. Room-temperature electromagnets are used in these to reach a maximum energy of 1.4 GeV inside the BOOSTER and 26 GeV at the Proton Synchrotron (PS). From the Super Proton Synchrotron (SPS) they are injected with an energy of 450 GeV into the LHC, where they reach their final collision energy.

The trajectory of the two beams is focalised and made to intersect at an angle at four points in the LHC, where each of the LHC experiments is located. They are CMS [29], ATLAS [30], ALICE [31] and LHCb [32]. They were built with different geometries and using detector systems purposely designed for their different physics analysis goals.

ATLAS (*A Toroidal LHC ApparatuS*) and CMS (*Compact Muon Solenoid*) are two general purpose detectors, situated in opposite sides of the LHC ring. Their physics objectives are very diverse, with focus on discovering and later studying the Higgs boson, Beyond Standard Model (BSM) physics searches (searches for evidence of supersymmetry, extra dimensions or dark matter, among others), high precision measurements of the SM processes and exploring the physics at the TeV scale in general. The two experiments performing their analyses independently allows the cross-checking of physics results.

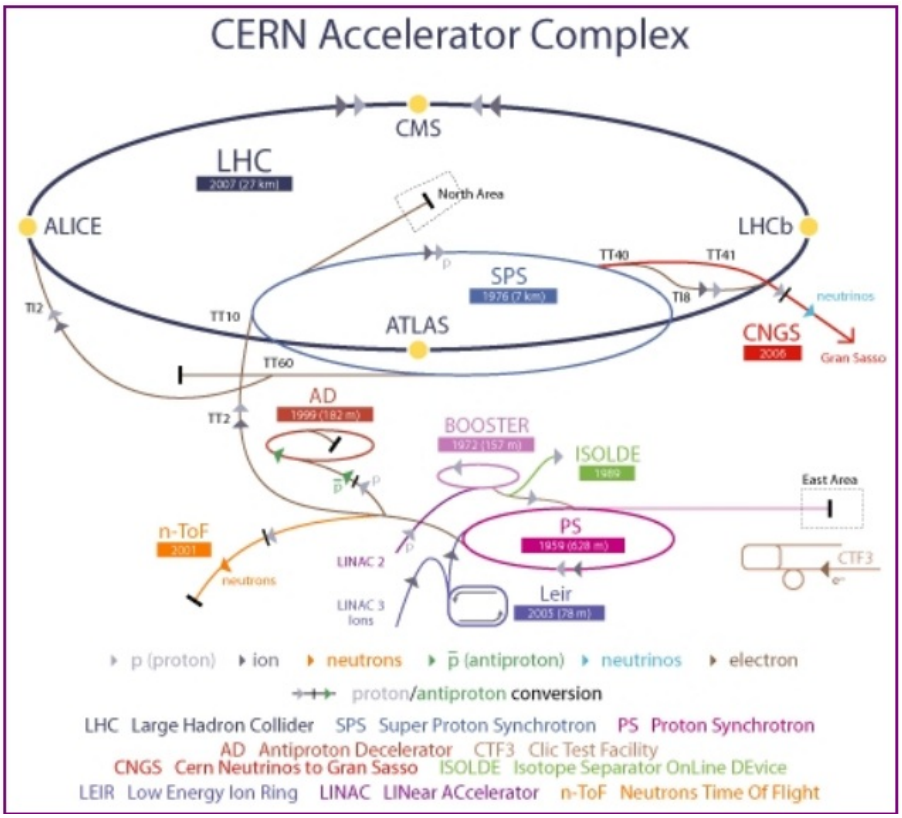


Figure 2.1: The CERN accelerator complex. The sequential accelerating steps towards the LHC are shown.

The LHC also accelerates and collides beams of heavy ions like lead or xenon, with specialised proton-ion or ion-ion collision runs. The high energy densities create a situation where quarks inside the hadrons are freed from their bond with gluons in a state known as quark-gluon plasma (QGP). All LHC experiments have a heavy ion physics program, but among them ALICE (*A Large Ion Collider Experiment*) is the one focused primarily on these collisions and the study of the properties of QGP.

The LHCb (*LHC beauty*) experiment focuses on CP-violation and the matter-antimatter asymmetry of the universe by studying the collisions in which a hadron containing a bottom quark is created. It has a different angular coverage due to its asymmetric configuration, which allows it to perform complementary analyses to those done by CMS and ATLAS.

2.1.1 Luminosity

The *instantaneous luminosity* $\mathcal{L}_{inst}$ is one of the most important parameters of the accelerator, which determines the amount of collisions that are produced per unit of time and of area. It is characterised by several parameters of the accelerator, according to its design and operation:

$$\mathcal{L}_{inst} = f\frac{n_1 n_2}{4\epsilon\beta^*}\ , \tag{2.1}$$

where those parameters are the number of particles per bunch n_1 and n_2, the frequency of bunch crossings f, the transverse emittance ϵ and the amplitude function β^*. The frequency is 40 MHz, and the number of particles in each bunch is n= 1.15×10^{11}. The nominal instantaneous luminosity is of the order of 10^{34} cm^{-2}s^{-1}, with a maximum obtained during Run 2 of 2.06 $\times10^{34}$cm$^{-2}s^{-1}$.

The other two parameters describe the beam quality, considering the bunch preparation and the accelerator magnet configuration. The emittance of a beam ϵ reflects how close two protons in a bunch are in the position-momentum phase space. On the other hand, the amplitude function β^* represents how much the beam squeezes in a short length. A low value in these parameters is needed to achieve high luminosities. They are both related to the size of the bunches, since they are directly proportional to the bunch sizes in the transverse directions, $\epsilon\beta^* = \pi\sigma_x\sigma_y$. Typical values for these parameters during Run 2 were ϵ =3.75 mm μrad and β^* =0.55 m, but they are slightly tuned during the data taking periods.

The number of events expected for a process with a cross section σ is given by

$$\frac{dN}{dt} = \sigma\mathcal{L}_{inst}, \tag{2.2}$$

and therefore, a high luminosity is needed in order to study physical processes with a low production cross section. The total number of collisions over a time period is proportional to the *integrated luminosity*, which is obtained separately for every run period:

$$\mathcal{L}_{int} = \int \mathcal{L}_{inst}dt \tag{2.3}$$

The instant luminosity is measured in the hadronic forward calorimeters and calibrated using van der Meer scans [33]. The scans are performed by varying the separation of the beams, and the resulting profile of luminosity as a function of that separation is fitted in order to determine the beam overlap width, from which the luminosity is calculated. Figure 2.2 shows the total luminosity recorded by the CMS experiment corresponding to the dataset of three years that is used in the H $\rightarrow$ WW analysis.

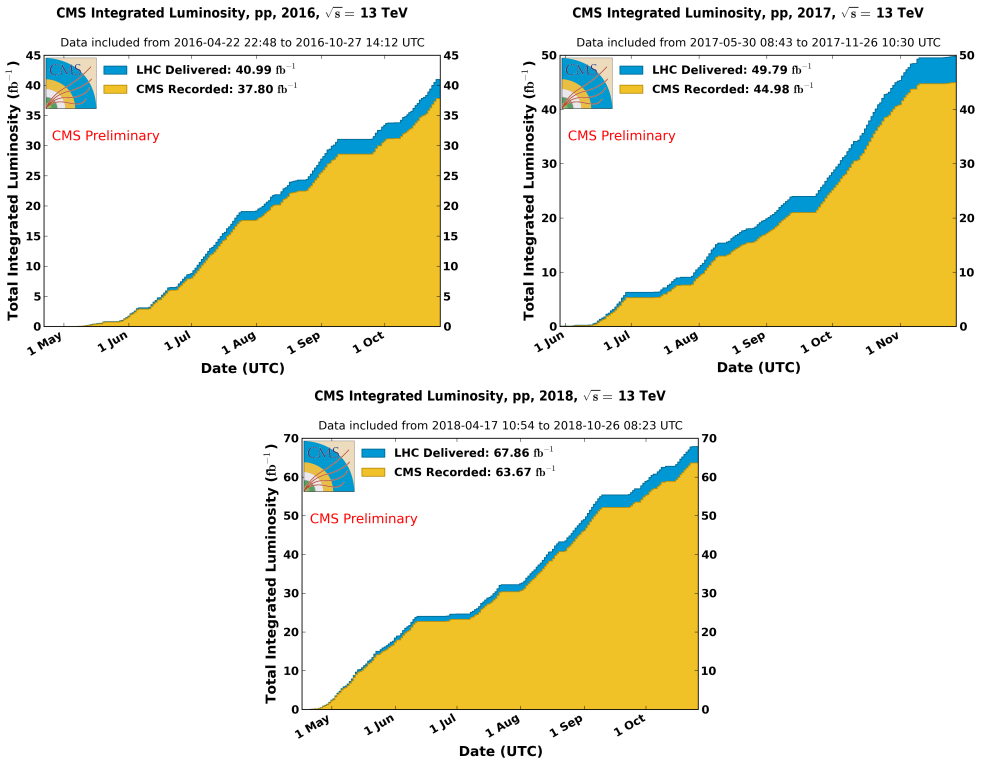


Figure 2.2: The luminosity recorded by CMS during 2016, 2017 and 2018 data taking. Obtained from [27].

2.2 The Compact Muon Solenoid experiment

The central element of the CMS detector [29] is its solenoid, a superconducting cylindrical coil that creates a homogeneous magnetic field of 3.8 T inside it, bending the trajectories of the particles that are produced at the collisions. Specialised particle detectors (subdetectors) are situated in concentric layers around the interaction point in a compact and hermetic manner to maximise the detection of the collision products.

The central zone of the detector is divided in five cylindrical slices along the Z axis, as shown in Fig. 2.3. They are called *wheels*, and are numbered from -2 to 2, where 0 is the central one where the interaction point is located. Almost all subdetectors are inside the wheels, with the exception of the forward calorimeters and forward muon chambers. The region corresponding to $|\eta|< 1.479$ is called *barrel* and the two regions on the sides with $|\eta|> 1.479$ are the *endcaps* sections. Different detector technologies are chosen for each of the detector regions, since the experimental conditions in terms of particle flux and radiation levels change between them.

2.2.1 Coordinate system in CMS

The coordinate system adopted by CMS is centered at the collision point inside the experiment. The Y-axis points vertically upward, the X-axis points toward the center of the LHC and the Z axis points along the beam direction, as it is shown in Fig. 2.4. Given the cylindric configuration of the detector, it is useful to use the radial distance R, azimuthal angle ϕ in the XY plane and the polar angle θ measured from the Z axis.

Particles produced in the collisions tend to be boosted along the beam direction, and angle differences in θ are not invariant under those boosts. We define the *rapidity* y as:

$$y = \frac{1}{2} \ln \left(\frac{E + p_z c}{E - p_z c} \right) , \tag{2.4}$$

where E is the energy of the particle and p_z its longitudinal momentum. Rapidity differences are invariant under boosts in the Z direction.

Using the angle ϕ, the rapidity y and the momentum in the plane orthogonal to the beam line (*transverse momentum*) p_{T}, the momentum of a particle can be written as:

$$p^{\mu} = (E, p_x, p_y, p_z) = (m_T \cosh y, p_T \sin \phi, p_T \cos \phi, m_T \sinh y) , \tag{2.5}$$

where the *transverse mass* is defined as $m_T = \sqrt{p_T^2 + m^2}$.

Experimentally, rapidity is hard to measure precisely for highly relativistic particles since both the total energy and the transverse momentum are needed. However, for ultra-relativistic particles approaching the limit of a negligible mass with respect to its momentum, the rapidity can be approximated with a quantity called the *pseudorapidity* or η, defined as:

$$\eta = - \ln \left[\tan \frac{\theta}{2} \right] , \tag{2.6}$$

which does not depend on the energy or the momentum of the particle and is therefore easier to measure. Low values of the pseudorapidity correspond to particles with almost all of their momentum in the transverse plane, whereas high values correspond to particles moving in the forward direction, close to the beam line, as seen in Fig. 2.5.

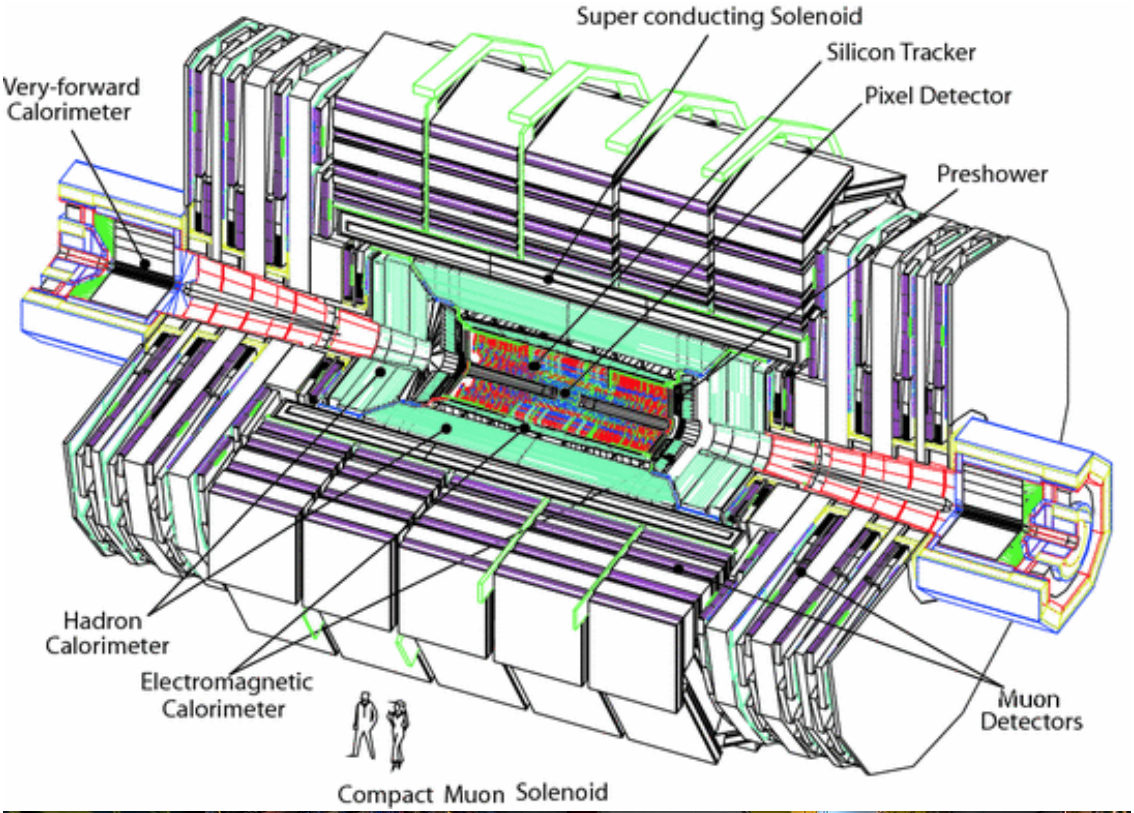


Figure 2.3: Above: A schematic of the detector in the cylindrical RZ plane. Below: A view of one of the external wheels of the CMS experiment, with the beam pipe and the hadronic calorimeter in display.

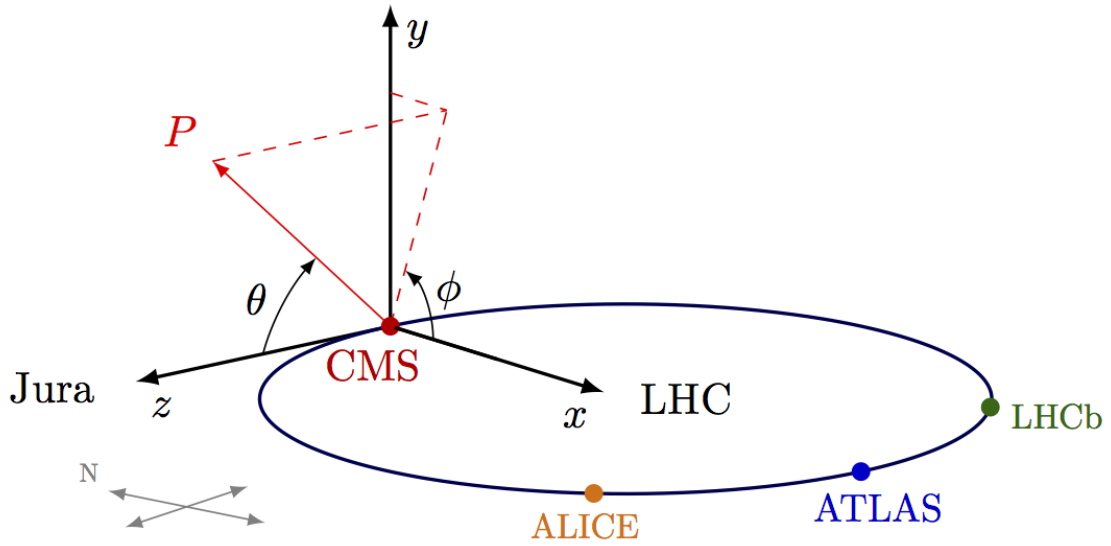


Figure 2.4: The LHC coordinate system at CMS.

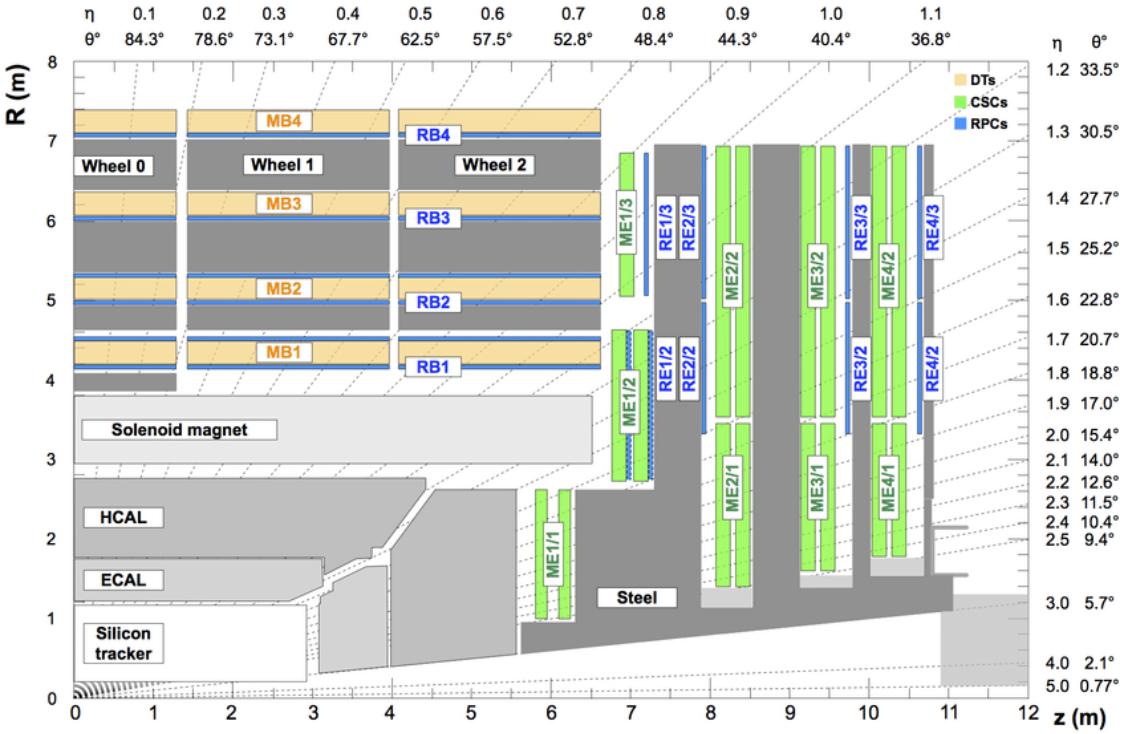


Figure 2.5: A quadrant of the CMS detector, where the positions are given in terms of both the polar angle θ and the pseudorapidity η. Picture taken from [34].

2.2.2 The Tracker

The innermost part of the CMS detector is the tracker system [35, 36], shown in Fig. 2.6. It is built with silicon, as it resists well the high radiation levels produced in the collisions and provides very precise measurements. The tracker consists of two main detectors: a pixel detector, covering the region from 4 to 15 cm in radius, and with |Z|<49 cm, and a strip detector, covering the region from 25 to 110 cm in radius, and with |Z|<280 cm.

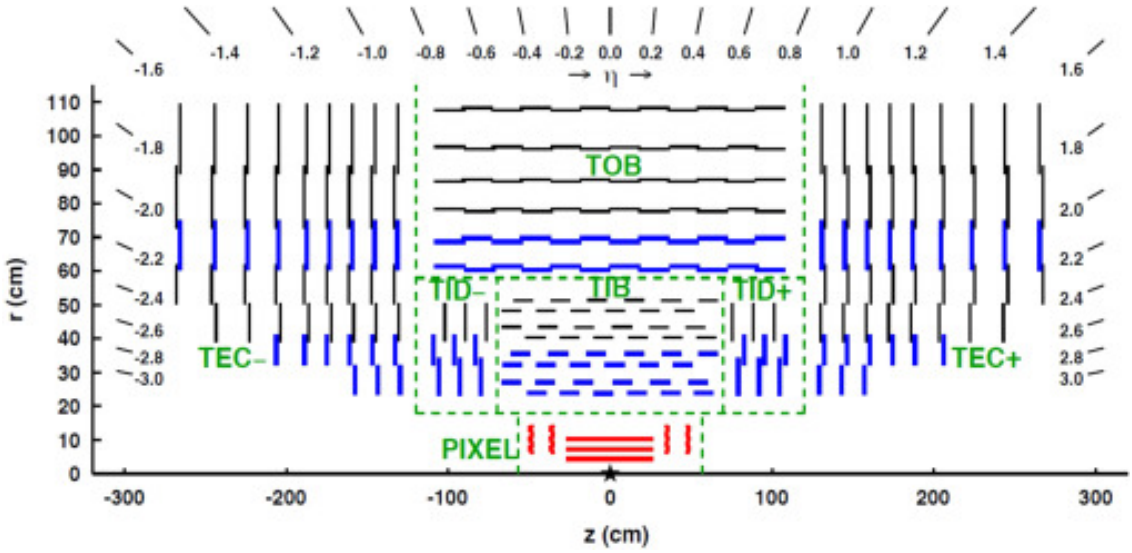


Figure 2.6: Layout of the CMS tracker system in the 2016 configuration. The silicon pixel detector is in red and the silicon strips in black or blue. Each line represents a module, and the blue strips indicate back-to-back modules that also measure the position in the orthogonal coordinate.

The CMS silicon pixel detector has 66 million active elements instrumenting a surface area of about 1 m^2. The detector consists of three concentric cylindrical barrel layers and four fan-blade disks which close the barrel ends. The barrel layers are located at average radii of 4.3, 7.3, and 10.2 cm and the endcap disks instrument the regions between radii 4.8 and 14.4 cm at mean longitudinal distances of 35.5 and 48.5 cm from the interaction point. The structure provides efficient three-hit coverage in the region of pseudorapidity $|\eta|$ <2.2 and efficient two-hit coverage in the region $|\eta|$ <2.5. The active elements are 100 μm × 150 μm pixels which are oriented with the smaller pitch in the azimuthal direction in the barrel and in the radial direction in the disks. These pixelated detectors produce 3-D measurements along the paths of the charged particles with single hit resolutions between 10 and 20 μm.

The CMS silicon strip detector has 9.3 million active elements instrumenting a surface area of 198 m^2. The detector consists of three large subsystems. The Tracker Inner Barrel and Disks (TIB/TID) extend in radius to 55 cm and are composed of four barrel layers, supplemented by three disks at each end. The TIB/TID delivers up to four R-ϕ measurements on a trajectory using 320 μm thick silicon microstrip sensors, which have their strips oriented parallel to the beam axis in the barrel and oriented radially in the disks. The strip pitch is 80 μm in the inner pair of TIB layers and 120 μm in the outer pair of TIB layers. In the TID, the mean pitch varies between 100 μm and 141 μm. The TIB/TID is enclosed within the Tracker Outer Barrel (TOB), which has an outer radius of 116 cm. The TOB consists of six barrel layers of 500 μm thick microstrip sensors with strip pitches of 183 μm in the first four layers and 122 μm in the last pair of layers.

The TOB extends to 118 cm in Z. Beyond this Z range, the Tracker EndCaps (TEC) instrument the region 124<|Z|<280 cm and 22.0<R<113.5 cm. Each TEC is composed of nine disks that are instrumented with up to seven rings of radial-strip silicon detectors. The sensor thicknesses are thin (320 μm) in the inner four rings and thick (500 μm) in the outer three rings. The

inner two layers of the TIB and TOB, the inner two rings of the TID and TEC, and the fifth ring of the TEC include a second microstrip detector module that is mounted back-to-back at a stereo angle of 100 mrad and that enables a measurement of the orthogonal coordinate. Assuming fully efficient planes and not counting hits in stereo modules, there are from 8 to 14 high precision measurements of track impact points for $|\eta|$ <2.4.

The main purpose of the tracker detectors is to provide a measurement of the momentum of charged particles with excellent resolution. Due to its high segmentation, the pixel detector not only forms high quality seeds for the track reconstruction algorithm offline, but is also used to do fast tracking online for primary vertex reconstruction, electron/photon identification, muon reconstruction, tau identification and b-tagging.

The average track reconstruction efficiency for promptly-produced charged particles with transverse momenta of p_T >0.9 GeV is 94% for pseudorapidities of $|\eta|$<0.9 and 85% for 0.9<$|\eta|$<2.5. The inefficiency is caused mainly by hadrons that undergo nuclear interactions in the tracker material. In the central region, tracks with 1<p_T < 10 GeV have a resolution in p_T of approximately 1.5%. The resolution in their transverse (longitudinal) impact parameters improves from 90 μm (150$^a\mu$m) at p_T = 1 GeV to 25 μm (45 μm) at p_T = 10 GeV. For isolated muons, the corresponding efficiency for muons with p_T >0.9 GeV is essentially 100%. Isolated muons of p_T = 100 GeV emitted at $|\eta|$<1.4 have a p_T resolution of approximately 2.8% and 10 μm (30 μm) in the transverse (longitudinal) impact parameters. The position resolution achieved for reconstructed primary vertices is 10–12 μm in each of the three spatial dimensions.

The pixel tracker was upgraded during the technical stop at the end of 2016 and its current configuration was used during 2017 and 2018. The 'Phase-1' [37] version of the pixel tracker barrel has four layers of 124 million pixels in total, corresponding to about 2 m^2 total area. The inclusion of a fourth layer with a minimal radius improved the pattern recognition and track reconstruction [38].

Before the start of the HL-LHC [39] both the strip tracker and the Phase-1 pixel detector will be replaced due to the significant damage and performance degradation they suffered during operation of the LHC, and to cope with the more demanding operational conditions[40].

2.2.3 The Electromagnetic Calorimeter (ECAL)

The next subdetector is the electromagnetic calorimeter or ECAL [41, 42]. It is a homogeneous calorimeter made of 61200 lead tungsten ($PbWO_4$) scintillating crystals that absorb the photon and electrons produced in the collision. Using the location and magnitude of the energy deposits, the direction and energy of photons and electrons are measured.

The ECAL is divided in three regions, as it is shown in Fig. 2.7. A central region or barrel (EB), with pseudorapidity coverage of $|\eta|$<1.48, two endcaps (EE) that increase the coverage to $|\eta|$<3 and a preshower detector (ES) in front of the endcaps (1.65 <$|\eta|$<2.6), which is made of a lead absorber and silicon strip sensors. It is used to identify single energetic photons in the endcap region, separating them from close pairs of photons from neutral pion decays.

The crystals have a truncated pyramidal shape and are positioned in a way that minimises inter-crystal gaps aligned with the trajectories expected for the particles. EB uses 23 cm long crystals with front face cross sections of around 2.2 cm × 2.2 cm, whereas the EE uses 22 cm long crystals with front face cross sections of 2.86 cm × 2.86 cm. The ES has two planes of silicon strips, the first one vertically aligned and the second horizontally, to provide accurate position measurement in both coordinates.

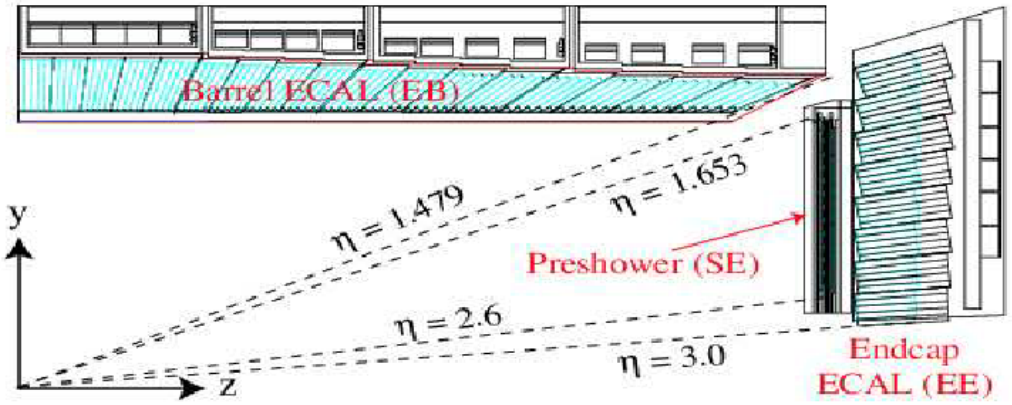


Figure 2.7: Longitudinal view of the three parts of the ECAL, with their respective pseudorapidity coverage. Picture taken from [43].

The electromagnetic cascades produced inside the crystals by the passing particles are absorbed and the scintillation light of the crystals is registered using photodetectors. Avalanche photodiodes (APDs) [44, 45] and vacuum phototriodes (VPTs) [46] are used as photodetectors in the EB and EE, respectively. The high-density (8.28 g/cm^3), short radiation length (X_0 = 0.89 cm), and small Molière radius (R_M = 2.2 cm) of $PbWO_4$ allow the construction of a compact calorimeter with fine granularity.

The energy resolution for electrons and photons in the ECAL is estimated from Z boson decays to electron pairs and to muon pairs plus a photon, respectively. It is better than 2% in the central region of the ECAL barrel (for pseudorapidity $|\eta|$ <0.8) and is 2–5% elsewhere. For photons the energy resolution for photons from 125 GeV Higgs boson decays varies across the barrel from 1.1% to 2.6% and from 2.2% to 5% in the endcaps. The calibration of the absolute energy is determined from $Z \rightarrow ee$ decays to a precision of 0.4% in the barrel and 0.8% in the endcaps.

The ECAL alignment and position resolution measurement is performed with isolated electrons from W boson decays using both the ECAL and tracker information. The achieved position resolution in EB (EE) is 3 (5) mrad in ϕ and 10^{-3} (2×10^{-3}) units in η.

2.2.4 The Hadronic Calorimeter (HCAL)

The last subdetector inside the solenoid as seen from the interaction point is the hadronic calorimeter or HCAL [47, 48]. It is a sampling calorimeter, with brass and steel as the passive material, and plastic scintillator tiles serving as active material. In the dense layers of the passive material the hadrons interact, and in the plastic scintillator a rapid light pulse is produced after a particle passes through. This light is collected using optic fibers and delivered to readout boxes where photosensors integrate the signal and send it to the trigger and data acquisition systems. The total amount of light produced by a shower is a measure of the energy of the particle that initiated that shower.

The HCAL consists in four different subsystems, as it is shown in Fig. 2.8. The inner barrel (HB) covers the pseudorapidity range $|\eta|$<1.4, with the endcaps (HE) covering 1.3 <$|\eta|$<3.0. Both the HB and the HE are divided in 18 wedges, each covering a 20 degree angle in ϕ. The passive material in them is brass, which in the HE takes the form of disks.

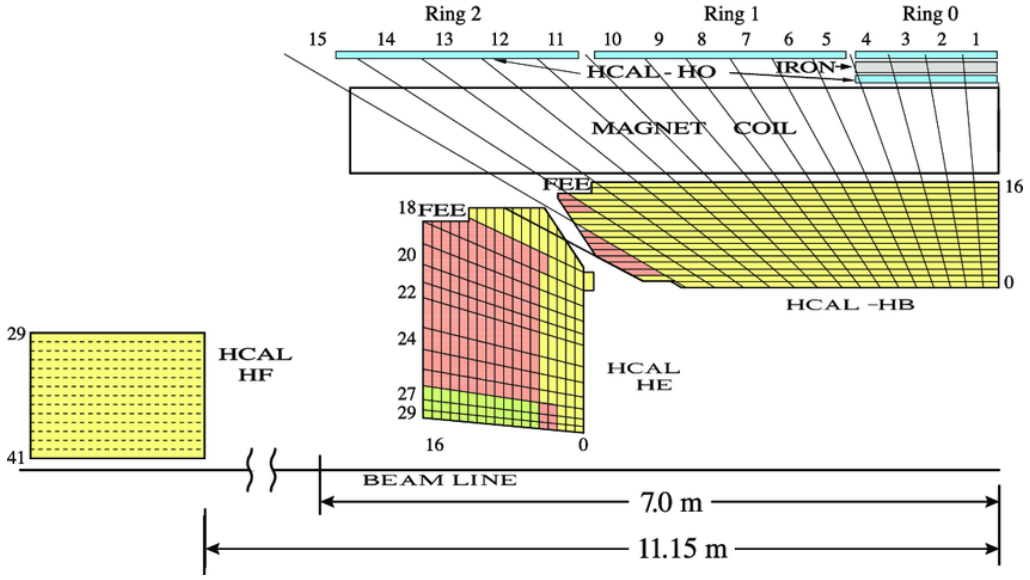


Figure 2.8: A quarter of the CMS HCAL, with the position of its four subsystems. Picture taken from [49].

Because of its limited space due to the magnet cryostat, the HB thickness is limited to 5.8 hadronic interaction lengths at $\eta = 0$ and increases to 10 interaction lengths at $|\eta| = 1.2$. To control and measure a possible energy leakage from HB, there are additional layers of scintillators outside the solenoid cryostat forming the outer barrel (HO) that use the magnet itself as passive material. About 5% of hadrons above 100 GeV deposit energy in HO, which covers the $|\eta|$ range between 0 and 1.2 forming a five ring structure that follows the wheels in the barrel. In the HO polyvinyl toluene (PVT) scintillators are used as the active material and there are two layers instead of one in the central ring where more leakage is expected.

The quartz fiber and steel hadron forward calorimeter (HF) [50] is located outside the detector wheels and covers a high η region, $3.0 < |\eta| < 5.2$. The active material in the HF are quartz fibers embedded in the steel, which emit Cherenkov light when the particles pass through. Long fibers (about 10 interaction lengths) measure the total signal coming from the full material length, whereas short fibers measure the energy deposition after 22 cm of steel. This allows the separation of showers generated by electrons and photons from those generated by hadrons. Since the magnetic field intensity in the HF is significantly smaller than in the HB and HE, conventional photomultipliers are used to register the Cherenkov light.

2.2.5 The Muon System

The muon system [51] is located outside the solenoid, inserted among the layers of the steel magnetic flux-return yoke. There are three subsystems that use different techniques, all based on gas ionisation chambers. The muons travelling through the chambers cause ionisation that is directed to an anode and converted into an electric signal. The three subdetectors are the *drift tube chambers* (DT), *resistive plate chambers* (RPC) and *cathode strip chambers* (CSC), and their position is depicted in Fig. 2.9 . The combination of all these systems provide a highly efficient muon detector that covers the $|\eta|$ region up to 2.4. The spatial resolution is 200–350 μm in the DT, 40–150 μm in the CSC, and 0.8–1.4 cm in the RPC, and the time resolution is 3 ns or better per chamber for all 3 systems. The efficiency for reconstructing hits originating from muons traversing the muon chambers is in the range of 94–99% and the segment reconstruction efficiency obtained by the muon track segments sections of the DT and CSC is over 96%.

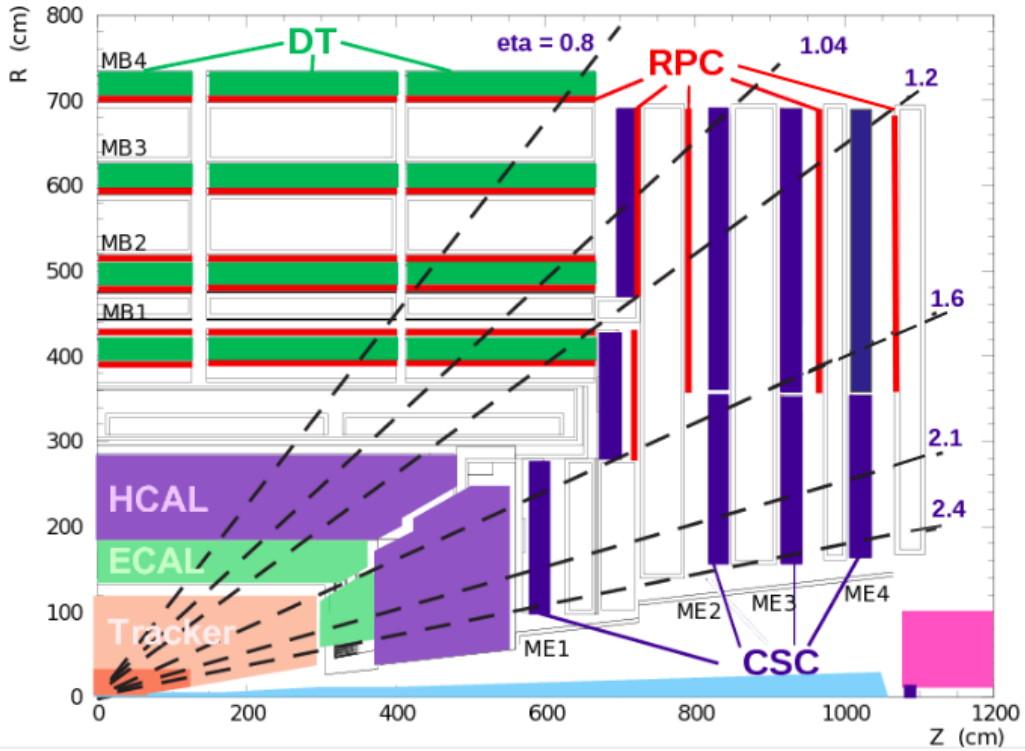


Figure 2.9: The CMS detector, with its position in cylindrical coordinates. Only one quadrant of the CMS detector is drawn here, but the rest is symmetrical with respect to the R and Z axes. The interaction point is at the lower left corner. The muon systems are shown in colours: drift tubes (DT) in green, resistive plate chambers (RPC) in red and cathode strip chambers (CSC) in blue. The grey areas represent the solenoid and the steel flux-return disks.

Drift Tubes

The drift tube chambers are located in the barrel wheels and cover the $|\eta|$ region up to 1.2. The basic element of the chambers is a drift tube cell, depicted in Fig. 2.10. The DT cell has a central wire at a high voltage (3600 V), two electrodes (cathodes) on the sides at -1800 V, and two electrodes above and below the wires at +1800 V. This setup combined with the gas mixture it is filled with (85% of Ar and 15% of CO_2, slightly overpressured by 2-10 mbar [52]) provides a saturated electron drift velocity of about 55 μm/ns. Considering the size of the cell (42 mm), the maximum drift time is almost 400 ns.

The chambers consists of twelve layers of cells, grouped as shown in Fig. 2.11. Those twelve layers are in turn divided into three *superlayers*, where the first and the third are oriented in the ϕ direction and the second is orthogonal to them (in the Z direction). The chamber structure in Fig. 2.11 shows the position of the three superlayers with their staggered layers, as well as the honeycomb spacer between superlayers two and three, which is a structural element.

Fig. 2.12 shows a transverse view of the DT barrel detector, divided in twelve sectors in azimutal angle. In each sector there are four concentrical chambers or stations, labeled MB1 to MB4, starting from the inside (MB stands for *muon barrel*). Besides, the MB4 stations in sectors 4 and 10 at the top and bottom of CMS are made up from two chambers each. These additional chambers are referred to as sectors 13 and 14. There are in total 250 chambers (5×4×12 + 5×2). The dimensions of each chamber depend on the sector and station, and they range from 1.99 m × 0.29 m × 2.54 m in the smallest MB1 chambers to 4.19 m × 0.29 m × 2.54 m in the largest MB4 chambers. In the outer MB4 stations there is no Z superlayer in the chambers, just those intended to measure the ϕ variable.

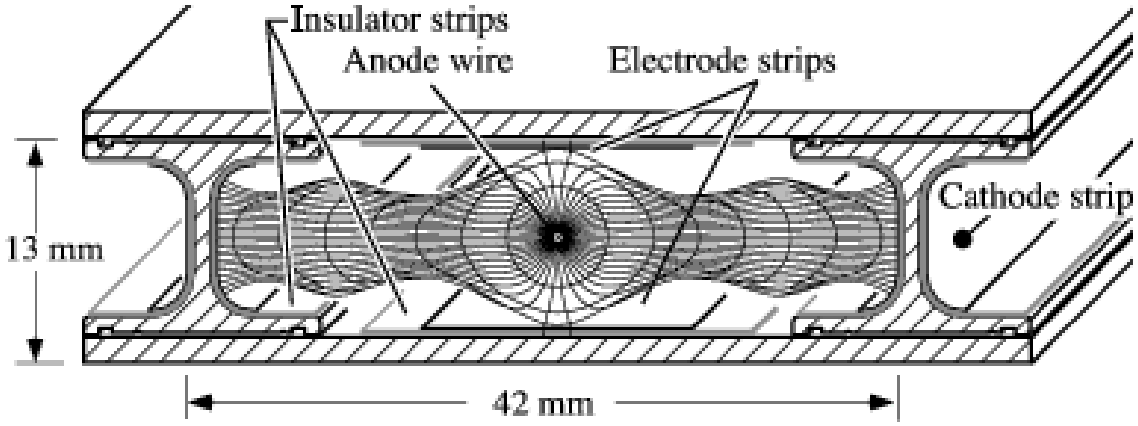


Figure 2.10: The basic structure of the DT chambers, a cell. The anode wire at the centre is set to a high voltage and the cathodes strips on the sides and the electrode strips at the top and bottom are set to lower voltages.

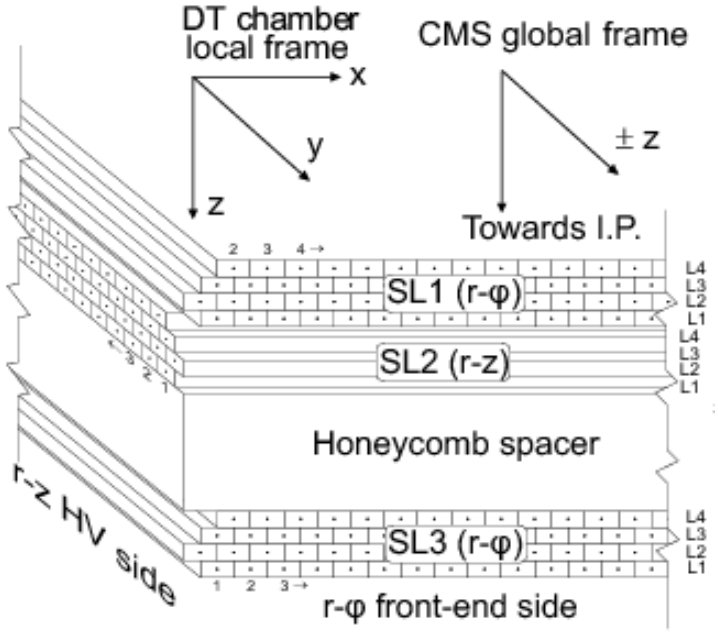


Figure 2.11: Layout of a DT chamber. Twelve layers (L) are located in three superlayers (SL), measuring in two orthogonal directions. Each of the pointed units in a layer is a cell with its wire. There is a honeycomb spacer between SL1 and SL2.

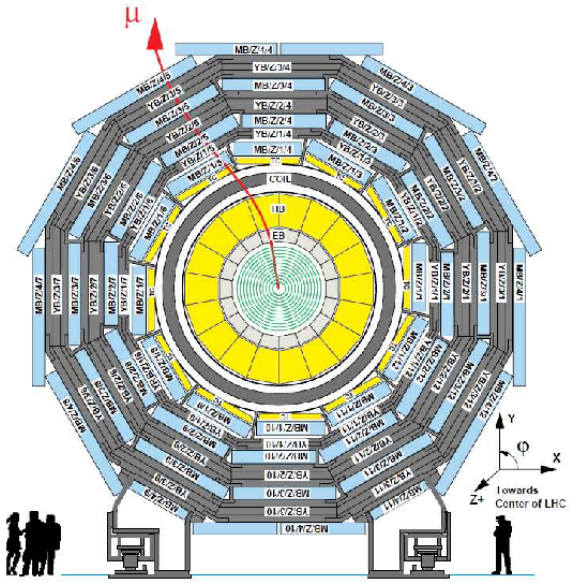


Figure 2.12: The muon DT chambers in a CMS barrel wheel. The chambers are denoted as MBZ/N/S, where Z is the barrel wheel number between -2 and +2, N the station number between 1 and 4, and S is the sector number between 1 and 12.

Cathode Strip Chambers

In the endcap regions of CMS the muon rates and background levels are higher, and the magnetic field is strong and non-uniform. For this reason, cathode strip chambers (CSC) are installed since they have fast response time, resulting from a short drift path, they can be finely segmented, and they tolerate better the non-uniformity of the magnetic field.

The CSC have a trapezoidal shape and they are arranged in a series of concentric rings centered on the beam line. The stations are separated by the iron disks of the flux return yoke, as depicted in Fig. 2.9. They increase the coverage of the $|\eta|$ region from 0.9 up to 2.4. Each CSC chamber consists of 6 staggered layers, sandwiched between cathode panels, which independently measure the muon position in 2 coordinates, R and ϕ. The cathode strips are oriented radially to measure the muon position in the bending plane (R-ϕ), whereas the anode wires provide a coarse measurement in R. The intersection point of the cathode strips and the anode wires with a signal determines the hit positions.

The CSC have four stations located in a direction perpendicular to the beam, labeled ME1-ME4. The first one is composed of three rings of chambers in the radial direction and the others are composed of two. In the inner rings of stations 2, 3, and 4, each CSC chamber subtends a ϕ angle of 20 degrees, whereas the other CSC subtend an angle of 10 degrees (18 and 36 chambers in each ring, respectively). The outermost chambers are about 3.4 m long and 1.5 m wide. All chambers use a gas mixture of 50% CO_2, 40% Ar, and 10% CF_4. The ME1/1 chambers are operated at an anode voltage of 2900 V and all others at 3600 V. The ME1/1 chambers, the closest to the collision are different, since the magnetic field in them is stronger. Therefore, the sagitta measurement at the first station is crucial and motivates more stringent requirements on the resolution and alignment in this station than in the other stations. This is done by using narrower strips and dividing ME1/1 into 2 regions so that the region closest to the beam line can trigger and be read out independently of the outer region.

Resistive Plate Chambers

CMS includes a subdetector system with excellent time resolution dedicated to reinforce the measurement of the correct beam crossing time at the highest LHC luminosities. The resistive plate chambers (RPC) are mounted in both the barrel and endcap regions, and they provide a fast, independent trigger with a lower p_{T} threshold over a large portion of the pseudorapidity range ($|\eta|<1.9$).

The RPC are double-gap chambers operated in avalanche mode and are primarily designed to provide timing information for the muon trigger. An RPC gap is made by two parallel bakelite plates placed at a distance of 2 mm and filled with a gas mixture of 96.2% $C_2H_2F_4$, 3.5% C_4H_{10} and 0.3% of SF_6. A charged particle passing through the RPC produces an avalanche of electrons in the gap between the two plates. This charge induces a signal on an external strip readout plane to identify muons from collision events with a precision of a few ns. Since a muon passing through the RPC induces a signal on more than one strip, the center of the cluster determines the position of the RPC hit.

They are organised in stations attached to the DT and the CSC. As Fig. 2.9 shows, there are four stations in the barrel (RB1-RB4), and three stations in the endcaps (RE1-RE3), with the innermost barrel stations containing two layers of RPC at both sides of the DT chambers. The rest of the chambers are single-layer only.The RPC strips are used to measure the coordinate in the bending plane. This means that the strips are oriented parallel to the wires of the DT chambers in the barrel, and they are oriented parallel to the CSC strips in the endcaps. The RPC are grouped in wheels/sectors or in rings just as the corresponding DT and CSC chambers.

2.3 The CMS Trigger System

The high luminosity at which the LHC operates yields a high rate of data being produced. The nominal rate of collisions is 40 MHz, but not every collision is interesting for study, with only a fraction of them containing events that could be used for BSM searches or SM precision studies of the CMS physics program. Moreover, the rate of collected events has to be lowered down significantly in order to be stored. The filtering is done by the *Trigger System* [53], which greatly reduces the information received to an amount that can be stored, processed and analyzed. The system functions in two steps: the Level 1 Trigger (L1T), which is hardware-based, and the High Level Trigger (HLT), which is software-based.

2.3.1 The Level 1 Trigger

The L1T takes information from the subdetectors and selects events at a rate of 100 KHz from the incoming 40 MHz, making decisions within 4 ns after the collision. It uses the input from the three muon detectors, as well as the ECAL and HCAL to select events containing candidate objects, such as ionisation deposits consistent with a muon or energy clusters consistent with an electron, photon, tau lepton or jet.

Figure 2.13 shows the schematic structure of the L1T. In the muon detectors, first segments are formed in the CSC and DT using the hits in the chambers (local trigger). Then, the information goes to the track finders that consider other chambers an apply pre-established pattern recognition algorithms to reconstruct the path, obtain muon candidates and measure their transverse momentum from the bending on the transverse plane. The position in a chamber in the outer stations is extrapolated from the information of the inner stations and if it agrees within a margin with a detected impact on the outer stations, a muon candidate is formed. The number of candidates is sorted and reduced, taking into account duplicated tracks from the same muon.

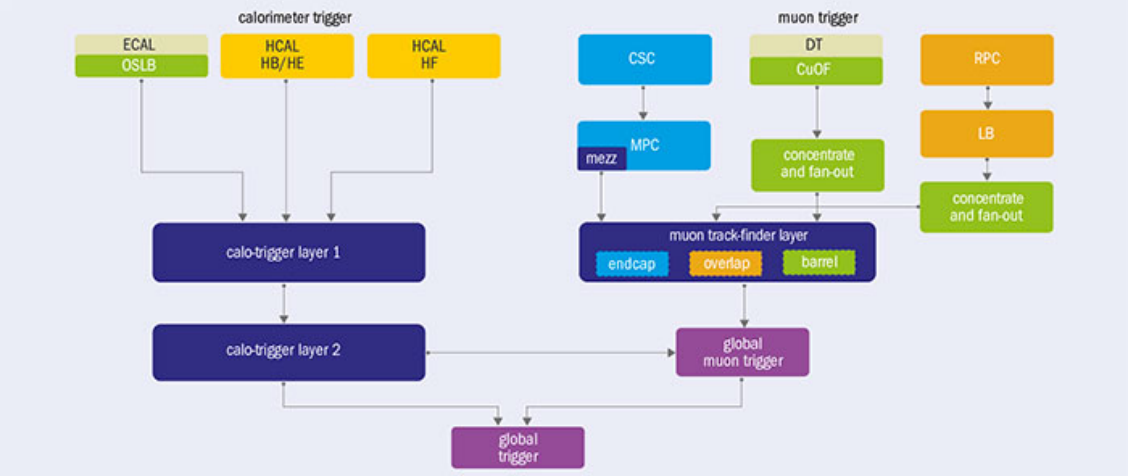


Figure 2.13: The schematic structure of the Level 1 Trigger. The muon systems find tracks in the barrel and endcaps separately and then combine their information. The ECAL, HCAL and HF trigger on events with electrons, photons, jets and missing energy. The combination of all that provides the global L1 trigger acceptance signal.

The RPC provides an additional trigger with a very good timing resolution that complements the other two muon subdetectors in finding tracks. It does not form segments, but the hits are used for recognising muon candidates in different chambers, then eliminating duplicates and selecting those with the best quality and the largest transverse momentum. Each candidate is assigned a p_{T}, a charge and a position in (ϕ, η).

The Global Muon Trigger (GMT) receives the input of eight candidates from the Barrel Muon Track Finder (BMTF), eight candidates from the Endcap Muon Track Finder (EMTF) and eight from the Overlap Muon Track Finder (OMTF). It merges candidates, removes duplicates, assigns an optimised momentum to the muons and ranks them according to the quality of the tracks and momentum. The eight best candidates are then sent to the L1 Global Trigger (GT).

The crystals in the ECAL are grouped in arrays into 'towers', 5×5 cell arrays in the barrel and more complex arrays in the endcaps, whose energy deposits determine which events fulfil the trigger requirements for electrons and photons. In the HCAL the calorimeter cells are grouped in towers as well, of granularity 0.087 in η and 0.087 rad in ϕ in the barrel, and 0.17 in η and 0.17 rad in ϕ in the endcaps. The ECAL and HCAL towers constitute the basic element of the calorimeter trigger.

The ECAL, HCAL and HCAL Forward (HF) deliver their information to the first trigger subsystem, the CALO Layer 1. It makes use of information such as the HCAL energy over the ECAL energy to identify e/γ candidates. At L1T level there is no information from the tracker, implying that at this stage the electrons and photons are the same kind of object, represented by ECAL deposits. To reconstruct objects from energy deposits in the calorimeters (towers), the position of the trigger candidate can be computed from an energy-weighted average centered on the main tower. Since electrons (and converted photons) tend to loose some of their energy through Bremsstrahlung emission when passing through the tracker material, a dynamic clustering technique is used to add neighbouring ECAL towers to form the e/γ candidates.

The CALO Layer 2 has the results of CALO Layer 1 as input and uses the information of the towers to form jets and tau lepton candidates. In the case of hadronically decaying tau leptons, many energy clusters associated with each decay product may be produced, so a dynamic clustering is used to reconstruct individual clusters, which can be subsequently merged. For jets, the reconstruction is based on a window centered around the largest tower, which corresponds to

the 0.4 cone radius used for the offline anti-kT reconstruction algorithm, as it is described in section 3.2.3 [54]. CALO Layer 2 computes other global quantities with full calorimeter granularity, such as the total transverse energy, the transverse energy of the jets alone, the missing transverse energy E_{T}^{miss} and the jet-based equivalent of the E_{T}^{miss}. It also estimates the pileup of an event and subtracts the energy in the calorimeter objects that comes from it. From the CALO Layer 2 twelve electron/photon objects, twelve jets and eight tau leptons are sent as candidates to the L1 Global Trigger.

The Global Trigger collects all the candidates and forwards them to the High Level Trigger if they pass some selection criteria, which define a *trigger path*. Each trigger path requires a certain object or group of objects to possess a minimum transverse momentum or surpass a threshold in another variable, like pseudorapidity, the angle between two objects or their invariant mass. Technical trigger paths are also defined independently on the event to be used for calibration.

2.3.2 The High Level Trigger

The HLT filtering is carried out by a processor farm that uses more complex algorithms to bring the 100 KHz rate to just about 1 KHz. The events are filtered using algorithms that involve several candidates and are more closely related to the offline reconstruction criteria. One of the main differences with respect to the L1T is that there is information from the tracker as well, which means that tracking and vertex finding is very important for reconstruction at this level. A tracking algorithm helps the particle reconstruction by improving the momentum resolution of muons, adding a tracking-based isolation criteria, and identifying jets originating from b quarks.

The overall output rate of the L1T and HLT is adjusted by prescaling the number of events that pass the selection criteria of specific algorithms. A set of concrete algorithms together with the set of prescales completely specifies the trigger selection. Even though the algorithms and the thresholds of the variables of the input objects are hard-coded in the firmware, changing to a different prescale value allows the adjustment of the trigger rate during a run by modifying the frequency of the affected trigger algorithms.

The event data are stored locally on disk and eventually transferred to the CMS Tier-0 computing center for offline processing and permanent storage. Events are grouped into a set of non-exclusive streams according to the HLT decisions. In addition to the primary physics streams, monitoring and calibration streams are also written for data validation purposes.

2.3.3 Fine Synchronisation of the DT Trigger

The subdetectors have to be well synchronised with the LHC machine clock in order to properly assign particles detected in different systems to the same event. The LHC clock works in 25 ns intervals, corresponding to the 40 MHz frequency of bunch crossings (BX), and correctly assigning a BX to each particle produced is crucial for the event description [55, 56]. In the muon barrel the position in each DT cell is computed by dividing the drift time of the electrons in the chamber by the constant drift velocity. The time obtained with the Time-to-Digital Converter are the sum of various contributions:

$$t_{TDC} = t_{TOF} + t_{prop} + t_{L1} + t_{drift} + t_0 = t_{pedestal} + t_{drift} \ , \tag{2.7}$$

where t_{TOF} is the time-of-flight of the muon, between the interaction point and the DT cell, t_{prop} is the time of the signal propagation along the wire, t_{L1} is the latency of the Level 1 trigger, t_{drift} is the drift time of the electrons inside the cell and t_0 represents the synchronisation time used to equalise the response of all the channels at the level of each chamber. All of these times except for the drift time would be added as a constant pedestal, resulting in a TDC time distribution for a given superlayer with a shape like that shown in Fig. 2.14.

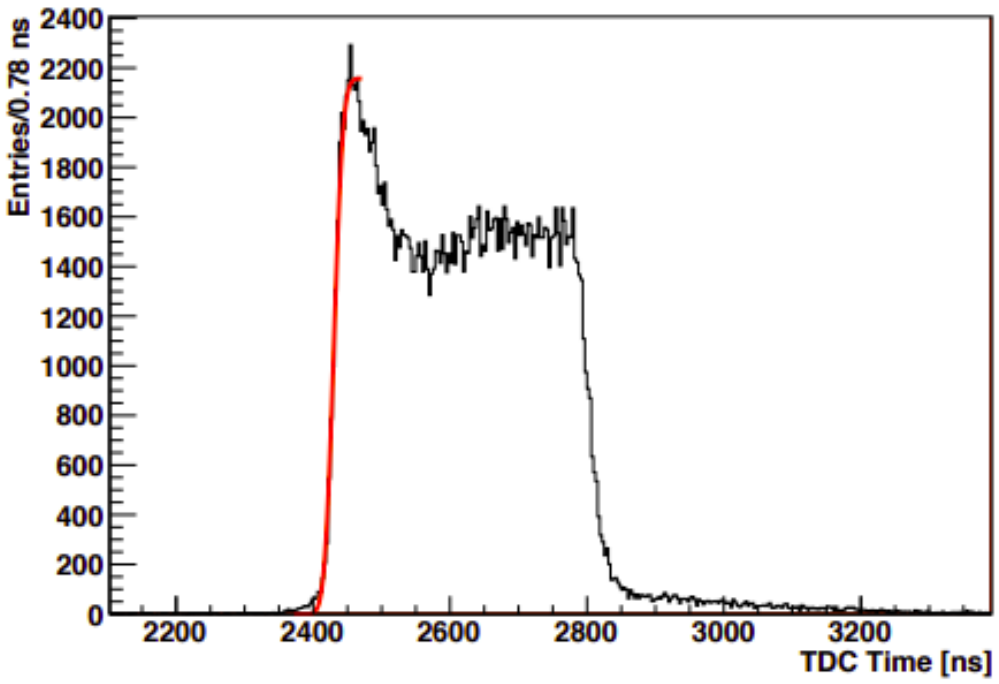


Figure 2.14: Time distributions in a DT superlayer. The red line indicates the pedestal time and subtracting that value the drift time is obtained. The position in the DT cell is obtained by dividing the drift time by the constant drift velocity. Picture taken from [57].

Since the muon time of flight in CMS varies from 12 ns for the closest muon station to 35 ns in the farthest one, the trigger electronics start to process the signal with a delay depending on the station. The time that takes the signal to propagate along the wire also depends on the different cable lengths connecting to the trigger electronics in each chamber, which therefore has its own delay with respect to the absolute clock time of the machine.

The Trigger and Timing Control system (TTC) provides the machine clock distribution so that any local trigger signal has a BX number identification assigned. To align the triggers of each of the chambers, a delay of up to 25 ns is required in order to assign all the chamber triggers of an event to the same BX, which has that duration. That amount of delay is what needs to be determined, to provide the value for t_0 in Eq. 2.7.

Time equalisation is achieved with the hardware, using cables with adequate length to connect between the components of the trigger chain. With this, delays of the order of one ns can be introduced as corrections for the synchronisation. Each muon station can be considered intrinsically synchronous, and they are all equipped with one TTC receiver, which means that the best delay setting for the TTC receivers must be determined in order to have correct time measurement and correct parent event identification.

A first approximation (rough synchronisation) would match the arrival of the muons to the chamber sometime within the 25 ns interval (the correct BX), but only doing this is not optimal. In this case, it is possible that the muon arrives at a layer outside the time interval and a hit of the superlayer segment could be lost, resulting in a reconstructed segment with worse quality. To prevent this, it is more appropriate to match the arrival of the muon with the beginning of the interval, which is the moment where it is most likely to be registered and least likely to be lost, as shown in Fig. 2.15.

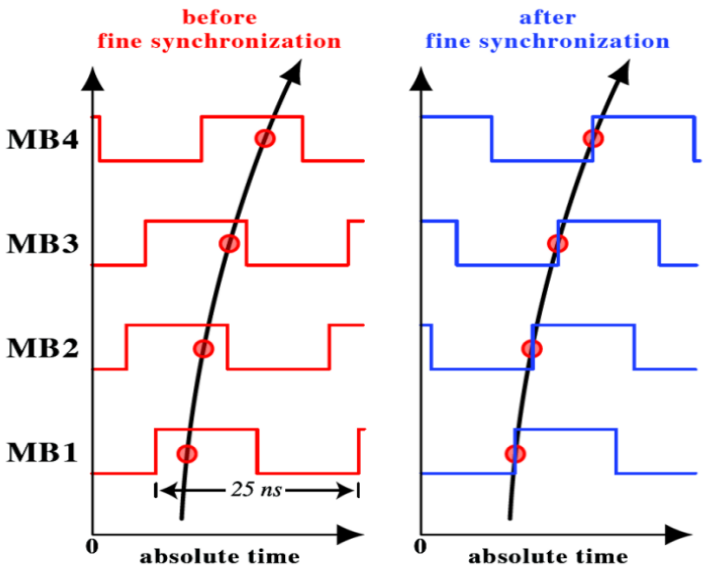


Figure 2.15: Difference between the first approximation and a finer synchronisation. After the synchronisation the muon arrives at the beginning of the 25 ns interval, the optimal time to be registered.

The synchronisation method

Since triggering at the times corresponding to the beginning of a BX is the optimal choice, and those times are separated by 25 ns, we expect to see a time periodicity in the loss of hits due to misidentifying the BX crossings as well. This is studied by looking at the segment quality in the chambers. The best quality possible for a segment is HH, meaning *high quality* in the first ϕ superlayer (4 hits, one per layer) and also high quality in the second ϕ superlayer of the chamber. If a hit is lost, for that superlayer the quality would become L or *low*, so the total quality for the segment would be HL or LH (7 hits in total).

This means that calculating the ratio of segments with quality HL over HH, and looking at how this ratio evolves over time, it is possible to estimate the time where this quantity would be maximal and therefore the trigger would have the worst efficiency. Consequently, the times associated to the maximum efficiency could be obtained from the lowest efficiency times just by taking into account the 25 ns periodicity of the BX, which means that the maximum and minimum are separated by half a period (12.5 ns). It is easier to identify the time corresponding to the maximum efficiency in this way, since the minimum efficiency in the HL/HH distribution appears as a peak and the maximum efficiency as a plateau.

In order to determine the time with the best ratio of HH segments we used cosmic muon data taken during the year 2016. Using cosmic muon data we cover all the cycle dependence since those muons arrive randomly and homogeneously during the 25 ns interval. Because of their almost vertical incidence angle they traverse more frequently the horizontal chambers than the vertical ones. The data were taken using muon triggers associated with bottom sectors only (numbered 8 to 12), meaning they would need to at least go through one bottom sector to be detected.

Figure 2.16 shows the measured ratio of segments HL/HH, where there is a clear 25 ns periodicity with respect to the chamber delay time, written in the X axis as t_0. The exact position of the peaks varies with the chamber. Because we were using the triggers in the bottom sectors, those sectors had around 0 ns delay. For top sectors (numbered 2 to 6), if the muons came from the surface they would cross them before the triggering bottom sector, and therefore the registered time for them would be negative, i.e. they crossed those chambers before triggering.

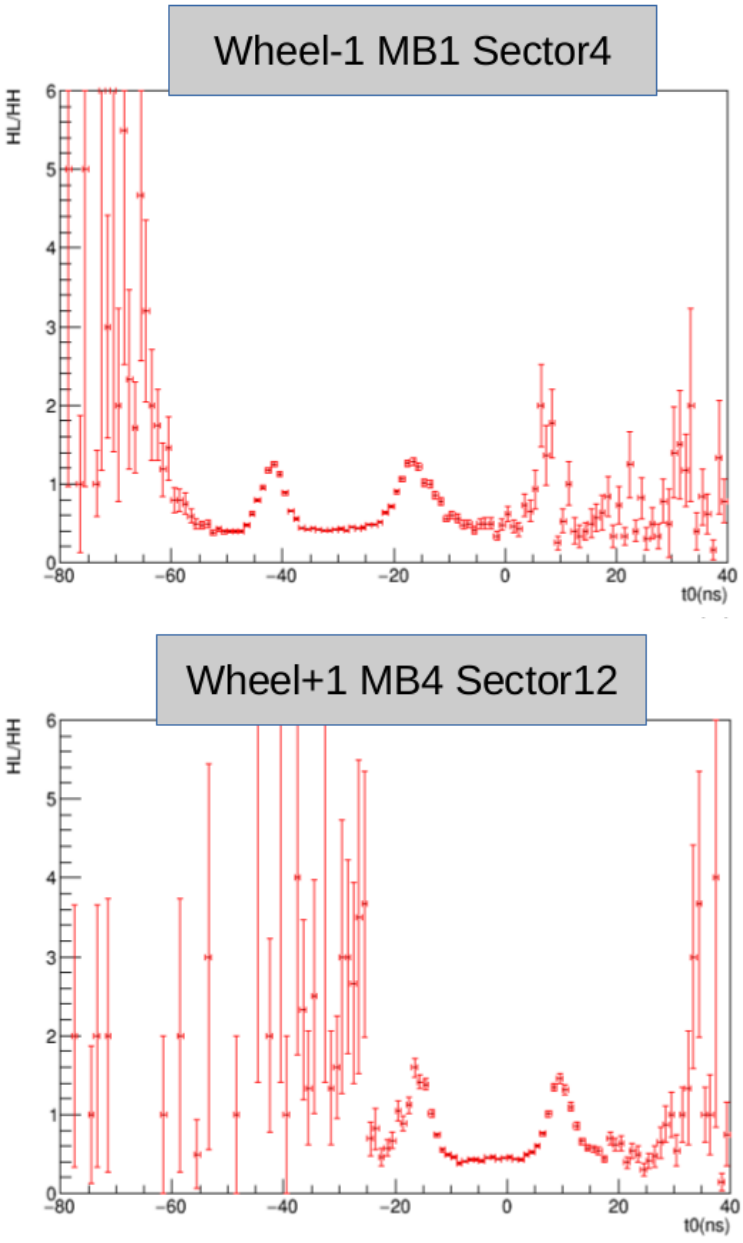


Figure 2.16: HL/HH ratio, obtained from 2016 cosmic muon data. This shows the 25 ns periodicity of the ratio as well as the different peak positions for the top sector 4 (top plot) and the bottom sector 12 (bottom plot).

In the HL/HH histograms the time is reduced to an interval between 0 to 25 ns, making use of their periodicity. A single peak is obtained, as shown in the example of a chamber in Fig. 2.17, which is fitted with a gaussian function plus a constant background. The maximum value corresponds to the time delay with the worst bunch crossing assignment.

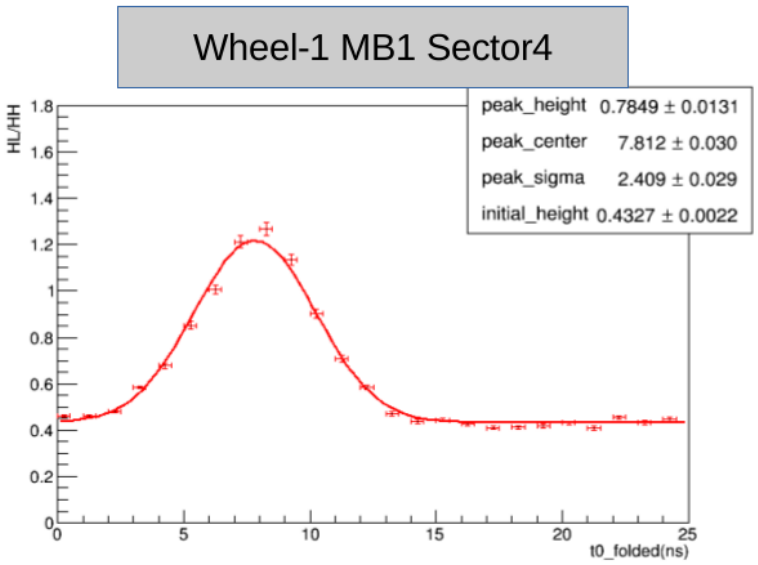


Figure 2.17: Example of HL/HH ratio after folding into a 25 ns interval, where the peak that represents the minimum efficiency phase is obtained from a fit.

This process is repeated for the 250 chambers of the barrel muon, and Fig. 2.18 shows the result of the fits in the 250 chambers classified by wheels, sectors and stations. The time that maximises the ratio of HH/HL is computed by adding 12.5 ns to the times obtained. It is important to note that for MB1 stations there is a difference between the external wheel chambers and the rest, because the drift velocity is not the same in those inner stations due to the stray magnetic field present in them.

Additional possible delay sources must be checked as well and included in the corrections obtained by this method. For each chamber a time distribution for the t_0 during collisions should be centered around 0. Using collision data of 2016 and selecting muon events that fulfil strict identification requirements (a *tight muon* definition described later in section 3.2.1) the segment t_0 distribution is checked for all chambers. This check of the t_0 distribution is necessary to perform, because it has been seen during Run 2 that whenever the high voltage (HV) of the central wires is lowered, the average value of t_0 increases. The reason for this behaviour is that with a lower voltage at the anode it takes a few ns more for the charges to accumulate in the wire and the signal to be produced. Figure 2.19 is an example of a chamber that had the HV lowered from 3600 V to 3550 V in October 2016, and the corresponding shift of the t_0 average value in that case.

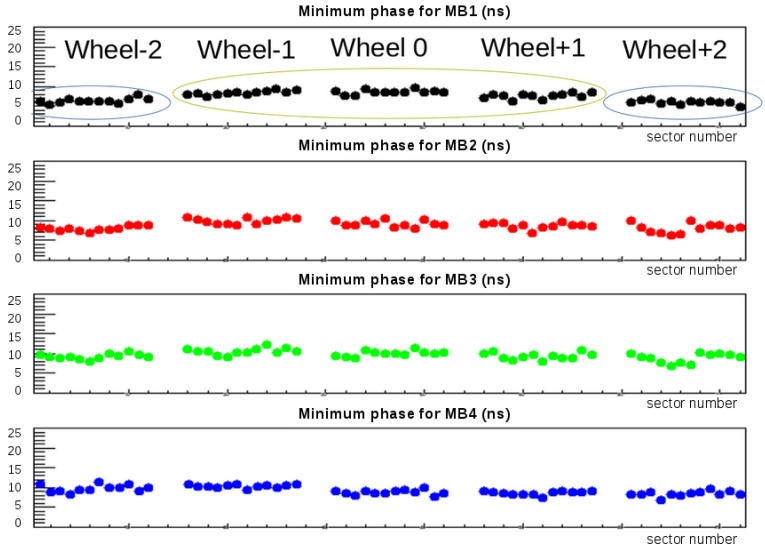


Figure 2.18: Values for the worst HH/HL segment fraction time delays in all the stations (colours), wheels (groups of points) and sectors (X axis) in 2016. For the MB1 there are two distinct averages, one for the external wheels and another for the rest.

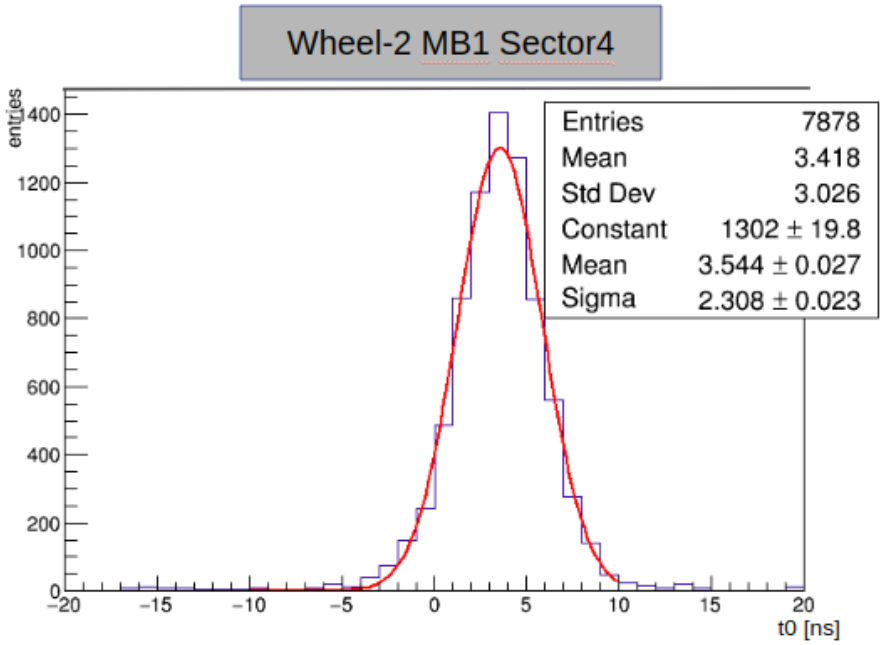


Figure 2.19: Example of a t_0 distribution for a chamber in October 2016 which had a slightly lower HV of 3550V, before synchronisation.

The result of synchronisation

The fine delay that was finally implemented into the hardware is obtained with the peak values obtained with the HH/HL ratios and the additional possible delay introduced in the average t_0 due to other factors in this simple way:

$$\begin{aligned} \text{correction} &= < t_0 > - \text{peak(worst quality)} + 12.5 \text{ ns} \\ &= < t_0 > - \text{peak(best quality)} \end{aligned} \tag{2.8}$$

The proposed corrections obtained in 2016 are presented in Fig. 2.20, with an average value of 4 ns in a wide distribution between 1 and 7 ns. We expect a correlation between the chambers that needed larger corrections and the BX assigned by them. Figure 2.21 shows this correlation, since chambers with a higher proposed delay are identified as previously having a higher average BX (BX=0 meaning here the correct BX of the LHC machine). In general, however, the differences in the average BX between the chambers are small because they are very close to 0 since the BX was already correctly assigned in most cases.

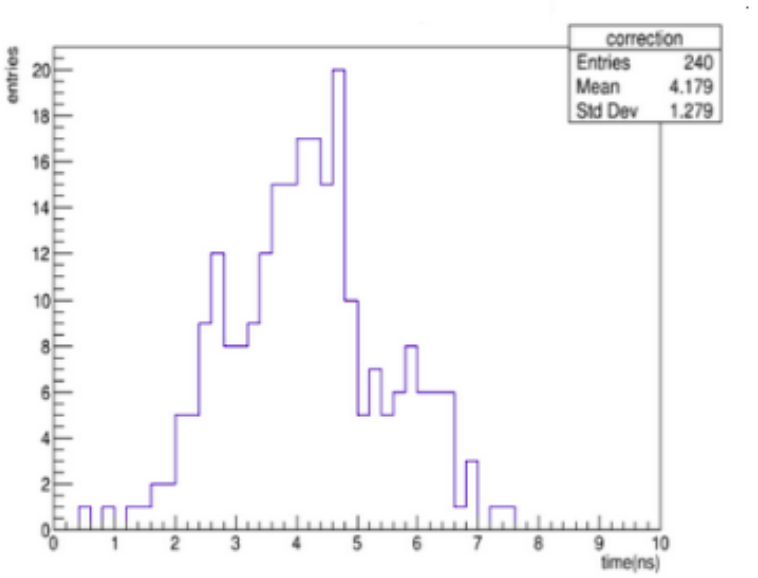


Figure 2.20: Proposed fine corrections after the June 2016 study. Every entry in this distribution corresponds to one DT chamber.

After the corrections were implemented in June 2016, there were some improvements in segment quality and BX identification. Data from before and after the changes are compared in Fig. 2.22, resulting in an increase in the amount of segments that were assigned to the correct BX as well as an increase in the number of HH quality segments.

In 2017 this study was performed again to check if changes in the chamber conditions related to the high voltage had an impact on the optimal synchronisation times. Figure 2.23 shows the average values for all the chambers in July 2017 and which of them had the HV changed at that time. The clear difference of a few ns delay between the chambers is observed and it was later corrected by adjusting the t_0 of those chambers by 2 ns. In the future such changes should become more common, as the detector shall have to deal with the higher luminosity of the next runs and lowering the voltage could help with the longevity of the chambers.

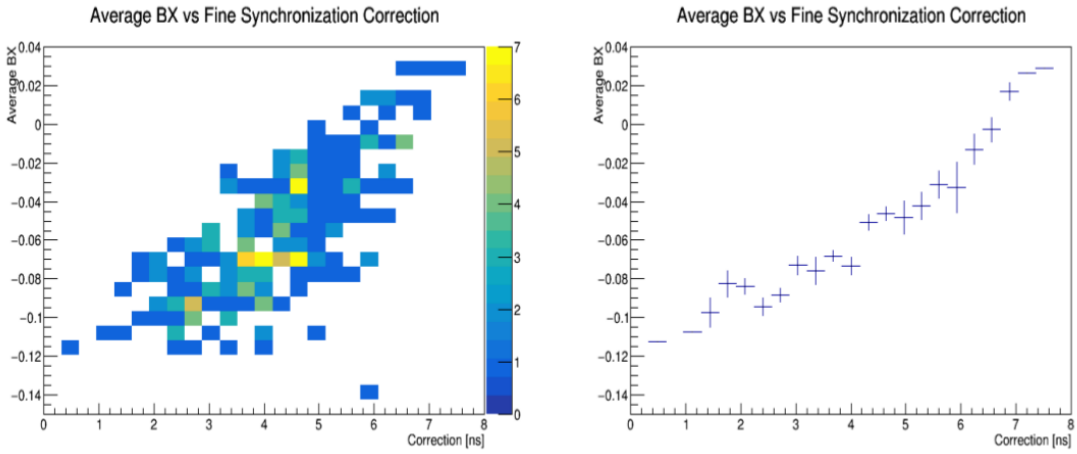


Figure 2.21: Corrections vs average BX in 2016. A correlation is observed for which the biggest proposed corrections correspond to the biggest average BX.

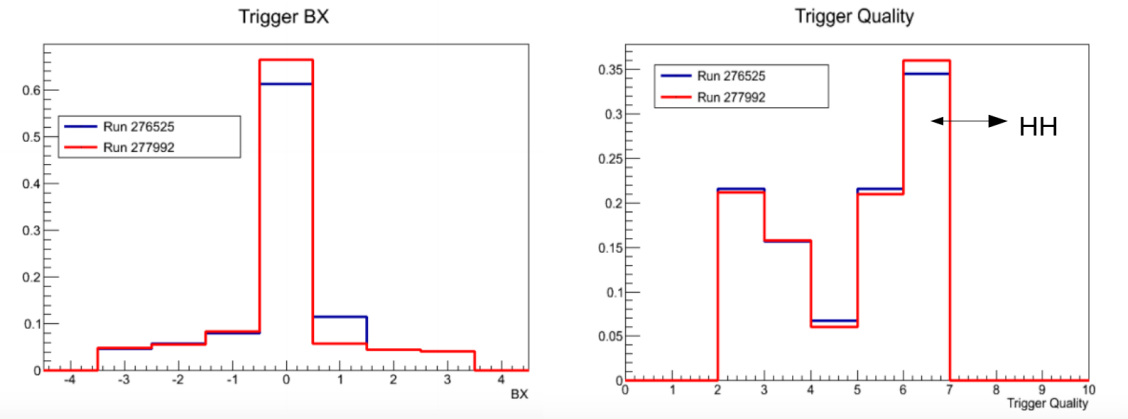


Figure 2.22: Before (blue) and after (red) the 2016 trigger fine synchronisation corrections were implemented. There was an increase in the correct assignment of BX as well as of the HH segment fraction (last bin in second plot).

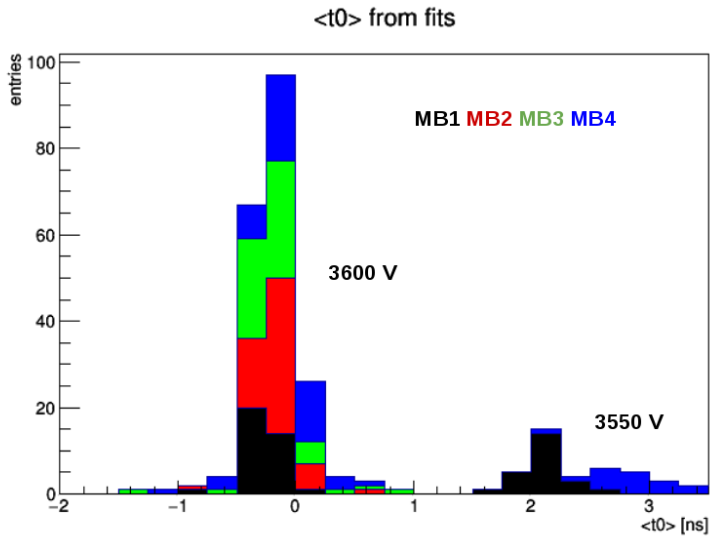


Figure 2.23: The average t_0 distribution for all the chambers in July 2017, sorted by stations. HV is written for each case.

2.4 CMS computing systems

After the trigger filtering of the events, data are stored in tape to be analyzed. Even considering the factor 4000 in data reduction due to the trigger system, the amount of information is large enough to require petabytes of storage capacity per year. To overcome this, CMS makes use of the interconnection between the more than 100 countries contributing to the LHC, using the World-wide LHC Computing Grid (WLCG). This helps to spread the work and resources necessary between the research institutes and universities in order to both store and analyze the data.

The CMS offline grid computing system [58] is arranged in four different layers or *tiers*, geographically distributed according to the CMS collaboration structure. It starts from a single Tier-0, which is connected to the output of the Trigger System and the Online Data Acquisition System of CMS, at CERN, which performs a prompt reconstruction of the raw data to create datasets with the physical objects. Then, the data is distributed to the seven Tier-1 centers, located in several countries with many members in the collaboration, such as PIC in Barcelona, Spain. These centers are in charge of data processing, performing reconstructing and calibration. They provide the datasets for more than 50 Tier-2 centers and *skim* the data, reducing their size. Tier-2 centers, such as CIEMAT in Madrid, provide storage and computing capacity for analysis and simulation of Monte Carlo (MC) events. Finally, there are numerous Tier-3 centers with smaller capacity that provide resources for local groups.

MC-based simulations produce events that try to mimic the physical processes happening at the collisions. For this reason they need not only to reproduce the topology and kinematics of the processes obtained from the theoretical model interaction, but also the behaviour of the detector. The response of the detector, which has to also consider its geometry, is performed by using the Geant-4 package [60].

The analysis in CMS are performed in a software framework that is common and available to the whole collaboration, called *CMSSW*. It is a constantly evolving object-oriented structure accessible with C++ and python code and developed for each data taking period. It contains all the services needed by the simulation, calibration and alignment, and reconstruction modules that process event data. The CMS Event Data Model (EDM) is centered around the concept of an *event*. An event is a C++ object container for all raw and reconstructed data related to a particular collision.

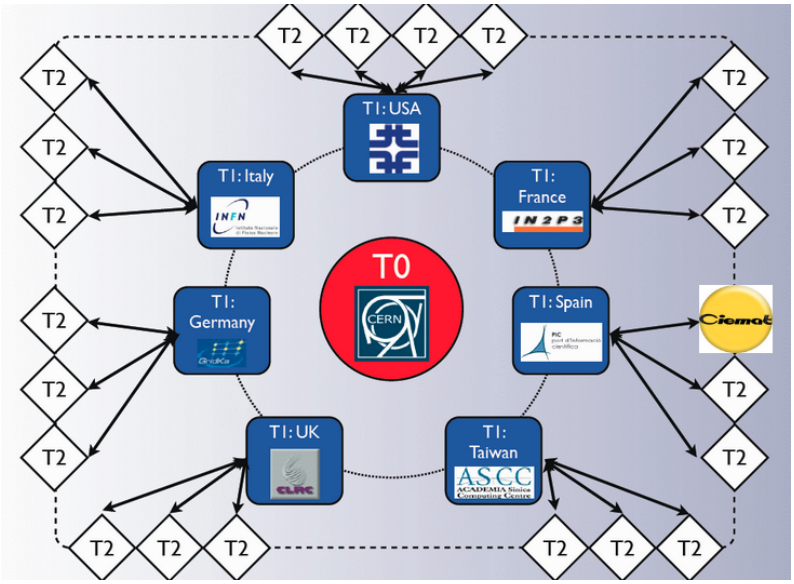


Figure 2.24: The CMS computing model, arranged in Tiers. After the data processing done in Tier-0 at CERN, data are distributed to one of the seven Tier-1 centers such as PIC in Spain and skimmed to be used by the collaboration members. Tier-2 centers such as CIEMAT help with the computing requirements of analysis and simulations. Picture taken from [59].

There are various data formats which are used. The first one is the RAW format, coming from the Tier-0, which has all the event information and requires more storage capacity. Then there are reconstructed (and re-reconstructed) formats with the event information processed by the Tier-1 and Tier-2 centers. Finally, the data format which contains all the information needed for CMS analyses with all the reconstructed objects and a smaller size is the Analysis Object Data (AOD). Similar formats reduced in size such as nanoAOD, the one used in this work, contain the important physics information of the event while having a reduced size. The way in which the data is stored and analysed is via ROOT [61] files, which contain all the event object information and kinematics, forming a tree-like structure with the variables as branches and the events as histogram entries.

Chapter 3

Particle identification and reconstruction

3.1 The Particle Flow Algorithm

All of the subdetectors mentioned before are suited to measure certain types of particles. The way all of that information comes together in order to completely describe everything that happens in the event is known as the Particle Flow algorithm (PF) [62]. With the PF the signals registered in the different parts of the detector translate into known particles or jets, which are the objects in the final states that is used in the analysis. By summing up the energy of the objects in the event we can also infer if neutrinos or other non-interacting particles were present somewhere in the collision.

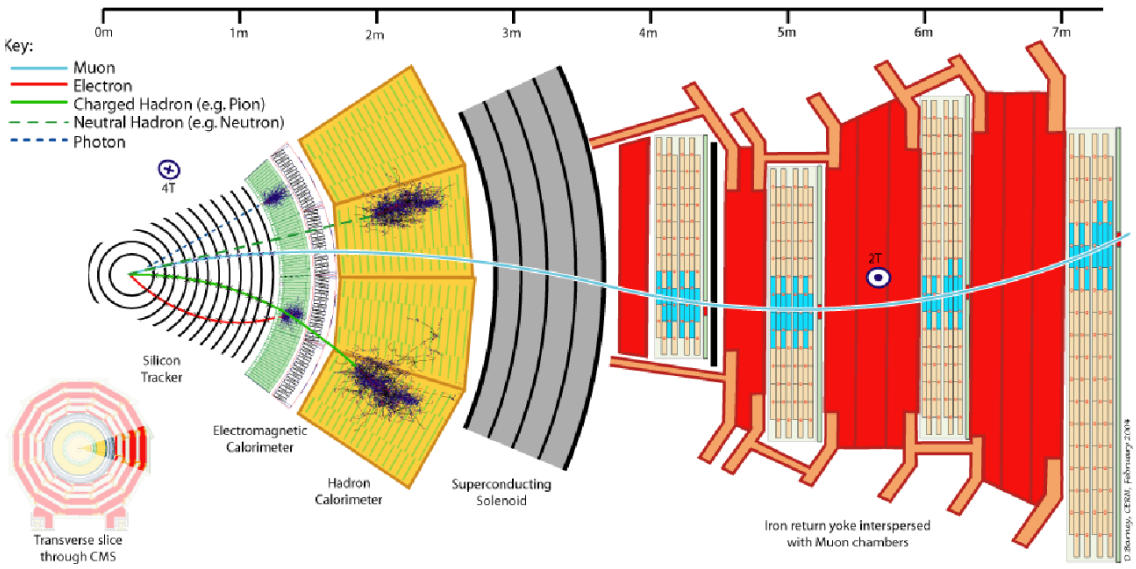


Figure 3.1: An example of the path taken by different kinds of particles inside the CMS detector.

Figure 3.1 shows an example of the signatures of some of the particles produced. Electrons and photons are identified in the electromagnetic calorimeter. They leave energy deposits in clusters after producing electromagnetic showers that reproduce the energy of the original particle when added. Whether or not they left a track at the silicon tracker determines if they were charged (electrons) or neutral (photons). Muons are identified by tracks at the muon chambers and tracks at the silicon tracker. They leave energy deposits in the electromagnetic calorimeter if they cause ionisation, but they are minimal.

Charged hadrons also leave a track in the central tracker and then they fully deposit their energy in the form of clusters in both of the calorimeters. Neutral hadrons are reconstructed from energy clusters separated from the extrapolated position of tracks. After identifying electrons, photons, muons and hadrons, the resulting PF candidates in the event are clustered into jets.

Finally, the missing energy at the plane transverse to the beam line, referred to as E_{T}^{miss} from now on, is used to confirm when a particle that has been produced at the event escaped the detector unnoticed. Since the colliding particles only carried momentum in the direction longitudinal to the beam axis, the sum of the momenta of all the particles in the final state should also be longitudinal.

This is how neutrinos are inferred at CMS, and therefore in analyses like $\mathrm{H} \to \mathrm{WW} \to 2\ell 2\nu$ only events with sufficient E_{T}^{miss} are considered. For new particle searches, the treatment of the E_{T}^{miss} and the neutrino background modelling are essential, since particles from BSM physics or dark matter candidates would be non-interacting as well.

3.1.1 Tracking

The first step of the reconstruction process is the local reconstruction at the tracker. The track reconstruction consists of the clustering of signals above specified thresholds in pixel and strip channels into hits, which are then used to estimate the cluster positions in the plane of each sensor and their uncertainties [63]. Candidate vertices are obtained with an algorithm using the collection of tracks if at least two of their associated tracks are incompatible with originating from other vertices.

The track finding is done in an iterative way. It starts with an initial seed consistent with a particle originated at the interaction region with a minimum transverse momentum. Three hits at the pixel, strips or both, or two hits plus a vertex determine an initial estimate of the track candidate. After that, Kalman filters [64] are used to extrapolate the trajectory and search for additional hits in the track candidate. With every additional hit, a fit to the trajectory is performed, improving the parameters of the trajectory with every point. Low quality tracks with a low number of hits or missing hits in layers crossed by the extrapolated trajectory are discarded. At the end of this process there are a group of track candidates that combine with other subsystems to characterise the different charged particles.

3.2 Final state objects

3.2.1 Muons

Hit, segment and track reconstruction

The muon reconstruction is performed by using the silicon tracker and the muon system. Tracker candidates are combined with the tracks from the muon systems to reconstruct the muons. The muon subdetectors, described in section 2.2.5, select muon candidates using the hits in the muon chambers [34]. The CSC and DT chambers are multi-layer detectors where hits are reconstructed in each layer. From the reconstructed hits, straight-line track *segments* are built within each CSC or DT chamber.

Depending on the information received from the detector, muons are divided into three groups:

- Each inner track with p_{T} >0.5 GeV and a total momentum larger than 2.5 GeV is extrapolated to the muon system. The tracks that are loosely matched to an inner muon chamber hit qualify as **tracker muon tracks**.

- **Standalone muons** are created from groups of DT and CSC segments, that act as the initial seeds for the pattern recognition in the muon spectrometer. The DT, CSC, and RPC hits along the muon trajectory are gathered and the result of the final fitting is called a standalone-muon track.
- The best quality tracks are obtained when, extrapolating a standalone-muon track to the tracker, a matching track is found. Those are **global muon tracks**, and their trajectory is the result of performing a combined fit using the inner and the standalone-muon tracks.

Punch-through particles could be misidentified as tracker-only muons, since they would reach the innermost muon chambers, and muons from cosmic ray showers could be detected as standalone muons as they cross the muon detectors. Information from both tracker and muon systems is therefore needed to precisely identify muons from the collision and assign them a momentum as accurate as possible.

Particle Flow muons

From global and tracker muons, PF muon candidates are selected according to some isolation and track reconstruction criteria. Isolated global muons are first selected by considering additional inner tracks and calorimeter energy deposits with a distance ΔR to the muon direction smaller than 0.3. The distance is defined in (η,ϕ) coordinates as $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. If the sum of the p_{T} of the tracks and of the E_{T} of the deposits does not exceed 10% of the muon p_{T}, the muon is considered a PF muon.

This isolation criterion alone is sufficient to adequately reject hadrons that would be misidentified as muons, hence no further selection is applied to these muon candidates. For nonisolated global muons, a stricter selection is applied, the *tight muon ID*. In addition, for nonisolated global muons it is required either three or more matching track segments in the muon detectors, or calorimeter deposits associated with the track that are compatible with a muon.

Muons that do not satisfy the tight muon selection due to a poorly reconstructed inner track are still considered if the standalone muon track fit is of high quality and is associated with a large number of hits in the muon detectors (23 out of 32 DT hits or 15 out of 24 CSC hits). Muons failing the selection due to a poor global fit are still considered if a high-quality fit is obtained with at least 13 hits in the tracker and the associated calorimeter clusters are compatible with a muon hypothesis.

Muon identification

To identify muons, a set of variables measured in the subdetectors is used, such as the number of hits per track, the degree of matching between tracker and standalone muon tracks for global muons or the track fit χ^2. Different identification criteria are defined according to these values to be used later in the physics analyses.

- **Loose muon identification** (ID) is the least restrictive muon definition and it selects prompt muons originating at the primary vertex as well as muons from light and heavy flavor decays, maintaining a low enough rate of the misidentification of charged hadrons as muons. A loose muon is a PF muon candidate and should be at least either a tracker or a global muon.
- **Medium muon ID** is a more optimised definition for prompt muons and heavy flavor decay muons. We define a medium muon as a loose muon with a tracker track that uses hits from more than 80% of the inner tracker layers it traverses. In this case the muon must be reconstructed as a tracker or a global muon, but if it is just a tracker muon the segment compatibility requirement is higher. The segment compatibility is tuned to target an overall efficiency of 99.5% for muons from simulated W and Z events.

- **Tight muon ID** aims to suppress muons from in-flight decays and from hadronic punch-through. A tight muon is defined from a loose muon when it is associated with a tracker track that uses hits from at least six layers of the inner tracker, including at least one pixel hit. It must be reconstructed as both a tracker muon and a global muon, and the inner track must be matched to at least two segments in the muon stations. A tight muon must also be compatible with having originated in the primary vertex, i.e. having a transverse impact parameter |dxy|<0.2 cm and a longitudinal impact parameter |dz| <0.5 cm. These parameters measure distances between the primary vertex and the origin point of the lepton projected into the transverse (XY) plane or the Z direction, respectively. Table 3.1 summarises the tight muon requirements.

- Particular criteria are used in specific analyses, such as **soft muon ID** for low p_{T} muons for B-physics and quarkonia analyses and **high momentum muon ID** for muons with p_{T} >200 GeV in high-mass particle searches.

Table 3.1: Identification criteria for tight muons

Tight muon requirements
- The candidate is reconstructed as a Global Muon
- Goodness-of-fit per degree of freedom (χ^2/dof) of the global muon fit < 10
- At least one muon chamber hit included in the global muon fit
- Muon segments in at least two muon stations
- Transverse impact parameter $\lvert dxy \rvert < 2$ mm
- Longitudinal impact parameter $\lvert dz \rvert < 5$ mm
- At least one pixel hit in the tracker track
- Number of tracker layers with hits > 5

Muon isolation

An important variable to discriminate between prompt muons and those from heavy flavour weak decays within jets is the hadronic activity around the muon. One way to evaluate this activity is through the muon isolation. We define the isolation of a muon as the momentum of all the PF particles around it in a cone of ΔR <0.4. Their relative momentum with respect to that of the muon must be less than 0.4 for the muon to satisfy the minimal isolation requirement.

$$PF_{relativeISO} = \frac{1}{p_{\mathrm{T}}(\mu)}\Big(\sum_{char.had.}^{PV} p_{\mathrm{T}} + \max\Big(0, \sum_{neut.had.}^{All} E_T + \sum_{photons}^{All} E_T - \frac{1}{2} \sum_{char.had.}^{PU} p_{\mathrm{T}}\Big)\Big) \quad (3.1)$$

The relative isolation is determined as the sum of the momentum of all hadrons in the cone originated in the primary vertex, divided by the muon p_{T}. The contribution from the neutral particles is the sum of all neutral particles corrected by the pileup neutral particles, which is estimated using simulations as being approximately equal to half the charged hadron contribution.

There are other common isolation definitions, such as the tracker-based isolation for muons. It is defined as the sum of the p_{T} of the tracker tracks in a cone of ΔR <0.3 around the muon. For the muon to satisfy this isolation condition, that sum should be less than 10% of the muon p_{T}. This is the isolation definition used by the PF algorithm to define isolated muons.

Muon efficiency

In the process of identifying a muon, there are sources of inefficiency due to the tracking, identification, isolation and trigger criteria:

$$\epsilon_{\mu} = \epsilon_{\text{tracking}} \times \epsilon_{ID_{tight}} \times \epsilon_{ISO_{tight}} \times \epsilon_{trigger} \tag{3.2}$$

- The tracking efficiency $\epsilon_{\text{tracking}}$ (for tracker and global muons)
- The ID efficiency $\epsilon_{ID_{tight}}$
- The isolation efficiency $\epsilon_{ISO_{tight}}$
- The trigger efficiency $\epsilon_{trigger}$ (for muon triggers)

The four efficiencies are computed individually using a tag-and-probe method, using pairs of muons coming from a known mass resonance such as the Z boson in both a data and a simulated sample enriched in those events. One of the muons, the *tag*, fulfils the criteria of the corresponding selection. The *probe* is the one we use to measure the efficiency of that selection. We search for the latter as a second muon in the event, requiring the condition that combining it with the tag muon they have an invariant mass close to the Z boson mass.

The background of tag-and-probe pairs not coming from the Z resonance is accounted for by performing a simultaneous fit to the invariant mass spectra for passing and failing probes with identical signal shape and appropriate background shapes [65]. The efficiency is then computed from the normalisation of the signal shapes in the two spectra.

The efficiency measurement is performed using only tag-and-probe pairs for which a single probe is associated with the tag, to prevent a high background in low p_{T} regions due to high track multiplicity. The measurement is performed in the same way in data and simulation and the results agree within 1% for loose muons and 3% for tight muons.

We estimate the trigger efficiency by computing a simple ratio of events that pass the trigger divided by the total events. No fitting has been performed as with the ID/isolation efficiencies, since the background contamination under the Z peak is very low when both muons are required to satisfy the full ID and isolation requirements. Double lepton triggers need two leptons to satisfy the requirements and hence two parts, each of them called a "leg". Trigger efficiencies are computed for each leg of the HLT path and then the efficiencies of each leg are combined according to the type of event we have (e+e, e+μ or μ+μ).

The efficiency is obtained for both data and MC and scale factors are computed using the ratio of efficiencies of data and MC as a function of the lepton p_{T} and η. These scale factors are used in the physics analyses to scale the MC and they correct for the difference in efficiencies between data and MC. The pileup reweighting is also applied on MC during the computation of efficiencies in order to properly describe the dependence on the number of pileup collisions.

Muon momentum determination

The default algorithm used by CMS to determine the muon momentum is the **Tune-P algorithm** [65]. For each muon, the Tune-P algorithm selects the p_{T} measurement from one of the four possible refits based on goodness-of-fit information to reduce tails in the momentum resolution distribution due to poor quality fits.

- The **Inner-Track** fit obtains the momentum using only information from the inner tracker. This is usually the optimal one for most muons with $p_{\text{T}} < 200$ GeV.
- The **Tracker-Plus-First-Muon-Station** fit, which also uses the innermost station. That station is the best at measuring the momentum within the muon system, because of reduced sensitivity deeper in the muon system due to possible showering.

- High energetic muons may produce a cascade inside a chamber, resulting in multiple hits. To deal with the high occupancy of the chamber the **Picky** fit is useful, since it uses the hits from the global muon track compatible with the extrapolated trajectory.
- When there is a significant trajectory bending due to energy losses of the muon, the **Dynamic-Truncation** fit adds hits from the segments in the muon chambers in succession if they are close to the extrapolated trajectory and starting from the innermost chamber, propagating outward.

The Tune-P algorithm was validated using muons from proton collisions and cosmic rays as well as MC simulations. A check for possible biases introduced by the algorithm was performed and no bias in the momentum, curvature or invariant mass distributions was identified. The PF algorithm [62] refines the information from Tune-P, exploiting information from the full event, by selecting refits that significantly improve the balance of missing p_{T} and by using an additional algorithm designed to select genuine missing energy events. The PF momentum assignment was also validated using MC simulations and muons from proton-proton collisions.

3.2.2 Electrons

Electron track reconstruction

Electrons are reconstructed using the information from the silicon tracker and of the electromagnetic calorimeter. The tracker obtains electron candidates in the same way that it does for muon candidates, but for electrons there are potential energy losses. As they cross the mechanical structure, services and the cooling system inside the tracker before reaching the ECAL, electrons lose energy via bremsstrahlung. The amount of radiation depends on the amount of tracker material traversed which varies strongly with η, rising from 0.3 X_0 at central pseudorapidities to 1.5 X_0 towards the edge of the barrel and falling back to 0.7 X_0 in the endcaps. Almost half of the electrons reduce their energy to 50% of their original value before reaching the ECAL, and about 10% lose more than 95% of their energy [66].

Inside the electromagnetic calorimeter the magnetic field spreads out the charged particles of the electromagnetic shower in many crystals, usually in the ϕ direction and not so much in η because the magnetic field points to the Z direction [67]. Approximately 94% of the incident energy of a single electron or photon is contained in 3×3 crystals, and 97% in 5×5 crystals [66]. Clustering algorithms are used to sum together energy deposits in adjacent crystals.

The clustering algorithm proceeds first with the formation of basic clusters, corresponding to local maxima of energy deposits. The basic clusters are then merged together to form a "supercluster", which is extended in ϕ, to recover the radiated energy. Because of the differences between the geometric arrangement of the crystals in the barrel and endcap regions, a different clustering algorithm is used in each region.

In the endcaps, the *island* algorithm is used. An example of it is shown in Fig. 3.2. Starting from a seed with more than 1 GeV, the cluster expands in the ϕ direction providing that the next crystal has an energy above a certain threshold and less energy than the previous crystal. Then it expands in the η direction and continues the process until no more crystals satisfy the requirements.

In the barrel a *hybrid* algorithm is used instead, since it exploits the geometry of the ECAL barrel and properties of the shower shape, collecting the energy in a small window in η and an extended window in ϕ. The clustering algorithm used in EB starts from the crystal with the maximum energy deposit and it adds next to it arrays of crystals if they surpass an energy threshold of 0.1 GeV. They have a rectangular shape, 1×3 or 1×5 if the central crystal contains an energy over 1 GeV. An example of a cluster formation using the hybrid algorithm is shown in Fig. 3.3.

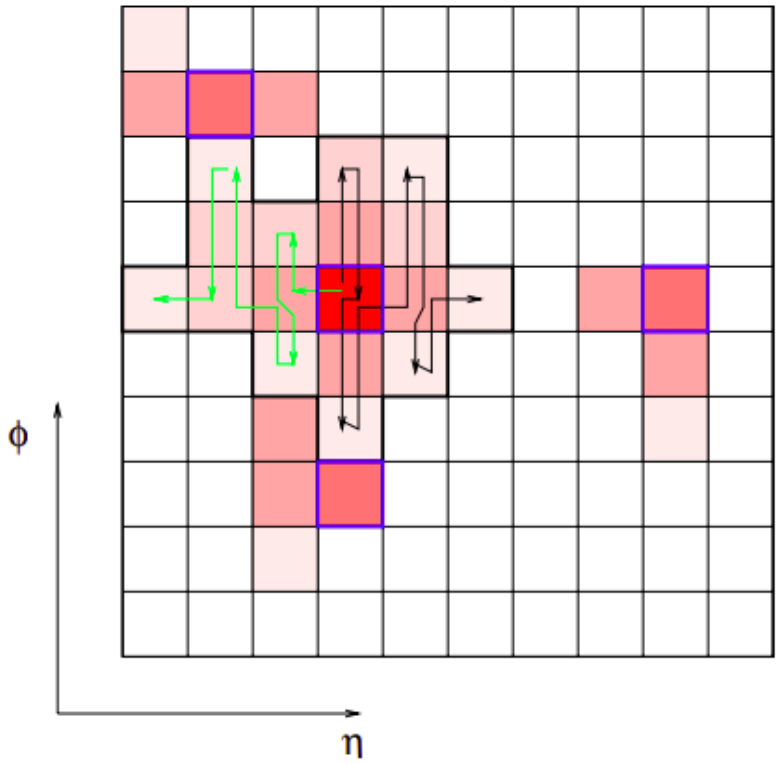


Figure 3.2: An example of a cluster formation using the islands algorithm used in EE. The black lines are the cluster boundaries and the blue lines mark the initial cluster seeds. Picture obtained from [68].

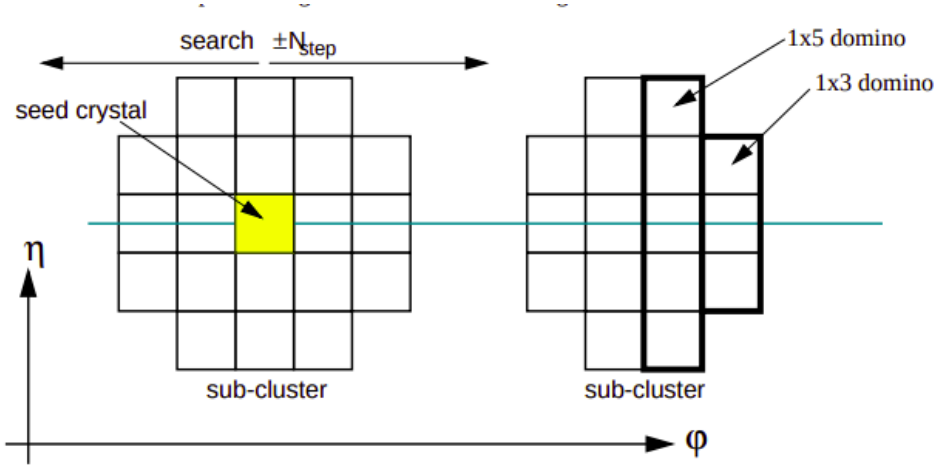


Figure 3.3: An example of a cluster formation using the hybrid algorithm used in EB. 1×3 or 1×5 dominoes are added in both ϕ directions at the side of the cluster seed. Picture obtained from [68].

The clusters are then grouped into superclusters if they satisfy a low energy threshold requirement. These superclusters are based on a cluster seed, to which other nearby clusters may be added. In the EE non-overlapping island clusters are grouped into a supercluster if they are close in a narrow window in η and a bit wider window in ϕ. The hybrid clustering algorithm in the EB inherently produces superclusters, providing the summed energy of the crystals in the cluster is above a certain minimum. The position of the superclusters is extrapolated to the electromagnetic calorimeter preshower, and energy deposits close to it are added to the supercluster total energy. The supercluster position is measured as the energy-weighed mean of all cluster positions.

Once the hits at the tracker are collected, the energy losses in each layer are approximated by a mixture of Gaussian distributions, with probabilities associated with each step. Two estimates of track properties are usually exploited at each measurement point that correspond either to the weighed mean of all the components, or to their most probable value. This procedure of track building and fitting, called Gaussian-sum filter or GSF [69], yields electron tracks that are extrapolated to the ECAL and used to extract track parameters at the surface of the ECAL.

The PF clustering of electrons is driven by GSF tracks. The PF cluster corresponding to the electron in the ECAL surface is the one that matches to the track at the exit of the tracker. To recover the bremsstrahlung photons emitted in the tracker, a straight line is extrapolated from each of the tracker layers to the ECAL, tangent to the electron track, and each matching PF cluster is added to the electron PF cluster. Additionally, to prevent converted photons from being added to the cluster, a specific procedure selects displaced Kalman-filter tracks through a dedicated multivariate analysis algorithm, and kinematically associates them to the PF clusters.

Electron identification and isolation

Several strategies are used to identify prompt electrons and to separate them from the background sources, which are generally photon conversions or electrons coming from hadron decays, mostly from those containing a b or c quark. Algorithms have been developed to apply sequential selections on a set of discriminants, which are then tested on simulated samples to optimise the electron ID and evaluate the corresponding efficiencies. Electron identification is based on many variables, which are grouped into these categories:

- **Calorimetric observables**, used to discriminate between genuine electrons (signal electrons or electrons from photon conversions) and jets with large electromagnetic components which might be misidentified as electrons. It is based on the transverse shape of electromagnetic showers in the ECAL and exploits the fact that electromagnetic showers are narrower than hadronic showers. The energy fractions deposited in the HCAL (expected to be small, as electromagnetic showers are essentially fully contained in the ECAL), as well as the energy deposited in the preshower and in the endcaps are used. Some of these variables are the ratio of energy in the HCAL with respect to the ECAL (H/E), the difference between the position of the extrapolated inner track and the supercluster position in η and ϕ ($\Delta\eta_{in}$ and $\Delta\phi_{in}$) or the weighed cluster RMS along η inside a 5×5 region of supercluster (5×5 $\sigma_{\eta\eta}$).
- **Isolation variables**, used to reduce the background due to misidentified jets or to genuine electrons within a jet resulting from semileptonic decays of hadrons containing b or c quarks. In both cases, the electron candidates have significant energy flow near their trajectories, and requiring electrons to be isolated from such nearby activity greatly reduces these sources of background. The PF isolation for muons written in Eq. 3.1 is defined for electrons with a cone size of 0.3 around the electron direction, relative to the electron p_{T}, with a threshold of ISO <0.06.
- **Tracking quality variables**, such as the number of expected missing hits in pixel layers, the number of layers with hits in the track trajectory, the χ^2 value of the reconstructed track, χ^2 of the GSF track and the number of tracker layers with hits in the trajectory of the GSF track.

- **Conversion rejection variables** used to reject secondary electrons produced from photon conversions. Photon conversions inside the tracker usually do not leave a first hit in its innermost layer, unlike prompt electrons, so that pattern can be used to efficiently identify conversion tracks. Other variables that distinguish between electrons from conversions are the impact parameters of the electron track, which are smaller when originated from the vertex of interest.

We apply a set of requirements on the variables mentioned above to identify true electrons and to separate them from fake ones. One way to achieve this is to define our electron ID using fixed selection criteria that would only change between electrons found in the ECAL barrel or in the endcaps. These IDs would be *loose*, *medium* and *tight* electrons, analogue to the muon working points described in section 3.2.1. The requirements which define these selections depend on various factors including the running conditions, which means that every time the conditions change, the selection needs to be tuned again.

An alternative approach is to use instead a selection based on a multivariate analysis (MVA). In this case a single discriminator variable is used, computed based on multiple variables of the electron object. It provides separation between the signal and backgrounds using a Boosted Decision Tree (BDT) algorithm. Then, a threshold is set on the discriminator value or use the full distribution for a shape-based statistical analysis. This option is explored in the electron ID optimisation explained in section 4.2.

Electron efficiency

To measure the electron ID and isolation efficiencies we use a similar tag-and-probe fitting method explained for muons in 3.2.1. This method uses a known mass resonance (J/Ψ or Z boson) to select electron pairs, and probe the efficiency of a particular selection criterion. In this case, the *tag* is a well identified and isolated electron which also satisfies the electron trigger requirement. The *probe* is another object reconstructed as an electron that satisfies some kinematical selection and that combined with the *tag* has an invariant mass close to the considered resonance. A simultaneous fit to the invariant mass spectra for passing and failing probes with identical signal shape and appropriate background shapes is performed and the efficiency is computed from the normalisations of the signal shapes in the two spectra.

We require the *probe* to satisfy the isolation requirement or the ID selection and compute the efficiency. This is done for both data and MC. The scale factors, which are the ratio of efficiencies of data and MC, are calculated and used to scale the MC to correct for the difference in efficiencies. The pileup reweighting is also applied on MC during the computation of efficiencies.

The electron trigger efficiencies are also calculated with a tag-and-probe method using the simpler counting method instead of fitting, since the amount of background contamination under the Z peak is very low in comparison to the number of Z boson events. Double-electron trigger efficiencies are computed combining the efficiencies of both the single electron HLT paths.

Electron charge determination

Bremsstrahlung followed by photon conversions affects the measurement of the charge of the electron. When the bremsstrahlung photons convert upstream in the detector, they lead to very complex hit patterns, and the contributions from conversions may be wrongly included in the fitting of the electron track. There are three ways to estimate the sign of the charge, and to reduce misidentification the electron charge is defined by the sign shared by the majority (2) of those estimates. They are the following:

- The first charge estimation comes from the sign of the Gaussian-sum-filter (GSF) track curvature.
- The second estimate is the sign of the associated Kalman-filter track matched to a GSF track when at least one hit is shared in the innermost region.
- The third one is evaluated using the supercluster position, and defined as the sign of the difference in ϕ between the vector connecting the beam to the supercluster position and the vector connecting the beam and the first hit of the electron GSF track.

Electron momentum determination

The electron momentum is estimated using a combination of the tracker and ECAL measurements. As expected, this quantity is particularly sensitive to the emission of bremsstrahlung photons and their conversions. Therefore, the electrons are classified according to their bremsstrahlung pattern, using observables sensitive to the emission and conversion of photons along the electron trajectory. They are divided in the following categories:

- *Golden* electrons, those with little bremsstrahlung and consequently with the best momentum estimation. They are defined by a supercluster formed by a single cluster and their energy loss via bremsstrahlung in the tracker is less than half their original energy.
- *Big-brem* electrons, which lost in a single step a large amount of energy by radiating. They are also defined by a supercluster formed by a single cluster, but they lost more than half their energy by photon emission.
- *Showering* electrons, which emitted plenty of radiation along their trajectory and are defined by a supercluster consisting of multiple clusters.
- *Bad track* electrons, the least populated category which identifies electrons with a poorly fitted track in the innermost part of the trajectory. Electrons with an energy loss fraction at least 15% higher in the ECAL than in the tracker fall into this category.

The supercluster energy is obtained by adding the energy of the individual crystals, and then corrected and calibrated to take into account effects such as energy leakage out of the supercluster, into gaps between crystals and modules or the transition region between barrel and endcaps. Energy loss in interactions in the material before the ECAL and additional energy from pileup interactions are also accounted for. An MVA technique is used to obtain the correction for these effects, validated with a comparison between corrected data and simulated Z→ee events.

The combination of supercluster energy and tracker momentum depends on the electron category, but in general at low energies (less than 15 GeV) or for electrons near gaps in detectors, the track momentum is expected to be more precise than the ECAL energy, and for high-energy electrons (250 GeV), the combination is dominated entirely by the energy measurement in the ECAL, and only the supercluster energy is used.

For the more general case, a regression technique is used, that defines a weight multiplying the track momentum in linear combination with the estimated supercluster energy: $p_{combination} = w \times p_{tracker} + (1-w) \times E_{SC}$. The corrected supercluster energy, its relative uncertainty, the track momentum and its relative uncertainty are the main input observables, but the electron class and the position in the barrel or endcaps are also included, providing information of the quality and amount of emitted bremsstrahlung.

3.2.3 Jets

Jet reconstruction

Jets are the experimental signatures of quarks and gluons, an ensemble of particles resulting from a hadronic shower. Since jets are objects composed of neutral and charged hadrons, electrons, photons and muons, they are the objects that benefit the most from the PF algorithm, which combines the information from all CMS subdetectors, and removes other identified objects before reconstructing the jets from the remaining particles.

At CMS, PF jets are reconstructed with the anti-k_T algorithm [54], that clusters particles reconstructed with the PF algorithm with a size parameter for the cone of R = 0.4 [70]. The PF jet reconstruction applies Charged Hadron Subtraction (CHS) [71] to reduce pileup, which removes charged particles not compatible with the reconstructed primary vertex before clustering.

Apart from PF jets, there are other types of ways to reconstruct jets at CMS, which differently combine individual contributions from subdetectors to form the inputs to the jet clustering algorithm, but are used only for specialised cases. These are calorimeter-only jets, track jets or Jet-Plus-Track jets [72], but they are not used in the analysis.

Jet energy calibration

The jet energy calibration (JEC) [73] is a procedure to obtain the energy of the corresponding final state parton or jet from the measured energy in the detector. The difference between them arise from the nonlinear and nonuniform response of the calorimeters, as well as sources of unwanted energy, the noise caused by the electronics or additional pileup interactions.

The PF jets are corrected in three steps. There is an *offset* correction that reduces the jet energy eliminating electronics noise and pileup contributions, a *relative* correction that removes variations in the jet response that depend on the jet η, and an *absolute* correction that guarantees that the variations in the jet response with respect to the jet p_T are small. The JEC can be summarised as:

$$E_{Corrected} = (E_{Uncorrected} - E_{offset}) \times C_{Rel}(\eta, p_T^{''}) \times C_{Abs}(p_T^{'}) \tag{3.3}$$

where $p_T^{''}$ denotes the transverse momentum of the jet corrected for offset and $p_T^{'} = p_T^{''} \times C_{Rel}(\eta, p_T^{''})$ is the transverse momentum of the jet corrected for offset and pseudorapidity dependence.

The jet energy correction factors are determined using MC samples of simulated events and data samples of proton-proton collisions. For the offset correction the ratio between the momentum of the generated and reconstructed jets is determined as a function of jet p_T and η and from that ratio the correction values are computed. Alternatively, the noise contribution from the offset correction is obtained by using collected data of minimum bias events. By selecting data from early runs the contribution from additional pileup collisions is accounted for, since those runs have a low percentage of one-interaction-only events.

In back-to-back dijet events the momentum of both jets is expected to be balanced in the transverse plane, and thus this topology is used to measure the relative correction. The correction for the η dependence is computed choosing events where one of the jet lies in the central region and the other jet has an arbitrary η coordinate. The central region is chosen as reference because of the uniformity of the detector and because it has a higher jet p_{T} reach.

Finally, for the absolute correction, we make use of γ+jets events, where there is a balance between the momentum of the photon and the jet in the event. The energy of the recoiling photon is used as a reference, since the photon energy is accurately measured in the ECAL. Z+jets events where the Z boson decays into a muon or electron pair are also used in a similar way for this purpose.

There is one additional correction that is applied to jets to solve an issue in 2016 and 2017 data periods. Due to a miscalibration, highly energetic readout in the ECAL endcap from jets, photons and electrons were mistakenly assigned by the L1 Trigger to the previous bunch crossing by an issue that is known as *prefiring*. The trigger fired to the previous bunch crossing rather than the bunch crossing where the highly energetic object came from, resulting in a loss of events. This issue has been studied in detail and correction factors have been derived. The amount of data lost to prefiring is about 3%, but applying prefiring corrections leads to a better data/MC agreement, specially in the most affected region of $2.5<|\eta|<3.0$.

b jet tagging

Jets originating from heavy-flavour decays are an important component of the LHC physics program. Special characteristics of b-hadrons are exploited to identify b-quark jets. The long lifetime of hadrons containing a b quark (b-hadrons) produced at LHC energies allow them to travel around 100-200 μm before decaying. Distances can be measured inside the tracker due to its excellent vertex reconstruction capability, leading to the identification of a secondary vertex separated from the main interaction vertex. The information about that secondary vertex, together with other characteristics based on track information or event kinematics, is used in a b-tagging algorithm [74] that provides a discriminant that evaluates how likely it is for the jet to have originated from a b quark.

3.2.4 Missing transverse energy

The missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,miss}$ is defined as the negative sum of the transverse momentum of all PF candidates in an event.

$$\vec{p}_{\mathrm{T}}^{\,miss} = -\sum_{PF} \vec{p_T} \tag{3.4}$$

The modulus of the $\vec{p}_{\mathrm{T}}^{\,miss}$ is usually called p_{T}^{miss} or E_{T}^{miss}. There is a physical component of the E_{T}^{miss} and an instrumental one. The physical sources of E_{T}^{miss} consists of particles that are not identified by any of the subdetectors, escaping particles whose presence we can infer by using energy and momentum conservation principles. The instrumental source of E_{T}^{miss} is caused by inefficiencies in the detection and measurement of the momentum of all the particles, which contribute to a low background of E_{T}^{miss} [75].

Due to the compactness and hermeticity of the detector, no particle originating at the collision should be able to pass through it unnoticed unless its nature is associated with a non-interacting behaviour, as it is the case for neutrinos, dark matter candidates or some BSM elusive particles. An excess in energy in a direction inside the transverse plane (E_{T}^{miss}) may be explained by a missing particle with equal energy escaping in the opposite direction, since the initial particles carried no momentum in the transverse plane and that should be conserved.

Three types of corrections are applied to PF E_{T}^{miss} [75]. The first correction is a direct result of the jet energy scale corrections being propagated to E_{T}^{miss}, and the other two are a mitigation for the degradation of the E_{T}^{miss} reconstruction due to the pileup interactions. To remove charged hadrons originating from the vertices for pileup interactions, the charged hadron subtraction (CHS) is applied to E_{T}^{miss}. In addition to CHS, the pileup correction to the E_{T}^{miss} removes an estimate of neutral pileup contributions as well.

The third correction is also a mitigation for the pileup effects. The distribution of physical E_{T}^{miss} is independent of ϕ because of the rotational symmetry of the collisions around the beam axis. However, the reconstructed E_{T}^{miss} does have a sinusoidal dependence on ϕ, caused by anisotropic detector responses, inactive calorimeter cells, detector misalignment or displacement of the beam spot. The amplitude of this modulation is reduced by shifting the origin of the coordinate in the transverse momentum plane.

Due to the higher pileup conditions in Run 2, sophisticated methods to mitigate the pileup effects in the E_{T}^{miss} computation were developed. The pileup per particle identification (PUPPI) algorithm [76] computes the E_{T}^{miss} from the PF candidates, assigning a weight to their momentum corresponding to their probability to originate from the primary interaction vertex. This algorithm assigns a weight to each particle in the event before clustering, which is used to rescale the particle momentum. Ideally, this would assign a value of zero to pileup particles, thus effectively removing them without having to correct for the pileup later.

Anomalous high-p_{T}^{miss} events may appear due to a variety of reconstruction failures, detector malfunctions or non-collision backgrounds. In the ECAL, spurious energy deposits may appear due to noisy sensors in the photodetectors or due to showers with a noncollision origin, such as interactions of the beam protons upstream of the detector. Dead cells in the ECAL can also lead to artificial p_{T}^{miss}. In the HCAL, energy can arise from noise in the photodiode or the readout box electronics, as well as from particle interactions with the light guides and photomultiplier tubes. All of these sources have been studied during Run 1 and Run 2, and algorithms were developed to identify and remove these events. Event filters are designed to identify more than 85–90% of the spurious high-E_{T}^{miss} events with a mistagging rate less than 0.1% and suppress them [77].

Chapter 4

Analysis of the Higgs boson production in its decay into a W boson pair

This chapter describes the components necessary for the measurement of the Higgs boson properties in the H $\rightarrow$ WW decay channel using the data collected with the CMS experiment between 2016 and 2018 at a center of mass energy of 13 TeV. Even though most of what is included in this chapter can be applied to any of the H $\rightarrow$ WW analyses, this work focuses on the characteristics of the gluon fusion production mode, with a final state of opposite-sign leptons with different flavour and E_{T}^{miss}.

The datasets and MC simulated samples are presented in section 4.1, object selection and the strategy followed are described in section 4.2, section 4.3 studies the descriptions of background processes, and all the associated uncertainties of the analysis are discussed in section 4.4.

4.1 Data and simulated samples

4.1.1 Datasets

The data used in this analysis were recorded in the years 2016, 2017, and 2018, with corresponding integrated luminosities of 35.9 fb^{-1}, 41.5 fb^{-1}, and 59.7 fb^{-1}, respectively [78, 79, 80]. Data were selected online by requiring one or two isolated leptons (muons or electrons) with high transverse momentum. Table 4.1 summarises the trigger requirements. In the case of single electron triggers, in 2016 a p_{T} of at least 25 GeV is required if the electron is detected in the central region ($|\eta| < 2.1$) and this is raised to 27 GeV if it is in the forward region ($2.1 < |\eta| < 2.5$). For single muon triggers the requirement is $p_{\mathrm{T}} \geq 24$ GeV in all the detector coverage ($|\eta| < 2.4$). In 2017 and 2018 the p_{T} thresholds for the single lepton triggers were raised because the increase in instantaneous luminosity led to a stricter filtering of the events. The p_{T} threshold for the single-electron triggers is $p_{\mathrm{T}} > 35$ GeV (32 GeV) in 2017 (2018). For muons, $p_{\mathrm{T}} > 27$ GeV (24 GeV) is the single muon trigger requirement for 2017 (2018).

For the double lepton triggers, a lower p_{T} threshold is required for the leading lepton. The dielectron trigger p_{T} threshold is set at 23 GeV for the leading electron and at 12 GeV for the subleading electron, and the dimuon trigger required values are 17 GeV and 8 GeV. Different flavour double lepton triggers require $p_{\mathrm{T}} > 23$ GeV for the leading lepton and $p_{\mathrm{T}} > 12$ GeV for the subleading lepton. In 2016 the threshold for subleading muons is 8 GeV.

Table 4.1: Triggers used in the H $\rightarrow$ WW $\rightarrow 2\ell 2\nu$ analysis

Single lepton triggers for 2016
Electron with $p_T > 25$ GeV, $\|\eta\| < 2.1$
Electron with $p_T > 27$ GeV, $2.1 < \|\eta\| < 2.5$
Muon with $p_T > 24$ GeV, $\|\eta\| < 2.4$
Single lepton triggers for 2017
Electron with $p_T > 35$ GeV, $\|\eta\| < 2.5$
Muon with $p_T > 27$ GeV, $\|\eta\| < 2.4$
Single lepton triggers for 2018
Electron with $p_T > 32$ GeV, $\|\eta\| < 2.5$
Muon with $p_T > 24$ GeV, $\|\eta\| < 2.4$
Double lepton triggers for all years
Electron with $p_T > 23$ GeV and electron with $p_T > 12$ GeV
Muon with $p_T > 17$ GeV and muon with $p_T > 8$ GeV
Electron with $p_T > 23$ GeV and muon with $p_T > 8$ GeV (2016 only)
Electron with $p_T > 23$ GeV and muon with $p_T > 12$ GeV
Muon with $p_T > 23$ GeV and electron with $p_T > 12$ GeV

Beside requiring a minimum p_T for the lepton, the online selection also included some identification and isolation requirements. In double lepton triggers these correspond to a minimum 'very loose' tracker isolation for muons and a 'loose' identification and isolation criteria for electrons, using both calorimetric and tracker information. The selection is stricter for single lepton triggers, with a more restrictive isolation requirement for muons and a tighter identification for electrons.

4.1.2 Simulations of signals and backgrounds

Several samples of simulated events, generated by MC methods are used in this analysis to estimate the expected distributions for the observables of all the background and signal processes and to calculate the efficiency of our selection. Different simulated samples are generated for each of the three years to take into account changes in the detector configuration and the pileup conditions as well as updates of the reconstruction software. These MC samples describe SM processes at next-to-leading order (NLO) or next-to-next-to-leading order (NNLO) accuracy in perturbative QCD. For the parton showering, hadronisation and the simulation of underlying event the generators describing the hard processes are interfaced with PYTHIA 8.2 [81]. The parton distribution functions (PDF) used and the underlying event (UE) tunes are common to all samples simulated for a given year.

All processes are generated using the NNPDF3 [82] parton distribution functions, with the accuracy matching that of the matrix element calculations. The PDF set used is NNPDF3.0 in 2016 [83] and NNPDF3.1 in 2017 and 2018 samples [84]. For the UE simulation the CUET8PM1 tune [85] is used for 2016 and the CP5 tune [86] is used for 2017 and 2018. To estimate a systematic uncertainty related to the choice of the UE tune, samples for the signal processes and the WW background are also generated with alternative tunes, which are representative of the uncertainties in the tuning parameters.

The samples of Higgs boson signals through the different production mechanisms describe the perturbative QCD properties of the process at NLO. They are generated with POWHEG [87, 88, 89] for all the production modes but bbH, and with the MADGRAPH5aMC@NLO [90] generator for the latter. The mass of the Higgs boson in all the generated samples is 125.09 GeV, using the value of the combined measurement of CMS and ATLAS using Run 1 data [91]. The current most precise measurement of the Higgs boson mass is 125.38 $\pm$ 0.14 GeV [92], but this analysis is not sensitive to the difference in precision between those two values. Events produced with POWHEG through the gluon fusion process are reweighted in order to match the Higgs boson p_T and number of jets distributions at NNLO accuracy [93], using the NNLOPS scheme [94].

The decay of the Higgs boson into a pair of W bosons that then decay leptonically is modelled by PYTHIA in most cases, and also by the generator JHUGen [95]. JHUGen is used for ggH and VBF in all years and for ZH and WH in some years. The simulated signal samples are normalised using cross sections [96] and decay rates [97] computed by the LHC Higgs cross-section Working Group. The full list of samples is included in Table 4.2.

In this work the signal corresponds to the WW decay channel where the W bosons decay leptonically. Higgs boson direct decays into tau lepton pairs which then decay into muons and electrons are considered as well and treated as signal in the inclusive analysis, but their contribution in the signal region is very small compared to $H \to WW$. Table 4.3 contains the list of these samples.

Table 4.2: $H \to WW$ simulated signal samples.

Process	Year	Production mode and decay	MC generators	σ (pb)
Gluon fusion	all	ggH $H \to WW \to 2\ell 2\nu$	POWHEG+JHUGen+ PYTHIA8	1.09
VBF	all	VBF $H \to WW \to 2\ell 2\nu$	POWHEG+JHUGen+ PYTHIA8	0.0854
W^+ associated	2016,2017	HW^+ $H \to WW$	POWHEG+ PYTHIA8	0.180
	2018	HW^+ $H \to WW$	POWHEG+JHUGen+ PYTHIA8	0.180
W^- associated	2016,2017	HW^- $H \to WW$	POWHEG+ PYTHIA8	0.114
	2018	HW^- $H \to WW$	POWHEG+JHUGen+ PYTHIA8	0.114
Z associated	2016	qqHZ $H \to WW$	POWHEG+ PYTHIA8	0.190
(qqZH)	2017, 2018	qqHZ $H \to WW \to 2\ell 2\nu$	POWHEG+JHUGen+ PYTHIA8	0.0199
Z associated	2016	ggZH $H \to WW$	POWHEG+ PYTHIA8	0.0264
(ggZH)	2017, 2018	ggZH $H \to WW \to 2\ell 2\nu$	POWHEG+ PYTHIA8	0.00277
ttH	all	ttH	POWHEG+ PYTHIA8	0.212
bbH	all	bbH $H \to WW \to 2\ell 2\nu$	MADGRAPH + PYTHIA8	0.011

Table 4.3: $H \to \tau\tau$ simulated samples.

Process	Year	Production mode and decay	MC generators	σ (pb)
Gluon fusion	all	ggH $H \to \tau\tau$	POWHEG+ PYTHIA8	3.04
VBF	all	VBF $H \to \tau\tau$	POWHEG+ PYTHIA8	0.237
W^+ associated	all	W^+H $H \to \tau\tau$	POWHEG+ PYTHIA8	0.0525
W^- associated	all	W^-H $H \to \tau\tau$	POWHEG+ PYTHIA8	0.0333
Z associated (qqZH)	all	ZH $H \to \tau\tau$	POWHEG+ PYTHIA8	0.0552
Z associated (ggZH)	all	ZH $H \to \tau\tau$	POWHEG+ PYTHIA8	0.00768

Samples of events for the main background processes are also generated at NLO using the generators POWHEG, MADGRAPHaMC@NLO or MCFM [98, 99, 100]. The backgrounds estimated with MC are non-resonant WW production, Drell-Yan (DY), top quark pair and single top and other multiboson processes (WZ, Wγ^*, ZZ, VVV), as listed in Table 4.4.

For WW production, samples of WW events produced through gg $\to$ WW and qq $\to$ WW are simulated separately. qq $\to$ WW events are produced with POWHEG and they are reweighted so that they match the p_T distribution of WW at NNLO plus next-to-next-to-leading logarithm (NNLL) accuracy in QCD [101, 102]. gg $\to$ WW events are generated with MCFM at LO, but a scaling factor k=1.4 is applied to scale the cross section from LO to NLO [103]. EWK WW production processes, which have two additional final state jets and are relevant for the analysis in the topology with 2 or more jets in the final state, are generated with MADGRAPHaMC@NLO at LO.

Samples of top quark pair events are generated with POWHEG, as well as single (anti)top quark events in the t- and tW channels. Single (anti)top quark events in the s-channel are generated at NLO with MADGRAPHaMC@NLO. Samples of DY events are produced with MADGRAPHaMC@NLO with a dilepton invariant mass between 10 and 50 GeV at LO and for DY ($\tau\tau$) with $m_{\tau\tau} > 50$ GeV at NLO. Wγ simulated events are obtained with MADGRAPHaMC@NLO at LO, and the rest of the diboson samples (Zγ, WZ/γ^* and ZZ) are generated with POWHEG or MADGRAPHaMC@NLO at NLO. All the triple vector boson samples (WWW, WWZ, WZZ and ZZZ) are inclusive in their decays and are generated with MADGRAPHaMC@NLO at NLO. Additional auxiliary MC samples are used for some specific studies.

Table 4.4: MC samples for background modeling.

Process	Year	Decay mode	MC generators	σ (pb)
qqWW	all	WW $\to 2\ell 2\nu$	POWHEG + PYTHIA8	12.178
ggWW	all	WW $\to 2\ell 2\nu$	MCFM + PYTHIA8	0.5748
Electroweak WW	all	WWjj $\to 2\ell 2\nu$ jj	MADGRAPH + PYTHIA8	0.08875
$t\bar{t}$	all	leptonic decays	POWHEG + PYTHIA8	87.310
Single top, s-channel	all	leptonic decays	MADGRAPH + PYTHIA8	3.360
Single top, t-channel	all	inclusive decays	POWHEG + PYTHIA8	70.69
Single top, tW	all	inclusive decays	POWHEG + PYTHIA8	71.20
DY	all	DY($\ell\ell$) $m_{\ell\ell}$ <50 GeV	MADGRAPH + PYTHIA8	18610
	all	DY($\tau\tau \to e\mu$)	MADGRAPH + PYTHIA8	255.630
Wγ	all	W$\gamma \to \ell\nu\gamma$	MADGRAPH + PYTHIA8	405.271
Zγ	all	Z$\gamma \to 2\ell\gamma$	MADGRAPH + PYTHIA8	58.83
WZ/γ^*	all	WZ/$\gamma^* \to 3\ell\nu$	POWHEG + PYTHIA8	58.59
WZ ($2\ell 2q$)	all	WZ $\to 2\ell 2q$	MADGRAPH + PYTHIA8	5.595
ZZ ($2\ell 2\nu$)	all	ZZ $\to 2\ell 2\nu$	POWHEG + PYTHIA8	0.564
ZZ ($2\ell 2q$)	2016	ZZ $\to 2\ell 2q$	POWHEG + PYTHIA8	3.220
	2017,2018	ZZ $\to 2\ell 2q$	MADGRAPH + PYTHIA8	3.220
ZZ (4ℓ)	all	ZZ $\to 4\ell$	POWHEG + PYTHIA8	1.212
ZZZ	all	ZZZ $\to$ any	MADGRAPH + PYTHIA8	0.01398
WZZ	all	WZZ $\to$ any	MADGRAPH + PYTHIA8	0.05565
WWZ	all	WWZ $\to$ any	MADGRAPH + PYTHIA8	0.16510
WWW	all	WWW $\to$ any	MADGRAPH + PYTHIA8	0.18331

The detector response is modelled using a precise description of the CMS detector, for which the GEANT4 package is used [60]. The effect of the additional pileup interactions is simulated adding minimum bias events generated with PYTHIA8. The number of events depends on the year and follows the profile of the number of interactions produced in the collisions, using information on the instantaneous luminosity provided by the accelerator. The average number of pileup interactions was 27 for 2016 and 37 for 2017 and 2018, as presented in Fig. 2.2.

4.2 HWW analysis objects

The final state objects in this analysis are a pair of different-flavour opposite-sign leptons ($e^+\mu^-$ or μ^+e^-) from the decay of the two W bosons and a significant missing energy in the transverse plane indicating the presence of neutrinos in the event.

4.2.1 Muons

Based on the general CMS muon identification criteria, an optimisation study was carried out, deciding as a starting point between tight ID and medium ID as possible choices (defined in 3.2.1). As a figure of merit, we evaluated the signal significance, calculated using signal (S) and background (B) yields as: $\textit{significance}= \frac{S}{\sqrt{S+B}}$.

We used a simulated signal sample of gluon fusion events and simulated samples of all the main backgrounds, and applied the same selection cuts used in the signal region of the analysis (described in Table 5.1). Additional requirements on the transverse (d_{xy}) and longitudinal (d_z) impact parameters (IP) with respect to the primary vertex were evaluated. This IP requirement, stricter than the one required in the tight ID definition, helps to reduce the number of nonprompt leptons coming from pileup or from W+jets processes.

The highest significance after applying the full event selection was achieved by using the tight ID working point plus the stricter impact parameter requirement, as given in Table 4.5. An example with the results for the ggH 0-jet region with low subleading lepton p_T (p_{T2} <20 GeV) is included in Table 4.6. This region is particularly affected by nonprompt lepton contamination.

Table 4.5: List of additional requirements for the tight muon objects in the H → WW analysis

Additional H → WW analysis requirements
- $\lvert d_{xy}\rvert$ <0.01 cm for p_T <20 GeV and $\lvert d_{xy}\rvert$ <0.02 cm for p_T >20 GeV
- $\lvert d_z\rvert$ <0.1 cm
- ttH MVA >0.8

Table 4.6: Number of expected events and signal significance in the 0-jet low p_{T2} region for the possible muon IDs, using MC samples for signal and backgrounds normalised to the luminosity of the 2017 dataset. Tight ID plus the impact parameter cuts has the best performance.

Selection	Signal HWW	WW	DY	tt	single top	W+jets	Significance
Medium ID	68	474	74	65	21	4038	**0.99**
Tight ID	68	470	33	65	21	951	**1.69**
Tight ID plus IP cuts	67	454	31	61	19	705	**1.83**

On top of the selection above, we apply an additional requirement to further reduce the nonprompt lepton background, using a multivariate analysis technique (MVA) [104] developed in the context of the analysis of the associated production of a Higgs boson with a top quark pair [105] as a discriminator between prompt and nonprompt leptons. It employs many variables such as the isolation of the lepton with respect to charged and neutral particles, the ratio of the p_T of the lepton to the p_T of the nearest jet, the b-tagging discriminant or the impact parameters of the lepton. The full selection that defines muons in our analysis combines the previous tight ID definition described in Table 3.1 and the impact parameter and MVA requirements, included in Table 4.5.

The muon relative isolation requirement used is $ISO_{tight} < 0.15$ in a cone of ΔR <0.4 around the muon (see 3.1). A less strict isolation requirement of $ISO_{loose} < 0.4$ is used to define a *loose* muon. Events with a third lepton with p_{T} >10 GeV that passes the loose muon criteria are rejected. This loose muon definition is also used later for the nonprompt lepton background studies, as described in section 4.3.5.

The muon momentum determination may be affected by instrumental effects such as detector misalignment, software reconstruction or uncertainties in the magnetic field. Its performance is studied in samples of $\mathrm{Z}/\gamma^* \to \mu\mu$ events [106], where the position and width of the mass peak in the reconstructed muon pair invariant mass distribution are compared with the value of the Z boson parameters in the PDG [26]. Corrections to the data and MC samples are derived as a function of the charge, p_{T}, η, and ϕ of the muons independently. The result after applying these corrections significantly improves the data/MC agreement, as can be seen in Fig. 4.1 and 4.2.

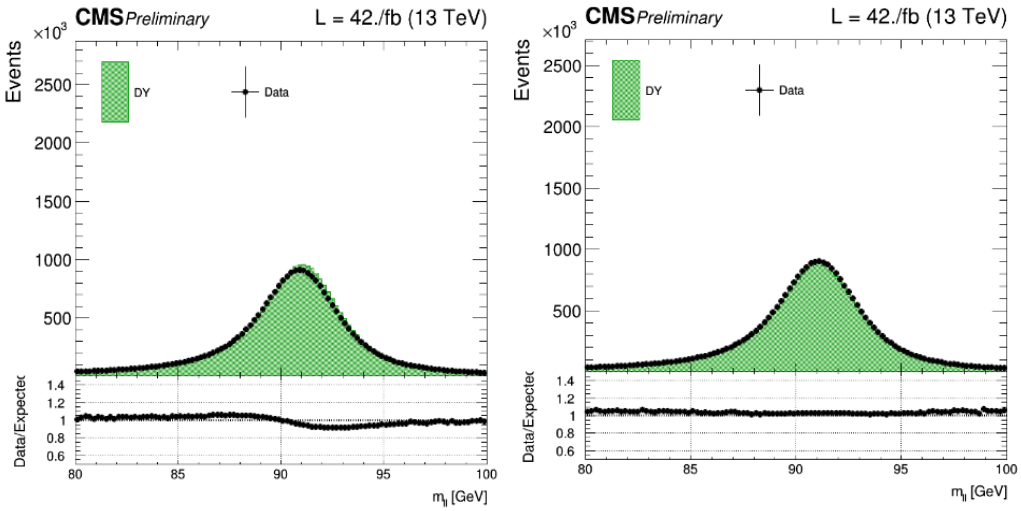


Figure 4.1: Comparison of the data/MC description of the dimuon invariant mass distribution in the DY control region before (left) and after (right) applying the muon momentum corrections to 2017 data and MC. The wiggle around the Z mass peak is flattened with the changes.

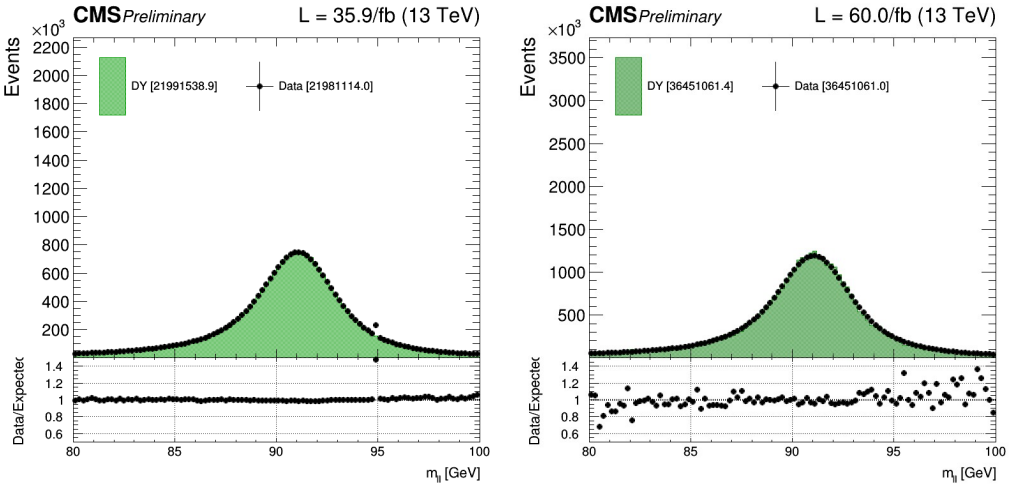


Figure 4.2: Dimuon invariant mass distribution in the DY control region for 2016 (left) and 2018 (right) after applying the muon momentum corrections.

4.2.2 Electrons

Electron identification criteria are based on the observables described in section 3.2.2. Setting requirements on the values for each of those variables, a *loose* and *medium* electron identification are defined. Tables 4.7 and 4.8 summarise the loose and a medium electron ID, respectively, with a list of all the relevant variables and their requirements. They correspond to the general recommendations for CMS analyses, with an additional impact parameter (d_{xy} and d_z) requirement to reduce the contribution of nonprompt leptons coming from pileup or from W+jets processes.

Table 4.7: Loose electron criteria for 2016

Loose ID cut-based requirements

ID variable	Barrel	Endcaps
$\lvert\Delta\eta_{in}\rvert<$	0.007	0.009
$\lvert\Delta\phi_{in}\rvert <$	0.15	0.10
full 5×5 $\sigma_{\eta\eta} <$	0.01	0.03
H/E <	0.12	0.10
$\lvert 1/E - 1/p\rvert <$	0.05	0.05
|PF relative isolation| <	0.15	0.15
Expected missing inner hits <=	1	1
Pass conversion veto	YES	YES

Additional stricter requirements in our analysis

ID variable	Barrel	Endcaps
$\lvert d_{xy}\rvert <$	0.05	0.1
$\lvert d_z\rvert <$	0.1	0.2

Table 4.8: Medium electron criteria for 2017 and 2018

Medium ID cut-based requirements

ID variable	Barrel	Endcaps
$\lvert\Delta\eta_{in}\rvert<$	0.004	0.007
$\lvert\Delta\phi_{in}\rvert <$	0.06	0.03
full5x5 $\sigma_{\eta\eta} <$	0.01	0.03
H/E <	0.12	0.10
$\lvert 1/E - 1/p\rvert <$	0.05	0.05
|PF relative isolation| <	0.15	0.15
Expected missing inner hits ≤	1	1
Pass conversion veto	YES	YES

Additional stricter requirements in our analysis

ID variable	Barrel	Endcaps
$\lvert d_{xy}\rvert <$	0.05	0.1
$\lvert d_z\rvert <$	0.1	0.2

The looser electron definition is used to reject events with three leptons and to perform nonprompt lepton background studies. We use the loose electron definition for 2016 and the medium electron definition for 2017 and 2018. They differ slightly, but the 2017 and 2018 definition being stricter for some variables allows for a better nonprompt background estimation later on. In general, medium is referred to as the 'looser' electron definition for 2017 and 2018.

The definition of the tight identification for electrons in the analysis follows an MVA approach. A single discriminant variable is computed with a Boosted Decision Tree (BDT) algorithm using as input the full list of identification observables. Two selections based on this MVA discriminant corresponding to 80% and 90% signal efficiency in a Z→ee MC sample are the starting point to define the analysis ID requirements.

To choose the optimal electron identification criteria a performance study was carried out comparing these two MVA selections and a few relative isolation requirements. Two quantities are used as a figure of merit to choose the best definition, *signal efficiency* and *background rejection*, which correspond to the following probabilities:

- Signal efficiency : Probability that an electron from a signal events passes the tighter criteria if it passed the looser selection.
- Background rejection : Probability that an electron from a background process does not pass the tighter criteria if it passed the looser selection.

In order to optimise the identification criteria we tried to maximise both signal efficiency and background rejection. The simulated Higgs boson events produced through gluon fusion were used to calculate the signal efficiency and W+jets, semileptonic $t\bar{t}$ and WW events were used to estimate the background rejection. The tight electron definition with the best performance is the one corresponding to the MVA requirement with 90% efficiency and a relative isolation of ISO_{rel}<0.06 in a cone of $\Delta R = 0.3$, as summarised in Table 4.9.

Table 4.9: List of our electron selection criteria used in the analysis

ID variable
MVA ID discriminant cut (maximum signal efficiency and background rejection)
|Relative PF Isolation| < 0.06

4.2.3 Jets

Higgs boson production through the gluon fusion process has no quarks or gluons in the final state at tree level considering a fully leptonic Higgs decay like $H \to WW \to 2\ell 2\nu$. However, there can be jets coming from the initial or final state radiation (ISR and FSR). Given the large background contribution from top pair production, events selected for the gluon fusion channel analysis are categorised based on the number of jets in the event. The division is in three categories: 0 jets, 1 jet and 2 jets or more, with the 0-jet category being the one that determines most of the sensitivity of the analysis. We use the standard CMS definition of PF jets clustered with the anti-k_T algorithm with a distance parameter of 0.4 after applying charged hadron subtraction for pileup mitigation. The jet energy is calibrated using the jet energy corrections described in section 3.2.3.

W bosons coming from top quark pair decays are always produced in association with b quarks. Therefore, requiring a low probability of being originated by a b quark for the jets by using b-tagging techniques is the most important tool to reduce the top quark pair background. CMS has developed sophisticated methods to identify jets coming from the hadronisation of b quarks (b jets). An MVA b-tagging algorithm [74] is used to assign to each jet a discriminant based on tracking and vertex information, and that discriminant is later used to reject events with b quarks. A jet is considered not to have originated from a b quark if the value of the b-tagging discriminant does not surpass a certain threshold. Several b-tagging algorithms have been tested, such as the Combined Secondary Vertex V2, Combined MVA v2, Deep Flavour and Deep CSV [107].

The signal significance was evaluated after rejecting events with b jets according to several b-tagging algorithm discriminants. The best performance was obtained for the Deep CSV discriminant with the loose b-tag working point. This corresponds in 2017 to an efficiency of correctly identifying a b-jet of 86.4% and a mistag rate of wrongly identifying a non-b jet of 10%, measured in a control sample of $t\bar{t}$.

4.2.4 Missing transverse energy

The missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,miss}$ is defined as the negative vectorial sum in the transverse plane of all PF particle candidate momenta and its modulus is called p_{T}^{miss} or E_{T}^{miss}. The PUPPI algorithm (see section 3.2.4) is used for pileup mitigation, weighing the momentum of each particle with the probability of that particle to originate from the primary interaction vertex. This was found to have a better resolution as well as a better data to MC agreement than the PF E_{T}^{miss} alone. The corrections applied to the jet energy scale are propagated to the E_{T}^{miss} calculation and, to avoid spurious high-E_{T}^{miss} events, appropriate E_{T}^{miss} filters are applied to MC and data for all years.

4.2.5 First validation

As a first test of the object definitions presented in this section, we checked the data/MC agreement in a Z control region, defined by requiring two opposite-sign leptons in the final state with an invariant mass between 60 and 120 GeV, where the invariant mass is defined as:

$$m_{\ell\ell} = \sqrt{(E_{\ell 1} + E_{\ell 2})^2 - |\vec{p}_{\ell 1} + \vec{p}_{\ell 2}|^2}\,. \tag{4.1}$$

Figures 4.3 and 4.4 illustrate the general comparisons of observation and prediction in the dielectron and dimuon final state. Simulated events are reweighted to match the p_{T} distribution measured in the data $\mathrm{Z} \to \mu\mu$ events [108]. The distributions show good agreement in both flavour channels in 2017 and 2018, but in 2016 the subleading lepton p_{T} and dilepton p_{T} distributions are slightly deviated at very high values. Since in this work we study final states with different-flavour leptons, we use another kind of samples to simulate the DY($\tau\tau$) process that do not require p_{T} reweighting, but that is explained in the following backgrounds section 4.3.

4.2.6 General event selection

Events are selected when they contain with two high p_{T} lepton candidates with opposite sign, originating from the primary vertex. These events are first selected online by single and double lepton triggers and depending on the flavour of the leptons they are categorised as dielectron, dimuon, or eμ pairs.

According to the criteria explained previously in this section, the leptons are required to be well identified and isolated. Leptons are identified in the angular region given by the detector acceptance, which corresponds to $|\eta|< 2.5$ for electrons and $|\eta|< 2.4$ for muons. The transverse momentum of the leptons is required to be higher than 13 GeV for electrons and higher than 10 GeV for muons, the difference in p_{T} threshold resulting from the trigger definitions. Based on the characteristics of the leptons from W boson decays, we select events with at least one lepton with $p_{\mathrm{T}} > 25$ GeV.

We consider PF jets with the aforementioned characteristics with $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 5.0$. For the analysis of the data recorded in the year 2017, to eliminate spurious jets caused by detector noise, all jets in the pseudorapidity range $2.5<|\eta|<3.0$ are excluded. For the purpose of vetoing events containing b-tagged jets, we tag jets in the region with $|\eta| <2.4$ and with a p_{T} of at least 20 GeV. All jets should be well separated from the isolated leptons, the distance between them should be at least $\Delta\mathrm{R} = 0.3$ for electrons and 0.4 for muons.

In the $\mathrm{H} \to \mathrm{WW}$ analysis, the final state has two neutrinos, which cannot be reconstructed independently. We measure the sum of the transverse momentum of the two neutrinos and therefore we do not have access to the full kinematics of the event. However, there are two observables that are measured with great discriminating power between signal and background processes. One of them is the **dilepton invariant mass** $m_{\ell\ell}$ and the other is the **missing transverse mass**

$$m_{\mathrm{T}}^{\mathrm{H}} = \sqrt{2|\vec{p}_{\mathrm{T}}^{\,\ell\ell}||\vec{p}_{\mathrm{T}}^{\,miss}|\left[1 - \cos\Delta\phi(\vec{p}_{\mathrm{T}}^{\,\ell\ell}, \vec{p}_{\mathrm{T}}^{\,miss})\right]}, \tag{4.2}$$

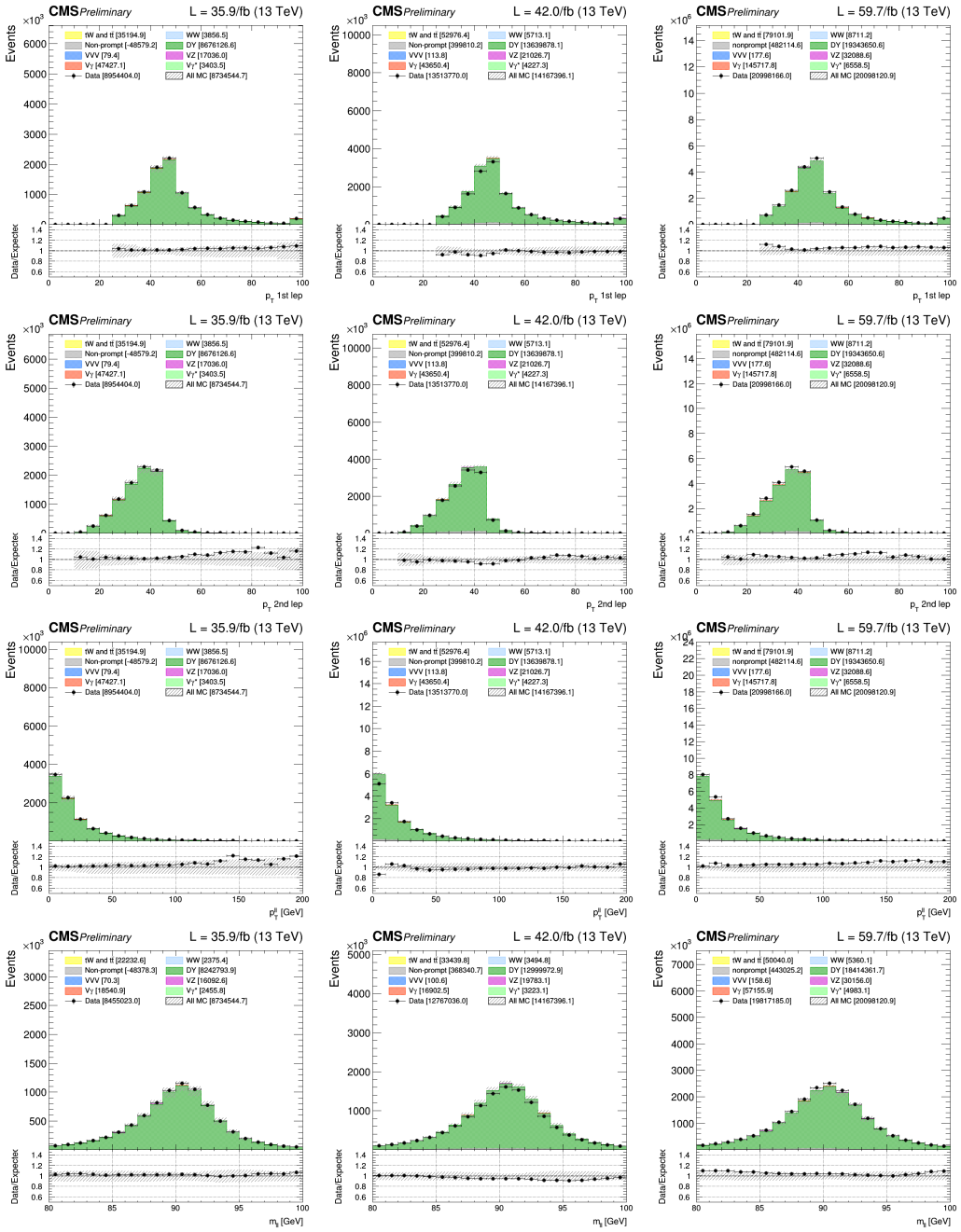


Figure 4.3: Comparison of observation and prediction in the Z→ *ee* dominated phase space from the 2016 (left), 2017 (center), and 2018 (right) data sets. From top to bottom: leading lepton p_{T}, subleading lepton p_{T}, dilepton p_{T} and dilepton invariant mass.

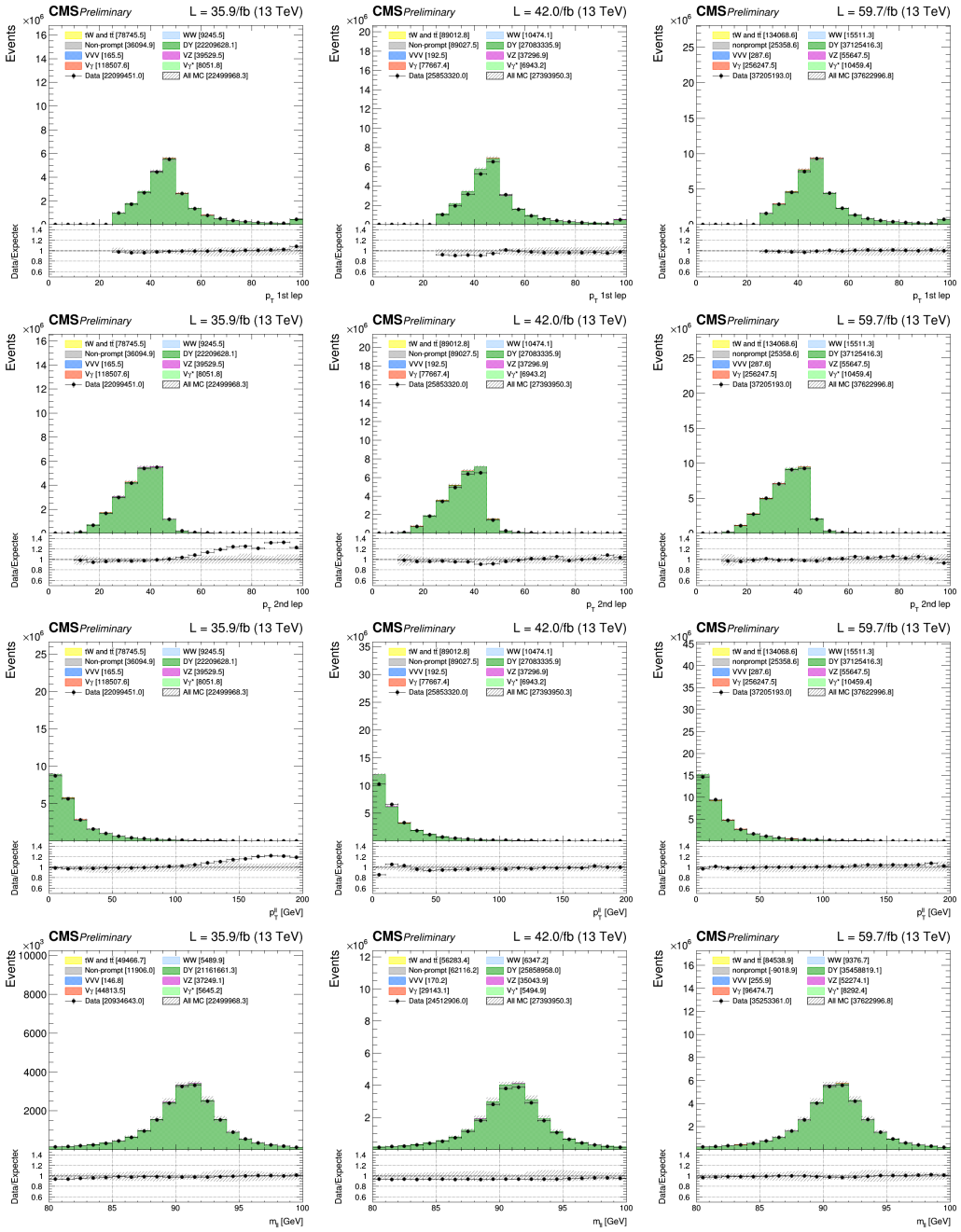


Figure 4.4: Comparison of observation and prediction in the Z$\rightarrow \mu\mu$ dominated phase space from the 2016 (left), 2017 (center), and 2018 (right) data sets. From top to bottom: leading lepton p_{T}, subleading lepton p_{T}, dilepton p_{T} and dilepton invariant mass.

where $\vec{p}_{\mathrm{T}}^{\ell\ell}$ is the sum of the transverse momenta of the two leptons and $\Delta\phi$ is the azimuthal opening angle between $\vec{p}_{\mathrm{T}}^{\ell\ell}$ and $\vec{p}_{\mathrm{T}}^{miss}$. This definition uses the maximum information available from the E_{T}^{miss} and the leptons to compute a magnitude similar to the invariant mass, but without the longitudinal component of the momenta. Both $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ are the basis for the two-dimensional distributions that are used for the Higgs boson signal extraction fit.

4.3 Background reduction and estimation

Many SM processes produce events with a final state containing some of the characteristics of the signal process explained in section 4.2. Feynman diagrams depicting the main background processes are shown in Fig. 4.5.

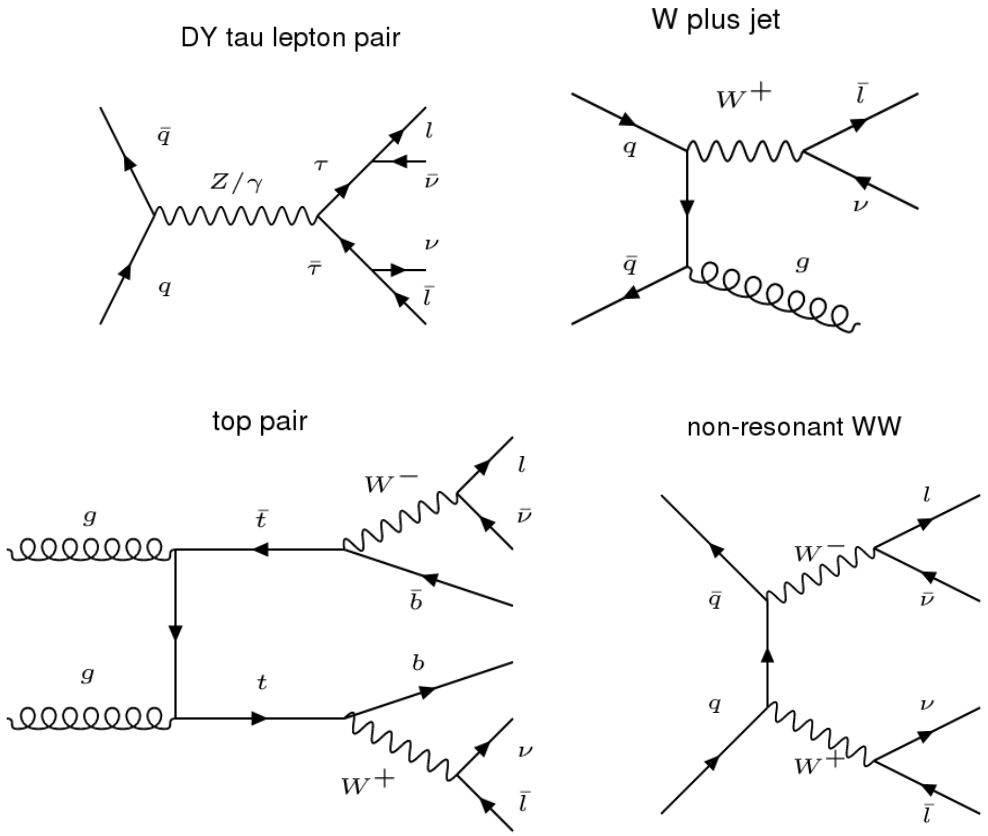


Figure 4.5: Feynman diagram examples for common background processes of Drell-Yan, W plus jets, top quark pair and non-resonant WW production.

This section describes the main backgrounds of the H → WW analysis, the selection applied to reduce them and the method used to estimate the remaining contribution, with some checks to validate those estimations. The contribution is obtained from simulated MC samples for most of these processes. However, for the background which involves a jet misidentified as a lepton, also called *fake lepton* or *nonprompt lepton* background, a more involved method using data control samples is presented. Several control regions enriched in these backgrounds are defined to check the validity of these estimations. To define these regions we require a kinematic selection different from the one applied to define our signal region, targeting the appropriate process.

4.3.1 Drell-Yan background

The production of two high-p_{T} leptons through DY is one of the most relevant background processes in the H $\rightarrow$ WW analysis. The two isolated leptons in the final state partially mimic our requirements. In our final state there is a muon and an electron, whereas the DY process produces same-flavour lepton pairs. However, in DY($\tau\tau$) with $\tau\tau \rightarrow \ell\nu\ell\nu$ the leptons can have different flavour as well. Moreover, neutrinos are also produced in the event, but the E_{T}^{miss} produced alongside leptons in W boson decays is larger than the quantity produced by tau lepton decays. Similarly, the dilepton transverse momentum $p_{\mathrm{T}}^{\ell\ell}$ is expected to be smaller. Signal event selection requires $E_{\mathrm{T}}^{miss} > 20$ GeV and $p_{\mathrm{T}}^{\ell\ell} > 30$ GeV to reduce the tau lepton background.

For events that satisfy the electron+muon trigger requirements, the DY($\tau\tau$) background contribution in our analysis is estimated with a data-driven technique called *tau-embedding* [109]. It consists on selecting Z$\rightarrow \mu\mu$ events from data, which are then removed from the event and replaced with simulated tau leptons with the same kinematic properties. The hybrid events produced in this way rely on simulation just for the decay of the tau leptons. To simulate the DY events that do not pass the electron-plus-muon triggers but pass any of the other triggers, DY($\ell\ell$) and DY($\tau\tau$) simulated samples are used, which are reweighted to match the dilepton p_{T} spectrum measured in data Z $\rightarrow \mu\mu$ events. In the signal region most of DY events come from the embedded samples.

The dilepton invariant mass $m_{\ell\ell}$ and the transverse mass $m_{\mathrm{T}}^{\mathrm{H}}$ (Eq. 4.2) distributions for signal events and DY($\tau\tau$) events are typically different. The DY $\rightarrow \tau\tau$ populates the phase space in the low $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ region. A control region enriched in DY $\rightarrow \tau\tau$ is thus defined by selecting events with $m_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV and $30 < m_{\ell\ell} < 80$ GeV. Distributions for some relevant variables in the 2018 DY($\tau\tau$) control region are shown in Figs. 4.6-4.8. As expected from a DY control region, the DY process is the biggest contributor. There is a generally a good data/simulation agreement in both shape of the distributions and normalisation, shown in the ratios summarised in Table 4.10, but in the 1-jet category of 2016 there is a slight deficit of data with respect to simulation. The distributions for the 2016 and 2017 control regions can be found in appendix A.1.1.

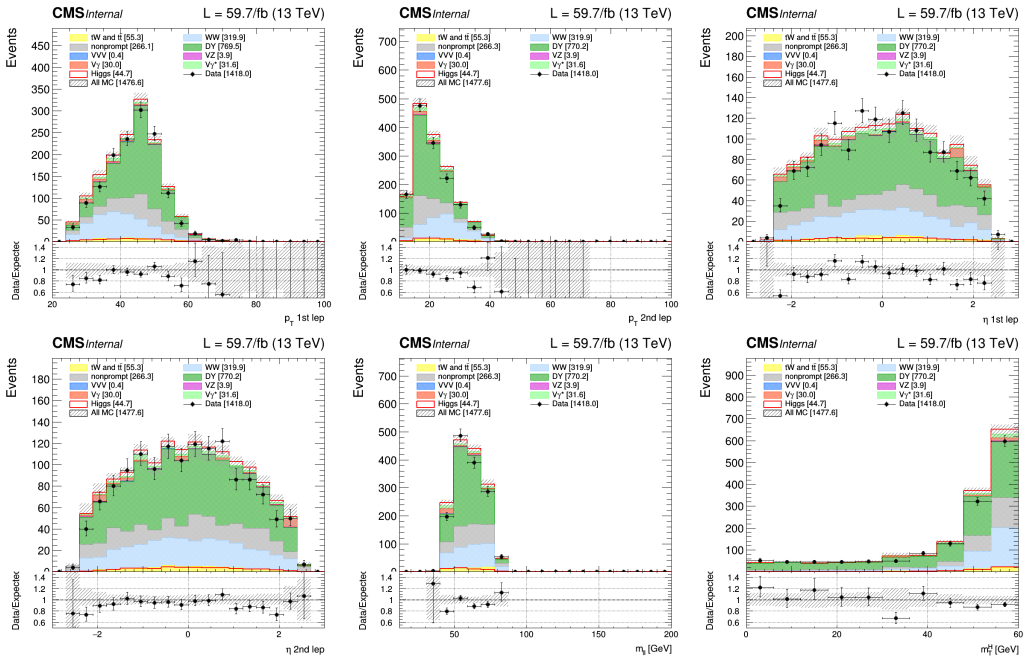


Figure 4.6: Distributions of the 2018 0-jet DY control region. The p_{T} and η of the two leptons are displayed, as well as the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

Table 4.10: Data-to-simulation scale factors for the DY background.

Category	2016	2017	2018
0-jet	0.97±0.05	0.92±0.04	0.96±0.04
1-jet	0.94±0.02	1.01±0.02	1.05±0.02
2-jet	0.97±0.04	1.02±0.04	1.03±0.03

The normalisation of the DY contribution in the signal region is estimated in these data control regions using simulated DY events, and independently for the 0-, 1- and ≥2-jet categories. The ratio of data/simulation computed in these control regions is used to constrain the DY background in the signal region for the signal extraction. Extrapolating from the control region to the signal region has an associated uncertainty in the signal region/control region event ratio.

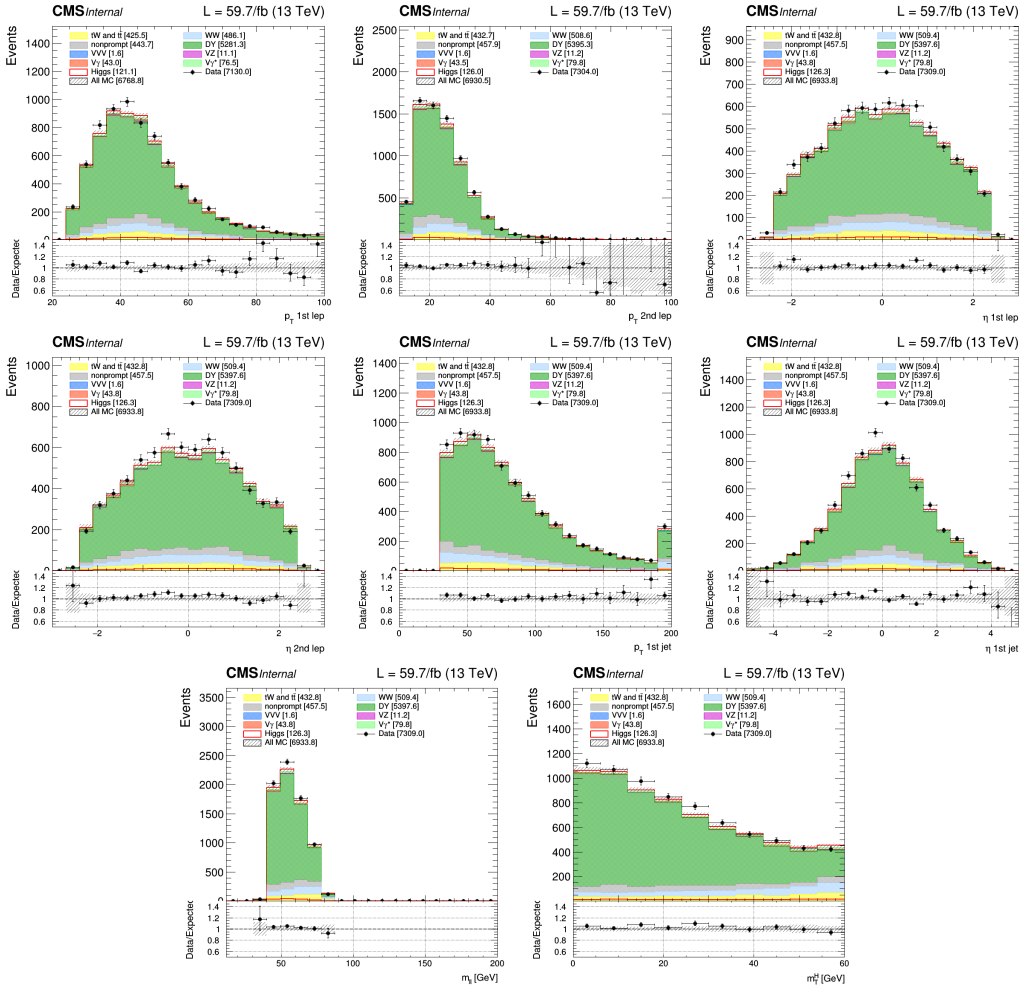


Figure 4.7: Distributions of the 2018 1-jet DY control region. The p_T and η of the two leptons and the jet are displayed, as well as the discriminating variables $m_{\ell\ell}$ and m_T^H.

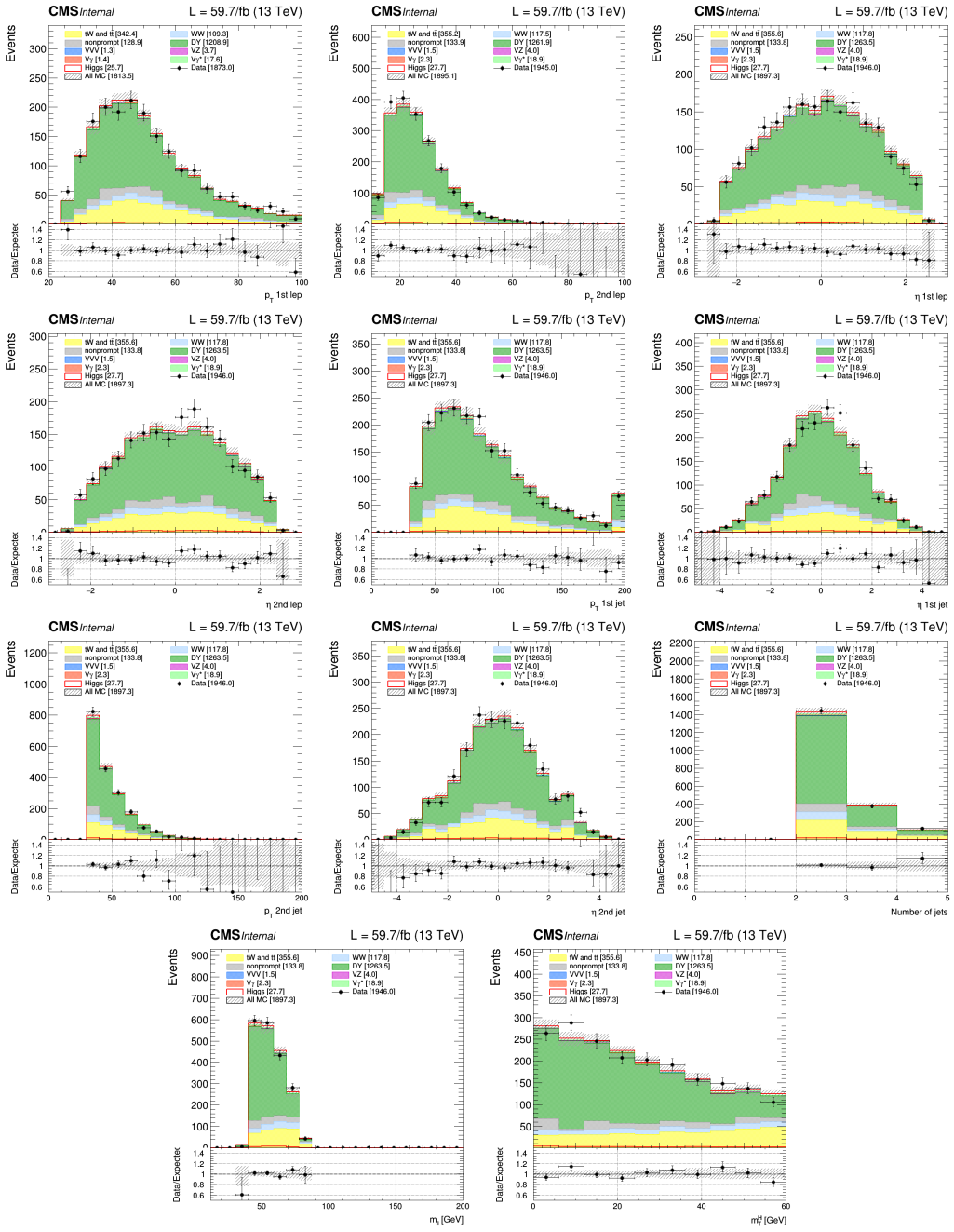


Figure 4.8: Distributions of the 2018 $\geq$ 2-jet DY control region. The p_{T} and η of the two leptons and the first two jets are displayed, as well as the number of jets in the event and the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

4.3.2 Top quark background

The top quark background is the sum of the top quark pair, associated tW and single top quark production processes, which have large cross sections at 13 TeV. 99.8% of the top quarks decay into a W boson plus a b quark. Top quark events could resemble a H $\rightarrow$ WW event if it were not for the presence of one or more b jets in the final state. The two important differences between this background and the H $\rightarrow$ WW signal are thus a higher jet multiplicity of the background and the fact that those additional jets originated from b quarks. B-tagging is therefore the most important tool to discriminate between signal and top production, and identifying b jets becomes essential in the 1- and $\geq$2-jet categories.

The shape of the distributions of the top quark background are modelled from MC, and the normalisation is obtained from data control regions separately defined for the 0-, 1- and $\geq$2-jet categories. The regions are selected with similar requirements to the signal regions with an additional b-tagging requirement.

The number of jets in the channel refers to jets with p_{T} above 30 GeV, but for b-tagging any jets with p_{T} above 20 GeV are considered. Therefore the requirement for the 0-jet control region is that a jet between 20 GeV and 30 GeV is b-tagged. In the 1-jet channel the only jet above 30 GeV should be b-tagged and in the $\geq$2-jet channels at least one jet should be identified as a b jet. In addition, the dilepton invariant mass is required to be higher than 50 GeV to reduce contribution from other processes in the top quark control regions, such as DY($\tau\tau$). $E_{\mathrm{T}}^{miss} > 20$ GeV and $p_{\mathrm{T}}^{\ell\ell} > 30$ GeV also reduce the DY($\tau\tau$) background in this region.

Figures 4.9-4.11 show some distributions of lepton and jet variables in the top quark control regions, where the this background dominates. In fact, for the categories with more than one jet, almost all events that pass the selection come from the top quark production. The shape of the distributions looks good, but not so much the normalisation, as shown in Table 4.11. There is generally less data than MC in these control regions, especially in the categories with more jets. This is why the normalisation of the top background is free to float in the signal extraction fit. The same behaviour is observed in the 2016 and 2017 control regions, which can be found in appendix A.1.2.

Table 4.11: Data-to-simulation scale factors for the top quark background.

Category	2016	2017	2018
0-jet	0.96±0.03	0.89±0.02	0.98±0.02
1-jet	0.91±0.01	0.87±0.01	0.92±0.01
2-jet	0.95±0.01	0.92±0.01	0.905±0.004

The top quark momentum in $\mathrm{t\bar{t}}$ MC events is reweighted using simulated samples to NNLO QCD plus NLO EWK precision [110]. This is done because the top quark p_{T} distribution is observed to be softer in data than in the simulated events and it is considered to be, at least partially, an effect of missing higher-order calculations. A weight is applied to $\mathrm{t\bar{t}}$ simulated events,

$$Weight(\mathrm{t\bar{t}}) = \sqrt{T(p_{\mathrm{T},top}) \cdot T(p_{\mathrm{T},antitop})}\,, \tag{4.3}$$

where the p_{T} reweighting function $T(p_{\mathrm{T}})$ used for 2017 and 2018 is

$$\begin{aligned} T_{CP5}(p_{\mathrm{T}}) = \exp[&-0.20 + 1.10 \times 10^{-4} \cdot p_{\mathrm{T}} - 1.30 \times 10^{-7} \cdot p_{\mathrm{T}}^2 + \\ &+ 58.35/(p_{\mathrm{T}} + 196.25)]\,. \end{aligned} \tag{4.4}$$

For 2016, the simulation of top quark pair production uses a different underlying event tune (CUET instead of CP5), which means that the same function cannot be used. First, a reweighting from CUET to CP5 is applied in that case,

$$T_{CP5 \rightarrow CUET}(p_{\mathrm{T}}) = 1.05 + 0.05 \cdot \tanh\left(-1.72 + 2.52 \times 10^{-3} \cdot p_{\mathrm{T}}\right)\,, \tag{4.5}$$

and the reweighting for 2016 is obtained by multiplying the two previous weights

$$T_{CUET}(p_{\mathrm{T}}) = T_{CP5}(p_{\mathrm{T}}) \cdot T_{CP5 \to CUET}(p_{\mathrm{T}}) \, . \tag{4.6}$$

The top quark contribution in the signal region is estimated in data control regions using simulated top quark production events, independently for the 0-, 1- and $\geq$2-jet categories. The ratio of data/MC computed in these control regions is used to constrain the top quark background in the signal region for the signal extraction. Extrapolating from the control region to the signal region has an associated uncertainty in the signal region/control region event ratio.

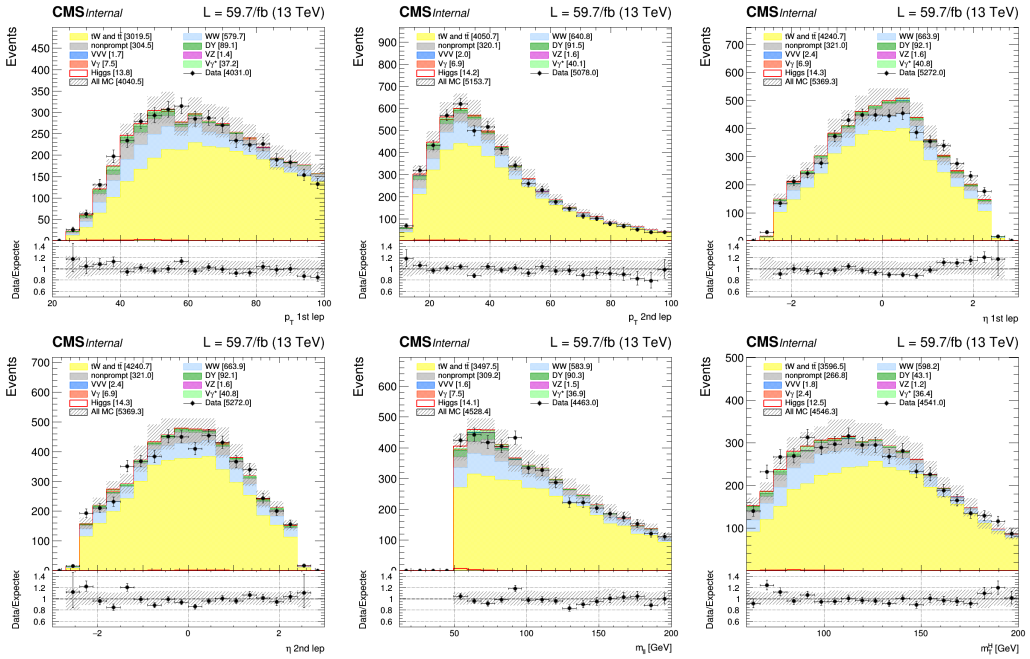


Figure 4.9: Distributions of the 2018 0-jet top control region. The p_{T} and η of the two leptons are displayed, as well as the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

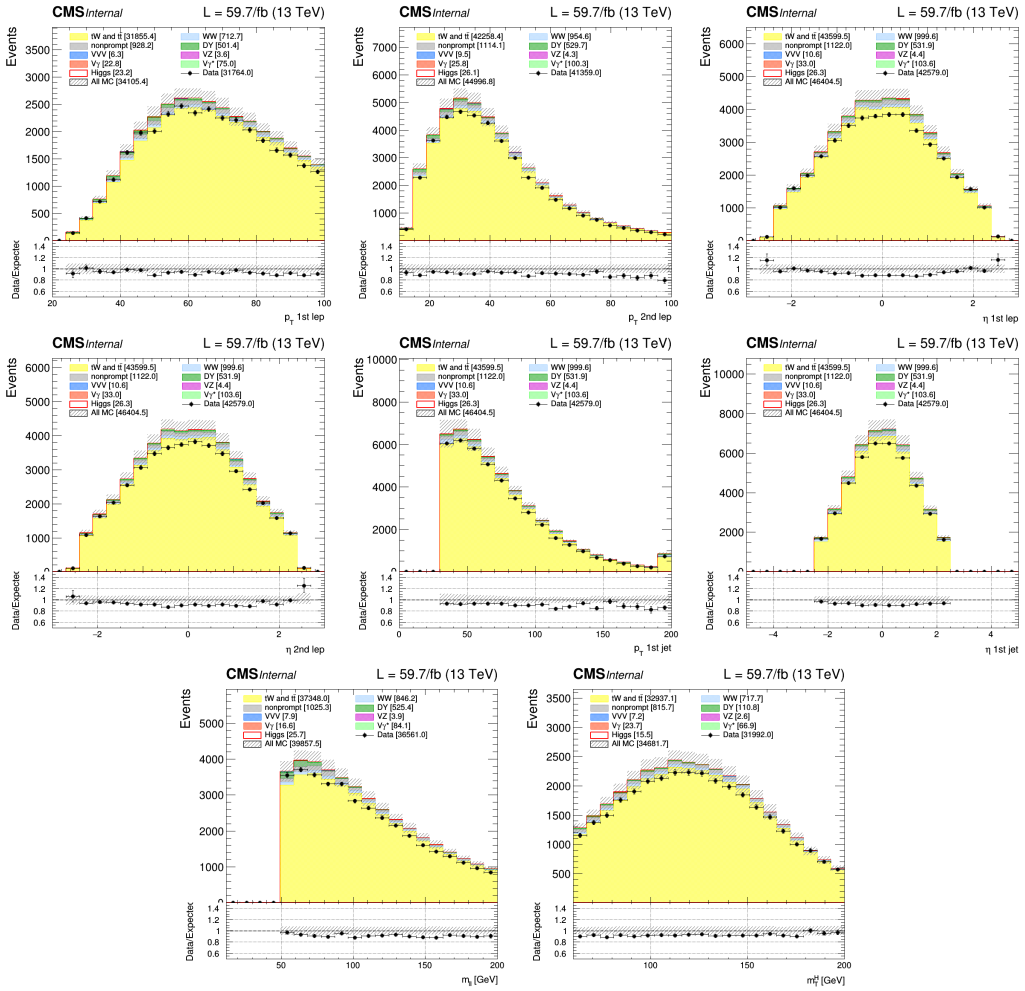


Figure 4.10: Distributions of the 2018 1-jet top control region. The p_{T} and η of the two leptons and the jet are displayed, as well as the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

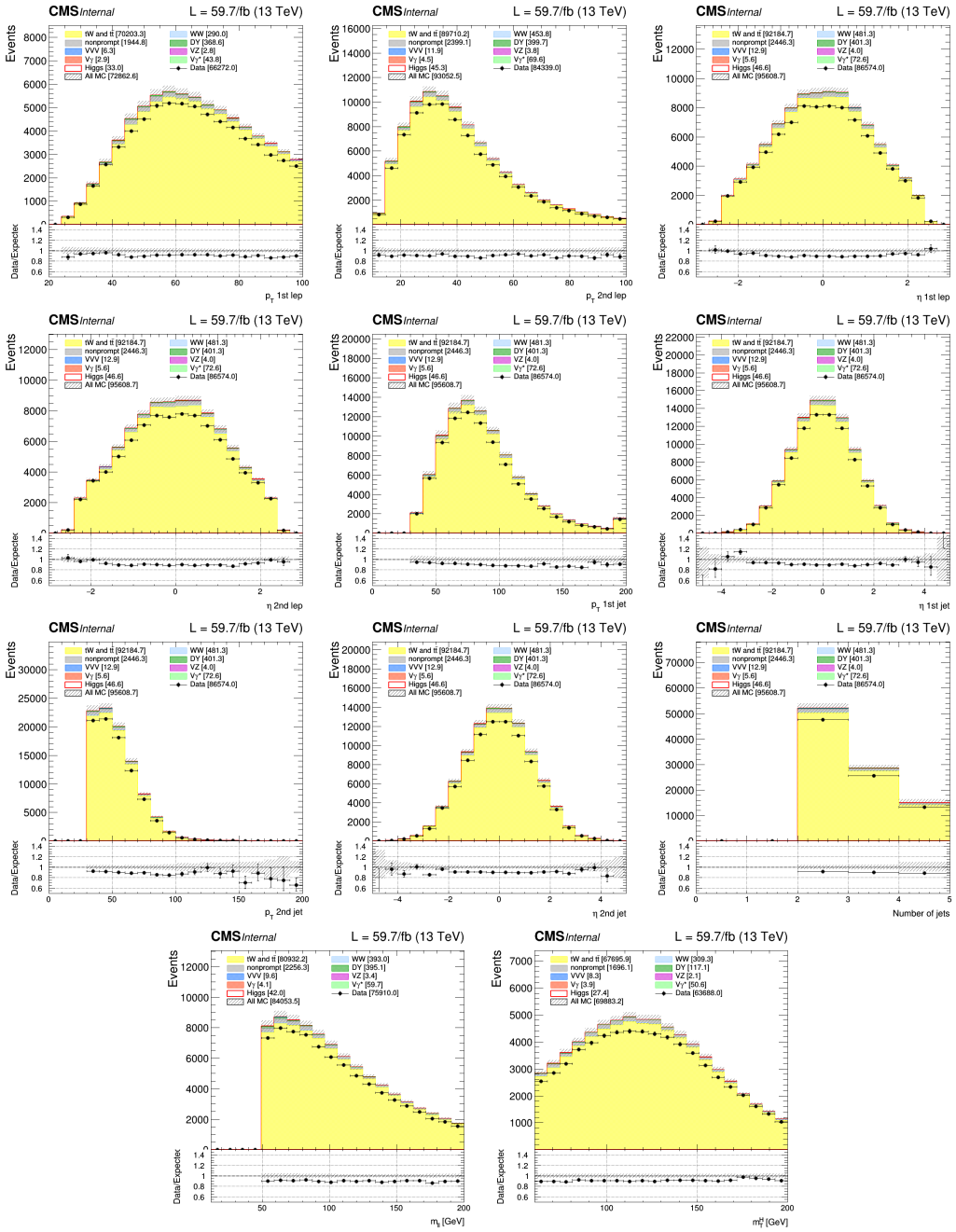


Figure 4.11: Distributions of the 2018 $\geq$ 2-jet top control region. The p_T and η of the two leptons and the first two jets are displayed, as well as the number of jets in the event and the discriminating variables $m_{\ell\ell}$ and m_T^H.

4.3.3 WW background

The nonresonant WW background is an irreducible background that has the same final state as the H → WW signal. However, since there is no resonance, the WW background is equally present at low $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ as it is for high $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$. The normalisation for this background is estimated in the fit procedure together with the signal, where it is left free to float. For validation purposes we also define a control region where the WW background dominates, by requiring $m_{\ell\ell} > 100$ GeV, $m_{\mathrm{T}}^{\mathrm{H}} > 60$ GeV and applying the rest of the signal region selection.

To reduce the DY background in this region we require $E_{\mathrm{T}}^{miss} > 20$ GeV and $p_{\mathrm{T}}^{\ell\ell} > 30$ GeV. In order to remove events where the second lepton does not come from a W boson, we compute the transverse mass $m_{\mathrm{T}}^{\ell 2}$ using the subleading lepton p_{T} and the $\vec{p}_{\mathrm{T}}^{miss}$:

$$m_T^{\ell 2} = \sqrt{2 p_{\mathrm{T}2} p_{\mathrm{T}}^{miss} \left[1 - \cos\Delta\phi\left(\vec{p}_{\mathrm{T}2}, \vec{p}_{\mathrm{T}}^{miss}\right)\right]} \tag{4.7}$$

and the requirement is that $m_{\mathrm{T}}^{\ell 2}$ should be greater than 30 GeV.

This control region distributions using 2018 samples are shown in Figs. 4.12-4.14. Unlike the previous control regions, this dedicated region is not fully dominated by the process it targets, WW. In fact, in the categories with a higher jet multiplicity the top quark contribution is larger than WW, even though events with b-tagged jets are vetoed. The results for 2016 and 2017 can be found in appendix A.1.3, and the data/MC ratio is shown in Table 4.12. The shape of the distributions looks good, but because of the top quark background normalisation difference observed before, there is also generally less data than MC in this control region too. This is especially noticeable in the categories with higher jet multiplicity that have a larger top quark contribution.

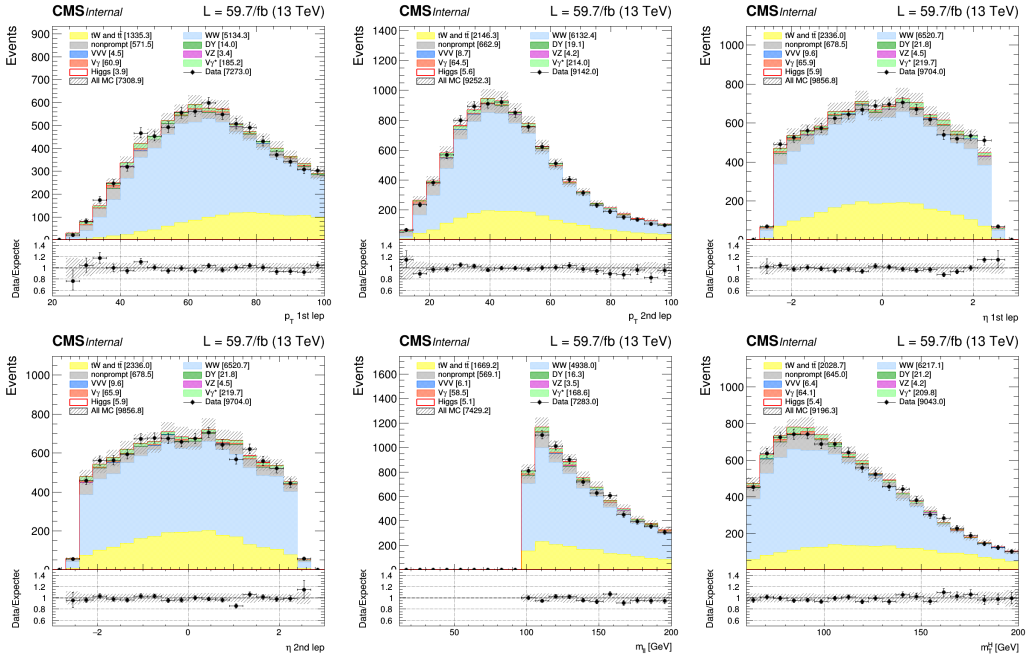


Figure 4.12: Distributions of the 2018 0-jet WW control region. The p_{T} and η of the two leptons are displayed, as well as the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

Table 4.12: Data-to-simulation scale factors for the WW background.

Category	2016	2017	2018
0-jet	0.99±0.02	0.96±0.02	0.98±0.01
1-jet	0.91±0.02	0.92±0.02	0.88±0.01
2-jet	0.91±0.03	1.02±0.03	0.89±0.02

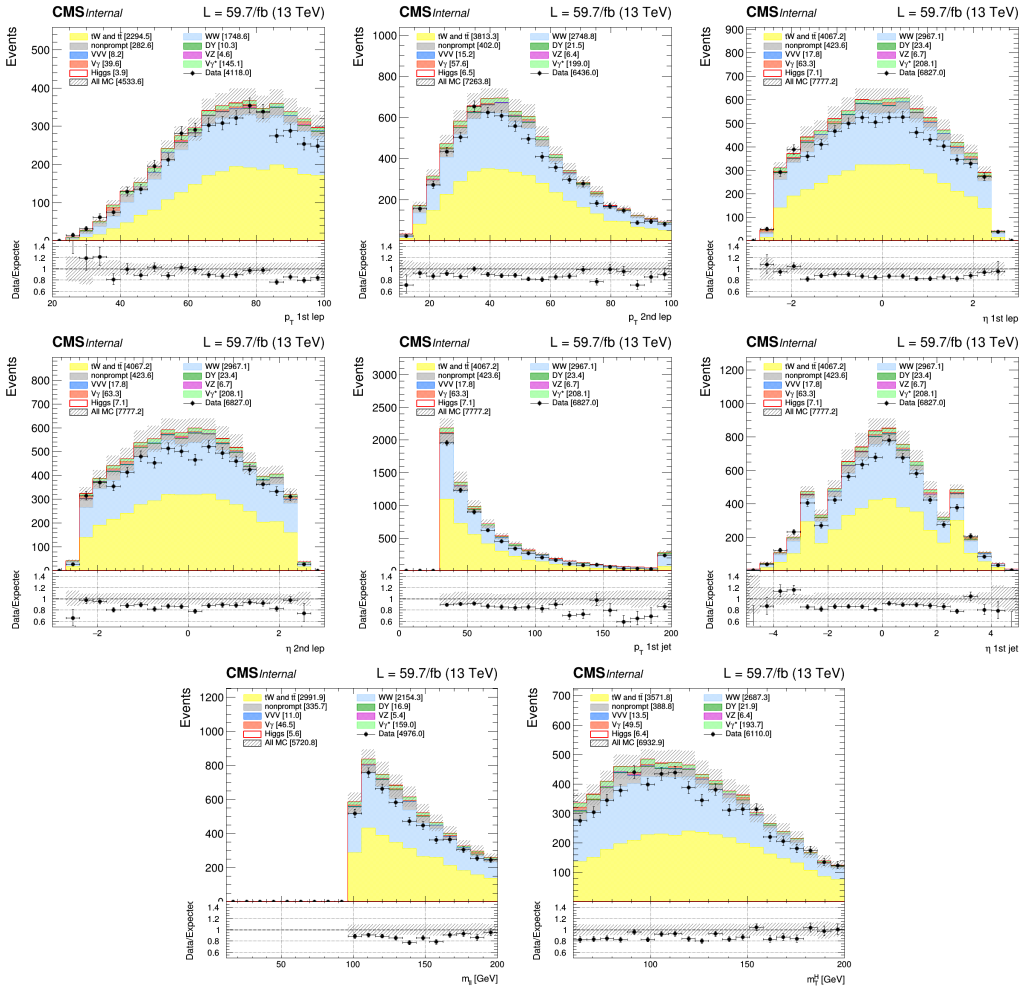


Figure 4.13: Distributions of the 2018 1-jet WW control region. The p_{T} and η of the two leptons and the jet are displayed, as well as the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

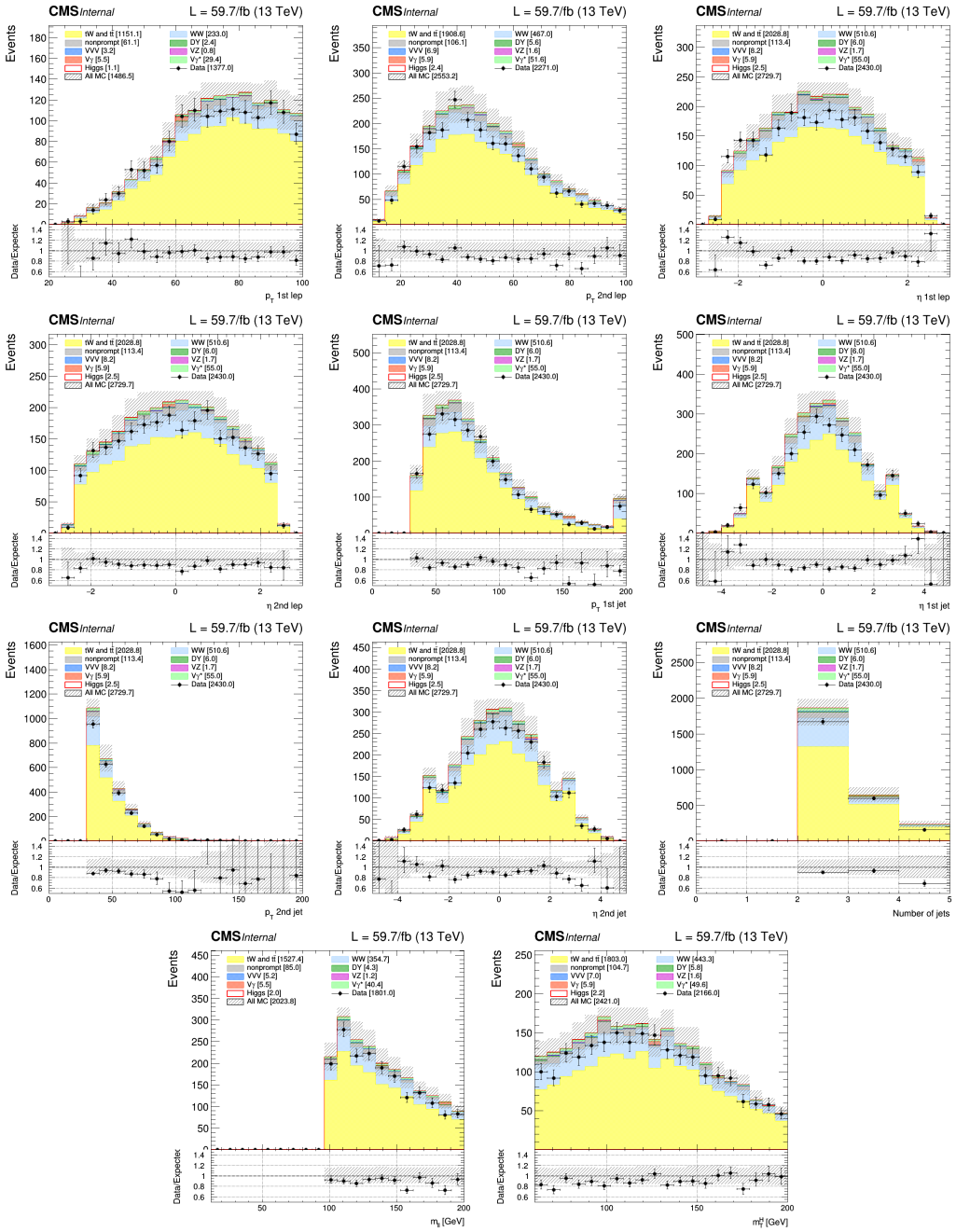


Figure 4.14: Distributions of the 2018 $\geq$ 2-jet WW control region. The p_{T} and η of the two leptons and the first two jets are displayed, as well as the number of jets in the event and the discriminating variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$.

4.3.4 Multiboson background

Other multiboson backgrounds could provide a small contribution in case one of the leptons in the final state is not identified. These backgrounds have a smaller impact because they have a small cross section and only contribute in the signal region due to inefficiencies or misidentification, since their final state contains more than two leptons. Normalisation scale factors for the WZ and Wγ^* backgrounds are obtained in two separate three-lepton control regions where these processes dominate. They require three leptons, two of them with the same flavour, opposite charge and an invariant mass close to the Z boson mass for WZ, and under 4 GeV for Wγ^*.

All backgrounds from diboson and triboson production aside from WW, WZ and Wγ^* are estimated according to their predicted theoretical cross sections and the shape is taken from MC simulation. These backgrounds are Wγ, ZZ, WWW, WWZ, WZZ and ZZZ.

4.3.5 The nonprompt lepton background

Prompt leptons are the direct decay products of particles originated in the collision or the products of the collision itself. In a fully leptonic H $\rightarrow$ WW decay there should be two prompt leptons in the final state from the decay of the two W bosons. However, there are other possible sources of leptons in the final state such as real leptons from heavy flavoured hadron decays or jets misidentified as isolated leptons. These cases are referred to as *nonprompt* leptons and also as the *fake* lepton background, since they are not leptons from the process of interest.

In the H $\rightarrow$ WW analysis, the primary source of background from lepton misidentification in the 0-jet category comes from misidentified W boson plus jets events. These events contain a real prompt lepton and missing energy from the W boson decay. If one of the jets is misidentified as a prompt lepton, the final state with two leptons and missing energy resembles the state that we are interested in. This is more likely to happen with lower p_{T} leptons, but is relevant in all the lepton phase space.

On the other hand, for the 1- and $\geq$2-jet categories, semileptonic top quark pair decays are the main source of lepton misidentification background. Other background processes such as QCD multijet events and fully hadronic top quark pair decays could produce nonprompt leptons as well, but since they involve two nonprompt leptons instead of one, they are less likely to contribute to the signal region background.

The nonprompt lepton contribution is not computed from MC because the simulations are not capable of precisely reproducing the instrumental behaviour. Instead, a data-driven method is used to determine the contribution of this background. The method used as well as the closure tests performed to validate it will be explained in this section.

The method

Let us start by discussing the method in a one-lepton final state case. The aim is to determine the number of events with a prompt (N_p) or nonprompt (N_f) lepton populating a sample of one-lepton events, selected using certain identification criteria similar to the ones of the H $\rightarrow$ WW analysis. To obtain them, we use the information on the number of events with one loose lepton (N_l). Those events either have a tight lepton (N_{t1}) or no tight leptons (N_{t0}). The following relations hold:

$$\begin{cases} N_\ell = N_p + N_f = N_{t1} + N_{t0} \\ N_{t0} = (1-p)N_p + (1-f)N_f \\ N_{t1} = pN_p + fN_f \end{cases} \tag{4.8}$$

where f and p are defined as follows:

- The **fake rate** (f) is the probability for a nonprompt lepton that passes the loose criteria to also pass the corresponding tight selection.

- The **prompt rate** (p) is the probability of a real prompt lepton that passes the loose criteria to also pass the corresponding tight selection. It is therefore very close to one and much higher than the fake rate.

These equations can be inverted to obtain the prompt and nonprompt contribution as a function of the probabilities that a lepton is identified as tight and the measured number of tight leptons:

$$\begin{cases} N_p = \frac{1}{p-f} \cdot [(1-f)N_{t1} - fN_{t0}] \\ N_f = \frac{1}{p-f} \cdot [pN_{t0} - (1-p)N_{t1}] \end{cases} \tag{4.9}$$

Given a data sample, the nonprompt contribution N_f is obtained with this expression, multiplying the number of events by a weight of $\frac{p}{p-f}$ if there is a loose but not tight lepton, or $\frac{1-p}{p-f}$ if there is a tight lepton. In section 4.2, the electron and muon definitions were detailed, where a 'tight' and a 'loose' definition was given for each particle. We now use those definitions to calculate the fake and prompt rates.

Eq. 4.8 can be generalised for the two-lepton final state:

$$\begin{cases} N_{\ell\ell} = N_{pp} + N_{fp} + N_{ff} = N_{t2} + N_{t1} + N_{t0} \\ N_{t0} = (1-p)^2 N_{pp} + (1-p)(1-f)N_{fp} + (1-f)^2 N_{ff} \\ N_{t1} = 2p(1-p)N_{pp} + (f(1-p) + p(1-f))\, N_{fp} + 2f(1-f)^2 N_{ff} \\ N_{t2} = p^2 N_{pp} + pf N_{fp} + f^2 N_{ff} \end{cases} \tag{4.10}$$

where $N_{\ell\ell}$ is the number of events with two leptons passing the loose criteria, N_{pp}, N_{fp} and N_{ff} are the number of events with zero, one or two nonprompt leptons, and N_{t0}, N_{t1} and N_{t2} are the number of events with zero, one or two tight leptons. These equations can again be inverted to extract the fake lepton contribution, relating the number of prompt and fake leptons with the number of tight and loose leptons:

$$\begin{pmatrix} N_{pp} \\ N_{fp} \\ N_{ff} \end{pmatrix} = \frac{f-p}{-(p-f)^3} \cdot \begin{pmatrix} f^2 & -f(1-f) & (1-f)^2 \\ -2fp & p(1-f) + f(1-p) & -2(1-p)(1-f) \\ p^2 & -p(1-p) & (1-p)^2 \end{pmatrix} \cdot \begin{pmatrix} N_{t0} \\ N_{t1} \\ N_{t2} \end{pmatrix} \tag{4.11}$$

This means that knowing the lepton fake and prompt rates we can relate zero, one or two tight lepton events with zero, one or two nonprompt leptons. We use this to obtain an estimation for the nonprompt background using the data events with two loose leptons passing the rest of the selection criteria and applying the corresponding weights calculated with Eq. 4.11.

The fake rate is computed in a control region enriched in nonprompt leptons. The method assumes that the fake rate in the signal region is the same as in the control regions where they are determined, which is not necessarily the case. The main difference is the detector response for leptons of different kinematical properties, therefore we compute the rates as a function of the lepton η and p_{T} to consider the possible differences between the processes.

Lepton fake rate determination

The procedure to calculate the fake rate starts from a sample of leptons satisfying loose isolation and identification criteria, some of which pass the tighter lepton requirements of the H $\rightarrow$ WW selection. We determine the lepton fake rate in a QCD control region. The control region selection focuses in back-to-back dijet processes where one of the jets is misidentified as a lepton. We select events with a high p_{T} jet and a lepton candidate well separated from the jet, with $\Delta\mathrm{R}(\ell,\mathrm{j}) > 1$.

The reason why we use dijet events is that the fake rate depends both on the p_{T} of the misidentified lepton and the p_{T} of the associated jet. Since the p_{T} of the jet associated with a lepton is rarely yielding the correct value, we make use of the topology of dijet events to use the away jet momentum as a proxy to select indirectly the p_{T} of the misidentified jet. The value for the p_{T} thresholds for the far away jet varies between 20 and 35 GeV, according to Table 4.13. The jet threshold for muons was optimised with a matching between muon isolation shapes of a W+jets and $\mathrm{t\bar{t}}$ MC samples and a dijet control region with same-sign leptons. This variable was chosen because it differentiates between loose and tight muons and therefore affects the fake rate. In the 0-jet category, the dijet control region with a jet threshold of 20 GeV has the best matching between W+jets and QCD [111]. For categories with a higher number of jets the corresponding threshold is higher, since the nonprompt lepton background comes predominantly from semileptonic $\mathrm{t\bar{t}}$ events, whose jets have higher momentum on average. The thresholds that approximate the best the muon isolation distribution with the QCD selection to the isolation shape obtained with a sample of combined W+jets and $\mathrm{t\bar{t}}$ events are 25 GeV for the 1-jet category and 35 GeV for the $\geq$2-jet category.

For electrons the difference between loose and tight identification criteria is not only based on isolation. Looking at electron isolation distributions after applying the identification selection to QCD events, there was no clear jet p_{T} threshold with an optimal matching between QCD and W+jets/$\mathrm{t\bar{t}}$. However, because of the limitations of the available triggers, a fixed value of 35 GeV for the jet p_{T} threshold was used in all jet categories. The choice of a certain jet p_{T} threshold for both muons and electrons is considered a source of uncertainty that will be estimated and discussed later.

Table 4.13: Leading jet p_{T} thresholds for the dijet control region selection, in the 0-, 1- and $\geq$2-jet categories.

	0-jet category	1-jet category	$\geq$2-jet category
electrons	35 GeV	35 GeV	35 GeV
muons	20 GeV	25 GeV	35 GeV

We select a sample enriched in QCD events from data selected online by the single lepton trigger paths listed in Table 4.14. Due to their low thresholds and the high instantaneous luminosity of Run 2, these triggers paths are heavily prescaled, which means that only a fraction of the events passing the trigger are recorded and the recorded luminosity is significantly smaller than the delivered one. The corresponding integrated luminosity is also given in the table.

These trigger paths require loose identification and isolation criteria that are looser than the offline selection. Given the usual high rate of electron triggers, they require the presence of a jet with $p_{\mathrm{T}} > 30$ GeV. This is a limiting factor for the offline requirement of the away jet p_{T} previously mentioned. We use two trigger paths per particle, one for the low p_{T} range (below 20 GeV) and another for the high p_{T} range. This is done to increase the amount of data of high p_{T} leptons.

Not all leptons with this selection are nonprompt, there are real leptons from W or Z leptonic decays as well. The contamination from Z events is reduced by requiring one and only one lepton in the event. The contribution of W events is reduced by requiring $E_{\mathrm{T}}^{miss} < 20$ GeV as well as $m_{\mathrm{T}}^{\mathrm{W}} < 20$ GeV, where $m_{\mathrm{T}}^{\mathrm{W}}$ is the transverse mass reconstructed from the lepton and the missing energy:

$$m_{\mathrm{T}}^{\mathrm{W}} = \sqrt{2p_{\mathrm{T}}^{\ell}E_{\mathrm{T}}^{miss}(1-\cos(\phi_{\ell}-\phi_{E_{\mathrm{T}}^{miss}}))}\,. \tag{4.12}$$

Table 4.14: Single lepton trigger paths used for selecting the enriched QCD sample.

Electron trigger paths	$\mathscr{L}_{\text{int}}$ [fb^{-1}]		
	2016	2017	2018
Electron with 8 GeV and PF jet with 30 GeV ($p_{\text{T}} < 25$ GeV)	//	0.00397	0.00641
Electron with 12 GeV and PF jet with 30 GeV ($p_{\text{T}} < 25$ GeV)	0.01485	//	//
Electron with 23 GeV and PF jet with 30 GeV ($p_{\text{T}} \geq 25$ GeV)	0.06281	0.04347	0.03891
Muon trigger paths	$\mathscr{L}_{\text{int}}$ [fb^{-1}]		
	2016	2017	2018
Muon with 8 GeV ($p_{\text{T}} < 20$ GeV)	0.00780	0.00290	0.00856
Muon with 17 GeV ($p_{\text{T}} \geq 20$ GeV)	0.21675	0.06594	0.04578

After applying the selections described above, there may be remaining contamination from prompt leptons from electroweak (W/Z+jets) and $\text{t}\bar{\text{t}}$ events. The contribution of fully hadronic top quark pair decays was found to be negligible. Remaining EWK contamination from W and Z events was estimated from MC. Using the theoretical cross section and the effective luminosity of the samples listed in Table 4.14, the remaining EWK contribution in the samples with a tight lepton and with a loose lepton is computed and subtracted.

The MC samples used to estimate the EWK contribution are W+jets with leptonic W decay, a DY($\ell\ell$)+jets sample with a dilepton invariant mass of 10 to 50 GeV and DY($\ell\ell$)+jets with a dilepton invariant mass above 50 GeV. They are generated using MADGRAPHaMC@NLO at LO for the low mass DY and MADGRAPHaMC@NLO at NLO for the others.

Figures 4.15 and 4.16 show the relative EWK contamination for electrons and muons. They represent the fraction of events corresponding to EWK processes, estimated with MC for a given effective luminosity (see Table 4.14) divided by the total data events in the control region. It is obtained separately for the loose lepton and tight lepton selection, that are the denominator and numerator of the fake rate computation. The contamination is increasingly important at higher p_{T}, where the contribution of real leptons is larger, as expected. It is also larger for the tight lepton sample than for the loose lepton sample.

The final expression for the calculation of the fake rate is:

$$fr = \frac{N_{tight}^{data} - N_{tight}^{EWK,MC}}{N_{loose}^{data} - N_{loose}^{EWK,MC}} \tag{4.13}$$

Figures 4.17 and 4.18 show the lepton fake rates of electrons and muons. The black and red lines correspond to the estimation before and after correcting for the EWK contribution, respectively. The correction has a significant effect for $p_{\text{T}} > 20$ GeV in both electrons and muons, since there are more leptons from W or Z decays in that region.

The fake rates are much lower for muons than for electrons, which is a consequence of the loose muon and tight muon definitions being very different. The stricter tight requirement for muons obtained with the MVA discriminant reduces the fake rate to about 10%, with similar results in the three data-taking periods. It also has a larger effect at small values of the muon p_{T}, where the MVA discriminates better between fake and prompt leptons, and therefore for p_{T} <12 GeV the muon fake rate is minimal. The distribution of the fake rate with respect to η is flat for muons, but there is a slight angular dependence in the electron fake rate.

The results for the three years are drawn together in Fig. 4.19. Electrons have a slightly higher fake rate in 2017 and 2018, because the medium ID definition was used in those years instead of the loose ID (as noted in section 4.2.2). For muons the same ID definition was used for all periods and the fake rates in the three years were consistent.

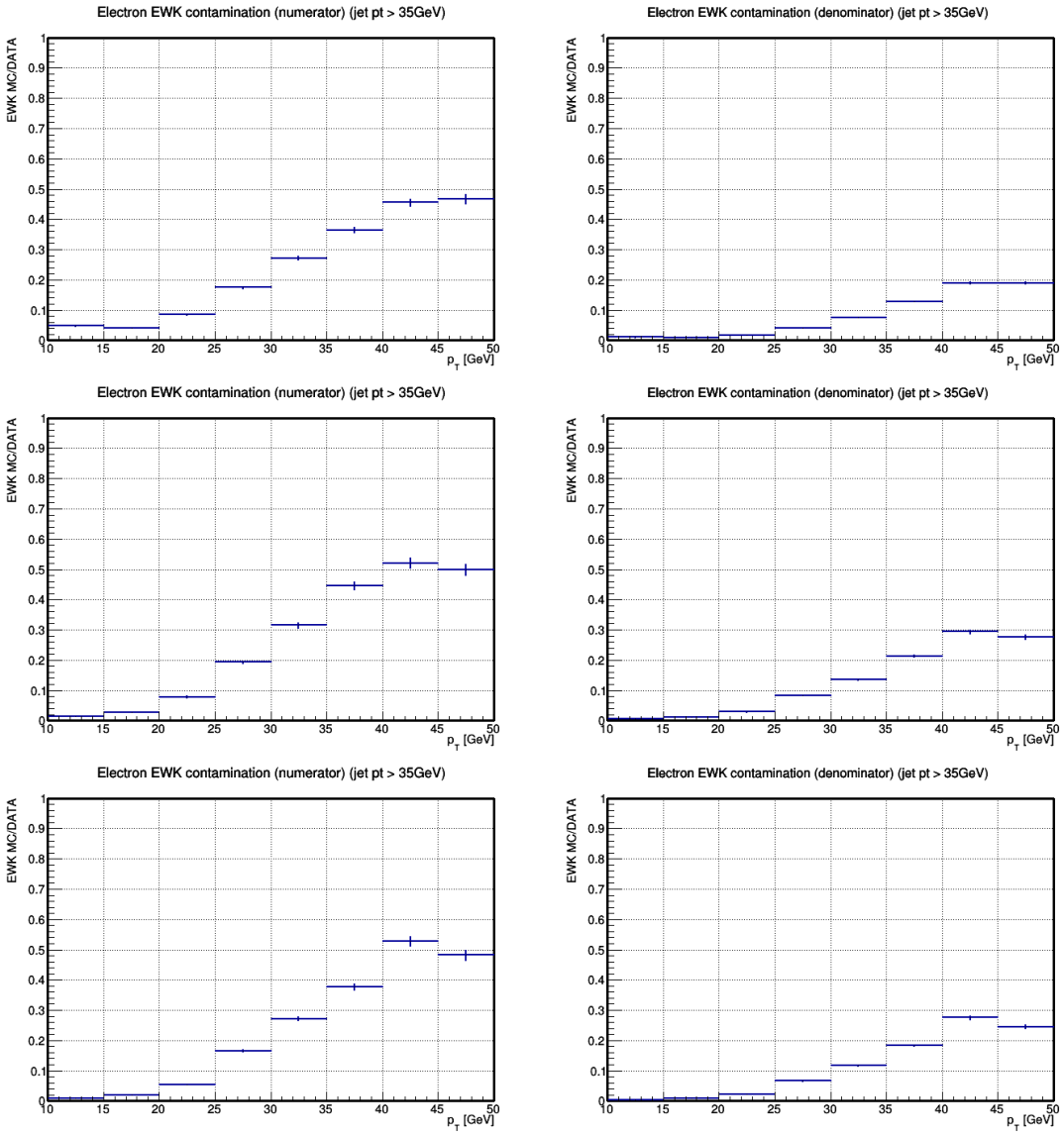


Figure 4.15: Relative EWK contamination for the tight electron sample (left) and for the loose electron sample (right) in p_{T} bins of the electron. It is computed in a control region obtained with a jet threshold of 35 GeV for the full Run 2: 2016 (top), 2017 (middle) and 2018 (bottom).

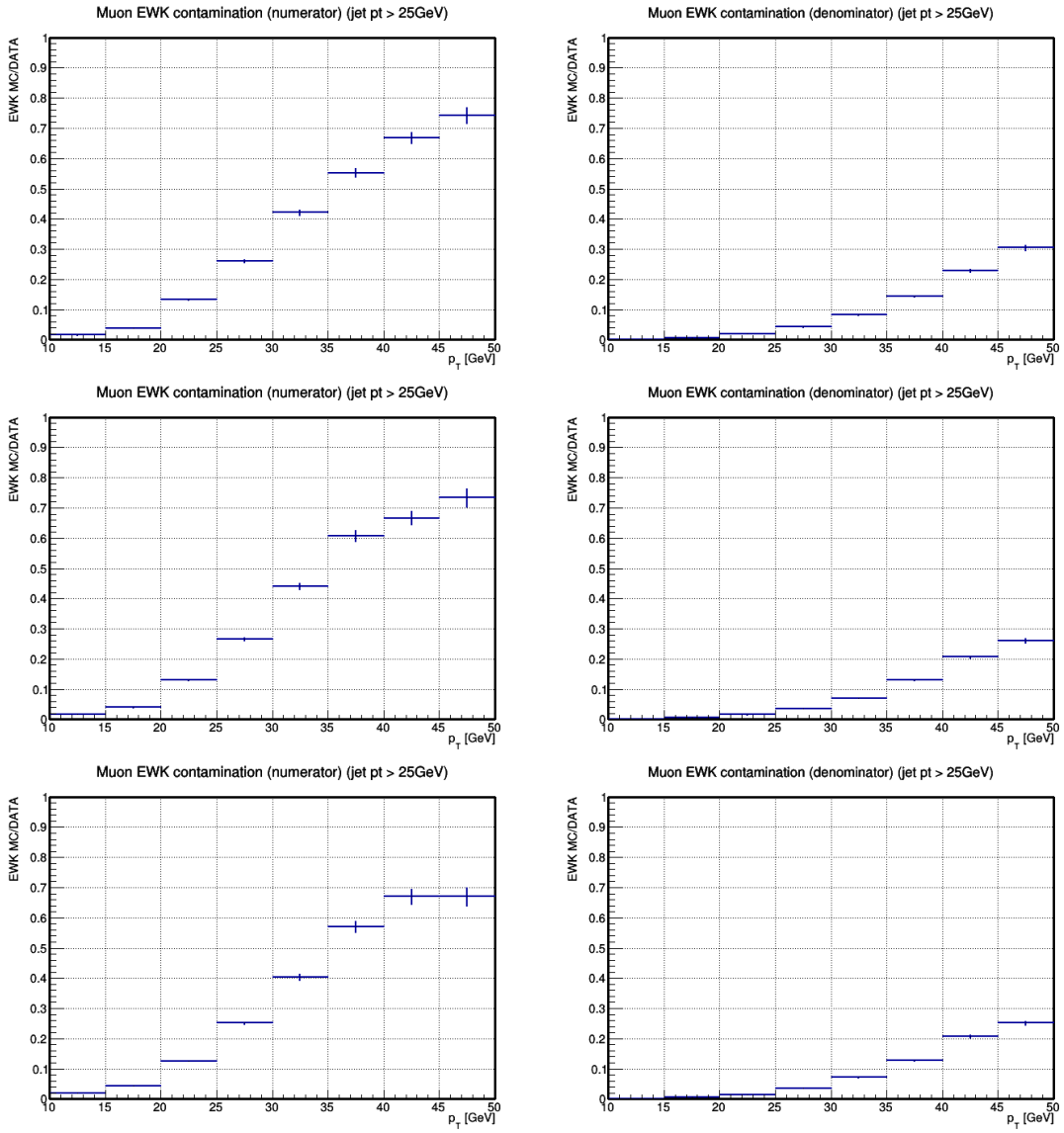


Figure 4.16: Relative EWK contamination for the tight muon sample (left) and for the loose muon sample (right) in p_{T} bins of the muon. It is computed in a control region obtained with a jet threshold of 25 GeV for the full Run 2: 2016 (top), 2017 (middle) and 2018 (bottom).

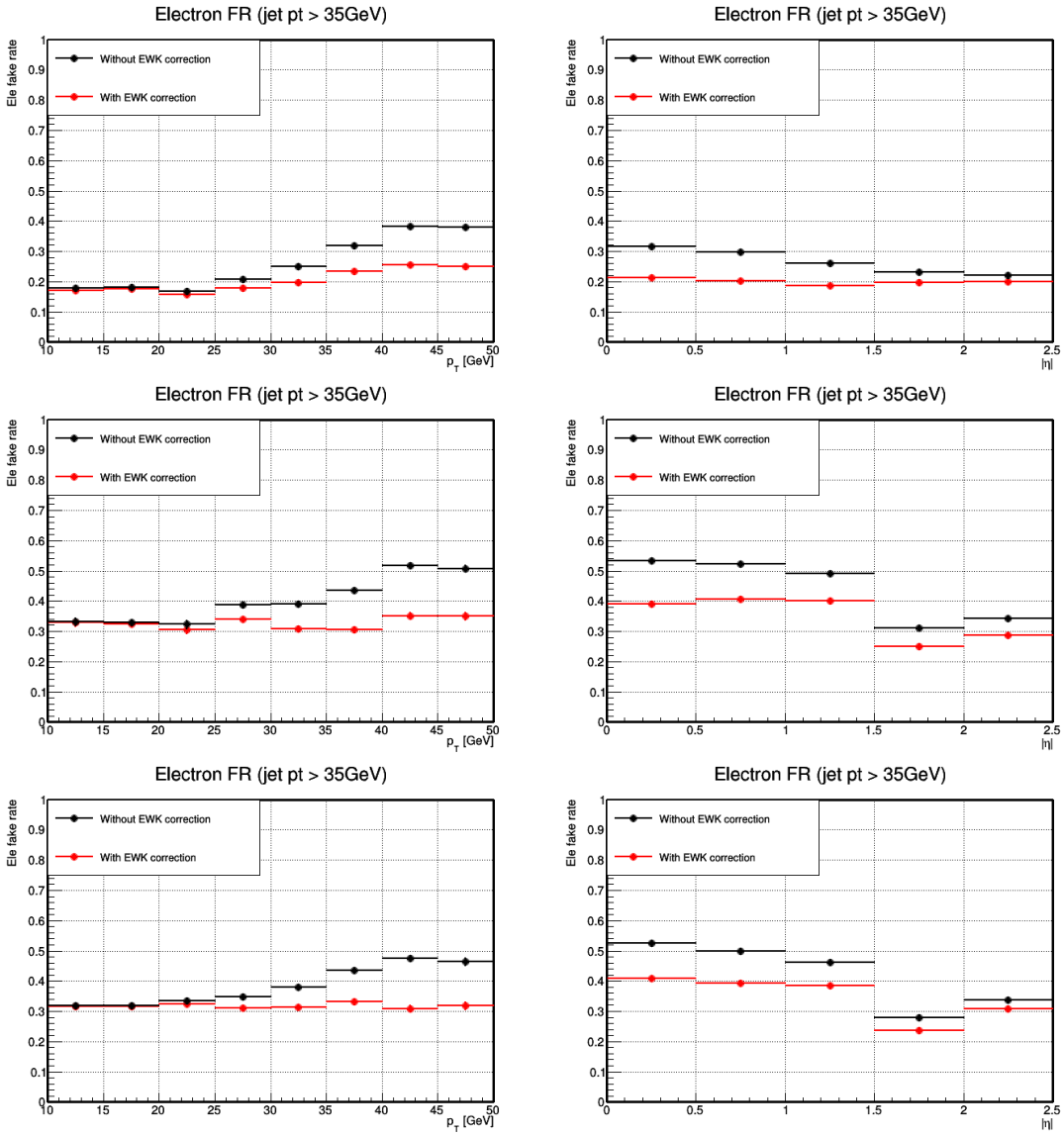


Figure 4.17: Electron fake rate as a function of the loose lepton p_{T} (left) and η (right), before and after applying the EWK correction (in black and red, respectively), from a control region sample obtained with a jet threshold of 35 GeV, and for the full Run 2: 2016 (top), 2017 (middle) and 2018 (bottom).

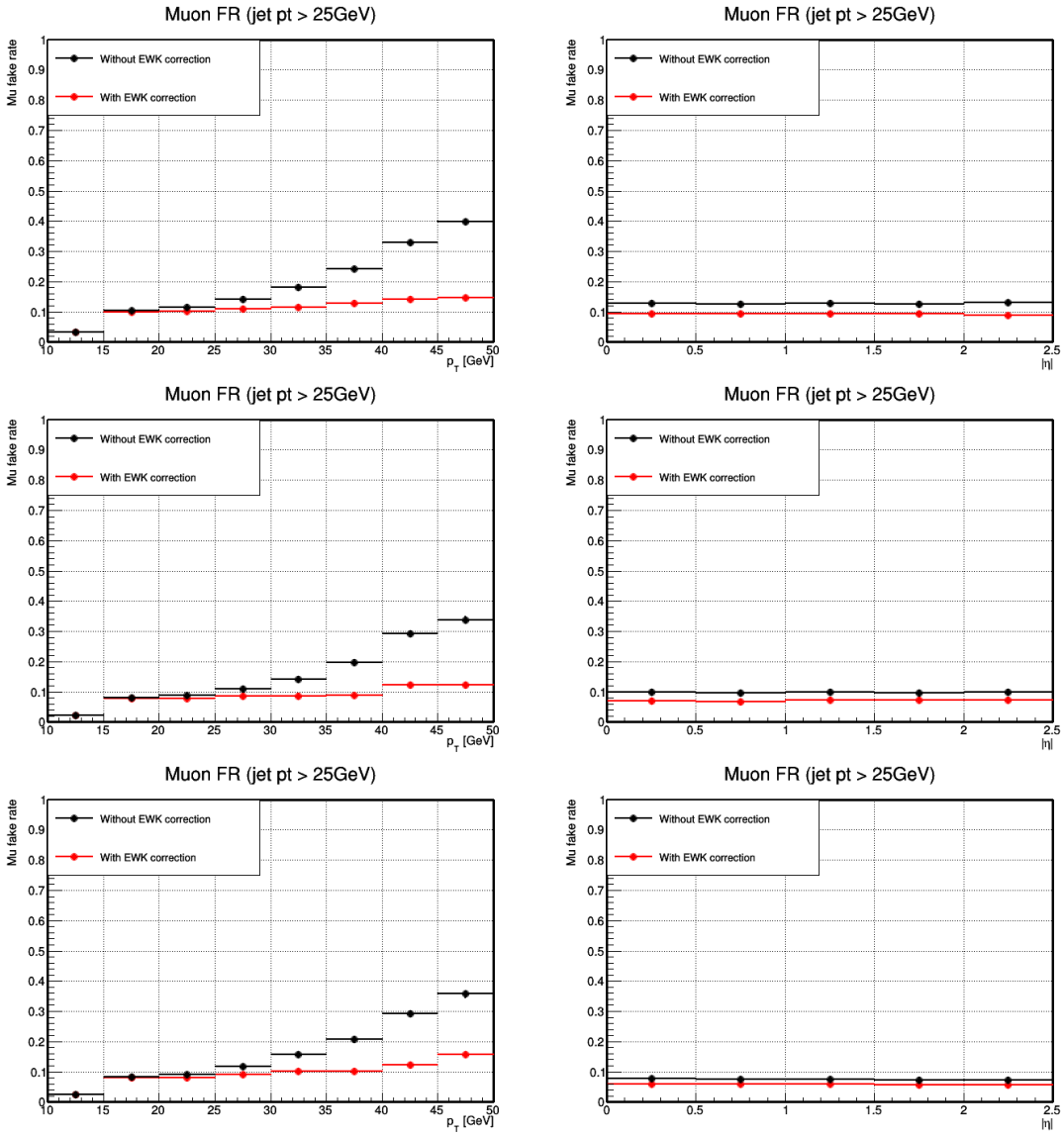


Figure 4.18: Muon fake rate as a function of the loose lepton p_{T} (left) and η (right), before and after applying the EWK correction (in black and red, respectively), from a control region sample obtained with a jet threshold of 25 GeV, and for the full Run 2: 2016 (top), 2017 (middle) and 2018 (bottom).

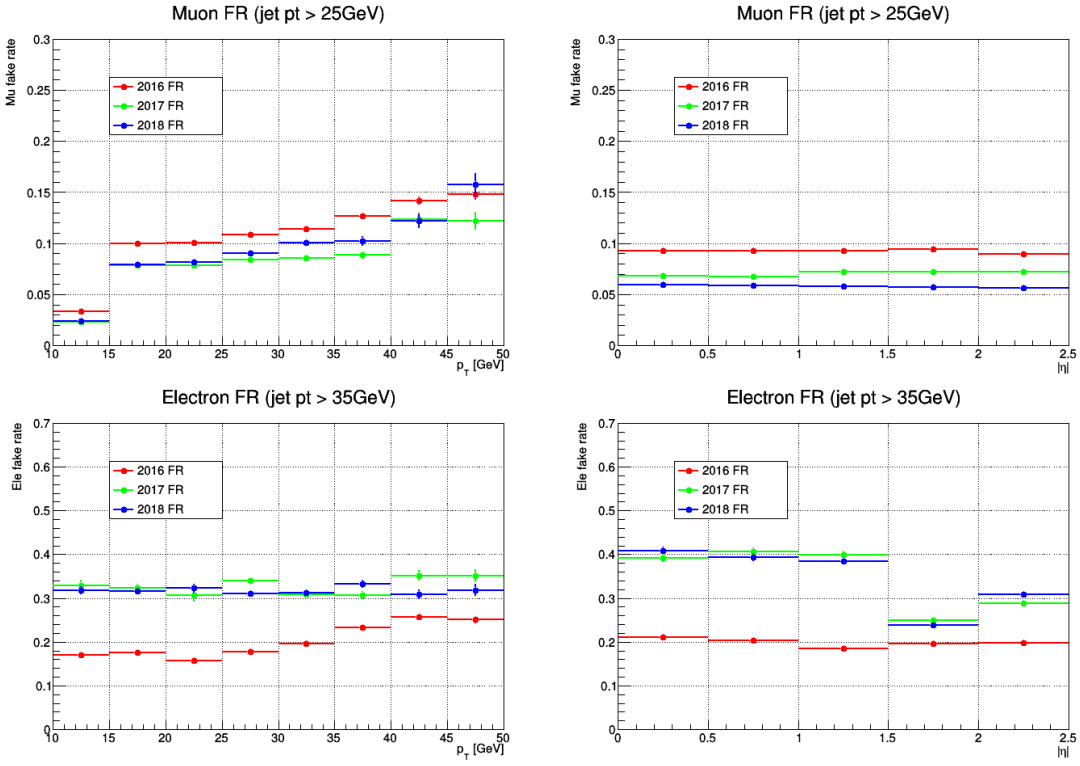


Figure 4.19: Muon fake rates (top) and electron fake rates (bottom) as a function of the loose lepton p_{T} (left) and η (right), obtained for the full Run 2.

Lepton prompt rate determination

The prompt rate, or ratio of tight-to-loose leptons for prompt leptons, is determined with the simulated samples of DY($\ell\ell$). The ratio of prompt loose leptons satisfying the tight requirements is obtained using a general tag and probe method. One of the two leptons from the Z boson decay is required to pass the tight lepton criteria and the ratio is computed from the other lepton compatible with a Z boson decay. In order to guarantee that they are compatible with a Z boson we select leptons pairs that fulfil the invariant mass constraint $|m_{\ell\ell}\text{-}m_{\mathrm{Z}}|< 15$ GeV.

The prompt rates for both muons and electrons are shown in Fig. 4.20. For electrons the prompt rate ranges from 0.8 to above 0.9, whereas for muons the difference between rates at lower and higher p_{T} is bigger. That is because the MVA discriminant used to define the tight muon selection is very restrictive at low p_{T}, in order to keep the fake rate low in that region, but it also reduces the prompt rate as a side effect. The prompt rate always increases with the transverse momentum, since it is easier for a high p_{T} prompt particle to be correctly identified as such. Over 90% of prompt muons and electrons with $p_{\mathrm{T}} > 40$ GeV pass our tight selection.

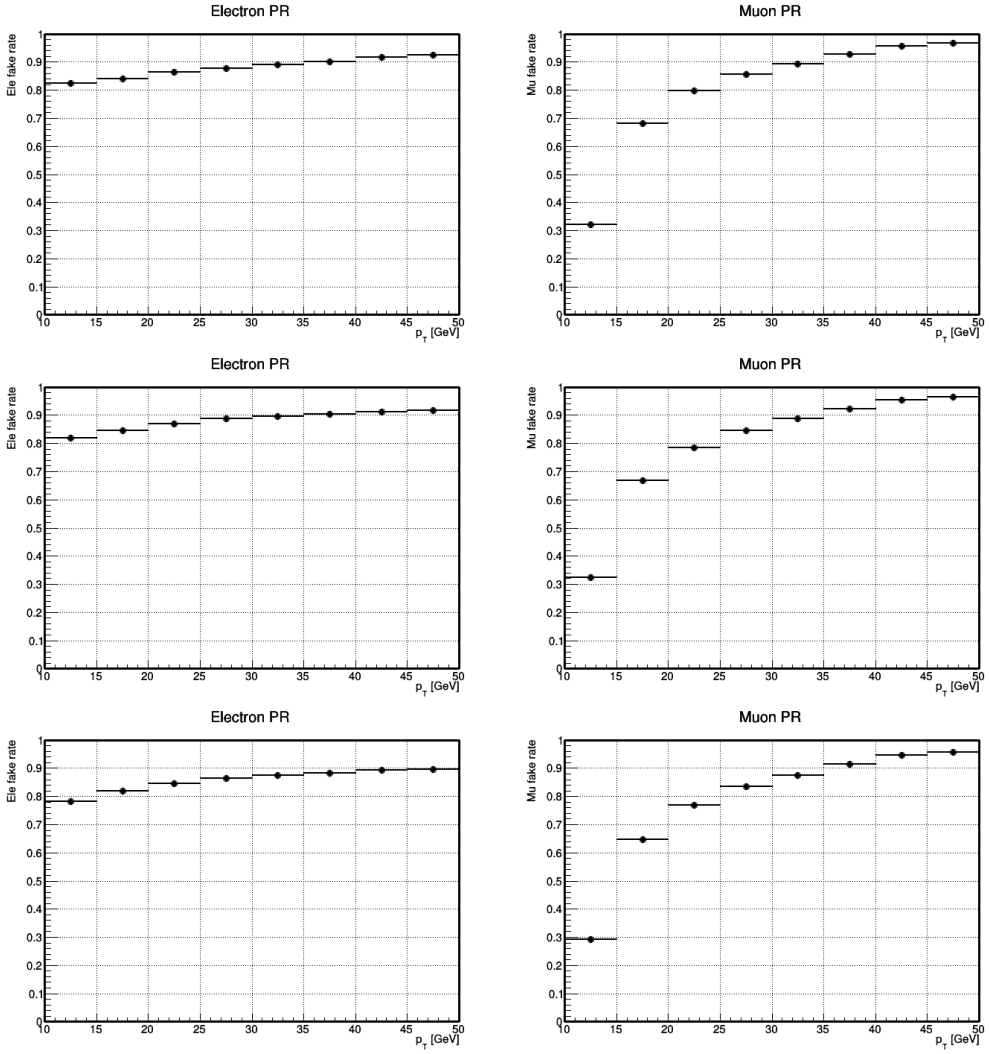


Figure 4.20: Prompt rate for electrons (left) and muons (right) in p_{T} bins, both estimated in a Z+jets simulated sample for the full Run 2: 2016 (top), 2017 (middle) and 2018 (bottom).

Validation of background modelling

The method of the nonprompt lepton determination has been tested in a control region with similar characteristics as our signal region. The signal region of the H → WW analysis has two leptons with opposite sign charges. We define a control region by keeping the same criteria as the signal selection, but requiring that the leptons have the same charge. This selects a region where the nonprompt lepton background dominates.

We apply the weights obtained with the expression in Eq. 4.11 to data events with two loose leptons and either zero, one or two tight leptons, to estimate the nonprompt lepton background contribution in this control region. The level of agreement between data and the estimated backgrounds in this same-sign control region is used as a validation of the data-driven method.

This closure test is performed in four independent control regions, analogous to the signal regions in the analysis. The events are classified depending on whether the second lepton (subleading) p_{T} is greater than 20 GeV and whether the subleading lepton is a muon or an electron. Given that the fake rates depend on the transverse momentum and the flavour of the lepton, dividing in four control regions makes for a more detailed check of the background estimation.

The results of the closure test on same-sign events are plotted in Figs. 4.21 to 4.23 for 2018 data. These plots are a combination of the $e\mu$+μe channels obtained for each number of jets. The main contribution in the control region comes from the fake lepton background, coloured in grey, same-sign electron pairs from Vγ/Vγ^* events shown in red and green, and fully leptonic top quark pair decays in yellow, more relevant in the higher jet number channels. The EWK $\mathrm{W}^{\pm}\mathrm{W}^{\pm}$ production is also a physical process contributing in the ≥2-jet category. The $m_{\ell\ell}$, $p_{\mathrm{T}}^{\ell\ell}$ and second lepton η distributions have a data/MC agreement better than 30%, as is summarised in Table 4.15. In the appendix A.1.4 2016 and 2017 validation results are also shown.

Table 4.15: Data/MC ratio in the same-sign control region

Data/MC ratio in the 2016 same-sign control region

Subcategory	Data	MC	ratio Data/MC
0 jets p_{T2} >20 GeV	920	832	1.11 ± 0.05
0 jets p_{T2} <20 GeV	612	509	1.20 ± 0.07
1 jet p_{T2} >20 GeV	990	950	1.04 ± 0.05
1 jet p_{T2} <20 GeV	319	302	1.06 ± 0.08
≥2 jets p_{T2} >20 GeV	775	728	1.06 ± 0.05
≥2 jets p_{T2} <20 GeV	186	170	1.09 ± 0.12

Data/MC ratio in the 2017 same-sign control region

Subcategory	Data	MC	ratio Data/MC
0 jets p_{T2} >20 GeV	1302	1212	1.07 ± 0.04
0 jets p_{T2} <20 GeV	877	1065	0.82 ± 0.04
1 jet p_{T2} >20 GeV	1365	1400	0.98 ± 0.04
1 jet p_{T2} <20 GeV	498	607	0.82 ± 0.05
≥2 jets p_{T2} >20 GeV	1141	976	1.17 ± 0.05
≥2 jets p_{T2} <20 GeV	303	310	0.98 ± 0.08

Data/MC ratio in the 2018 same-sign control region

Subcategory	Data	MC	ratio Data/MC
0 jets p_{T2} >20 GeV	2053	2147	0.96 ± 0.03
0 jets p_{T2} <20 GeV	1397	1591	0.88 ± 0.03
1 jet p_{T2} >20 GeV	2217	2287	0.97 ± 0.03
1 jet p_{T2} <20 GeV	734	887	0.83 ± 0.04
≥2 jets p_{T2} >20 GeV	1755	1645	1.07 ± 0.04
≥2 jets p_{T2} <20 GeV	440	392	1.12 ± 0.08

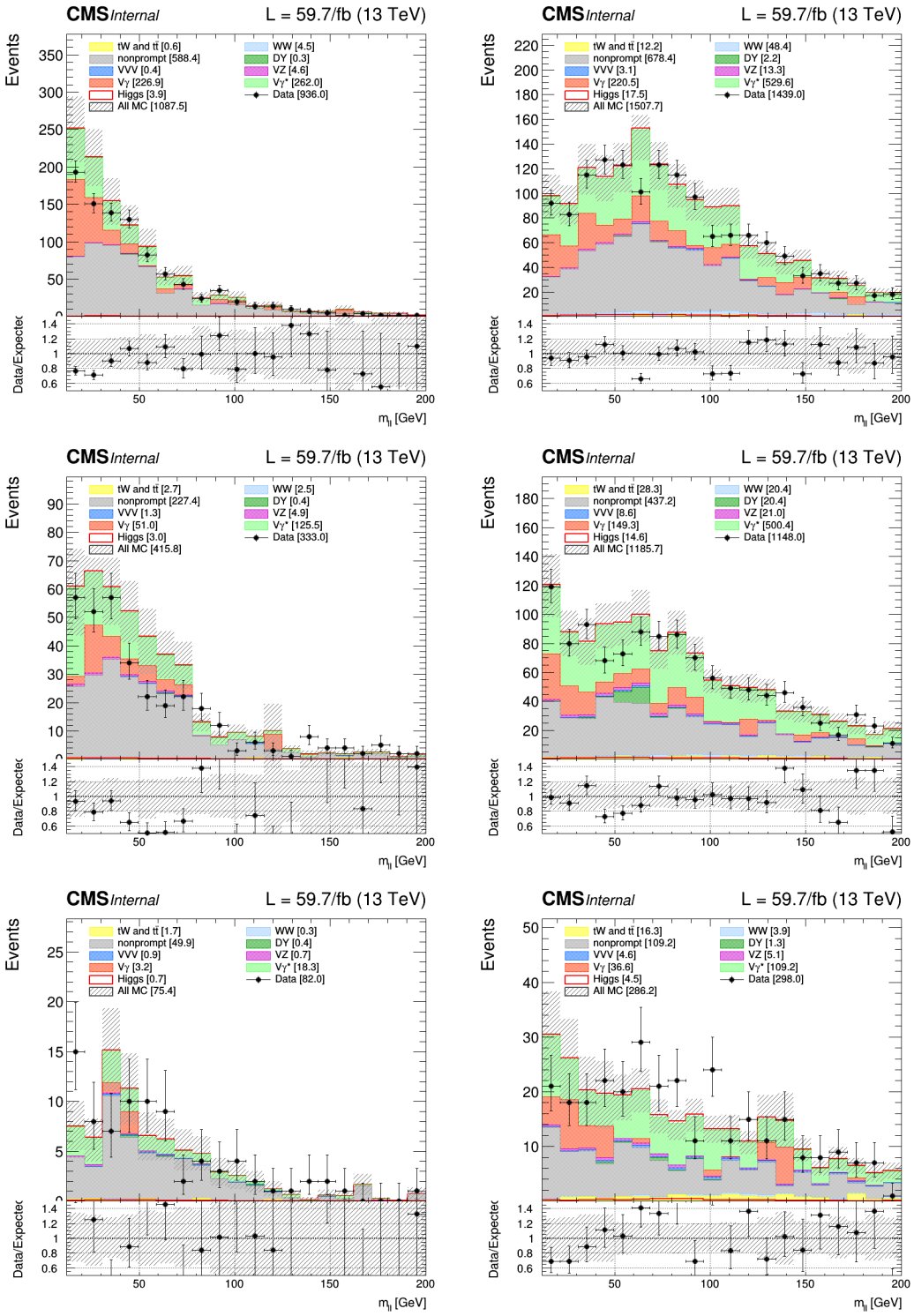


Figure 4.21: Dilepton invariant mass distributions in the same-sign control region for eμ events for the year 2018 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

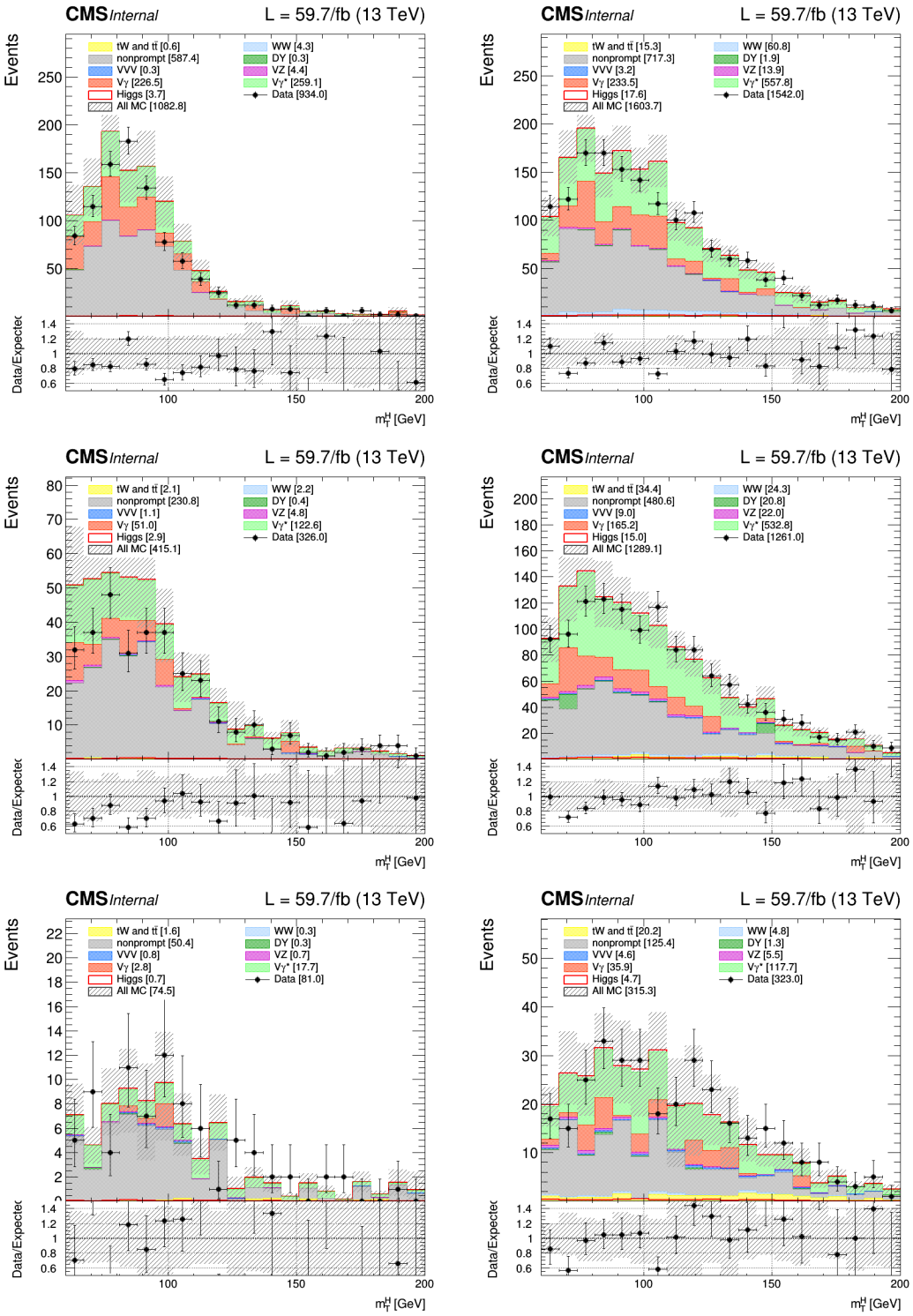


Figure 4.22: Trasverse mass distributions in the same-sign control region for eμ events for the year 2018 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

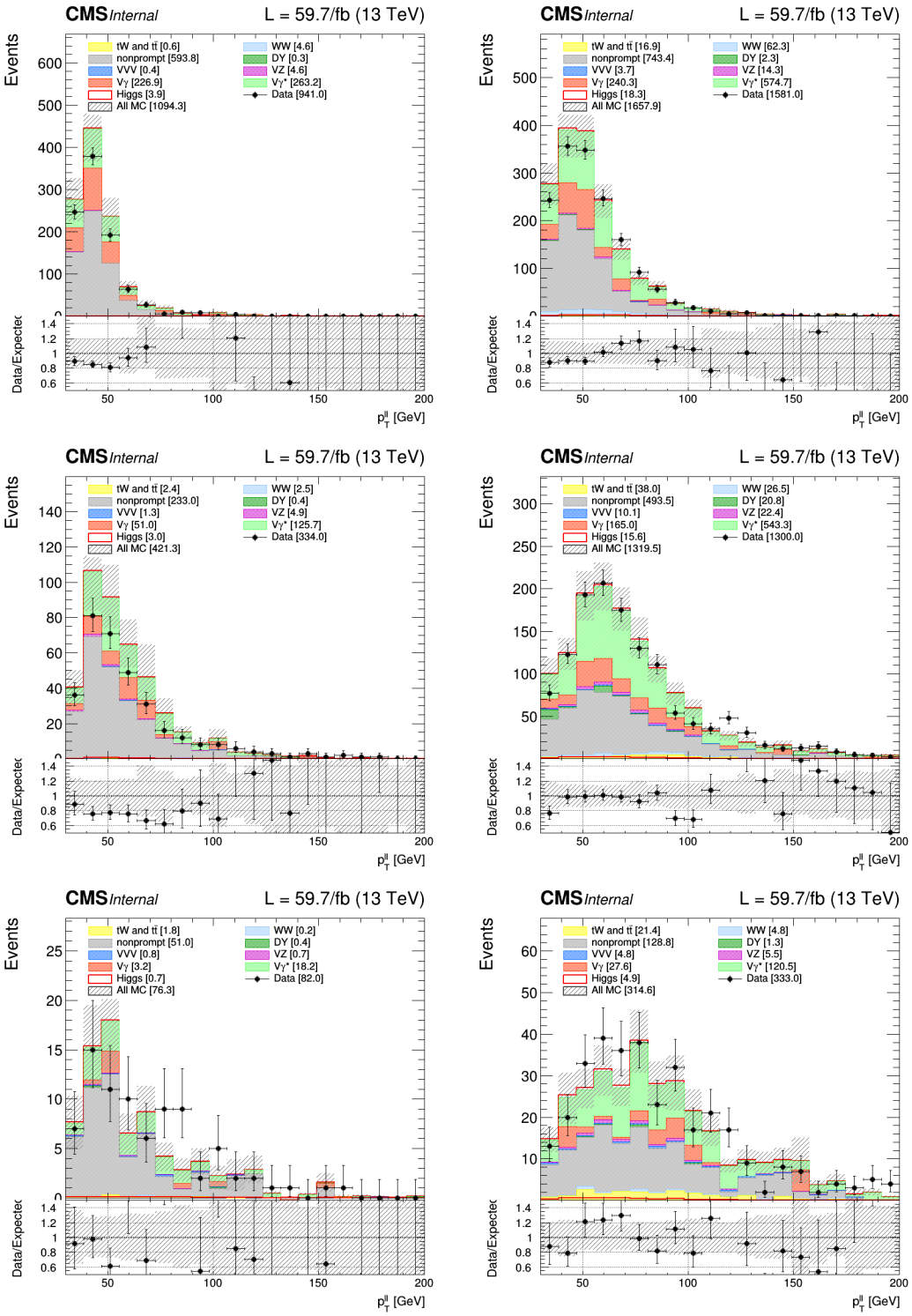


Figure 4.23: Dilepton transverse momentum distributions in the same-sign control region for eμ events for the year 2018 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

A study on the jet flavour composition

One of the basic assumptions of the method for the nonprompt lepton background determination is that the fake rates determined in a dijet QCD control region can be used in the signal region of the analysis. This would require the characteristics of the jets misidentified as leptons to be the same in QCD events and in the signal region background processes (W+jets and $t\bar{t}$), which may not be the case. To address this issue, the fake rates were computed as a function of the kinematic variables of the leptons (p_T and η).

Another important related issue is the origin of the jet being misidentified as a lepton. Jets originated from a heavy flavour quark produce nonprompt leptons from the decay of heavy flavoured hadrons, whereas misidentification of jets originated from light quarks have a pure instrumental origin. The definition of the tight and loose identification criteria is specially sensitive to these details and hence the determined fake rates are as well. In order to optimise our definition for the loose and tight electron selection in 2017 and 2018 described in section 4.2.2, a study on the flavour composition of jets that are misidentified as leptons was carried out.

The sources of fakes leptons for electrons and muons are not the same. We estimate the fake rates using a QCD-enriched control region, and if we look at the jet flavour in QCD MC events in Fig. 4.24, jets associated with a loose electron or a loose muon have different flavour distributions. The majority of the jets associated with muons come from b quarks, whereas for electrons there are more associated jets coming from light quarks and gluons.

We expect that the probability of a jet to be identified as a electron depends on the associated jet flavour, and the flavour composition of jets in a sample may change significantly depending on the chosen selection. Since for muons the b quarks are always the main contribution, the composition is therefore not significantly dependant on the choice of our selection, and thus only electrons are discussed during the rest of this study.

The W+jets contribution is estimated with the fake rates obtained with a dijet QCD control region. In Fig. 4.25 we compare the jet flavour distributions of jets associated to loose electrons in a W+jets MC sample with the QCD one. Each of the bins shows the fraction of jets corresponding to each flavour. There are differences in the fraction of light flavour quarks, b quarks and gluons. Electrons are associated to light quarks in W+jets samples, whereas in QCD dijet events they are more commonly associated to b quarks and gluons. These differences in jet flavour composition may induce an uncertainty in the nonprompt lepton background estimation method.

As an alternative, we also considered estimating the fake rate using a Z+jets control region instead of the dijet QCD one. In Fig. 4.25 the flavour distribution in a Z+jets MC sample is also shown, differing from the distributions observed in W+jets. There are more b jets associated to electrons for Z+jets and more associated c-quark jets in W+jets processes. Compared to the dijet QCD control region, electrons in Z+jets samples are more often produced in association with quark jets, whereas in QCD processes gluon jets are more common.

Similar distributions were obtained for different electron loose and tight selections, some of them showed in Fig. 4.26. It was found that a loose electron selection using a cut-based medium ID (Table 4.7) and a tight selection that included a stricter isolation cut on top of the MVA discriminant cut (Table 4.9) was the optimal combination to have a similar flavour composition, and thus they were the final electron ID criteria required.

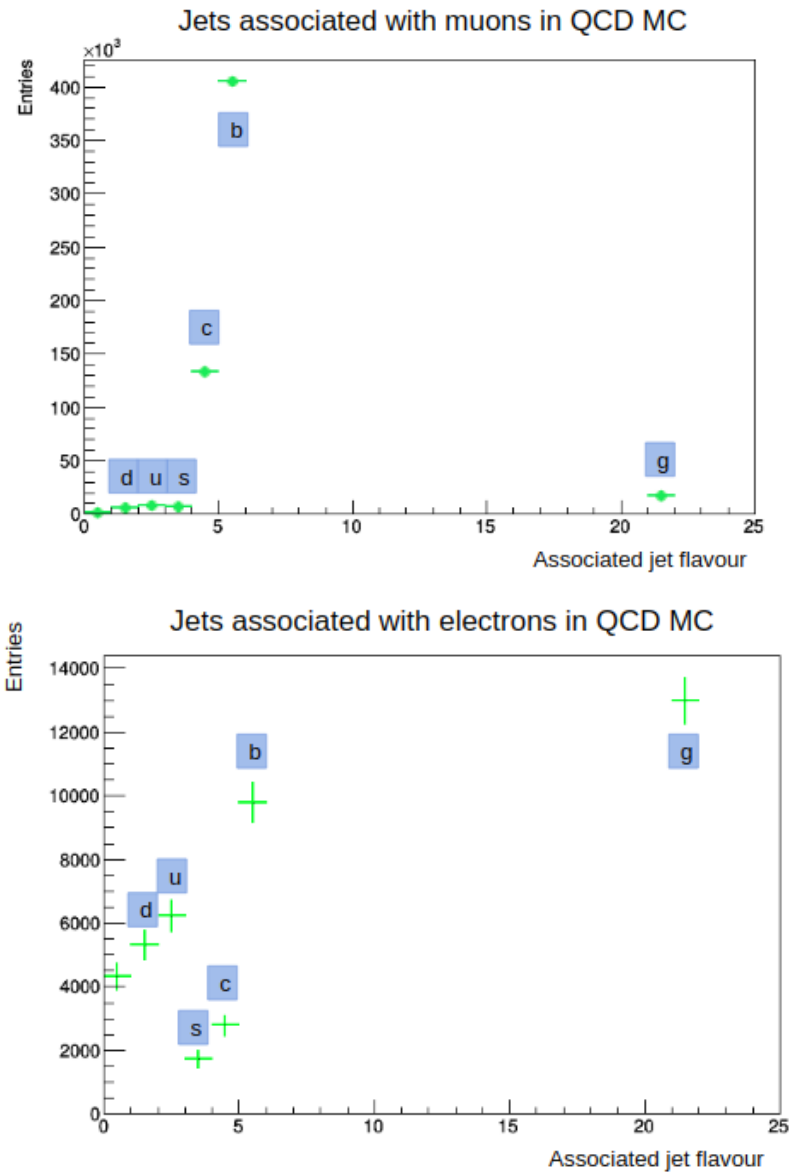


Figure 4.24: Flavour of the jets associated with loose muons (above) and loose electrons (below) in a QCD MC sample. The fake muons come mostly from b quark jets (PdgId=5) and the electrons from light quarks (PdgId<5) and gluons (pdgId=21). The PdgId=0 jets corresponds to events with unassigned flavour in the sample.

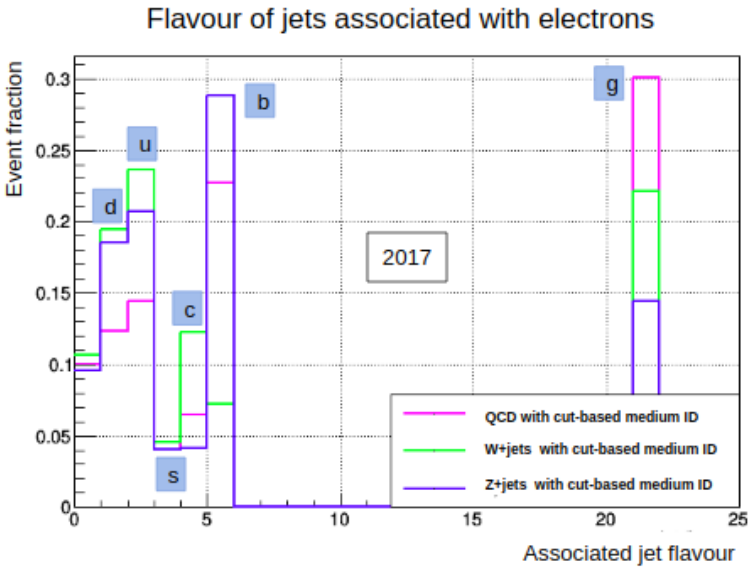


Figure 4.25: Comparison of the associated jet flavour normalised distributions for the chosen loose electron selection in all three control regions. On the X axis, the PDG Id value is shown (21 for gluons and 1–6 correspond to quarks, 5 being the bottom quark). The values at 0 are for jets with unassigned flavour.

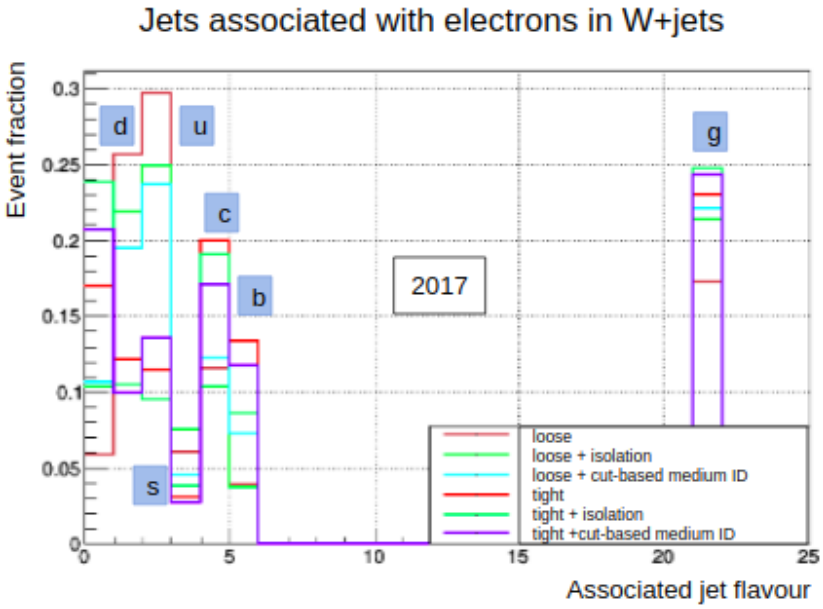


Figure 4.26: Several distributions related to possible loose and tight electron definitions. The chosen cut-based identification for loose electrons and additional isolation requirements for the tight electrons.

Systematic uncertainties of the nonprompt lepton background

Several sources of systematic uncertainties that may impact the estimation of the nonprompt lepton background have been evaluated. One of them is the p_{T} threshold of the jet recoiling against the possible fake lepton when defining the dijet control region. This uncertainty is assessed by varying up and down the recoiling jet p_{T} threshold with respect to the nominal values listed in Table 4.13. We estimate the systematic uncertainty by calculating the fake lepton contribution in the signal region for several jet p_{T} thresholds and obtaining the relative difference in events after increasing or lowering the threshold. This is done separately for each year and jet category, and the results are summarised in Tables 4.16 to 4.18.

The 'nominal' event yields correspond to the nominal jet p_{T} thresholds and the up and down jet p_{T} variations are obtained by adding or subtracting 10 GeV to the nominal threshold value. The event yields are computed by changing the jet p_{T} thresholds independently for electrons and muons. We also consider the statistical uncertainty of the fake and prompt rates determination as a systematic uncertainty of the fake lepton contribution. We vary the event yields within their statistical uncertainties, which is also shown on the tables.

Table 4.16: The nonprompt lepton background estimation for 2016. Up and down jet p_{T} threshold variations are shown, as well as the variations due to the statistical uncertainty of the lepton fake rate. The shown uncertainties are statistical only.

2016	0-jet category	1-jet category	≥2-jet category
Nominal	702±18	311±13	78±8
Electron jet p_{T} up	647±16	286±12	72±7
Electron jet p_{T} down	722±19	319±14	80±8
Electron statistical up	753±20	329±15	83±8
Electron statistical down	654±17	295±12	74±7
Muon jet p_{T} up	663±18	288±13	72±7
Muon jet p_{T} down	721±19	324±14	85±8
Muon statistical up	712±19	320±14	81±8
Muon statistical down	693±18	303±13	75±7

Table 4.17: The nonprompt lepton background estimation for 2017. Up and down jet p_{T} threshold variations are shown, as well as the variations due to the statistical uncertainty of the lepton fake rate. The shown uncertainties are statistical only.

2017	0-jet category	1-jet category	≥2-jet category
Nominal	1974±42	986±30	276±17
Electron jet p_{T} up	2066±44	1016±32	283±18
Electron jet p_{T} down	1939±41	972±30	273±17
Electron statistical up	2306±50	1144±37	320±21
Electron statistical down	1690±35	852±25	238±14
Muon jet p_{T} up	1909±41	945±30	263±17
Muon jet p_{T} down	2002±42	1023±30	292±17
Muon statistical up	1997±42	1007±30	284±17
Muon statistical down	1951±41	966±30	267±17

The systematic uncertainties, presented in Table 4.19, are obtained from the relative differences between the nominal event yields and the variations. For muons the jet threshold variation uncertainty is greater than the fake rate statistical uncertainty, and the statistical uncertainty is similar for the three years, as expected. For electrons, since the loose definitions are not the same in 2016 and the rest of the years, there are some differences. The 2016 period has less statistical uncertainty than the other years, but for 2017 and 2018 the jet p_{T} uncertainties are much smaller.

Table 4.18: The nonprompt lepton background estimation for 2018. Up and down jet p_{T} threshold variations are shown, as well as the variations due to the statistical uncertainty of the lepton fake rate. The shown uncertainties are statistical only.

2018	0-jet category	1-jet category	≥2-jet category
Nominal	3024±57	1403±41	338±23
Electron jet p_{T} up	3107±60	1457±44	350±24
Electron jet p_{T} down	2967±55	1370±40	330±22
Electron statistical up	3491±70	1629±51	396±28
Electron stat down	2630±47	1216±34	291±18
Muon jet p_{T} up	2927±56	1351±41	323±22
Muon jet p_{T} down	3063±57	1445±42	353±23
Muon statistical up	3056±57	1433±42	350±23
Muon statistical down	2993±57	1373±41	326±22

Table 4.19: Systematic uncertainties associated to the nonprompt lepton data-driven estimation in the eμ channel for the 0-, 1-and ≥2-jet categories.

2016	0-jet category	1-jet category	≥2-jet category
Electron jet p_{T}	+7.8/−2.8%	+8.0/−2.6%	+7.6/−2.0%
Electron statistical	+7.2/−6.9%	+5.6/−5.4%	+5.9/−5.6%
Muon jet p_{T}	+5.6/−2.6%	+7.6/−4.1%	+8.2/−9.5%
Muon statistical	+1.3/−1.3%	+2.9/−2.8%	+3.8/−3.7%
2017	0-jet category	1-jet category	≥2-jet category
Electron jet p_{T}	+4.7/−1.7%	+3.0/−1.5%	+2.7/−1.1%
Electron statistical	+16.8/−14.4%	+15.9/−13.6%	+16.1/−13.7%
Muon jet p_{T}	+3.3/−1.4%	+4.2/−3.7%	+4.6/−5.9%
Muon statistical	+1.2/−1.2%	+2.1/−2.1%	+3.1/−3.0%
2018	0-jet category	1-jet category	≥2-jet category
Electron jet p_{T}	+2.7/−1.9%	+3.8/−2.3%	+3.5/−2.4%
Electron statistical	+15.4/−13.0%	+16.2/−13.3%	+17.2/−13.8%
Muon jet p_{T}	+3.2/−1.3%	+3.7/−3.0%	+4.2/−4.6%
Muon statistical	+1.1/−1.0%	+2.2/−2.1%	+3.5/−3.4%

The total uncertainty considering those systematic uncertainties does not cover the discrepancies shown in the same sign control region in the eμ channel. This is a consequence of the remaining differences in the sources of fakes in the dijet control region and in the W+jets background that was previously discussed. The flavour dependence is accounted for by adding an additional 30% systematic uncertainty for the fake lepton background that is considered for all years.

4.4 Analysis uncertainties

The systematic uncertainties may affect the shape of the discriminating variable distributions of some processes and may also change their normalisation. According to the source, we divide them into background-related uncertainties, theoretical model uncertainties and experimental uncertainties. The theoretical uncertainties are in general correlated among the three years, whereas most of the experimental uncertainties are uncorrelated. A summary of the uncertainties is presented in Tables 4.20 to 4.23.

Background-specific uncertainties

The estimation of the normalisation and shape of the expected backgrounds is affected by several sources of uncertainty:

- In the DY($\tau\tau$) background the signal-to-control region event ratio has a 2% uncertainty, due to theoretical uncertainties arising from PDF uncertainties and scale variations. Aside from that, $\tau\tau$-embedded samples have an uncertainty on the event production method. Events in the $\tau\tau$-embedded samples are obtained by replacing muon pairs from a Z boson decay with tau leptons, but to prevent double counting, muons in the final state that come from τ decays are removed. There is a 10% uncertainty in the shape of the removed events in the other MC samples. All DY uncertainties are correlated among the three years.

- The signal-to-control region event ratio in the top background has an associated uncertainty of 1%, obtained computing the events that migrated between regions with the scale variations. Since the $\mathrm{t\bar{t}}$ and single top processes are treated as one single background and fitted together, we have to consider an uncertainty in the composition of the top background. That uncertainty is the theoretical uncertainty in the ratio of the cross sections of $\mathrm{t\bar{t}}$ and single top processes, which is of the order of 8%. The p_{T} of $\mathrm{t\bar{t}}$ events is reweighted to NNLO QCD plus NLO EWK precision, which also has an uncertainty with both a normalisation and shape component, computed by the difference between applying the reweighting or not. All the uncertainties of the top background are correlated among the three years.

- There are uncertainties associated to the theoretical description of the WW production modes. The WW production cross section has an uncertainty associated to the p_{T} reweighting in $\mathrm{qq} \rightarrow \mathrm{WW}$ samples to NNLO plus NNLL. It is assessed by variating the NNLL weights, shifting the renormalisation and factorisation scales and the scale below which QCD radiations are resummed. For gluon fusion produced WW there is a theoretical 15% uncertainty in the scale factor applied to correct for the differences between LO and NLO [103]. All these uncertainties are correlated among the three years.

- The nonprompt lepton background uncertainties have been described in section 4.3.5. There is a systematic uncertainty derived from the statistical uncertainty of the fake rate determination method and a systematic uncertainty obtained from the variations of the jet p_{T} threshold. They are uncorrelated among the three years and have values in the range of 1-10%, shown in previous tables. To account for the differences related to the flavour of the nonprompt lepton sources a conservative 30% uncertainty in the background normalisation is included, which is correlated among the three years.

- The Wγ^* and WZ backgrounds are predicted from MC, but a data-to-MC scale factor is computed for them using control regions with three leptons in the final state where they dominate. Wγ^* has a 25% uncertainty in that scale factor of 0.94, and WZ has a 16% uncertainty in its 1.14 scale factor. All of these uncertainties are correlated among the three years.

Uncertainty	Process	Type	Correlated
CR/SR acceptance	top and DY($\tau\tau$)	normalisation	yes
$\tau\tau$ veto in background samples	non-DY samples	normalisation	yes
Single top / $\mathrm{t\bar{t}}$ composition	top	shape	yes
Top quark p_{T} reweighting	top	shape	yes
qqWW NNLL resummation	qqWW	shape	yes
ggWW LO$\rightarrow$NLO factor	ggWW	normalisation	yes
Fake rate (statistical and jet p_{T})	nonprompt	shape	no
Fake rate (jet composition)	nonprompt	normalisation	yes
Wγ^* and WZ data/MC	Wγ^* and WZ	normalisation	yes

Table 4.20: Background-related uncertainties and their correlations among years.

Theoretical model uncertainties

Most of the theoretical uncertainties appear because of the limited knowledge of the processes, and are a consequence of the choices of PDF and the strong coupling constant α_s, the effect of missing higher-order corrections and the modelling of the UE and parton shower. These apply to both the signal samples and the simulated backgrounds, and are in most cases correlated among the three years.

- One source of uncertainty for the cross sections of the processes is the one associated with PDF sets. It affects the overall cross sections as well as the shape of the distributions. Since its effect on the shape of the distributions is small, it is considered as an effect on the acceptance of the event selection instead, i.e. an effect on the normalisation that accounts for the events that may not pass the signal selection due to variations in the PDF. This acceptance uncertainty is shown in Table 4.21 for the relevant processes, which are the Higgs boson signal processes and the ggWW and VZ backgrounds. The uncertainty on the PDF is estimated by using the several replicas of the PDF set with variations. For the Higgs boson signal processes simulated with NNPDF3, that uncertainty was computed by the LHC Higgs cross-section working group [97]. This uncertainty is correlated among the three years.

Sample	PDF variation	Scale variation
ggH	0.6%	—
qqH	0.2%	0.3%
WH	0.3%	1.0%
ZH	0.2%	1.5%
ggWW	0.6%	1.2%
WZ/Wγ^*	0.1%	—

Table 4.21: Uncertainties on acceptance for the PDF and scale variations.

- Missing higher-order terms in the perturbation series of the cross section calculations also affect the shapes and the overall cross sections. For the background processes, MC simulated events are generated with different renormalisation and factorisation scales, which have been varied up or down by a factor two. The uncertainty is taken from the envelope of the varied distributions, where the maximum deviation is used as the uncertainty. WZ and Wγ^* scale uncertainties are shape-only, since their normalisation was estimated in their own control regions.

  QCD scale uncertainties on Higgs boson production processes excluding ggH are split into ones on the overall cross section and on the shape. The theoretical uncertainties that affect the overall cross section in the Higgs boson signal samples have been computed by the LHC Higgs cross-section working group, and the shape uncertainties in the signal samples are small and are considered as a normalisation-type uncertainty in the event acceptance. The scale variations acceptance uncertainty is given in Table 4.21.

  The gluon fusion theoretical QCD scale uncertainty consists of nine sources that reflect on not just the overall scale, but also migrations between jet-multiplicity bins, different Higgs p_{T} bin migrations and an uncertainty associated to the missing higher order finite top quark mass corrections. These uncertainties are correlated among the three years.

- The uncertainty associated to the UE model is estimated by shifting within their uncertainties the respective Pythia8 tunes, CUET8PM1 for 2016 and CP5 for 2017 and 2018. This uncertainty is 1.5% and it is uncorrelated between 2016 and the other two years, but correlated between 2017 and 2018.

- The dependence on the parton shower model is estimated by the ISR and FSR scales, varying them up and down by a factor two. The systematic uncertainties related to ISR and FSR scale variations are uncorrelated between them, but correlated among the three years.

Uncertainty	Process	Type	Correlated
PDF	all processes	normalisation	yes
QCD scale	backgrounds and ggH other signals	shape normalisation	yes
Underlying event	WW, ggH and VBF	normalisation	yes (2017/2018)
Parton shower	WW, ggH and VBF	normalisation	yes

Table 4.22: Theoretical model uncertainties and their correlations among years.

Experimental uncertainties

Systematic uncertainties due to experimentally measured quantities are propagated to all signal and background simulations. These include uncertainties in the object reconstruction efficiencies, trigger efficiencies or in the momentum or energy measurement, among others.

- The uncertainty in the measured luminosity is determined from the luminosity calibration obtained from the van der Meer scans [33]. It is 2.5%, 2.3%, and 2.5% for the datasets of 2016 [78], 2017 [79], and 2018 [80], respectively. This uncertainty is partially correlated among the three years.

- Trigger efficiencies of the used triggers (see Table 4.1) were computed using a tag-and-probe method selecting Z boson decay events. The associated uncertainty is obtained varying the 'tag' selection and the Z boson mass window and is uncorrelated among the three years, estimated as less than 1% for each.

- Lepton reconstruction, isolation and identification efficiencies are also obtained from the tag-and-probe method. The uncertainty in the scale factors applied to correct for the difference in identification efficiency between data and MC is of the order of 1% for electrons and 2% for muons. These uncertainties are uncorrelated among the three years.

- Electron and muon momentum scale corrections are obtained from comparing the position and width of the Z boson mass peak in DY MC and data. We take as uncertainties in the lepton momentum scale 0.6-1% for electrons and 0.2% for muons, which correspond to differences between the MC and data in the position of the Z boson mass peak. Propagating these uncertainties to the signal region the effect is of the order of 1% for muons and 1-3% for electrons, uncorrelated among the three years.

- The jet energy corrections applied have an associated jet energy scale (JES) uncertainty. How it affects our measurement is estimated by shifting the JES within its uncertainty. There are different sources of the uncertainty related to the offset, absolute and relative corrections explained in section 3.2.3. This affects E_{T}^{miss} as well, because the changes in jet energy directly propagate to it. Some of the sources of uncertainty in the JES are correlated among years.

- Apart from the uncertainties in the E_{T}^{miss} due to the propagation of jet and lepton uncertainties, there are uncertainties in the p_{T}^{miss} unclustered component, which are a result of the pileup energy uncertainty and the energy scale of the unclustered particles. They are assessed by varying the momentum of each PF candidate that is not associated to a lepton or a jet, causing an effect on the rates in signal region of the order of 1% to 10%, depending on the process. This uncertainty is uncorrelated among the three years.

- The ratio between the normalised distributions of the number of inelastic collisions for each year and the MC pileup profile is used as an event weight to match the simulated PU profile to the actual one. This pileup reweighting uncertainty has a small effect on the normalisation of all simulated samples, thus it is only considered in the main backgrounds and signals. Since each year has a different pileup profile, this uncertainty is uncorrelated among the three years.
- The effect on our measurement caused by uncertainties in our knowledge of the b-tagging efficiency is estimated by changing the b-tagging scale factors according to their uncertainties. Variations of the scale factors related to unknown jet flavor compositions of the samples are correlated among years, whereas variations accounting for the statistical uncertainties of the measurement samples are uncorrelated.
- Inefficiencies due to the L1 trigger prefiring in 2016 and 2017 were corrected in MC simulations using additional trigger efficiency scale factors. The corresponding uncertainties are uncorrelated between the two years.

Uncertainty	Process	Type	Correlated
Integrated luminosity	all processes except WW, top, $\tau\tau$	normalisation	partially
Trigger efficiency	all processes	shape	no
Lepton ID efficiency	all processes	shape	no
Lepton p_{T} scale	all processes	normalisation	no
Jet energy scale	all processes	normalisation	no
Unclustered MET	all processes	normalisation	no
Pileup reweighting	WW, top, $\tau\tau$, ggH and VBF signal	normalisation	no
b-tagging scale factors	all processes	shape	partially
L1 trigger prefiring	all processes, only for 2016 and 2017	shape	no

Table 4.23: Experimental uncertainties and their correlations among years.

Statistical uncertainties

Additionally to the systematic uncertainties, the finite MC samples used to estimate signal and background contributions have their statistical uncertainties. To estimate them we use the method proposed by Barlow and Beeston [112], which consists of introducing a group of parameters multiplying the expected number of events from each source (each signal and background sample) in each bin and using those parameters in the fit. If we consider independent sources of uncertainty, then we can simplify and instead use a just a single parameter for each bin. That group of multiplicative values are constrained to 1 within total statistical uncertainty in the bin, computed by adding the uncertainties of the different contributions in quadrature and assuming the number of events in each bin follows a Poisson distribution .

Chapter 5

Analysis results for ggH in the HWW decay channel

5.1 Statistical Procedure

In the following section the statistical methodology used to extract the Higgs signal in the $H \to WW$ analysis is presented. This methodology was designed by the LHC Higgs Combination Group, which provides the guidelines for the ATLAS and CMS Higgs analyses [113].

Comparing the predictions for both signal and background yields with the observed data a *signal strength modifier* μ is obtained. It is defined as the ratio between observed and expected SM Higgs boson cross sections. This signal strength affects all production mechanisms by exactly the same scale, leaving the decay branching ratios unchanged. In this case the signal strength in this case is our *parameter of interest* during the fit. In the case of the STXS analysis we use several parameters of interest, corresponding to the signal strengths of each of the STXS bins.

The expected yields depend on the *nuisance parameters*, which represent the different uncertainties. Some uncertainties are correlated between them and therefore the nuisance parameters are as well. The likelihood function to observe data given a set of nuisances and a given signal strength can be written as a product of Poisson probabilities for each bin:

$$\mathcal{L}(data|\mu,\theta) = \prod_{i}^{bins} \frac{(\mu s_i(\theta) + b_i(\theta))_i^n}{n_i!} e^{-\mu s_i(\theta) - b_i(\theta)}\, p(\theta,\tilde{\theta}), \tag{5.1}$$

where b_i, s_i are the expected background and signal yields in the i-th bin and n_i are the observed yields in that bin. $\tilde{\theta}$ is the default value of the nuisances and $p(\theta,\tilde{\theta})$ represents the prior distribution functions for the nuisance parameters. We use a log-normal prior distribution, which is a good choice for positive-defined observables such as efficiencies, scale factors or luminosities.

To compare the compatibility of the data with a background-plus-signal hypothesis we can construct the *test statistic* for the signal-plus-background q_μ, based on the logarithm of the profile likelihood ratio:

$$q_\mu = -2\ln\frac{\mathcal{L}(data|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}\text{, with } \hat{\mu} \geq 0 \tag{5.2}$$

where $\hat{\mu}$ and $\hat{\theta}$ are the parameters that correspond to the global maximum likelihood and $\hat{\theta}_\mu$ is the set of nuisances that corresponds with the maximum of $\mathcal{L}$ given a certain signal strength μ.

To obtain the nuisance parameters that best describe the experimental data $\hat{\theta}^{obs}$ and the observed signal strength $\hat{\mu}^{obs}$ a binned maximum likelihood fit is performed. For one-dimensional measurements such as the signal strength the 68% and 95% confidence intervals are defined by the union of intervals for which q_μ <0.99 and q_μ <3.84, respectively. To compute the SM expected results, the observed data is replaced by an *Asimov dataset* [114], generated with all parameter values set to the SM expectations. An initial 'a priori' fit is performed with the Asimov dataset, where the nuisances are fixed at their central values in order to obtain the expected uncertainty in the signal strength modifier.

The discriminant distributions used in our analysis, $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$, are modeled as a Poisson random variable for each bin, with the mean value obtained by adding all the contributions of the signals and backgrounds in that bin. The maximum likelihood fit, from which the signal strength is extracted, is performed using two-dimensional distributions based on the $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ variables.

The systematic uncertainties that represent normalisation effects are considered as nuisance parameters that follow log-normal distributions like this:

$$p(\theta, \tilde{\theta}) = \frac{1}{\sqrt{2\pi}\ln(k)} \exp\left(-\frac{(\ln(\theta/\tilde{\theta}))^2}{2(\ln(k))^2}\right)\frac{1}{\theta} \tag{5.3}$$

with $\tilde{\theta}$ the default value and k a factor close to 1 that characterises the width of the distribution. This factor is the standard deviation of the distribution and corresponds to the given systematic uncertainty. The nuisance parameters overall affect the normalisation and the shape of the observable distributions in both signal and backgrounds.

To assess the importance of each nuisance parameter, their associated *impacts* are evaluated. How much the signal strength changes when a single nuisance parameter shifts by one standard deviation ($\theta_i \to \theta_i + \Delta\theta_i$) is reflected in the impact plots. They also show the direction in which the signal strength changes when increasing/decreasing the value of the given parameter.

The normalisation of the DY, top quark and WW backgrounds is left free to float in the fit. In order to better fit the DY and top quark backgrounds we make use of the number of events measured in the control regions that were described in section 4.3. That number of events is included in the maximum likelihood fit as well, and the normalisation of the corresponding backgrounds in those control regions is a common factor with those backgrounds in the signal region. Using the control region/signal region ratio of top quark or DY events and its uncertainty, the normalisation in the background control regions propagates to the signal region and constrains those backgrounds.

5.2 Inclusive ggH analysis results

5.2.1 Event selection

The final event selection in the signal region is summarised in Table 5.1, which includes the general event selection described in section 4.2.6. There must be two leptons with different flavour; events with a third loose lepton with $p_{\mathrm{T}} > 10$ GeV are discarded. The invariant mass of the lepton pair must be above 12 GeV, to reject low-mass resonances such as J/ψ. A minimum E_{T}^{miss} of 20 GeV is required, as well as $p_{\mathrm{T}}^{\ell\ell} > 30$ GeV.

Categories are defined depending on the number of jets with $p_{\mathrm{T}} > 30$ GeV. Jets are selected in $|\eta|< 5.0$ region for 2016 and 2018 and in $|\eta|< 2.5$ or $|\eta|> 3$ for 2017, to avoid the 'prefiring' described in section 3.2.3. All jets must be separated from the muons (electrons) by $\Delta\mathrm{R} \geq 0.4(0.3)$, and all jets above 20 GeV have to pass the b-jet rejection criteria. Finally, we require that $m_{\mathrm{T}}^{\ell 2}> 30$ GeV, as defined in Eq. 4.7, and $m_{\mathrm{T}}^{\mathrm{H}} > 60$ GeV (defined in Eq. 4.2).

Table 5.1: Selection requirements for the events in the gluon fusion different flavour category

Event selection
- A tight electron and a tight muon with opposite sign
- p_{T1} >25 GeV, p_{T2} >10(13) GeV for μ(e)
- Lepton $\|\eta\|$ <2.4(2.5) for μ(e)
- Lepton PF relative isolation <0.15(0.06) in a cone of $\Delta R = 0.4(0.3)$ for μ(e)
- No loose third lepton with p_{T3} >10 GeV
- $m_{\ell\ell}$ >12 GeV (using two leading leptons)
- $p_{\mathrm{T}}^{\ell\ell}$ >30 GeV (using two leading leptons)
- No b-jet with p_{T} >20 GeV and $\|\eta\|<$ 2.4
- PUPPI E_{T}^{miss} >20 GeV
- $m_{\mathrm{T}}^{\ell 2}$ >30 GeV
- $m_{\mathrm{T}}^{\mathrm{H}}$ >60 GeV

5.2.2 Event categorisation

The gluon fusion channel drives the sensitivity of the H $\rightarrow$ WW analysis. It is the Higgs production process with the highest cross-section (see section 1.2), and the absence of jets in the final state at tree level results in a clearer signal with less multijet background. We take advantage of this by splitting the signal region in 0-,1-, and 2 or more jet categories.

The 0- and 1-jets categories are divided in several subcategories, as it is shown in Table 5.2, each of them with its own background estimation. The samples are divided in four subcategories depending on the lepton charge and flavour of the leading and subleading leptons: $e^+\mu^-$, $e^-\mu^+$, μ^+e^- and μ^-e^+.

Dividing into subcategories depending on the charge of the leptons is done because final states with different charges have different signal-to-background ratios. That is because the H $\rightarrow$ WW signal has a neutral final state and is therefore charge symmetric, but some of the background processes are not. Particularly, this applies to the W boson plus jets production, which is the most common source of the nonprompt lepton background in the 0-jet category. W^+ bosons are produced more often than W^- at the LHC [115], which means that events where the leading lepton has a positive charge are less likely to be have nonprompt leptons.

Finally, an additional splitting in 0- and 1- jet categories is done to separate the cases where the subleading lepton p_{T} is higher or lower than 20 GeV. The lower p_{T2} region has more nonprompt lepton background but less WW and $\mathrm{t\bar{t}}$ background, and therefore a larger signal-to-background ratio. This splitting into a total of eight subcategories improved the analysis sensitivity by separating subcategories with better signal-to-background ratio than others. For the $\geq$2-jet category the number of events is not enough to split into subcategories and only one category is considered. The nonprompt background from W+jets does not play a relevant role in this category.

5.2.3 Signal region distributions

The Higgs boson signal is extracted using the $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ variables as discriminant observables. For each of the categories described in section 5.2.2 we obtain the distributions of those variables, which are shown in Figs. 5.1 to 5.4 for 2018. As expected, in general the dominant background process in the signal region is the WW production and the second highest background contribution comes from the top quark processes, which in the 1-jet categories have the largest number of events.

Table 5.2: Analysis subcategories.

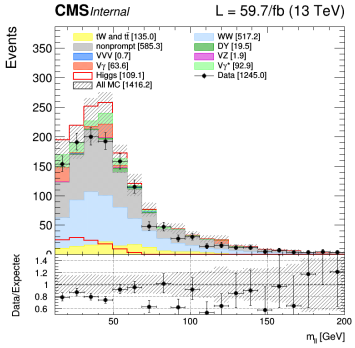
category	sub-categories	
0 jets ggH	$p_{T2} > 20$GeV	$e^+ \mu^-$ $e^- \mu^+$ $\mu^+ e^-$ $\mu^- e^+$
	$p_{T2} < 20$GeV	$e^+ \mu^-$ $e^- \mu^+$ $\mu^+ e^-$ $\mu^- e^+$
1 jet ggH	$p_{T2} > 20$GeV	$e^+ \mu^-$ $e^- \mu^+$ $\mu^+ e^-$ $\mu^- e^+$
	$p_{T2} < 20$GeV	$e^+ \mu^-$ $e^- \mu^+$ $\mu^+ e^-$ $\mu^- e^+$
$\geq$2 jets ggH	$e\mu + \mu e$	

The nonprompt lepton background contribution depends on the signal category, as explained in the previous section. There is a large difference in the number of nonprompt lepton events between categories where the second lepton is an electron or a muon, which is a consequence of their loose and tight definitions. The very strict tight muon requirement corresponded to a low muon fake rate, which consequently resulted in a very reduced nonprompt muon background, in comparison to the electron background. Therefore, in the low p_{T2} categories, which have a larger nonprompt contribution, the number of events is much higher when the second lepton is an electron. The agreement between simulation and data is good in the high p_{T2} categories. In the $p_{T2} < 20$ GeV region, the reduction in the number of events plus the relative increase in nonprompt leptons results in a worse agreement, particularly if the subleading lepton is an electron. The same behaviour is observed using 2016 and 2017 datasets, which can be found in the appendix A.2.

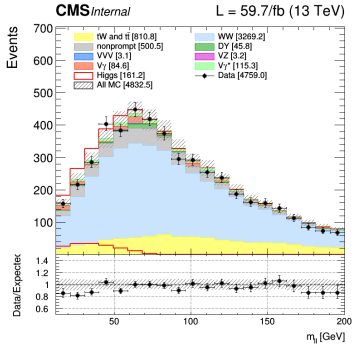
With those two variables we construct a two-dimensional binned histogram, using an optimised binning that depends on the statistics available in each subcategory. The binning is chosen in the following manner:

- For p_{T2} <20 GeV categories 6 bins are used for $m_{\ell\ell}$ [12, 25, 40, 50, 70, 90, 210] (in GeV) and 6 bins are used for m_T^H [60, 80, 90, 110, 130, 150, 200] (in GeV).
- For p_{T2} >20 GeV categories 9 bins are used for $m_{\ell\ell}$ [12, 25, 35, 40, 45, 50, 55, 70, 90, 210] (in GeV) and 8 bins are used for m_T^H [60, 80, 90, 100, 110, 120, 130, 150, 200] (in GeV).

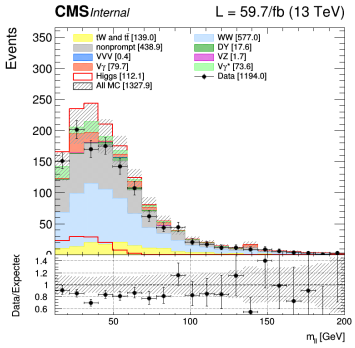
We use an 'unrolled' version of the (m_T^H, $m_{\ell\ell}$) distribution. We build a one-dimensional distribution where adjacent bins have identical m_T^H values and $m_{\ell\ell}$ runs from the first bin (60 to 80 GeV) to the last (90 to 210 GeV). After 6 (low p_{T2}) or 8 (high p_{T2}) consecutive bins corresponding to the same m_T^H interval, the next one corresponds to the following m_T^H interval. Figures 5.5 to 5.6 show the two-dimensional signal region distributions for 2018 in a one-dimensional 'unrolled' histogram, which share the characteristics of the one-dimensional histograms described previously. The 2016 and 2017 signal region distributions can be found in the appendix A.2.



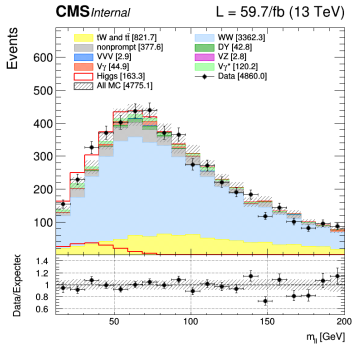
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



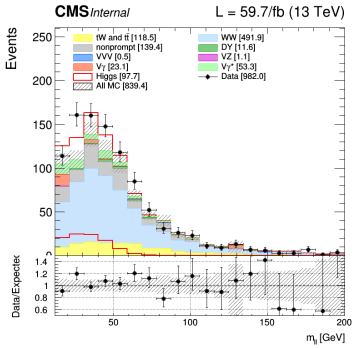
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



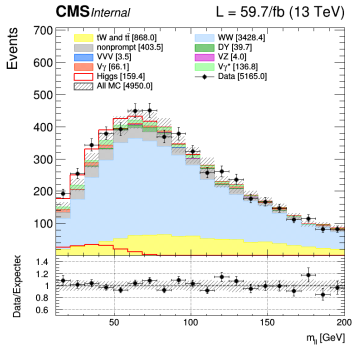
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



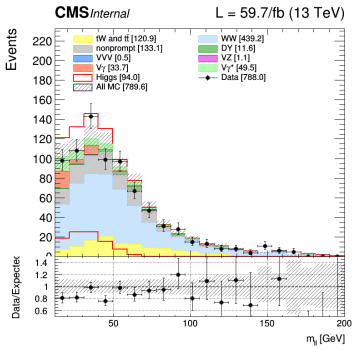
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



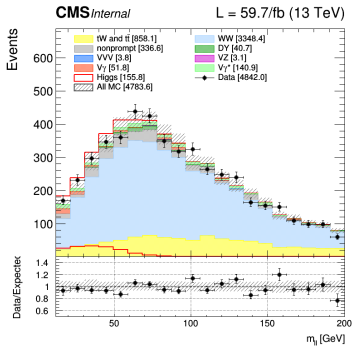
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

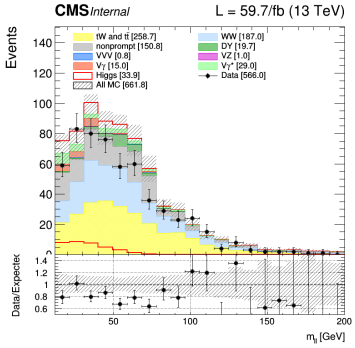


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH

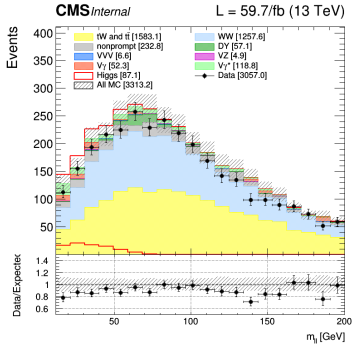


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

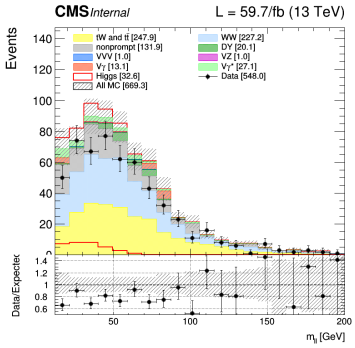
Figure 5.1: Dilepton mass distribution in the 0-jet ggH subcategories for 2018



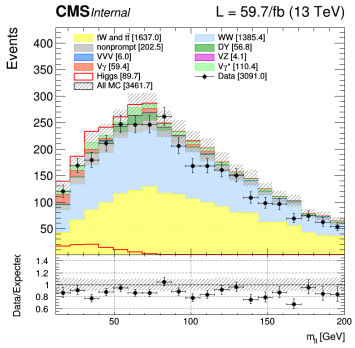
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



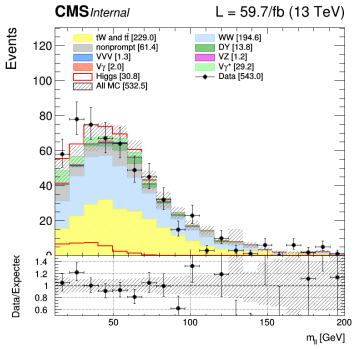
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



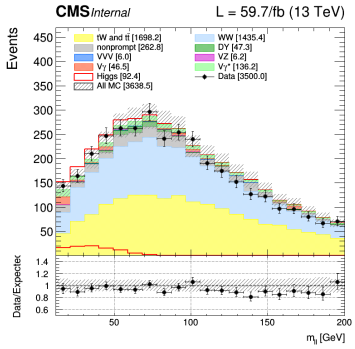
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



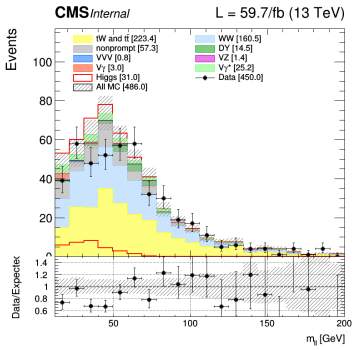
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



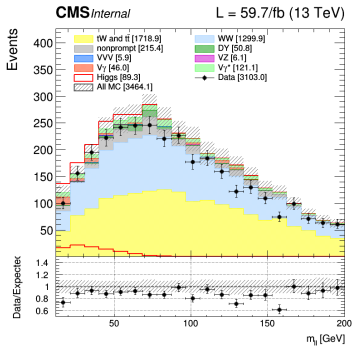
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

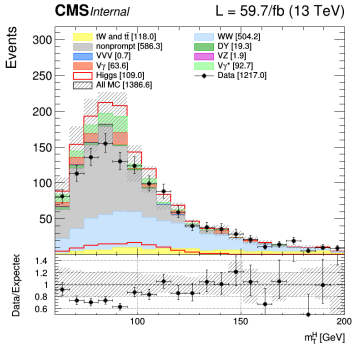


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

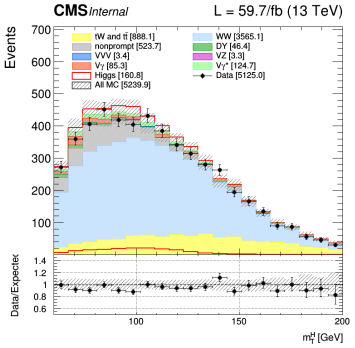


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

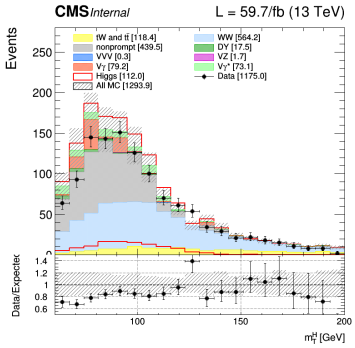
Figure 5.2: Dilepton mass distribution in the 1-jet ggH subcategories for 2018



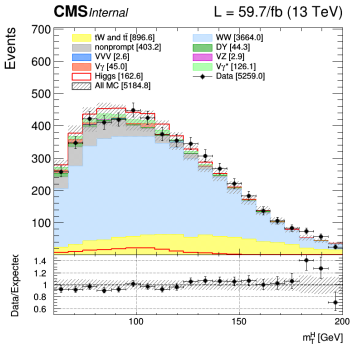
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



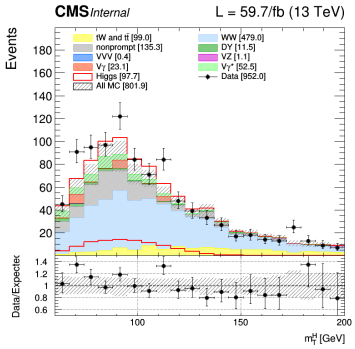
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



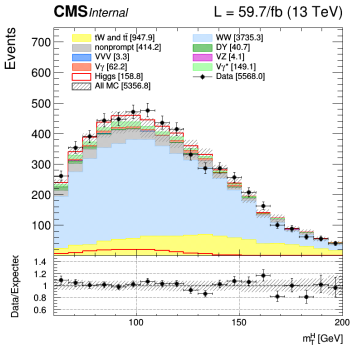
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



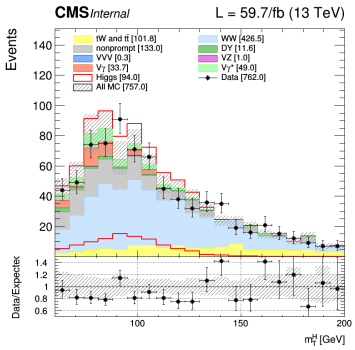
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



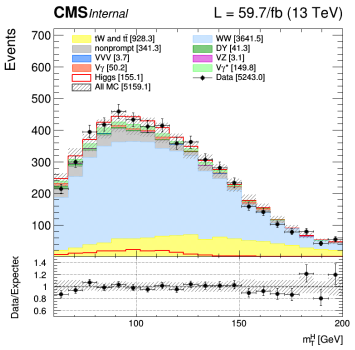
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

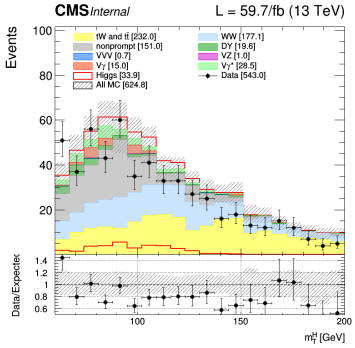


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH

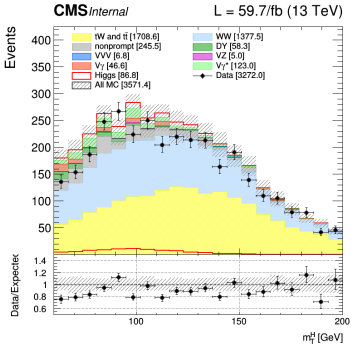


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

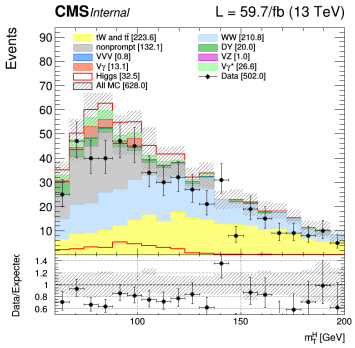
Figure 5.3: Transverse mass distribution in the 0-jet ggH subcategories for 2018



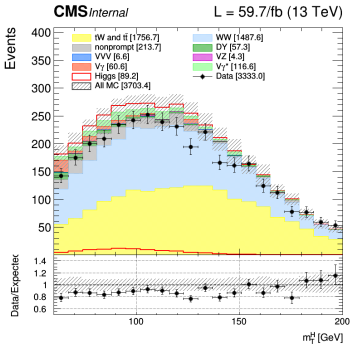
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



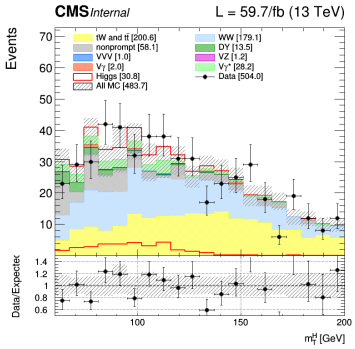
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



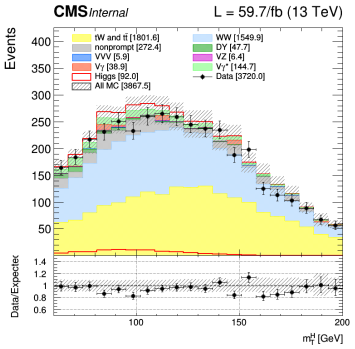
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



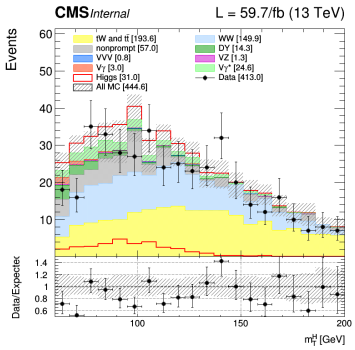
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



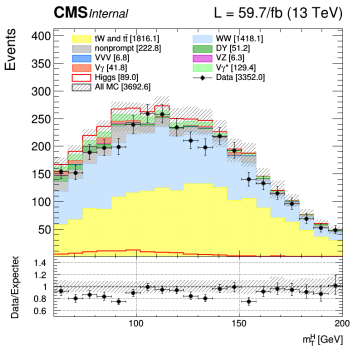
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

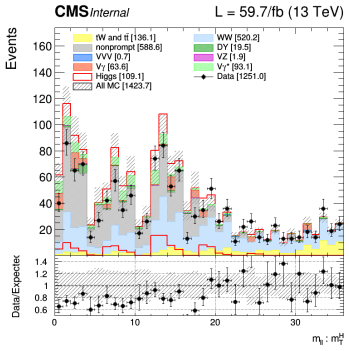


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

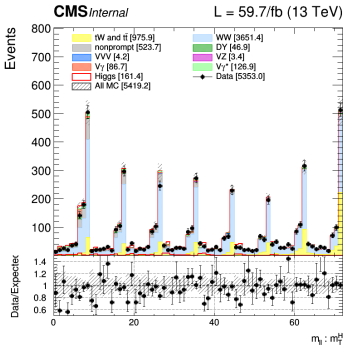


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

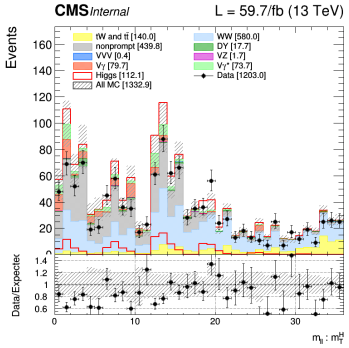
Figure 5.4: Transverse mass distribution in the 1-jet ggH subcategories for 2018



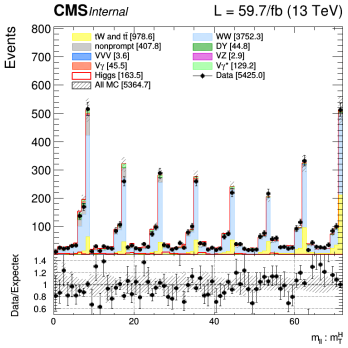
(a) $\mu^- e^+ p_{T2} < 20$ GeV 0-jet ggH



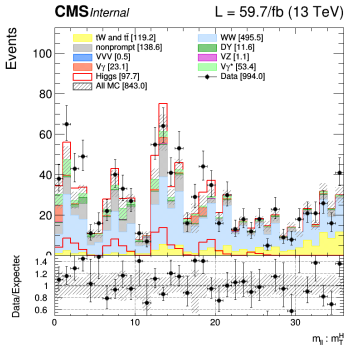
(b) $\mu^- e^+ p_{T2} > 20$ GeV 0-jet ggH



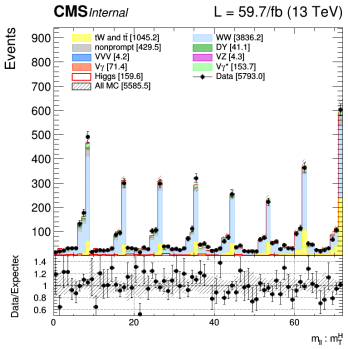
(c) $\mu^+ e^- p_{T2} < 20$ GeV 0-jet ggH



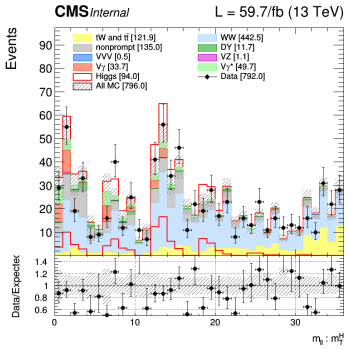
(d) $\mu^+ e^- p_{T2} > 20$ GeV 0-jet ggH



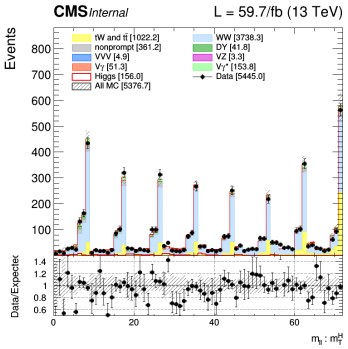
(e) $e^- \mu^+ p_{T2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{T2} > 20$ GeV 0-jet ggH

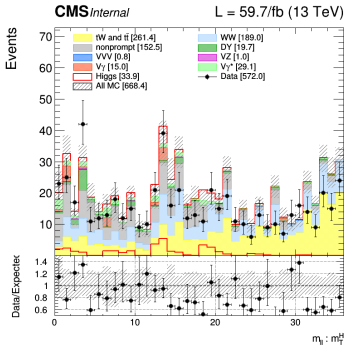


(g) $e^+ \mu^- p_{T2} < 20$ GeV 0-jet ggH

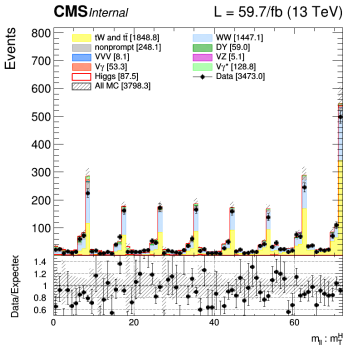


(h) $e^+ \mu^- p_{T2} > 20$ GeV 0-jet ggH

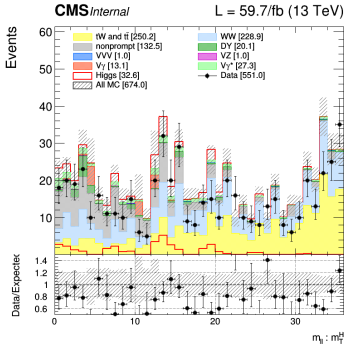
Figure 5.5: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 0-jet ggH subcategories for 2018. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.



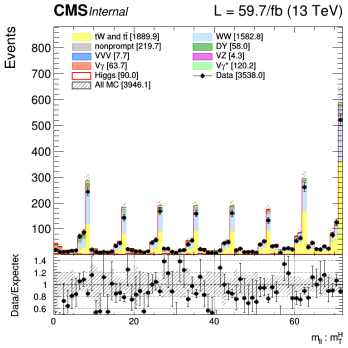
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



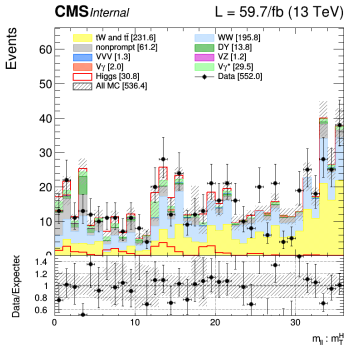
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



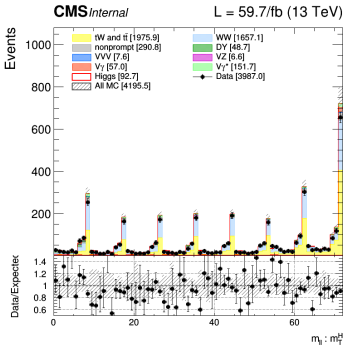
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



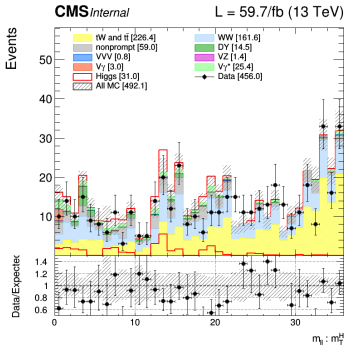
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



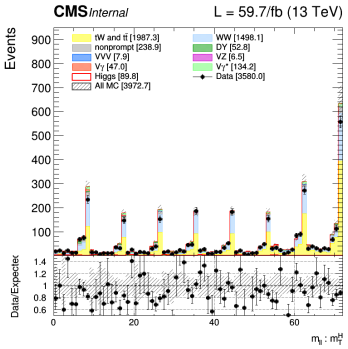
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

Figure 5.6: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 1-jet ggH subcategories for 2018. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.

Table 5.3: Expected yields of signal and background events (before fitting).

2016			
Process	0-jet	1-jet	≥2-jet
ggH (H → WW)	536 ± 18	208 ± 11	38.2 ± 7.6
qqH (H → WW)	5.68 ± 0.46	23.4 ± 1.0	3.00 ± 0.52
WH (H → WW)	10.57 ± 0.43	12.78 ± 0.49	3.61 ± 0.36
ZH (H → WW)	4.70 ± 0.27	4.43 ± 0.32	1.84 ± 0.27
H → $\tau\tau$	3.97 ± 0.16	3.19 ± 0.34	1.06 ± 0.31
DY	133 ± 108	194.3 ± 5.5	62.0 ± 4.8
Top quark	2644 ± 139	5374 ± 327	3009 ± 544
WW	8548 ± 200	3486 ± 125	603 ± 64
Nonprompt leptons	768 ± 89	351 ± 44	85 ± 20
VZ	11.93 ± 0.48	14.62 ± 0.51	3.65 ± 0.34
Vγ	289 ± 31	128 ± 28	37.3 ± 9.8
Vγ^*	409 ± 36	345 ± 254	81 ± 17
VVV	10.17 ± 0.55	16.42 ± 0.89	8.28 ± 0.73

2017			
Process	0-jet	1-jet	≥2-jet
ggH (H → WW)	674 ± 22	292 ± 14	49 ± 10
qqH (H → WW)	6.86 ± 0.33	28.3 ± 1.1	4.70 ± 0.35
WH (H → WW)	13.83 ± 0.47	16.77 ± 0.75	4.59 ± 0.40
ZH (H → WW)	4.90 ± 0.12	4.52 ± 0.13	2.00 ± 0.17
ttH (H → WW)	0.01 ± 0.00	0.19 ± 0.03	0.96 ± 0.22
H → $\tau\tau$	4.71 ± 0.22	4.25 ± 0.43	1.05 ± 0.11
DY	173.2 ± 6.1	276 ± 16	115 ± 48
Top quark	3246 ± 192	6164 ± 450	3302 ± 703
WW	11425 ± 242	4734 ± 169	723 ± 80
Nonprompt leptons	1994 ± 254	1007 ± 126	278 ± 67
VZ	12.79 ± 0.48	16.96 ± 0.59	3.96 ± 0.50
Vγ	311 ± 30	145 ± 36	30 ± 10
Vγ^*	541 ± 37	467 ± 34	94 ± 18
VVV	11.90 ± 0.72	23.47 ± 0.89	11.42 ± 0.76

2018			
Process	0-jet	1-jet	≥2-jet
ggH (H → WW)	1005 ± 33	430 ± 21	69 ± 15
qqH (H → WW)	10.47 ± 0.57	42.6 ± 1.4	6.76 ± 0.47
WH (H → WW)	19.81 ± 0.85	23.62 ± 0.84	6.88 ± 0.62
ZH (H → WW)	8.24 ± 0.38	6.96 ± 0.33	3.25 ± 0.31
ttH (H → WW)	0.02 ± 0.02	0.27 ± 0.04	1.37 ± 0.28
H → $\tau\tau$	5.99 ± 0.31	5.64 ± 0.92	1.43 ± 0.11
DY	239.0 ± 6.0	292 ± 14	86 ± 55
Top quark	4489 ± 317	8535 ± 617	4555 ± 930
WW	16824 ± 410	6825 ± 253	1049 ± 94
Nonprompt leptons	3040 ± 398	1440 ± 212	344 ± 93
VZ	19.34 ± 0.79	26.2 ± 1.0	6.14 ± 0.75
Vγ	452 ± 51	257 ± 167	26 ± 25
Vγ^*	822 ± 52	628 ± 42	138 ± 22
VVV	18.8 ± 1.0	34.9 ± 1.9	15.5 ± 2.0

Table 5.3 contains the expected signal and background yields in the signal region for each of the datasets, summed over the lepton flavour, p_{T2} and charge categories. Apart from the difference in total number of events resulting from the increasing luminosity, a difference in the relative contribution of the nonprompt lepton background is seen between 2016 and the other years, due to the difference in lepton identification criteria. The expected background events decrease for categories with larger number of jets, except for the top quark background and DY, VZ and VVV processes, which have more events in the 1-jet categories.

5.2.4 Expected signal strength uncertainty

In order to assess the performance of these channels, a simultaneous fit including all the aforementioned categories is performed. These uncertainties are determined with a fit using the Asimov dataset instead of data, where the signal strength values have been set to their SM expected values. The expected signal strength modifier uncertainties per-year and per-jet category are shown in Table 5.4, as well as the combination across years and jet categories. The result shows that the 0-jet category has the smallest uncertainty, between 15 and 18% with a combined uncertainty using all years of 10-11%. 1-jet categories have uncertainties between 22 and 28% with a combined uncertainty of 15-16%, and ≥2-jet categories have higher uncertainties, between 57 and 87% with a combined uncertainty of 40-46%. The precision improves between years because of the increase in luminosity. The uncertainties for each year combining the 3 categories range between 13 and 16% and the combined uncertainty for all years and categories is 9-10%.

Table 5.4: Expected signal strength uncertainties obtained by an Asimov dataset fit. The results are shown for each year and each jet category, as well as combined across jet categories (last row) and years (rightmost column). The cell on the bottom right is the result of the combination of all years and all categories.

	2016	2017	2018	Combination
0 jets	−0.17/+0.18	−0.17/+0.18	−0.15/+0.16	−0.10/+0.11
1 jet	−0.26/+0.28	−0.23/+0.25	−0.20/+0.22	−0.15/+0.16
≥2 jets	−0.79/+0.87	−0.74/+0.80	−0.57/+0.63	−0.40/+0.46
Combination	−0.15/+0.16	−0.15/+0.16	−0.13/+0.14	−0.09/+0.10

To evaluate which uncertainties most affect the result of the signal strength, impact plots are produced for each of the three years. Among the long list of uncertainties considered for each year, the thirty whose variations change the signal strength result the most are shown in Figs. 5.7, 5.8 and 5.9 for the 0- and 1- jet categories and the years 2016, 2017 and 2018, respectively. The impacts of the uncertainties in the signal categories with two or more jets are in appendix B.

The uncertainties are classified according to their impact, with the highest ranking being the ones which most affect the signal strength. The 'impacts', presented on the right side of the plots, represent the relative variation of the signal strength when a certain nuisance shifts by one standard deviation. The colour of the plot represents if the variation of the signal strength is correlated (red) or anticorrelated (blue) with the variation in the nuisance parameter. The left side of the impact plot represents the 'pulls' or how much the fitted value of the nuisance differs from the initial one in terms of its uncertainty, which may increase or decrease from their original values during the fit. Since these plots are using the Asimov dataset, the maximum likelihood fit result for each nuisance is equal to its initial value in all cases, and the pulls are all equal to 0.

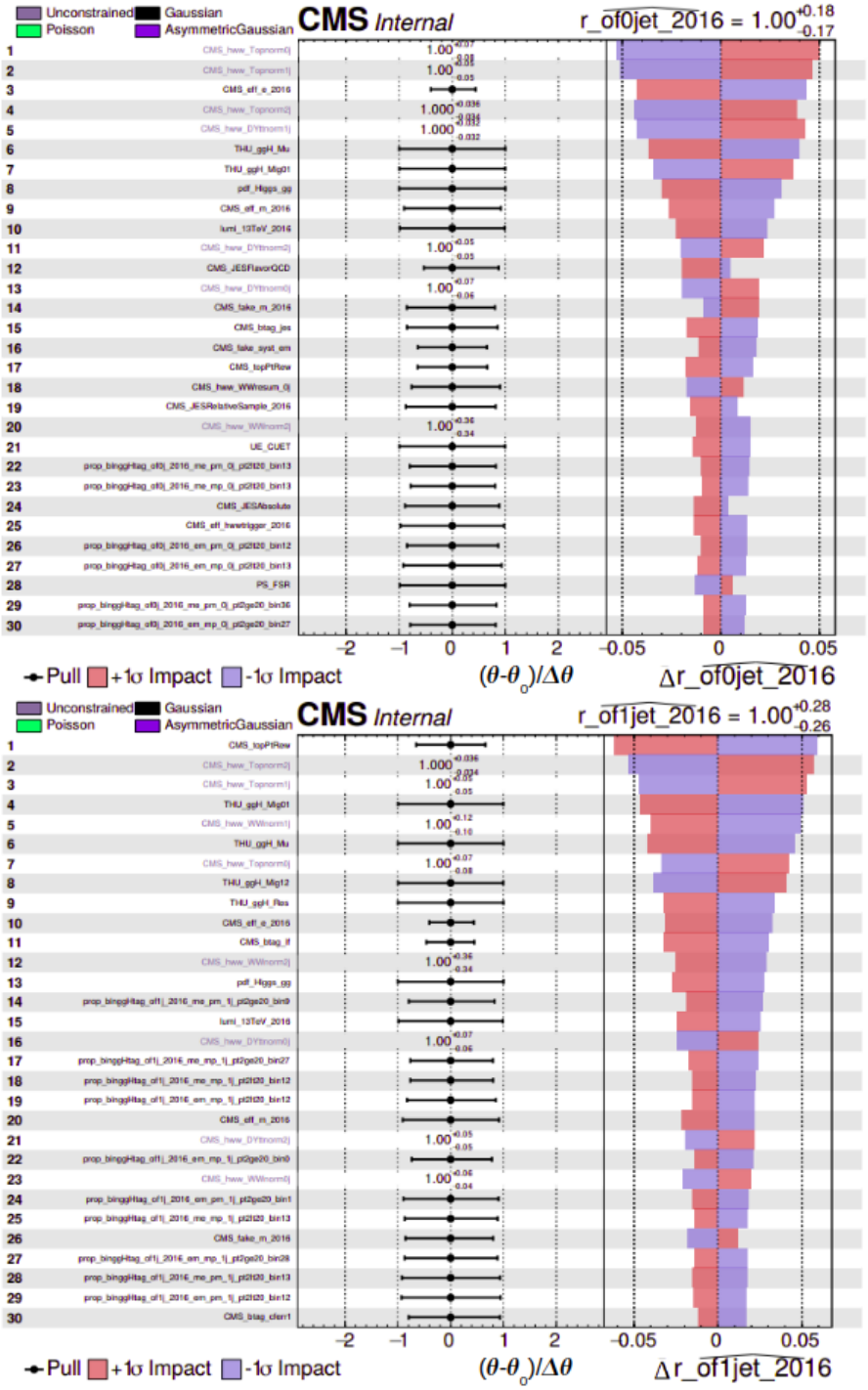


Figure 5.7: Impact of the first 30 uncertainties with highest ranking on the expected signal strengths for 0-jet (top), 1-jet (bottom), obtained with the 2016 dataset.

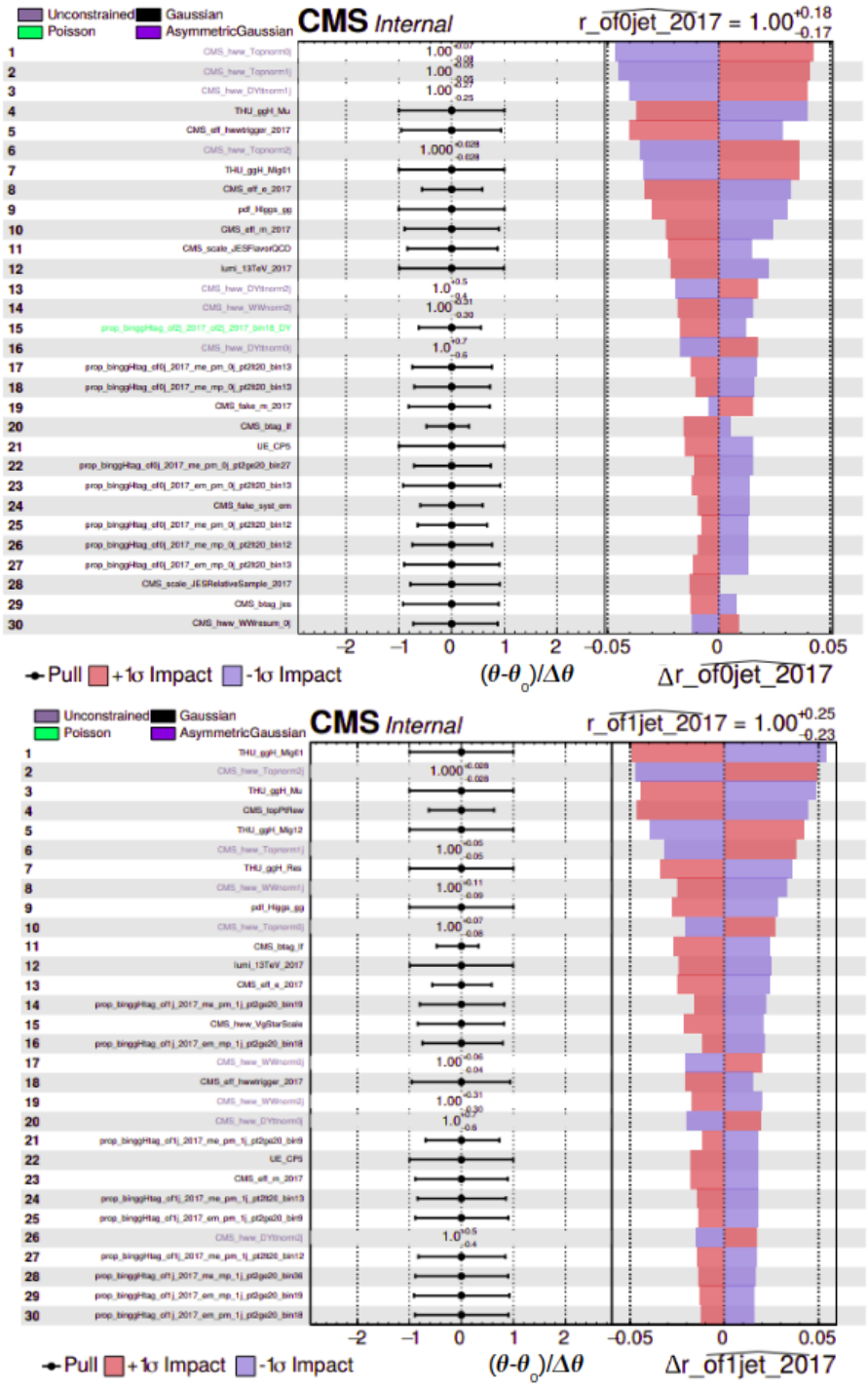


Figure 5.8: Impact of the first 30 uncertainties with highest ranking on the expected signal strengths for 0-jet (top) and 1-jet (bottom), obtained with the 2017 dataset.

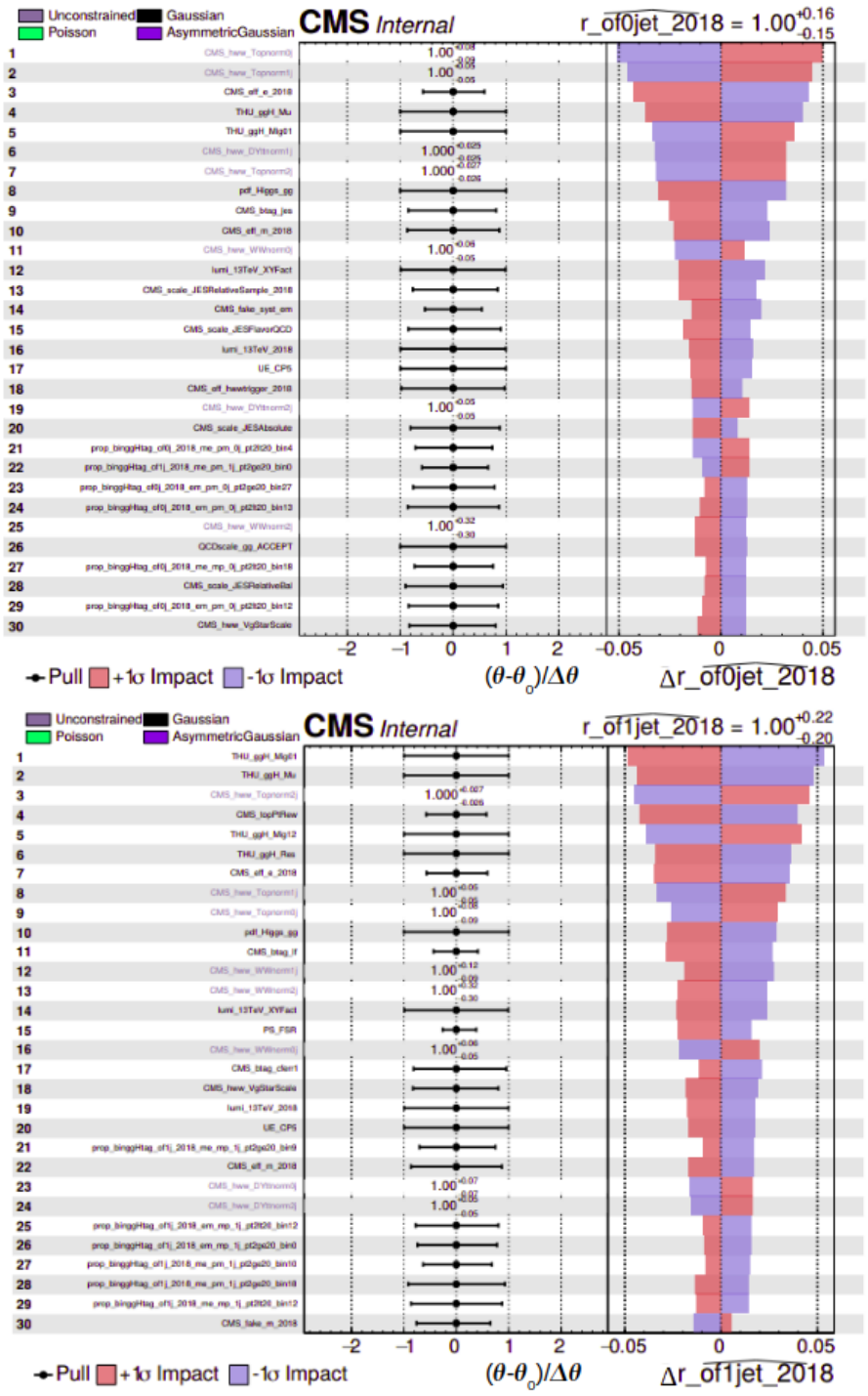


Figure 5.9: Impact of the first 30 uncertainties with highest ranking on the expected signal strengths for 0-jet (top) and 1-jet (bottom), obtained with the 2018 dataset.

In the 0-jet category, the parameter that changes the signal strength result the most is the normalisation of the top background, which was left free to float in the fit. These normalisation parameters are not nuisance parameters since they are not constrained within uncertainties, but their impact is also shown in these plots, along with the normalisation value with its uncertainty on the left side. In this case, using data Asimov all normalisations are equal to one. The impact shows that the normalisation of the top quark background may change the value of the signal strength up to 5%. The DY background normalisation is also very important in the 0-jet category, with a maximum 4% effect on the signal strength.

From the theoretical uncertainties, the ggH theoretical scale and PDF uncertainties are the most relevant ones, and the experimental uncertainties that are most impactful are the electron and muon efficiencies and the uncertainty in the luminosity. Similar uncertainties appear in the main impacts of the 1-jet category, with the main difference being that the top quark p_{T} reweighting is more relevant, as the top quark background modeling is even more important in the 1-jet category. The WW background normalisation is in this case more impactful than the DY background normalisation. Finally, there are other uncertainties that affect more this category with 1 jet, such as the theoretical ggH scale uncertainties reflecting migrations of jet-multiplicity bins and the experimental uncertainties for the jet b-tagging.

5.2.5 Observed signal strength

We determine the observed signal strength modifier through a maximum likelihood fit to the observed data, where the contributions of the different signal and backgrounds processes scale according to the signal strength and nuisance parameter values that maximise the likelihood function given the data. The contribution from all the signal samples will be multiplied by the same signal strength, using all the production modes and both the $\mathrm{H} \to \mathrm{WW}$ as well as the $\mathrm{H} \to \tau\tau$ signal samples. However, most of the signal contribution using our selection that targets the gluon fusion production mode comes from ggH $\mathrm{H} \to \mathrm{WW}$ events, as was shown in the yields of Table 5.3. The signal strength is determined separately for every year and for the 0- and 1-jet categories. We also obtain a signal strength per year using the combined 0- and 1- categories, and we combine the datasets to obtain results for each jet category. The obtained signal strength modifiers are summarised in Table 5.5 and Fig. 5.10.

Table 5.5: Observed signal strength obtained with a maximum likelihood fit. The results are shown for each year and each jet category, as well as combined across jet categories (last row) and years (rightmost column). The cell on the bottom right is the result of the combination of all years and all categories.

	2016	2017	2018	Combination
0 jets	$0.88^{+0.19}_{-0.17}$	$0.85^{+0.17}_{-0.17}$	$0.82^{+0.15}_{-0.14}$	$0.86^{+0.11}_{-0.10}$
1 jet	$1.36^{+0.31}_{-0.28}$	$1.09^{+0.25}_{-0.23}$	$0.72^{+0.20}_{-0.18}$	$0.99^{+0.17}_{-0.15}$
Combination	$1.01^{+0.17}_{-0.16}$	$0.93^{+0.16}_{-0.16}$	$0.79^{+0.13}_{-0.13}$	$0.90^{+0.10}_{-0.09}$

The combination of 0- and 1-jet categories with the 2016 dataset shows a good agreement with the SM expectation and so do the 2017 categories. 2018 is 1.6 σ away from the SM prediction. The final value for the Higgs boson global signal strength is $\mu_{global}=0.90^{+0.10}_{-0.09}$, compatible with the SM.

The uncertainty in the combined result of the signal strength can be separated into theoretical systematic, experimental systematic and statistical uncertainties:

$$\mu_{global} = 0.90^{+0.10}_{-0.09} = 0.90^{+0.04}_{-0.05}\ (\text{stat.})^{+0.06}_{-0.05}\ (\text{exp.})^{+0.07}_{-0.06}\ (\text{theo.}) \tag{5.4}$$

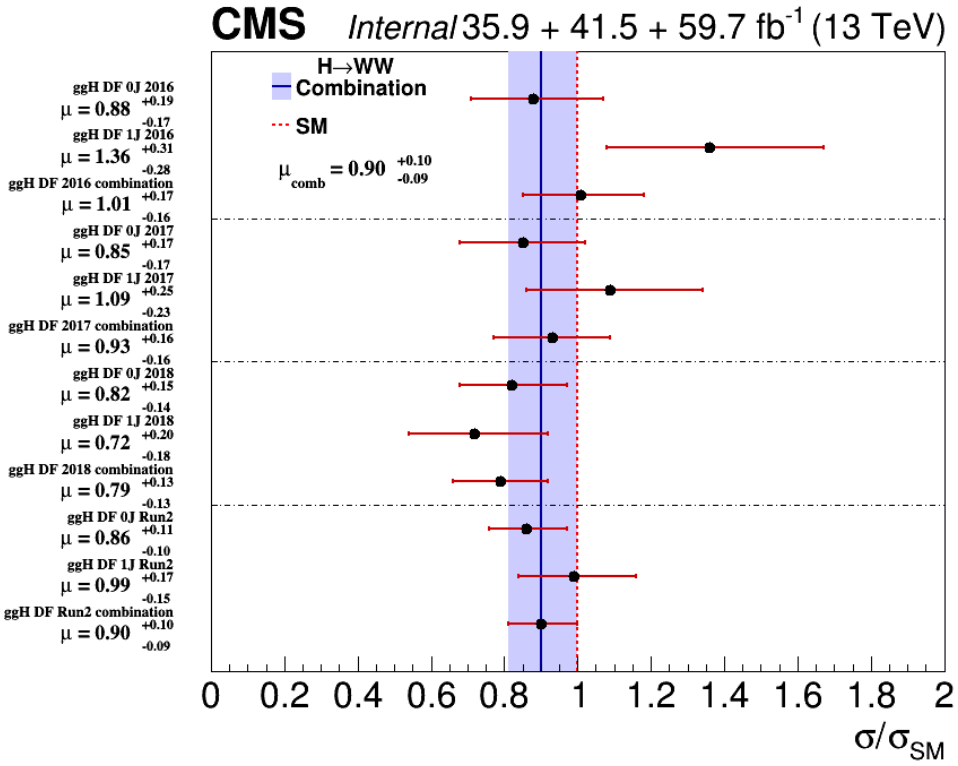


Figure 5.10: Measured signal strengths for the 0- and 1-jet categories and their combination, obtained with the three datasets. The final results of the Run 2 combination are shown below.

The previous result obtained with the 2016 dataset only [24] was μ=$1.28^{+0.18}_{-0.17}$, which was higher than the new result presented in this work, as seen in 5.6. The main difference comes from the 0-jet category. There are several possible reasons for this variation: since the previous analysis there have been changes in the lepton definitions, b-tagging algorithm employed, E_{T}^{miss} definition, PU identification, samples being used and other more subtle changes. A combination of all of those factors resulted in that shift in the central value.

Table 5.6: Comparison with the previous signal strength results for 2016

	0-jet	1-jet	Combination
Latest result	$0.88^{+0.19}_{-0.17}$	$1.36^{+0.32}_{-0.28}$	$1.01^{+0.17}_{-0.16}$
Previous result for 2016	$1.30^{+0.24}_{-0.23}$	$1.29^{+0.32}_{-0.27}$	$1.28^{+0.18}_{-0.17}$

In Tables 5.7 to 5.9 the pre-fit and post-fit yields of signals and backgrounds are compared for the three years in the 0- and 1-jet categories, where all the H → WW signal processes have been added into one contribution. The uncertainties given include statistical and systematic uncertainties. The uncertainties in the pre-fit yields are obtained with the initial value of the systematic uncertainties, whereas the post-fit ones are obtained with the fitted values of the nuisance parameters. This results in a reduction of the uncertainty in the yields of almost all processes after fitting. The processes that have the most change in yields are the top quark and WW production, since they are the main backgrounds, but the relative difference is larger in the nonprompt lepton background. That one is reduced post-fit in all cases, and in 2017 and 2018 that reduction is over 40%.

Table 5.7: Pre-fit and post-fit yields obtained with the 2016 dataset.

Category	Process	Pre-fit	Post-fit
0-jet ggH	$H \to WW$	557 ± 18	560 ± 26
	$H \to \tau\tau$	3.97 ± 0.16	4.05 ± 0.19
	DY	133 ± 108	119 ± 37
	Top quark	2644 ± 139	2381 ± 76
	WW	8548 ± 200	9161 ± 120
	Nonprompt leptons	769 ± 89	560 ± 39
	$V\gamma$	289 ± 31	247 ± 16
	$V\gamma^*/VZ$	421 ± 33	401 ± 23
	VVV	10.17 ± 0.55	9.99 ± 0.38
1-jet ggH	$H \to WW$	248 ± 11	263 ± 15
	$H \to \tau\tau$	3.19 ± 0.34	3.19 ± 0.28
	DY	194.3 ± 5.5	168.5 ± 2.9
	Top quark	5374 ± 327	4792 ± 150
	WW	3486 ± 125	3556 ± 147
	Nonprompt leptons	351 ± 44	256 ± 20
	$V\gamma$	128 ± 28	123 ± 18
	$V\gamma^*/VZ$	359 ± 112	332 ± 55
	VVV	16.42 ± 0.89	16.10 ± 0.59

Table 5.8: Pre-fit and post-fit yields obtained with the 2017 dataset.

Category	Process	Pre-fit	Post-fit
0-jet ggH	$H \to WW$	699 ± 20	633 ± 29
	$H \to \tau\tau$	4.71 ± 0.22	4.22 ± 0.21
	DY	173.2 ± 6.1	160.8 ± 7.2
	Top quark	3246 ± 192	3100 ± 88
	WW	11425 ± 242	11535 ± 140
	Nonprompt leptons	1994 ± 254	1113 ± 71
	$V\gamma$	311 ± 30	259 ± 18
	$V\gamma^*/VZ$	554 ± 34	589 ± 31
	VVV	11.90 ± 0.72	11.86 ± 0.52
1-jet ggH	$H \to WW$	342 ± 14	318 ± 16
	$H \to \tau\tau$	4.25 ± 0.43	3.95 ± 0.28
	DY	276 ± 16	286 ± 12
	Top quark	6164 ± 450	5868 ± 229
	WW	4734 ± 169	4481 ± 183
	Nonprompt leptons	1007 ± 126	582 ± 42
	$V\gamma$	145 ± 36	113 ± 12
	$V\gamma^*/VZ$	484 ± 32	504 ± 30
	VVV	23.47 ± 0.89	22.81 ± 0.65

Table 5.9: Pre-fit and post-fit yields obtained with the 2018 dataset.

Category	Process	Pre-fit	Post-fit
0-jet ggH	$H \to WW$	1044 ± 34	817 ± 39
	$H \to \tau\tau$	5.99 ± 0.31	4.76 ± 0.21
	DY	239.0 ± 6.0	240.8 ± 6.7
	Top quark	4489 ± 317	4338 ± 128
	WW	16824 ± 410	18026 ± 199
	Nonprompt leptons	3040 ± 398	1558 ± 78
	$V\gamma$	452 ± 51	368 ± 31
	$V\gamma^*/VZ$	841 ± 53	877 ± 46
	VVV	18.8 ± 1.0	17.91 ± 0.47
1-jet ggH	$H \to WW$	504 ± 21	385 ± 19
	$H \to \tau\tau$	5.64 ± 0.92	4.47 ± 0.24
	DY	292 ± 14	320.2 ± 6.4
	Top quark	8535 ± 617	7271 ± 209
	WW	6825 ± 253	7104 ± 208
	Nonprompt leptons	1440 ± 212	738 ± 45
	$V\gamma$	257 ± 83	177 ± 24
	$V\gamma^*/VZ$	654 ± 43	681 ± 40
	VVV	34.9 ± 1.9	32.96 ± 0.84

The impacts of the uncertainties in the signal strength are shown in Figs. 5.11 to Figs. 5.13 for the three datasets. The normalisation of the top quark and DY backgrounds, the theoretical ggH uncertainties and the muon and electron efficiencies are the uncertainties that affect the 0-jet category the most. In the 1-jet categories, top quark and WW background normalisations, theoretical ggH uncertainties, top p_T reweighting and b-tagging are the most impactful uncertainties. Comparing them to the expected impacts, there are not many differences in the type of uncertainties that have the highest ranking in each category, with a few exceptions. One of such is the statistical uncertainties (parameters of the Barlow-Beeston method) of two bins in the 2017 signal region, that have more effect than expected and a pull of over one standard deviation. They also seem to be very constrained by the fit. Other statistical uncertainties have large pulls in 2017 and 2018, but with less effect on the signal strength.

Impacts from the uncertainties affecting the combination result of the three years are shown for the 0- and 1-jet categories in Fig. 5.14 and combined in Fig. 5.15. Theoretical uncertainties in the ggH QCD scale and PDF have the biggest impacts in the signal strength, as well as normalisation of the top quark and DY background. Experimental uncertainties such as the electron efficiency, luminosity and the top quark p_T reweighting are also among the most relevant ones. The nonprompt background systematic uncertainty of 30% appears a bit pulled from the central value, which is probably a consequence of using a very conservative value for the uncertainty, which results in a much smaller value after the fit. The statistical uncertainties due to low populated bins of the two-dimensional distributions of 2017 are present in the combination as well.

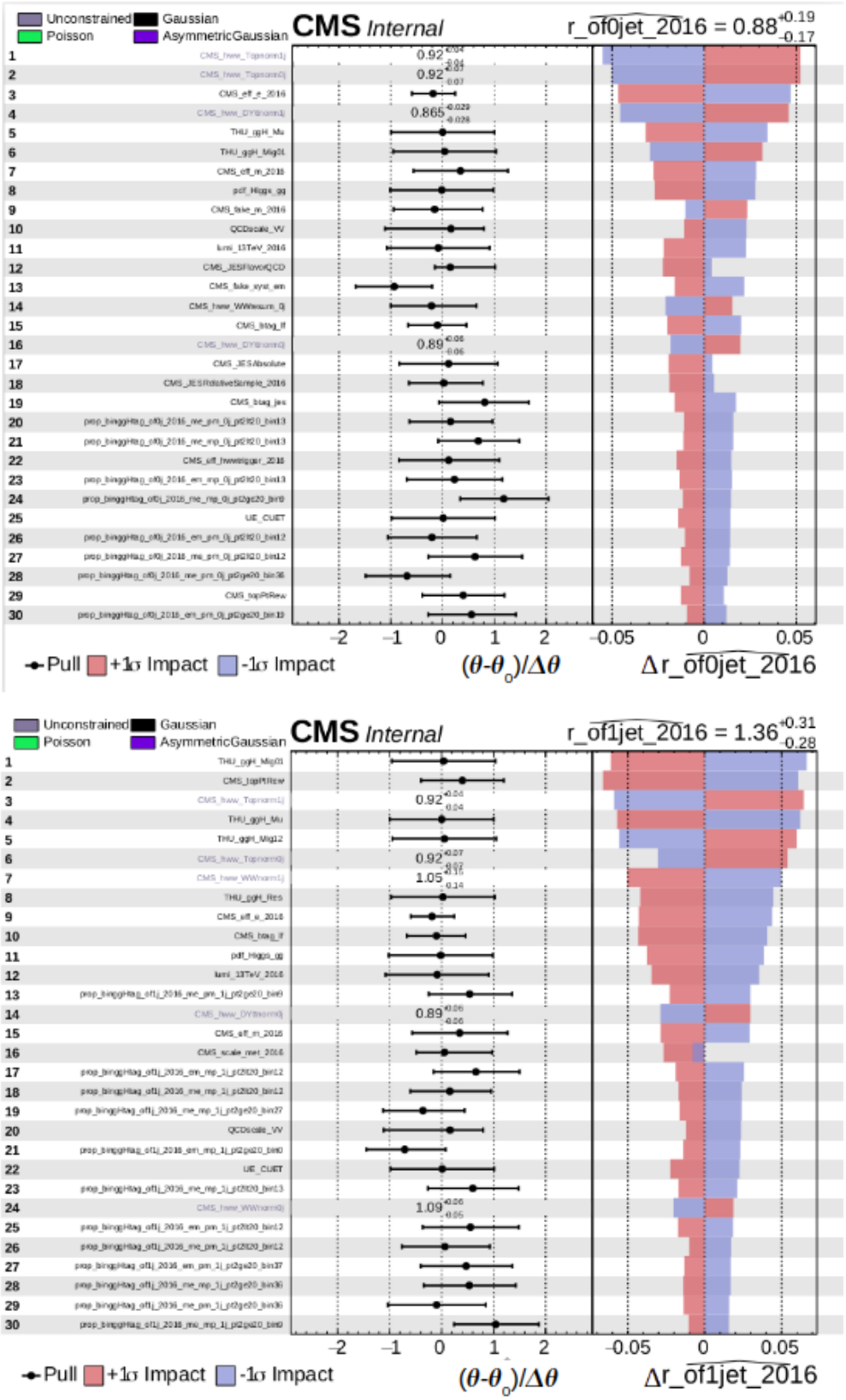


Figure 5.11: Impacts of the first 30 uncertainties with highest ranking on the measured signal strengths for 0-jet (top), 1-jet (bottom), obtained with the 2016 dataset.

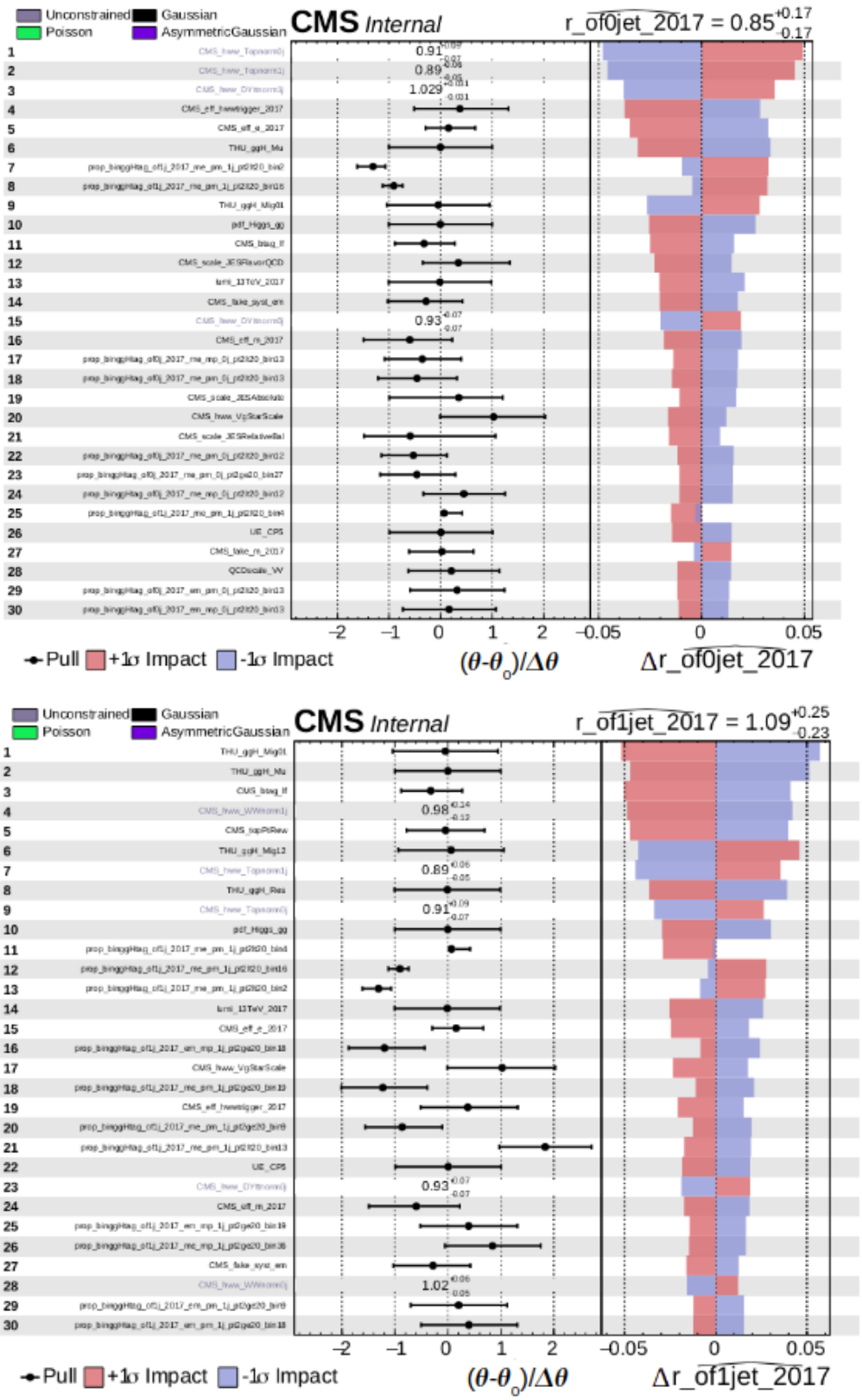


Figure 5.12: Impacts of the first 30 uncertainties with highest ranking on the measured signal strengths for 0-jet (top), 1-jet (bottom), obtained with the 2017 dataset.

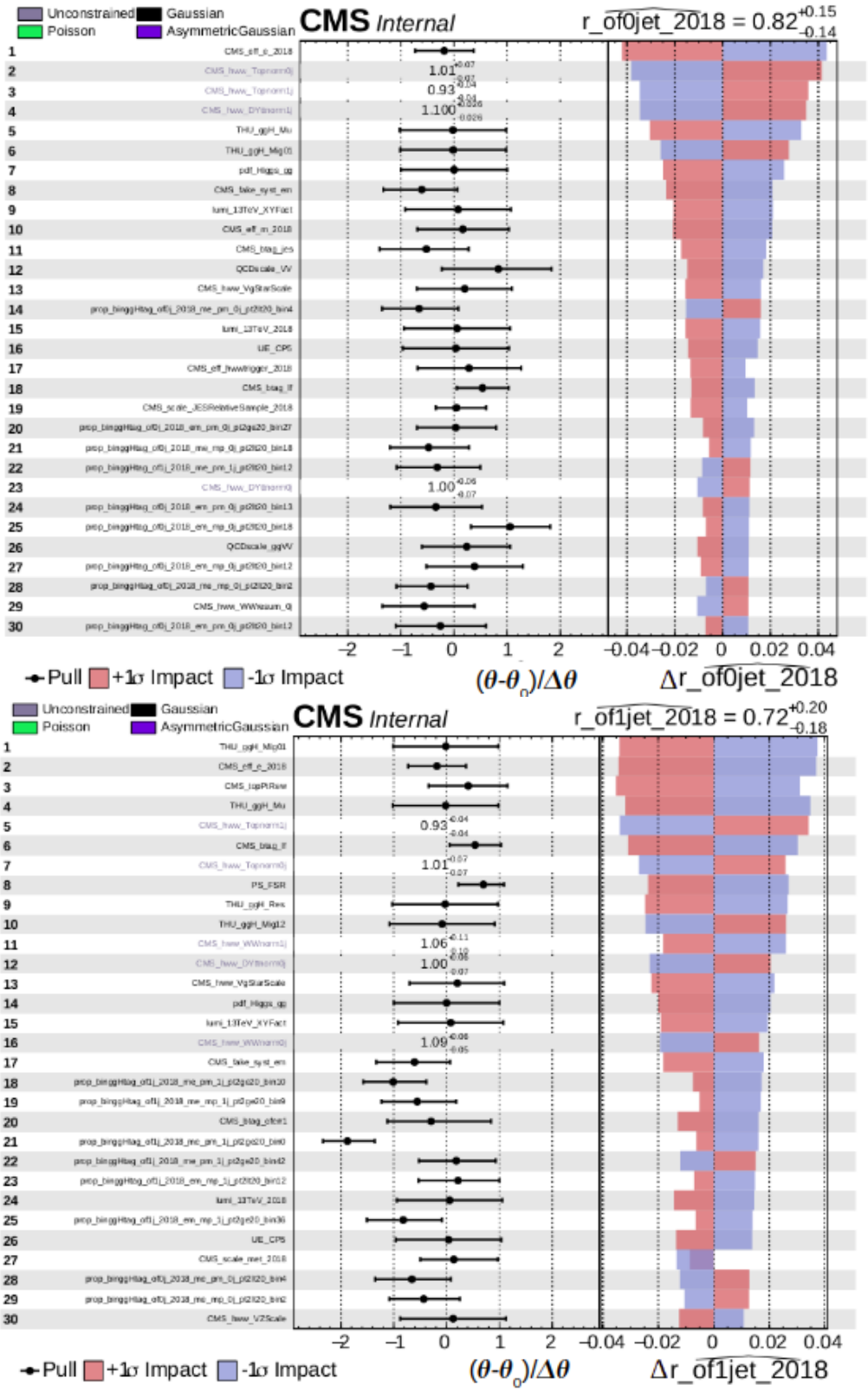


Figure 5.13: Impacts of the first 30 uncertainties with highest ranking on the measured signal strengths for 0-jet (top), 1-jet (bottom), obtained with the 2018 dataset.

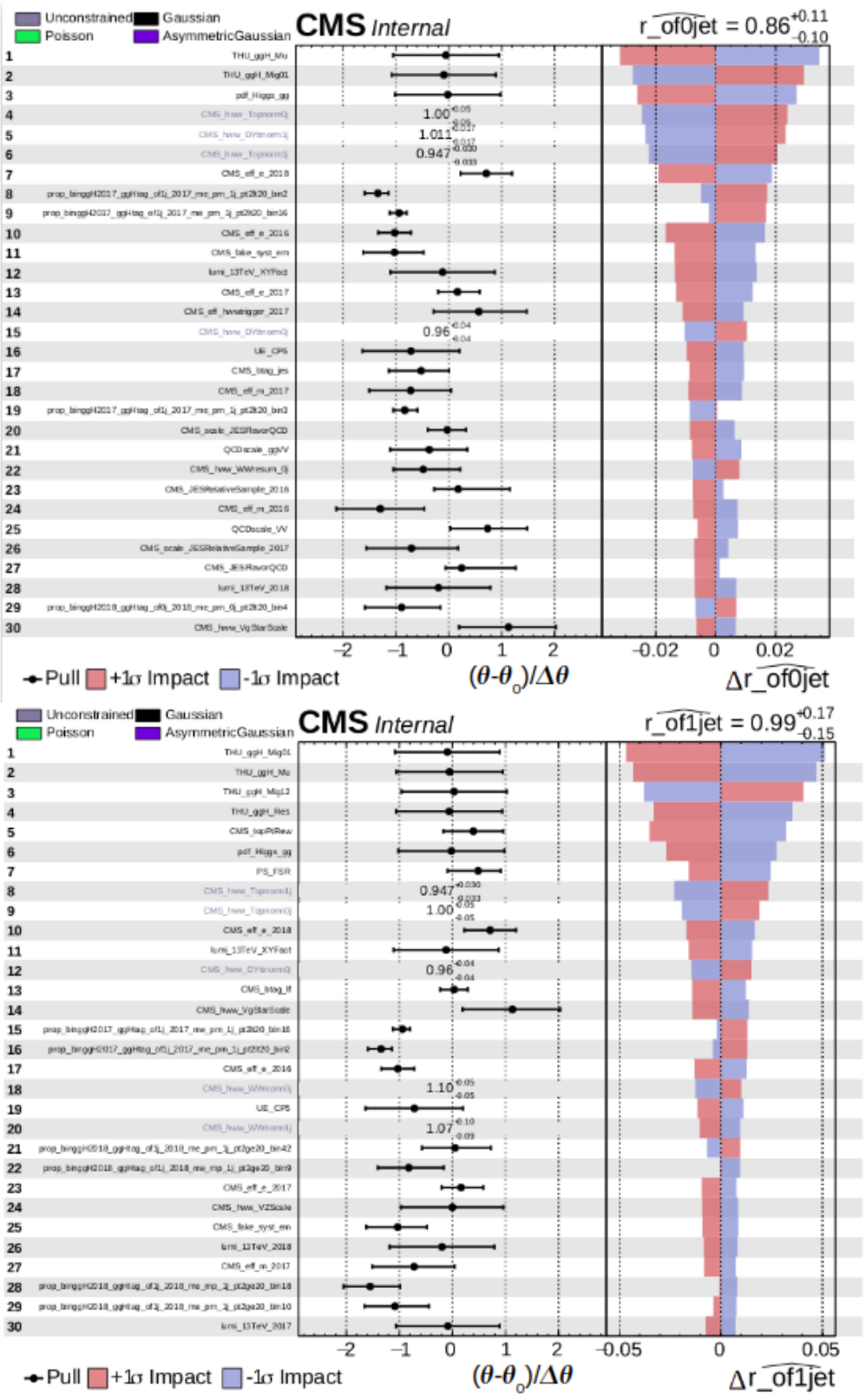


Figure 5.14: Impacts of the first 30 uncertainties with highest ranking on the measured signal strengths for 0-jet (top), 1-jet (bottom), obtained with the combination of the three datasets.

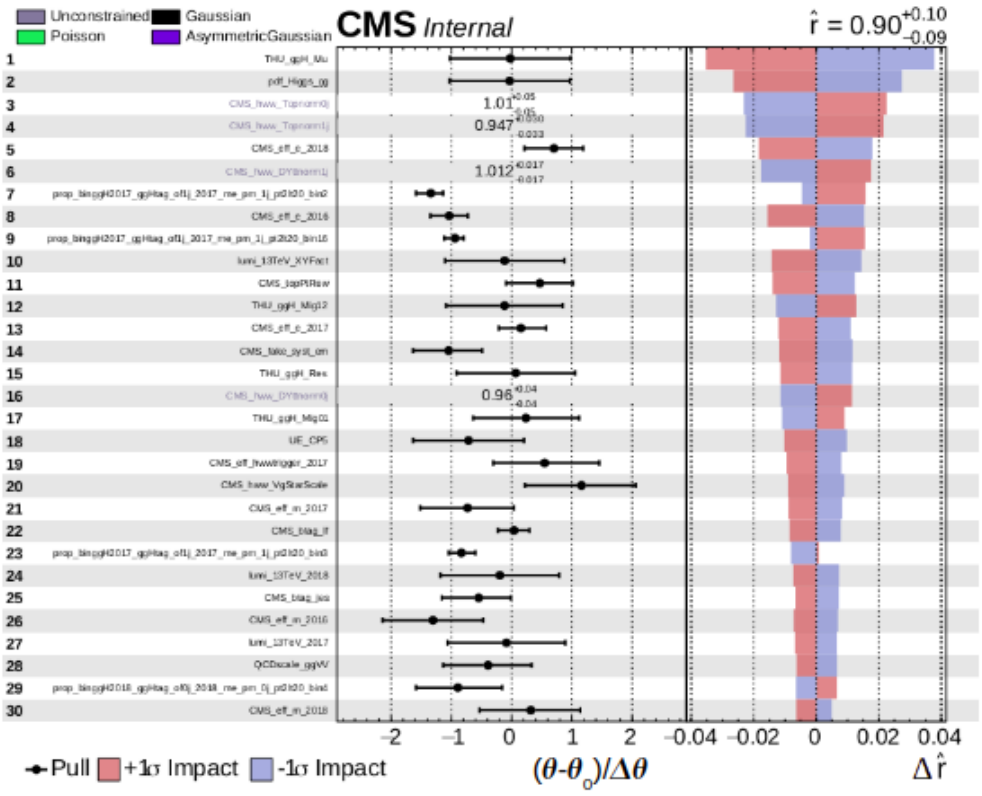


Figure 5.15: Impact of the first 30 uncertainties with highest ranking on the measured signal strengths for the combination of 0- and 1-jet categories, obtained with the combination of the three datasets.

5.3 Simplified Template Cross Section analysis results

5.3.1 The STXS framework

The Higgs boson analyses during the LHC Run 1 measured as their main results inclusive signal strength modifiers and multiplicative coupling modifiers. The so-called STXS framework was developed by the theorists and experimentalists of the LHC Higgs Cross Section Working Group to improve these measurements in Run 2 by scrutinising in more detail different kinematic regions. It aims to separate measurement and interpretation steps in order to reduce the theory dependencies that are contained in the measurements, including the dependence on the underlying physics model (SM or BSM). Moreover, they provide more information for theoretical interpretations by performing the measurement in finely-grained regions of phase space. Other CMS analysis also achieved some results using the STXS framework and Run 2 data, such as H→ZZ [116], H→ $\gamma\gamma$ [117] and H→ $\tau\tau$ [17], and the ATLAS experiment has some Run 2 STXS measurements [118], in agreement with the SM predictions.

The primary goals of the STXS framework are to maximise the sensitivity of the measurements while at the same time to minimise their dependence on the theory. Since these are competing requirements, a nontrivial compromise has to be achieved. Some of the defining features of the STXS are:

- Measure signal strength modifiers and cross sections in mutually exclusive regions of phase space and separated into production modes, which allows the isolation of possible BSM effects in certain regions.
- Analyses can use different optimised selection criteria for the definition of the STXS categories at reconstruction level.
- The kinematic regions are inclusive over the Higgs decays, which allows a global combination of the measurements in all decay channels.
- If an analysis has enough sensitivity, a finer binning is suggested in the framework.

The measured exclusive regions of phase space, called *STXS bins* for simplicity, are specific to the different production modes. STXS bins are defined by using generator-level information in the MC simulation about the Higgs boson and the jets in the event. All the bins require a Higgs boson rapidity to be less than 2.5. The high Higgs boson rapidity region of the phase space is not accessible because of the limited coverage of the detector in the forward region. Generator-level jets are defined as anti-k_T jets with a distance parameter of 0.4 and a minimum p_T of 30 GeV, built from all stable particles not coming from the Higgs boson decay. These generator-level jets also include charged leptons, photons and neutrinos.

The STXS bins have been progressively defined in stages corresponding to increasingly fine granularity thanks to the larger data samples available. The initial *stage 0* production bins simply corresponded to the different Higgs boson production mechanisms: ggH, qqH, $b\bar{b}H$, $t\bar{t}H$, tH and VH, which was divided into qqWH, qqZH and ggZH, with hadronic decays of the W or Z boson. To account for the evolving experimental sensitivity, a *stage 1.1* was defined, where measurements in each of the production modes were performed in bins divided by kinematic variables at generator level such as the number of generator-level jets or the p_T of the Higgs boson.

The *stage 1.1* definition further developed into the *stage 1.2* definition for the Higgs production modes that we study in the H → WW analyses. The proposed scheme for ggH and ggZH hadronic (i.e. gg→ ZH where Z→qq, H→WW) is shown in Fig. 5.16, the scheme for VBF and qq→ VH hadronic is shown in Fig. 5.17 and VH leptonic (i.e. VH where V → $\ell\ell/\ell\nu$, H → WW) is shown in Fig. 5.18. Solid lines in the coloured boxes determine the minimal proposed STXS bins, whereas optional divisions between bins that could be used if there was enough sensitivity are drawn as dashed boundaries. The *stage 1.2* bins for $b\bar{b}H$, $t\bar{t}H$ and tH production modes are not presented in this work because of their much smaller cross section.

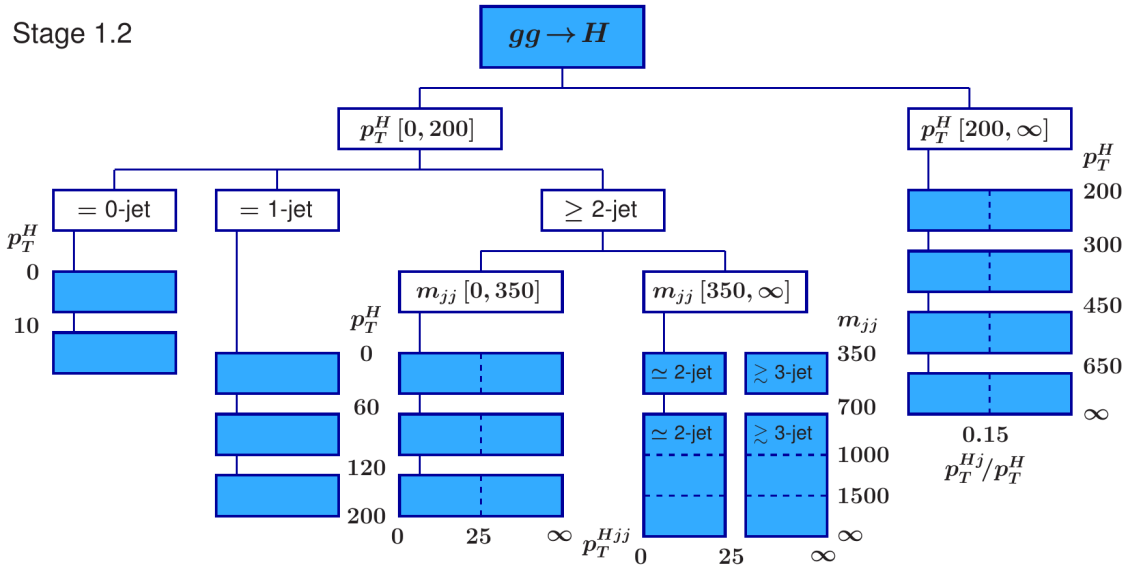


Figure 5.16: Simplified template cross-section scheme that defines the kinematic regions in the STXS stage 1.2 categorisation for the gluon fusion production mechanism, as well as gg→Z(→qq)H. The STXS bins are shown in blue boxes, where solid lines divide different bins and dashed boundaries represent additional bins that could be used if there was enough sensitivity. Picture taken from [119].

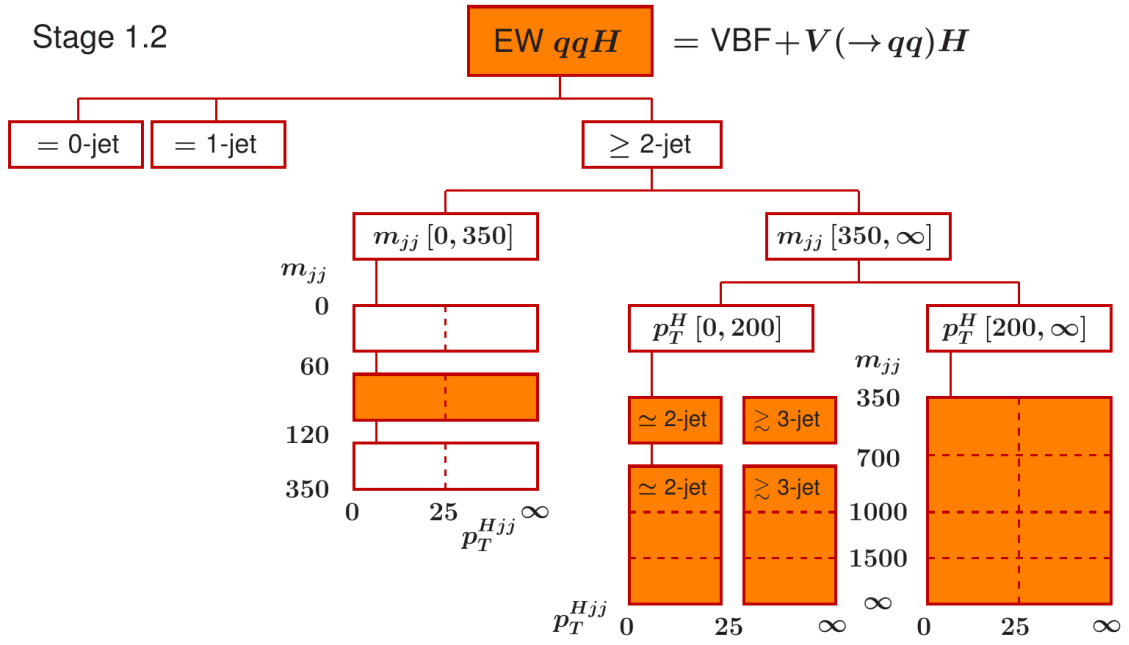


Figure 5.17: Simplified template cross-section scheme that defines the kinematic regions in the STXS stage 1.2 categorisation for the vector boson fusion and the associated VH production modes. The vector boson in VH has to decay hadronically. The bins for vector boson fusion and for VH hadronic production modes are shown in orange boxes, where solid lines divide different bins and the dashed boundaries represent additional bins that could be used if there was enough sensitivity. VH production is concentrated on the $60 < m_{jj} < 120$ GeV bin and VBF in the higher m_{jj} bins. Picture taken from [119].

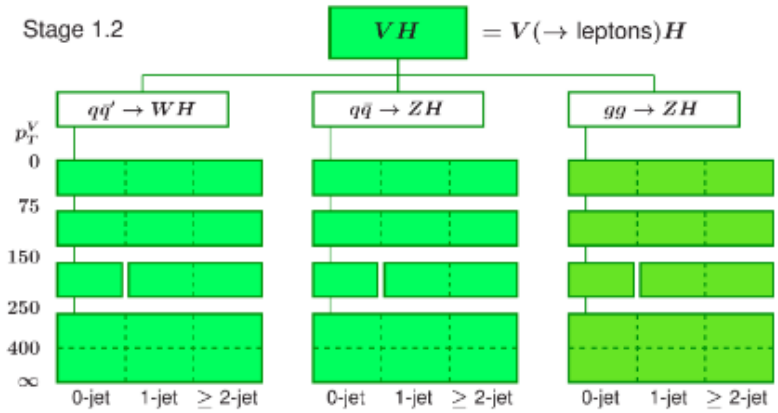


Figure 5.18: Simplified template cross-section scheme that defines the kinematic regions in the STXS stage 1.2 categorisation for the VH associated production, with the leptonic decay of the vector boson. The bins are shown in green boxes, where solid lines divide different bins and the dashed boundaries represent additional bins that could be used if there was enough sensitivity. Picture taken from [119].

The STXS bins defined in Figs. 5.16, 5.17 and 5.18 use as main kinematic variables the number of jets in the event and the Higgs transverse momentum $p_{\mathrm{T}}^{\mathrm{H}}$. At reconstructed level in H $\rightarrow$ WW that variable is calculated as the vectorial sum of the $\vec{p}_{\mathrm{T}}^{miss}$ and the transverse momentum of the leptons. In the bins with two jets or more, the dijet invariant mass m_{jj} and the transverse momentum of the two jets plus the Higgs transverse momentum $p_{\mathrm{Hjj}}^{\mathrm{T}}$ are also used to define the bin boundaries. In the leptonic VH STXS scheme the p_{T} of the associated vector boson $p_{\mathrm{T}}^{\mathrm{V}}$ is also used to define STXS bins.

For the ggH STXS bins, there is a first division in $p_{\mathrm{T}}^{\mathrm{H}}$, with four bins with $p_{\mathrm{T}}^{\mathrm{H}}$ higher than 200 GeV: 200<$p_{\mathrm{T}}^{\mathrm{H}}$ <300 GeV, 300<$p_{\mathrm{T}}^{\mathrm{H}}$ <450 GeV, 450<$p_{\mathrm{T}}^{\mathrm{H}}$ <650 GeV and $p_{\mathrm{T}}^{\mathrm{H}}$ >650 GeV. The region with $p_{\mathrm{T}}^{\mathrm{H}}$ <200 GeV is divided into three jet categories with 0, 1 and 2 or more jets. The 0-jet region is divided into two STXS bins depending on the Higgs boson p_{T}: $p_{\mathrm{T}}^{\mathrm{H}}$ < 10 GeV and $p_{\mathrm{T}}^{\mathrm{H}}$ > 10 GeV. The 1-jet region is divided into 3 STXS bins, with $p_{\mathrm{T}}^{\mathrm{H}}$ < 60 GeV, 60<$p_{\mathrm{T}}^{\mathrm{H}}$ <120 GeV and 120<$p_{\mathrm{T}}^{\mathrm{H}}$ <200 GeV. The $\geq$2-jet region has a first division depending on the value of the dijet invariant mass m_{jj}, there is a region with m_{jj}<350 GeV and another with m_{jj}>350 GeV. For the low dijet mass region there are three bins with $p_{\mathrm{T}}^{\mathrm{H}}$ < 60 GeV, 60<$p_{\mathrm{T}}^{\mathrm{H}}$ <120 GeV and 120<$p_{\mathrm{T}}^{\mathrm{H}}$ <200 GeV. Finally, there are four $\geq$2-jet high dijet mass bins, depending on m_{jj} and $p_{\mathrm{Hjj}}^{\mathrm{T}}$. There are two bins with m_{jj}<700 GeV and two bins with m_{jj}<700 GeV. In each of those cases, one of the bins corresponds to $p_{\mathrm{Hjj}}^{\mathrm{T}}$<25 GeV and the other to $p_{\mathrm{Hjj}}^{\mathrm{T}}$>25 GeV. The division at $p_{\mathrm{Hjj}}^{\mathrm{T}}$ > 25 GeV implicitly establishes a $\geq$3-jet category, since the transverse momentum of the Higgs boson plus two jets system has to be balanced with the existence of more jets.

There are different STXS binning schemes for ggH, qqH and VH leptonic, adapted to the kinematics of the corresponding production mode. This is the reason for having a division in 0- and 1-jet bins for ggH, but not for VBF and VH, where there is no signal in 0- or 1-jet categories. Dividing into low and high dijet mass events in the $\geq$2-jet in ggH allows to disentangle those bins from the VBF STXS bins, since the dijet invariant mass is expected to be higher in VBF events. The VH hadronic production bin is dominated by the contribution in the single bin of the qqH scheme with a dijet mass between 60 and 120 GeV, compatible with two jets coming from a W or Z boson decay. VH leptonic STXS bins are splitted in the production mode at generator level, distinguishing between qq $\rightarrow$ WH, qq $\rightarrow$ ZH and gg $\rightarrow$ ZH, as well as the vector boson p_{T}.

The ggH STXS bins at high $p_{\mathrm{T}}^{\mathrm{H}}$

In *stage 1.1* the high $p_{\mathrm{T}}^{\mathrm{H}}$ region for the gluon fusion production mode was represented by a single bin of $p_{\mathrm{T}}^{\mathrm{H}}$ >200 GeV. The decision to divide that high $p_{\mathrm{T}}^{\mathrm{H}}$ bin into four for *stage 1.2* (as shown in 5.16) is the result of a study performed by CMS and ATLAS analysis groups to check possible binning configurations in that region. A higher granularity in that kinematic region is interesting from a theoretical perspective because of the possibility of isolating a possible BSM physics effect, but the very low amount of data in that region means there is low sensitivity for the measurements performed in each bin if we divide the region too much.

In H $\rightarrow$ WW we tested the significance of a signal in three STXS bins defined by different $p_{\mathrm{T}}^{\mathrm{H}}$ boundaries in the $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV region: $200 p_{\mathrm{T}}^{\mathrm{H}} < 300$ GeV, $300 p_{\mathrm{T}}^{\mathrm{H}} < 400$ GeV and $p_{\mathrm{T}}^{\mathrm{H}} > 500$ GeV. Our aim was first to check the sensitivity of our analysis to those regions, and then see whether that sensitivity could improve if there was a splitting not only in $p_{\mathrm{T}}^{\mathrm{H}}$ but in the number of jets as well, or if the bin boundary depended on the p_{T} of the jets. We studied some possible STXS binning configurations by evaluating the signal significance for each of them, using a ggH signal and the main background samples (WW and top quark) with a total luminosity of 137.5 fb^{-1}.

Figure 5.19 shows some estimations of the signal significance using those three high $p_{\mathrm{T}}^{\mathrm{H}}$ bins (in different colours), dividing into the 3 jet categories and how much that significance could change if there was a jet p_{T} threshold. The jet p_{T} threshold is represented by a limit on the parameter r, which is the ratio between the jet p_{T} and the Higgs boson p_{T}. The first result from this study is that for H $\rightarrow$ WW the decrease in significance of the signal strength measured in the bins with a higher $p_{\mathrm{T}}^{\mathrm{H}}$ threshold meant that there was little sensitivity in the bins with $p_{\mathrm{T}}^{\mathrm{H}}$ >300 GeV. This is not an issue since we can merge the bins with $p_{\mathrm{T}}^{\mathrm{H}}$ >300 GeV for our analysis and the definitions of higher $p_{\mathrm{T}}^{\mathrm{H}}$ bins can be kept for other Higgs boson analyses that may be sensitive to that region.

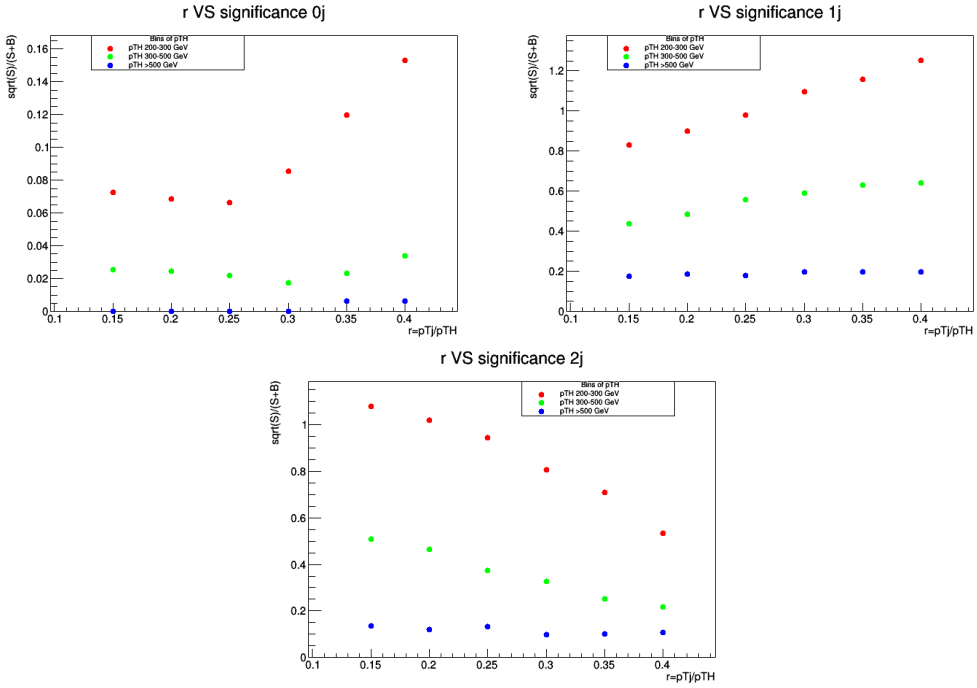


Figure 5.19: Signal significance for three STXS high $p_{\mathrm{T}}^{\mathrm{H}}$ bins (in colours) as a function of the threshold set for the ratio between the first jet p_{T} and the Higgs p_{T}. The 0-jet category has no sensitivity while the other two show a small dependence with respect to that ratio.

There was not much to gain from splitting in number of jets at high $p_{\mathrm{T}}^{\mathrm{H}}$, since almost all the events in that kinematic region have jets and the possible 0-jet bin would contain very low statistics (less than 1 expected event above 300 GeV). This is also the case even if we define the jet categories according a different jet p_{T} threshold than the 30 GeV that we use. Using a different jet p_{T} threshold for the bin boundaries was not useful for increasing the significance since the events migrate between jet bins without introducing meaningful changes to the significance.

The final decision to divide the high $p_{\mathrm{T}}^{\mathrm{H}}$ bin into four for the *stage 1.2* (as shown in 5.16) is the result of this study and similar ones performed by other CMS and ATLAS analyses. The most sensitive $200 p_{\mathrm{T}}^{\mathrm{H}} < 300$ GeV was kept, and for the higher $p_{\mathrm{T}}^{\mathrm{H}}$ region three bins were defined instead of two. This does not change the HWW analysis approach, since we do not have enough sensitivity and merge those bins into one, but for future combinations with other channels and with ATLAS analyses, these bins can be considered again.

STXS bin event fraction

We can obtain the fraction of events from the signal simulated samples that are included in each of the STXS bins. Those fractions are obtained without applying any selection requirements, but applying the theoretical reweighting for gluon fusion samples to NNLO. The results are summarised in Tables 5.10 to 5.13. Some differences are observed between samples of different years due to differences in PDF and UE tunes. However, for a combined measurement the reference cross sections should be the same for each year so that signal strengths can be compared. To account for this, the difference between years is treated as an uncertainty in the event fraction, taking the 2018 fractions as the central value.

In Table 5.10 we see that the 0-jet bins have about 57% of the events, compared to 22% for the 1-jet bins and less than 10% for the $\geq$2 bins, which is why we have more sensitivity in the lower number of jet categories. 45% of the events come from the 0-jet $10 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin alone. In 1-jet bins the number of events decrease for higher $p_{\mathrm{T}}^{\mathrm{H}}$ values, with the 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin having the largest contribution of the three. The high $p_{\mathrm{T}}^{\mathrm{H}}$ bins only correspond to about 1.4% of the events, with most of its contribution coming from the $200 < p_{\mathrm{T}}^{\mathrm{H}} < 300$ GeV bin. Events with a Higgs rapidity range of $|\eta| > 2.5$, which correspond to the region that cannot be measured, amount to 8-9% of the total events.

Table 5.10: Event fraction estimation for the STXS gluon fusion bins using the signal MC samples of the analysis corresponding to the three years. Differences are consequence of different PDF and UE tunes.

ggH bins	2016	2017	2018
0-jet $p_{\mathrm{T}}^{\mathrm{H}} \leq 10$ GeV	11.80%	12.60%	12.64%
0-jet 10 $<p_{\mathrm{T}}^{\mathrm{H}} <200$ GeV	45.76%	44.99%	44.97%
1-jet $p_{\mathrm{T}}^{\mathrm{H}} \leq 60$ GeV	13.01%	14.11%	14.12%
1-jet 60 $<p_{\mathrm{T}}^{\mathrm{H}} <120$ GeV	8.07%	8.21%	8.15%
1-jet 120 $<p_{\mathrm{T}}^{\mathrm{H}} <200$ GeV	1.34%	1.41%	1.43%
$\geq$2-jet $m_{jj} \leq$350 GeV, $p_{\mathrm{T}}^{\mathrm{H}} \leq 60$ GeV	1.94%	1.96%	1.98%
$\geq$2-jet $m_{jj} \leq$350 GeV, 60 $<p_{\mathrm{T}}^{\mathrm{H}} <120$ GeV	3.87%	3.72%	3.71%
$\geq$2-jet $m_{jj} \leq$350 GeV, 120 $<p_{\mathrm{T}}^{\mathrm{H}} <200$ GeV	1.91%	1.74%	1.74%
$\geq$2-jet 350 $<m_{jj} <700$, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} \leq 25$ GeV	0.45%	0.52%	0.51%
$\geq$2-jet 350 $<m_{jj} <700$, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} >25$ GeV	0.69%	0.59%	0.59%
$\geq$2-jet $m_{jj} >700$ GeV , $p_{\mathrm{T}_{jj}}^{\mathrm{H}} \leq 25$ GeV	0.22%	0.23%	0.23%
$\geq$2-jet $m_{jj} >700$ GeV , $p_{\mathrm{T}_{jj}}^{\mathrm{H}} >25$ GeV	0.30%	0.24%	0.25%
200 $<p_{\mathrm{T}}^{\mathrm{H}} <300$ GeV	1.10%	1.03%	1.01%
300 $<p_{\mathrm{T}}^{\mathrm{H}} <450$ GeV	0.33%	0.30%	0.30%
450 $<p_{\mathrm{T}}^{\mathrm{H}} <650$ GeV	0.06%	0.06%	0.06%
$p_{\mathrm{T}}^{\mathrm{H}} >650$ GeV	0.01%	0.01%	0.01%
Higgs $\lvert\eta\rvert >2.5$	9.14%	8.29%	8.30%

Table 5.11: Event fraction estimation for the STXS vector boson fusion bins using the signal MC samples of the analysis corresponding to the three years. Differences are consequence of different PDF and UE tunes.

VBF bins	2016	2017	2018
0-jet	7.49%	6.98%	6.96%
1-jet	33.26%	32.85%	32.87%
$\geq$2-jet $m_{jj} \leq 60$ GeV	1.38%	1.34%	1.39%
$\geq$2-jet 60 $<m_{jj} <120$ GeV	2.52%	2.48%	2.47%
$\geq$2-jet 120 $<m_{jj} <350$ GeV	11.99%	12.30%	12.34%
$\geq$2-jet $m_{jj} >350$ GeV, $p_{\mathrm{T}}^{\mathrm{H}} \geq 200$ GeV	3.91%	4.02%	3.95%
$\geq$2-jet 350 $<m_{jj} <700$, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} \leq 25$ GeV	9.63%	10.25%	10.21%
$\geq$2-jet 350 $<m_{jj} <700$, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} >25$ GeV	4.10%	3.86%	3.86%
$\geq$2-jet $m_{jj} >700$ GeV, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} \leq 25$ GeV	14.28%	14.97%	14.97%
$\geq$2-jet $m_{jj} >700$ GeV, $p_{\mathrm{T}_{jj}}^{\mathrm{H}} >25$ GeV	4.46%	4.21%	4.25%
Higgs $\lvert\eta\rvert >2.5$	6.96%	6.74%	6.73%

Table 5.12: Event fraction estimation for the STXS WH hadronic and leptonic bins using the signal MC samples of the analysis corresponding to the three years. Differences are consequence of different PDF and UE tunes.

WH hadronic bins	2016	2017	2018
0-jet	6.14%	4.85%	5.64%
1-jet	32.03%	29.51%	31.87%
$\geq$2-jet $m_{jj} \leq 60$ GeV	3.48%	2.92%	3.19%
$\geq$2-jet 60 $< m_{jj} <$120 GeV	28.70%	32.72%	30.29%
$\geq$2-jet 120 $< m_{jj} <$350 GeV	13.17%	14.20%	13.13%
$\geq$2-jet $m_{jj} >$350 GeV, $p_{\mathrm{T}}^{\mathrm{H}} \geq 200$ GeV	0.82%	0.83%	0.77%
$\geq$2-jet 350 $< m_{jj} <$700, $p_{\mathrm{T}Hjj}^{\mathrm{T}} \leq 25$ GeV	0.37%	0.38%	0.37%
$\geq$2-jet 350 $< m_{jj} <$700, $p_{\mathrm{T}Hjj}^{\mathrm{T}} >$25 GeV	2.01%	2.15%	1.95%
$\geq$2-jet $m_{jj} >$700 GeV, $p_{\mathrm{T}Hjj}^{\mathrm{T}} \leq 25$ GeV	0.06%	0.06%	0.06%
$\geq$2-jet $m_{jj} >$700 GeV, $p_{\mathrm{T}Hjj}^{\mathrm{T}} >$25 GeV	0.48%	0.54%	0.45%
Higgs $\lvert\eta\rvert >$2.5	12.75%	11.84%	12.28%
WH leptonic bins	2016	2017	2018
$p_{\mathrm{T}}^{\mathrm{V}} \leq 75$ GeV	46.08%	46.28%	46.37%
75 $< p_{\mathrm{T}}^{\mathrm{V}} <$150 GeV	29.02%	29.33%	29.27%
150 $< p_{\mathrm{T}}^{\mathrm{V}}$ 250 GeV 0-jet	5.18%	5.28%	5.24%
150 $< p_{\mathrm{T}}^{\mathrm{V}}$ 250 GeV $\geq$1-jets GeV	4.00%	3.89%	3.83%
$p_{\mathrm{T}}^{\mathrm{V}} >$250 GeV	2.95%	2.80%	2.81%
Higgs $\lvert\eta\rvert >$2.5	12.77%	12.43%	12.47%

Table 5.13: Event fraction estimation for the STXS ZH hadronic and leptonic bins using the signal MC samples of the analysis corresponding to the three years. Differences are consequence of different PDF and UE tunes.

qqZH hadronic bins	2016	2017	2018
0-jet	4.37%	5.62%	4.39%
1-jet	29.07%	31.71%	29.08%
$\geq$2-jet $m_{jj} \leq 60$ GeV	2.75%	3.20%	2.76%
$\geq$2-jet 60 $< m_{jj} <$120 GeV	34.39%	30.29%	34.40%
$\geq$2-jet 120 $< m_{jj} <$350 GeV	13.88%	13.09%	13.89%
$\geq$2-jet $m_{jj} >$350 GeV, $p_{\mathrm{T}}^{\mathrm{H}} \geq 200$ GeV	0.77%	0.79%	0.77%
$\geq$2-jet 350 $< m_{jj} <$700, $p_{\mathrm{T}Hjj}^{\mathrm{T}} \leq 25$ GeV	0.37%	0.38%	0.37%
$\geq$2-jet 350 $< m_{jj} <$700, $p_{\mathrm{T}Hjj}^{\mathrm{T}} >$25 GeV	2.08%	1.95%	2.07%
$\geq$2-jet $m_{jj} >$700 GeV, $p_{\mathrm{T}Hjj}^{\mathrm{T}} \leq 25$ GeV	0.06%	0.06%	0.06%
$\geq$2-jet $m_{jj} >$700 GeV, $p_{\mathrm{T}Hjj}^{\mathrm{T}} >$25 GeV	0.50%	0.46%	0.49%
Higgs $\lvert\eta\rvert >$2.5	11.76%	12.44%	11.73%
ggZH leptonic bins	2016	2017	2018
$p_{\mathrm{T}}^{\mathrm{V}} \leq 75$ GeV	15.93%	16.04%	15.88%
75 $< p_{\mathrm{T}}^{\mathrm{V}} <$150 GeV	42.98%	44.26%	43.47%
150 $< p_{\mathrm{T}}^{\mathrm{V}} <$250 GeV 0-jet	7.72%	9.43%	9.09%
150 $< p_{\mathrm{T}}^{\mathrm{V}} <$250 GeV $\geq$1-jets	21.70%	20.35%	20.51%
$p_{\mathrm{T}}^{\mathrm{V}} >$250 GeV	9.08%	7.07%	8.35%
Higgs $\lvert\eta\rvert >$2.5	2.59%	2.86%	2.70%
qqZH leptonic bins	2016	2017	2018
$p_{\mathrm{T}}^{\mathrm{V}} \leq 75$ GeV	45.46%	45.69%	45.81%
75 $< p_{\mathrm{T}}^{\mathrm{V}} <$150 GeV	29.90%	30.20%	30.16%
150 $< p_{\mathrm{T}}^{\mathrm{V}} <$250 GeV 0-jet	5.47%	5.37%	5.37%
150 $< p_{\mathrm{T}}^{\mathrm{V}} <$250 GeV $\geq$1-jets	4.18%	4.13%	4.12%
$p_{\mathrm{T}}^{\mathrm{V}} >$250 GeV	9.08%	7.07%	8.35%
Higgs $\lvert\eta\rvert >$2.5	11.92%	11.75%	11.69%

5.3.2 Event selection and categorisation in ggH STXS

The same event selection criteria of the ggH inclusive analysis described in section 5.2.1 is used in the STXS analysis. However, we do not follow the same categorisation as the inclusive analysis, since in STXS the phase space is divided in bins, which already limits the amount of data in each of them. Therefore, there is no eight-category splitting in flavour, charge and subleading lepton p_{T} to avoid large statistical uncertainties.

The STXS definition is done in terms of generator-level variables, but the definition of reconstructed categories (reconstructed bins) depends on the reconstructed-level variables. There are migrations between adjacent bins because of the limited resolution of the reconstructed $p_{\mathrm{T}}^{\mathrm{H}}$, and the bin boundaries at reconstructed level can be optimised to reduce those migration effects. Figure 5.20 shows the reconstructed-level ggH STXS bins and from which generator-level bin the events originate, assuming a reconstructed binning scheme identical to the generator-level one but using the reconstructed-level variables. Each column is normalised to one, meaning that all the generator-level contributions to a given reconstructed-level bin add up to one. The off-diagonal contributions show that there are migrations between the reconstructed bins. This generally happens between bins that share a boundary, since reconstructed-level quantities may differ from the generator-level ones and the bins are defined according to them.

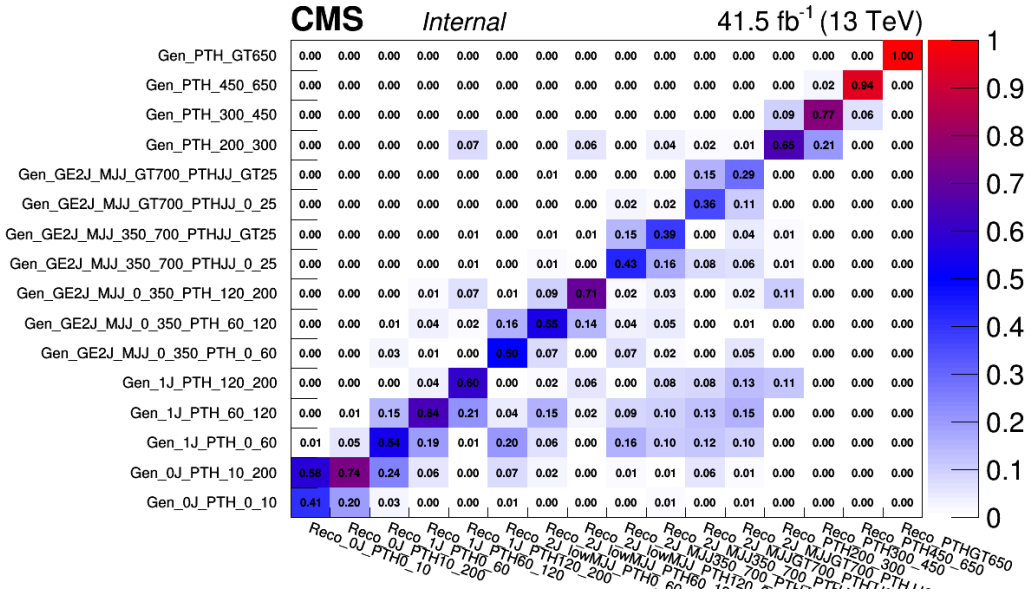


Figure 5.20: Migration matrix for 2017 using a binning scheme at reconstructed level equal to the generator-level one. On the vertical axis generator-level bins are displayed and on the horizontal axis their reconstructed-level counterparts. Each column is normalised to one.

We can also observe the bin migrations in Fig. 5.21, where the generator-level $p_{\mathrm{T}}^{\mathrm{H}}$ distributions are shown for the events corresponding to the different 0-, 1-, and $\geq$2-jet with m_{jj}<350 GeV reconstructed bins. The boundary at 10 GeV for 0-jet and the boundaries at 60 and 120 GeV for 1- and $\geq$2-jet bins, drawn in the figure, do not completely separate the reconstructed bin distributions. In particular, the 0-jet bins have a large overlap and that is why 64% of generator-level 0-jet $p_{\mathrm{T}}^{\mathrm{H}} < 10$ GeV are reconstructed in the 0-jet $10<p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin.

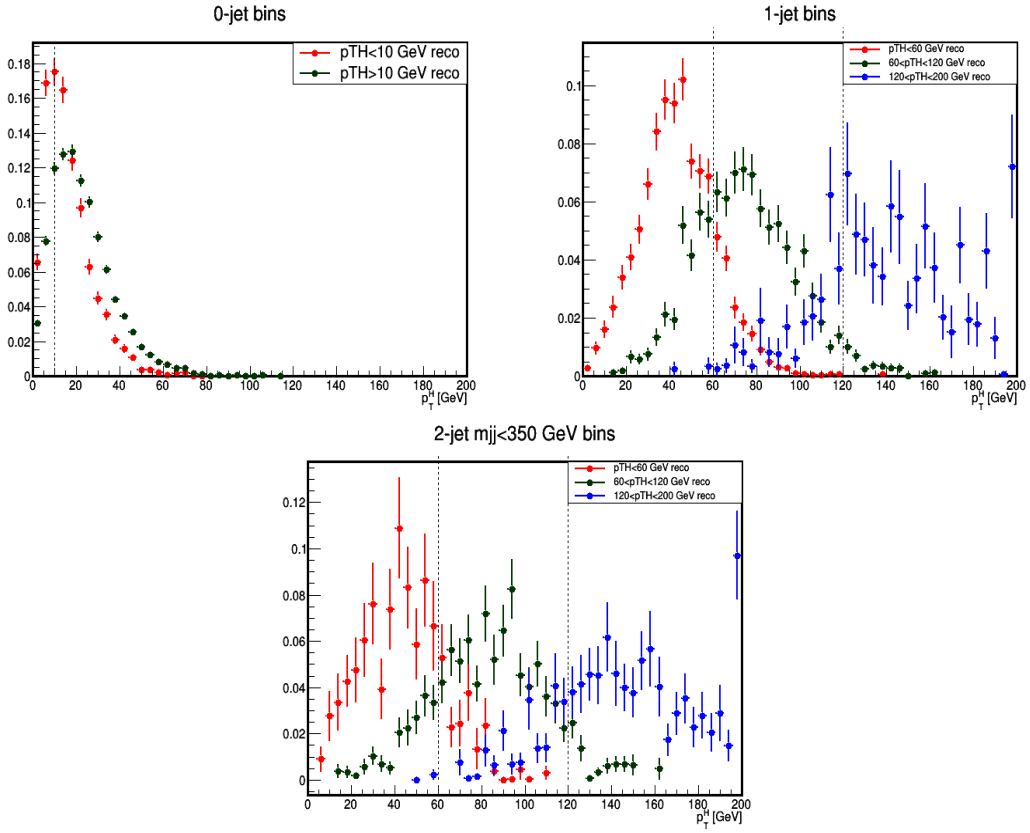


Figure 5.21: Generator-level $p_{\mathrm{T}}^{\mathrm{H}}$ distributions in the 2017 reconstructed-level bins.

A change in the 0-jet boundary at $p_{\mathrm{T}}^{\mathrm{H}}$ =10 GeV was attempted to reduce the migration between 0-jet reconstructed bins, but no significant decrease in the migration matrix nondiagonal elements was observed when changing that boundary from 10 to 20 GeV. We have also compared the uncertainty in the signal strength determined for the two 0-jet bins for the two different choices. By changing the reconstructed-level 0-jet bin definitions, the expected signal strength modifier uncertainty of those two bins obtained by the maximum likelihood fit using data Asimov increased by a few percent. Therefore, shifting the boundary of the reconstructed bins did not improve the measurement, and to avoid a large correlation between those bins a single signal strength modifier is obtained in the analysis for the two 0-jet bins merged together.

5.3.3 Uncertainties in ggH STXS analysis

The sources of uncertainty in this analysis are essentially the same ones described in section 4.4, with a few differences:

- The gluon fusion theoretical uncertainty is consists of nine parameters that reflect not just the overall scale, but also migrations between jet-multiplicity bins, different Higgs boson p_{T} bin migrations and an uncertainty in the $p_{\mathrm{T}}^{\mathrm{H}}$ distribution associated to the missing higher order finite top quark mass corrections. It is specially suited for the STXS analysis, since the migration uncertainties are irrelevant for measurements performed inside a certain STXS bin. However, in order to make sure that these uncertainties do not change the total cross section by changing the amount of events in each STXS bin, there is a renormalisation factor calculated for each bin, which keeps the total number of events constant. Since the STXS bins are independently renormalised, this has an effect on the theoretical uncertainties of the ggH signal. Those that were normalisation-only can effectively be removed, only keeping the uncertainties that affect the shape of the discriminant variable distributions. Theoretical uncertainties such as the PDF normalisation are therefore excluded, reducing the theory dependence.

- There is an additional QCD scale acceptance uncertainty due to the measurement in the STXS bins. This shape uncertainty is evaluated in each STXS bin by varying the renormalisation and factorisation QCD scales by a factor 2 up and down, and renormalising the event yield such that these uncertainties affect only the shape in each STXS bin, but not the total number of events, keeping the cross section unchanged. These uncertainties are uncorrelated among different years, but correlated between adjacent bins.

5.3.4 STXS expected signal strength uncertainties

For each of the STXS bins we use the distributions of the discriminant variables $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ as in the inclusive analysis. Figure 5.22 shows as an example the distributions for two most populated bins, 0-jet and 1-jet with $p_{\mathrm{T}}^{\mathrm{H}}$ <60 GeV, using the 2018 dataset. The binning used for the discriminant variables is the same as the one used for the $p_{\mathrm{T}2} < 20$ GeV categories in the inclusive analysis, since some of the STXS bins are also very statistically limited. The division in $m_{\ell\ell}$ corresponds to 6 bins [12, 25, 40, 50, 70, 90, 210] (in GeV) and the division in $m_{\mathrm{T}}^{\mathrm{H}}$ corresponds to 6 bins [60, 80, 90, 110, 130, 150, 200] (in GeV).

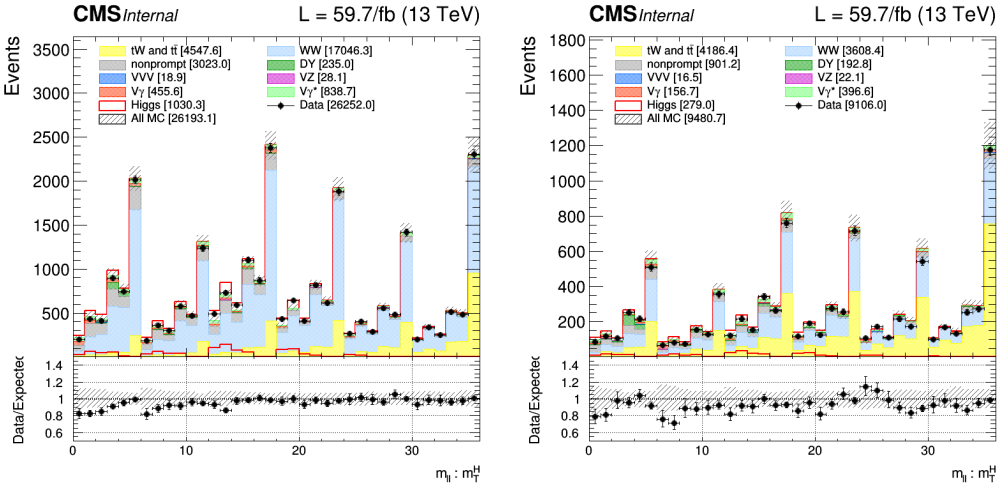


Figure 5.22: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 0-jet (left) and low $p_{\mathrm{T}}^{\mathrm{H}}$ 1-jet (right) ggH STXS bins for 2018. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.

Using the Asimov dataset, the expected signal strength uncertainties are measured in the ggH STXS bins. They are obtained with a simultaneous fit in each of the STXS bins, using a signal strength modifier for each bin. In this analysis only the ggH H $\rightarrow$ WW samples are considered as signal, and therefore the measured signal strengths will correspond to cross sections of that same process in different kinematic regions. This is different to the inclusive analysis described in the previous section, where a global signal strength was obtained using other Higgs boson production modes and decays. The result is presented in Fig. 5.23 for all three years and their combination. Some bins from the original scheme that had uncertainties over 200% are merged together, such as the $p_{\mathrm{T}}^{\mathrm{H}}$ >300 GeV bins and the $\geq$2-jet high m_{jj} bins that were separated by a $p_{\mathrm{Hjj}}^{\mathrm{T}}$ threshold. With the individual STXS bins a combination of all of the bins is added as the first entry in each figure.

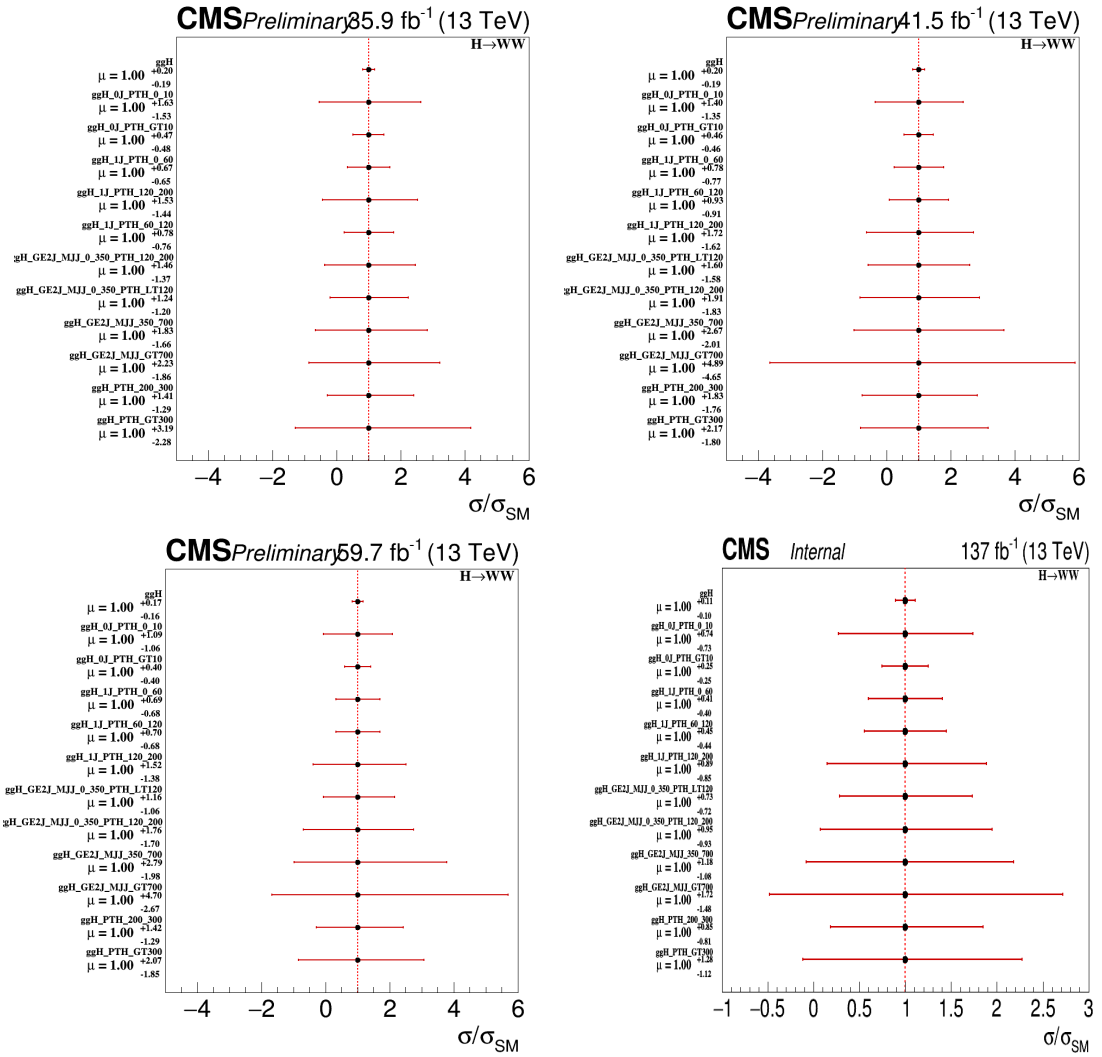


Figure 5.23: Expected signal strength for the gluon fusion channel stage 1.2 STXS bins for 2016 (top left), 2017 (top right), 2018 (bottom left) and the combination (bottom right). Some of the original bins that have low sensitivity have been merged.

The combined result has uncertainties smaller than 100% in most cases, but in the ≥2-jet bins with m_{jj} >350 GeV and in the $p_{\mathrm{T}}^{\mathrm{H}}$ >200 GeV bins it seems some additional merging is necessary to have a more precise result. Therefore, the two ≥2-jet m_{jj}> 350 GeV bins are merged into one, and the two high $p_{\mathrm{T}}^{\mathrm{H}}$ bins are merged into one as well. Additionally, even though the 0-jet bins have low expected uncertainties, because of the large correlations between them caused by reconstruction resolution effects they are also be merged into a single bin.

5.3.5 Production modes in the 2-jet STXS bins

The 0- and 1-jet ggH STXS bin definitions does not overlap with the other ggH and the VBF and VH STXS bin definitions. This is however not the case with the ≥2-jet bins, since ggH and qqH share the definition of some bins and in the m_{jj} <350 GeV bins the two production modes contribute. This poses the question of whether the separation between them is actually possible. If there is an orthogonal definition for the production modes, cross sections can be obtained by them independently, otherwise only the combination of ggH and qqH may be obtained. This was the approach followed in the H → ττ analysis [17].

The origin of the signal in the reconstructed bins in terms of production modes is shown in Fig. 5.24. Reconstructed bins using a ggH selection have some contribution from VBF generator-level bins, specially for m_{jj}> 700 GeV bins where ggH generator-level bins contribute to less than 40% of the total events. VBF reconstructed bins have some contribution from ggH, which in the low dijet mass bin corresponds to more than 50% of the events. The single VH hadronic reconstructed bin has very little contribution from VH generator-level bins (about 20%), being dominated by ggH. Mixing from VH in ggH and VBF only happens in the low dijet mass reconstructed bins and it contributes to less than 10% of the total events.

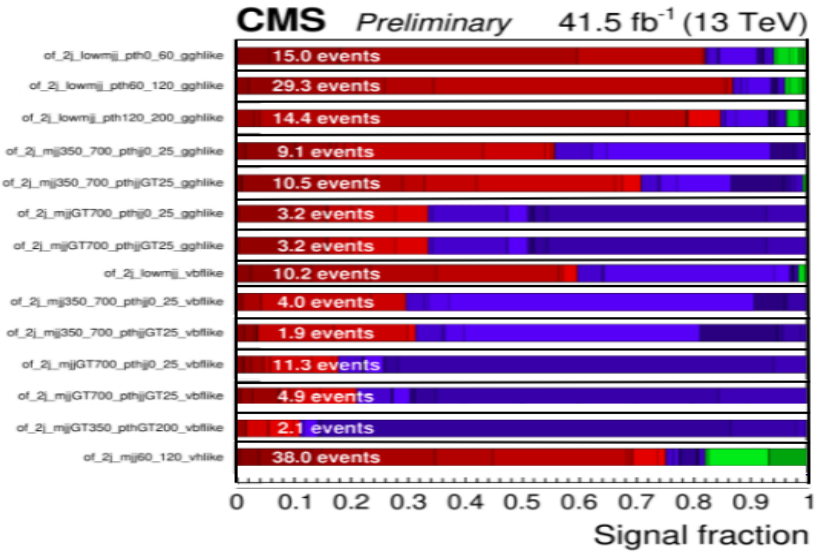


Figure 5.24: Signal fraction of ≥2-jet STXS reconstructed bins in 2017. In red, blue and green, ggH, VBF and hadronic VH bins at generator level are shown, respectively.

The correlations between ggH and VBF signal strengths translate into large uncertainties when combining the results of the two production modes. The signal strength modifier measured in ggH is the parameter with the highest impact in the VBF signal strength measurement and vice versa. To overcome this, a multiclass Deep Neural Network (DNN) is employed, which is trained with events from top quark and WW backgrounds, as well as VBF and ggH signals.

It uses a total of 27 input variables, corresponding to kinematic variables of leptons, jets and E_{T}^{miss}, as well as quark/gluon jet discriminators. The outputs of the DNN are four class discriminators, one for each type of process, where an event with the ggH discriminator larger than the other three discriminators would be "ggH-like". These four class discriminators are a better choice for the multidimensional fit to obtain the signal strength in ≥2-jet STXS bins than $m_{\ell\ell}$ and $m_{\mathrm{T}}^{\mathrm{H}}$ alone.

5.3.6 STXS signal strength results for 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins

At the time of writing this thesis, the analysis of the H → WW channel in other production modes is still in process, and therefore only results with 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins are presented, since the analysis of the ≥2-jet bins need to consider also the other production modes. We have determined the STXS signal strength modifiers in five STXS bins, one bin for 0-jets, three 1-jet bins and one bin for $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV. The results are summarised in Table 5.14.

Table 5.14: Measured signal strength modifiers in STXS bins.

ggH STXS bins	2016	2017	2018	All datasets
0-jet	$0.89^{+0.20}_{-0.20}$	$1.01^{+0.21}_{-0.26}$	$0.94^{+0.19}_{-0.18}$	$0.96^{+0.12}_{-0.12}$
1-jet < 60 GeV	$2.0^{+0.8}_{-0.7}$	$1.1^{+0.7}_{-0.7}$	$0.9^{+0.6}_{-0.6}$	$1.3^{+0.4}_{-0.4}$
1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV	$1.8^{+1.0}_{-0.9}$	$0.9^{+0.7}_{-0.7}$	$0.7^{+0.6}_{-0.6}$	$0.9^{+0.4}_{-0.4}$
1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV	$1.4^{+1.6}_{-1.4}$	$1.9^{+2.0}_{-2.3}$	$-0.3^{+1.1}_{-1.1}$	$0.7^{+0.8}_{-0.8}$
$p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV	$0.7^{+1.8}_{-1.7}$	$-0.8^{+2.1}_{-1.9}$	$0.6^{+1.2}_{-1.2}$	$0.8^{+0.9}_{-0.9}$
Combination	$1.08^{+0.18}_{-0.18}$	$1.00^{+0.18}_{-0.23}$	$0.88^{+0.16}_{-0.16}$	$0.99^{+0.11}_{-0.10}$

The measured signal strengths measured in the 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins for the three years are presented in Figs. 5.25 to 5.27. In all of the years the combination obtained with the five bins is included. All signal strength modifiers are consistent with 1 within uncertainties, but there are bins that are dominated by large statistical uncertainties. In 2017 for the $p_{\mathrm{T}}^{\mathrm{H}}$ >300 GeV bin and in 2018 for the 1-jet 120<$p_{\mathrm{T}}^{\mathrm{H}}$ <200 GeV bin the signal strength modifier has a negative value, because the observed data in those bins was less than the expected background as a consequence of statistical fluctuations in those low-populated bins. The result obtained with the combination of the STXS bins is also in good agreement with the SM prediction for all the years.

The three-year combination of the ggH signal strength modifier measured in the 5 STXS bins is shown in Fig. 5.28, as well as the result obtained using a single signal strength for all the bins. The data shows a good agreement with the SM predictions in all bins and the combination, and there is low correlation/anticorrelation between the results of the signal strength modifiers using these merged bins, as shown in Fig. 5.29. There is an uncertainty of the order of 10% for the 0-jet bin, a 40% uncertainty in the 1-jet bins with lower $p_{\mathrm{T}}^{\mathrm{H}}$, and an 80% and 90% uncertainty in the less populated 1-jet high $p_{\mathrm{T}}^{\mathrm{H}}$ and $p_{\mathrm{T}}^{\mathrm{H}}$ >200 GeV bins,respectively. The uncertainty in the combined result of the signal strength can be separated into theoretical systematic, experimental systematic and statistical uncertainties:

$$\mu_{ggH}^{STXS} = 0.99^{+0.11}_{-0.10} = 0.99^{+0.04}_{-0.04}\ (\text{theo.})^{+0.07}_{-0.07}\ (\text{exp.})^{+0.07}_{-0.07}\ (\text{stat.}) \tag{5.5}$$

To estimate those uncertainties, nuisance parameters of a group of uncertainties are fitted, whereas the remaining uncertainties are fixed to their value obtained by the maximum likelihood fit. The combined result is more limited by the systematic uncertainties, especially the experimental ones. There is a slight reduction in the theoretical uncertainty with respect to the inclusive analysis.

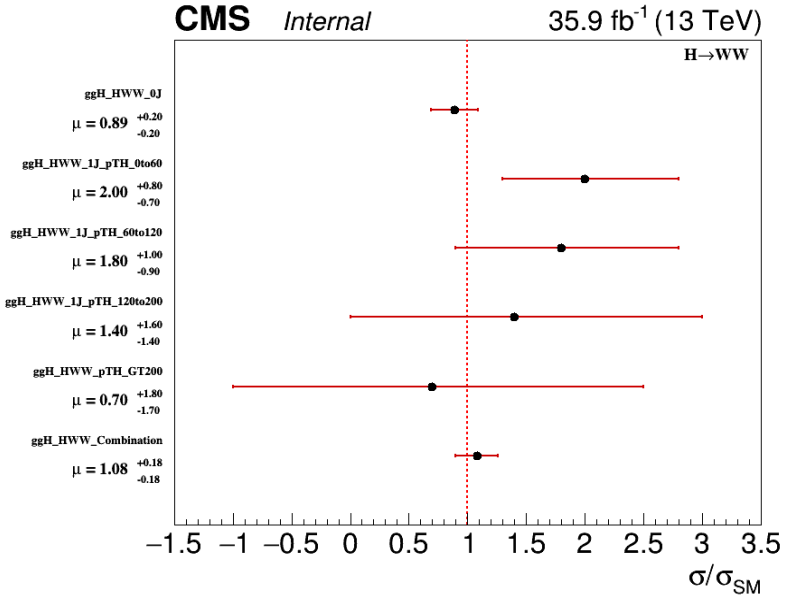


Figure 5.25: Measured signal strengths for the 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins, as well as their combination, obtained with the 2016 dataset.

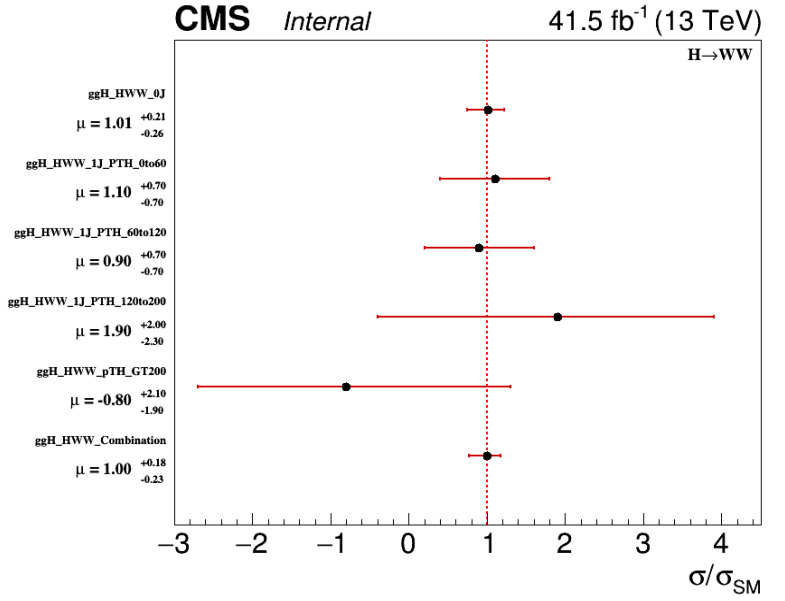


Figure 5.26: Measured signal strengths for the 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins, as well as their combination, obtained with the 2017 dataset.

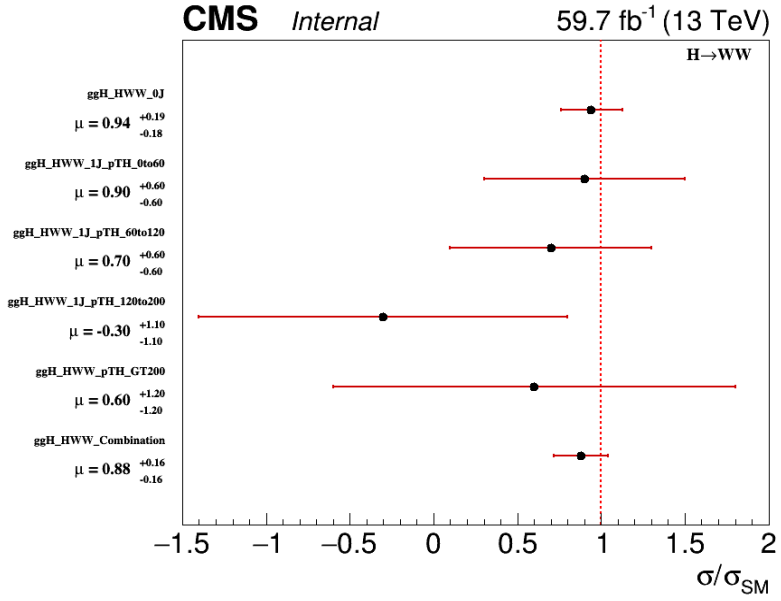


Figure 5.27: Measured signal strengths for the 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins, as well as their combination, obtained with the 2018 dataset.

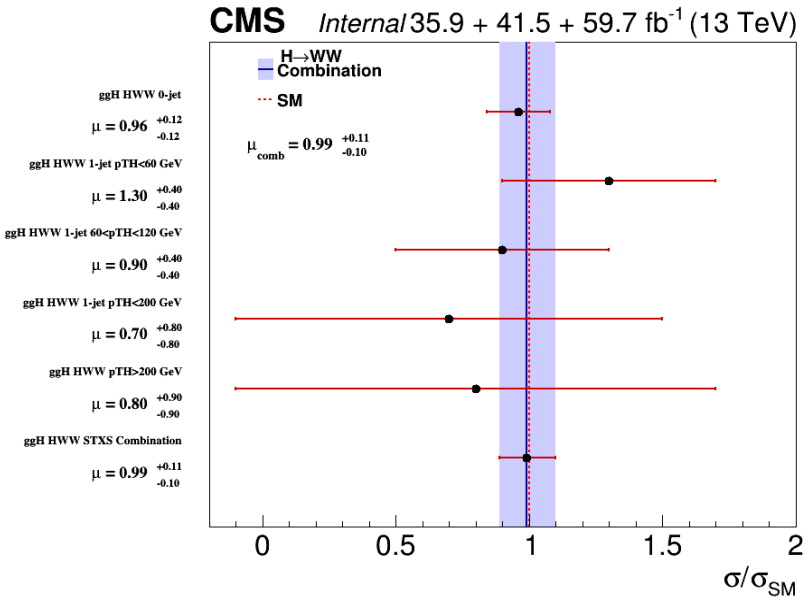


Figure 5.28: Measured signal strengths for the 0-jet, 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ bins, as well as their combination, obtained with the combination of the three datasets.

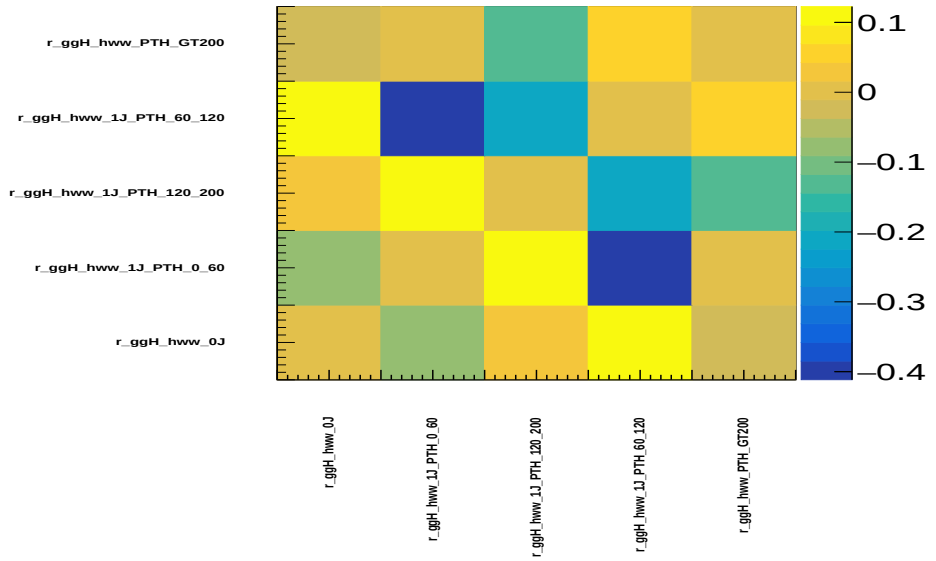


Figure 5.29: Correlation between the signal strength parameters of the 0-jet, 1-jet and $p_{\mathrm{T}}^{\mathrm{H}}$ >200 GeV bins, obtained using the combination of the three datasets.

To evaluate which uncertainties most affect the result of the signal strengths of each STXS bin, impact plots are produced for each of the three years and their combination. The thirty uncertainties that have the largest impacts in 2018 in the 0- and 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ STXS bins are shown in Fig. 5.30. The uncertainties are classified according to their impact, with the highest ranked ones having the greatest effect in the measured signal strength.

One advantage of the STXS analysis is that some of the theoretical uncertainties that have a high impact in the inclusive analysis are not as meaningful, but there is a trade-off with the statistical uncertainties. Compared to the categories in the inclusive analysis, the STXS bins have much fewer events in some cases, since most of them are in the 0-jet or 1-jet with $p_{\mathrm{T}}^{\mathrm{H}}$ <60 GeV bins (58% as was noted in Table 5.10). Consequently, statistical uncertainties appear in the highest ranked impacts for the low-populated STXS bins, such as $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV, even after merging the four high $p_{\mathrm{T}}^{\mathrm{H}}$ bins into a single one.

The systematic uncertainties that affect the signal strength modifier in each of the STXS bins change a lot depending on the bin. The 0-jet bin is similar to the 0-jet category in its uncertainties, with the normalisation of the top quark, DY and WW backgrounds and the electron efficiency being the most important. Jet energy scale uncertainties and the nonprompt background systematic nuisance appear next, followed by some statistical uncertainties. In this bin there are no large pulls, and the maximum value for the 0-jet bin impacts is 7%.

In the 1-jet bins the biggest impacts are higher than in the 0-jet bin and higher for the less populated bins. E_{T}^{miss} scale and b-tagging are important in the 1-jet bins, with the E_{T}^{miss} scale nuisance parameter having a large pull in the three bins. The WW, DY and top quark background normalisation is more important for the more populated low $p_{\mathrm{T}}^{\mathrm{H}}$ bin, whereas there statistical uncertainties play a more important role in the others. Top p_{T} reweighting is the second highest ranked impact in the intermediate $p_{\mathrm{T}}^{\mathrm{H}}$ bin, similar to the impacts in the inclusive analysis, but in this case it is quite pulled as well.

In the high $p_{\mathrm{T}}^{\mathrm{H}}$ STXS bin, the uncertainty from the top quark and DY background normalisation, the jet energy scale and the underlying event uncertainties have the highest impacts in the signal strength, but statistical uncertainties are very significant in this bin. In this case it is the UE impact the one with a large pull, with over 2 standard deviations.

Finally, the combination of the 5 bins has a list of main impacts similar to the dominant 0-jet bin, with background normalisation, electron efficiency, nonprompt background systematic and jet energy scale being the most relevant uncertainties. The impacts are smaller than 5% in most cases and the largest pulls correspond to top p_{T} reweighting and E_{T}^{miss} scale uncertainties of the 1-jet bins.

In appendix C the 2016 and 2017 impact plots are shown, with a similar discussion. The uncertainties that have the biggest impacts in the signal strength obtained with the combination of the three datasets are shown in Fig. 5.31. Most of the impacts are reduced considerably in the combination and generally the main uncertainties are correlated among the three years. The main uncertainty sources affecting the signal strength in the 0-jet bin are the background normalisations, the electron efficiencies for the three years and the PS model for FSR. There is a large pull for the nonprompt background systematic uncertainty, which can also be seen in the low $p_{\mathrm{T}}^{\mathrm{H}}$ 1-jet bin. Aside from that, there are two very constrained statistical uncertainties, of a 1-jet high $p_{\mathrm{T}}^{\mathrm{H}}$ bin in 2017 and a very high $p_{\mathrm{T}}^{\mathrm{H}}$ bin in 2018, resulting from low number of events in those bins.

Top p_{T} reweighting, b-tagging, background normalisations and E_{T}^{miss} scale are the most impactful systematic uncertainties in the 1-jet bins. The E_{T}^{miss} scale has large pulls in the three years in several bins and some constrained statistical uncertainties from individual years also appear in the combination.

The high $p_{\mathrm{T}}^{\mathrm{H}}$ bin is dominated by statistical uncertainties, and the systematic uncertainties with the largest impacts are top quark and WW background normalisation, b-tagging, jet energy scale and the STXS acceptance, that has a one-sided impact in 2016. The biggest impacts in the combined result of the five bins looks almost like the impacts in the 0-jet combination, since it is by far the most populated one.

Signal strengths modifiers can be translated into cross section values using the SM cross section and the event fractions computed in 5.10. Table 5.15 summarises the cross sections obtained in the STXS bins, with the SM cross sections computed using the event fractions and the SM gg→H → WW → $\ell\nu\ell\nu$ cross section value. The ggH production cross section equals to σ_{ggH}=$48.5^{+4.6}_{-6.7}$ pb [96], the branching ratio $\mathcal{B}$(H → WW)=0.2152±0.0033 and $\mathcal{B}$(WW→ $\ell\nu\ell\nu$)=9× $(0.108 \pm 0.009)^2$, thus the total cross section prediction for the SM equals to σ_{SM}=$1.09^{+0.11}_{-0.15}$ pb. The theoretical uncertainty has already been considered in the measurement.

Table 5.15: SM and measured cross sections in STXS bins.

ggH STXS bins	Event fraction	SM σ	Observed μ	Measured σ
0-jet	57.61% ± 0.05%	631.51 ± 0.60 fb	0.96 ± 0.12	606 ± 76 fb
1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV	14.1% ± 1.1%	154 ± 12 fb	1.3 ± 0.4	202 ± 64 fb
1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV	8.15% ± 0.08%	89.38 ± 0.91 fb	0.9 ± 0.4	80 ± 36 fb
1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV	1.43% ± 0.09%	15.63 ± 0.96 fb	0.7 ± 0.8	11 ± 13 fb
$p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV	1.38% ± 0.13%	15.1 ± 1.4 fb	0.8 ± 0.9	12 ± 14 fb

This signal strength modifier obtained with the combination of the five STXS bins corresponds to a gg→H → WW cross section of

$$\sigma_{ggH}^{STXS} = \mu_{ggH}^{STXS} \times \sigma_{SM} = 1.08^{+0.12}_{-0.11} \text{ pb} . \tag{5.6}$$

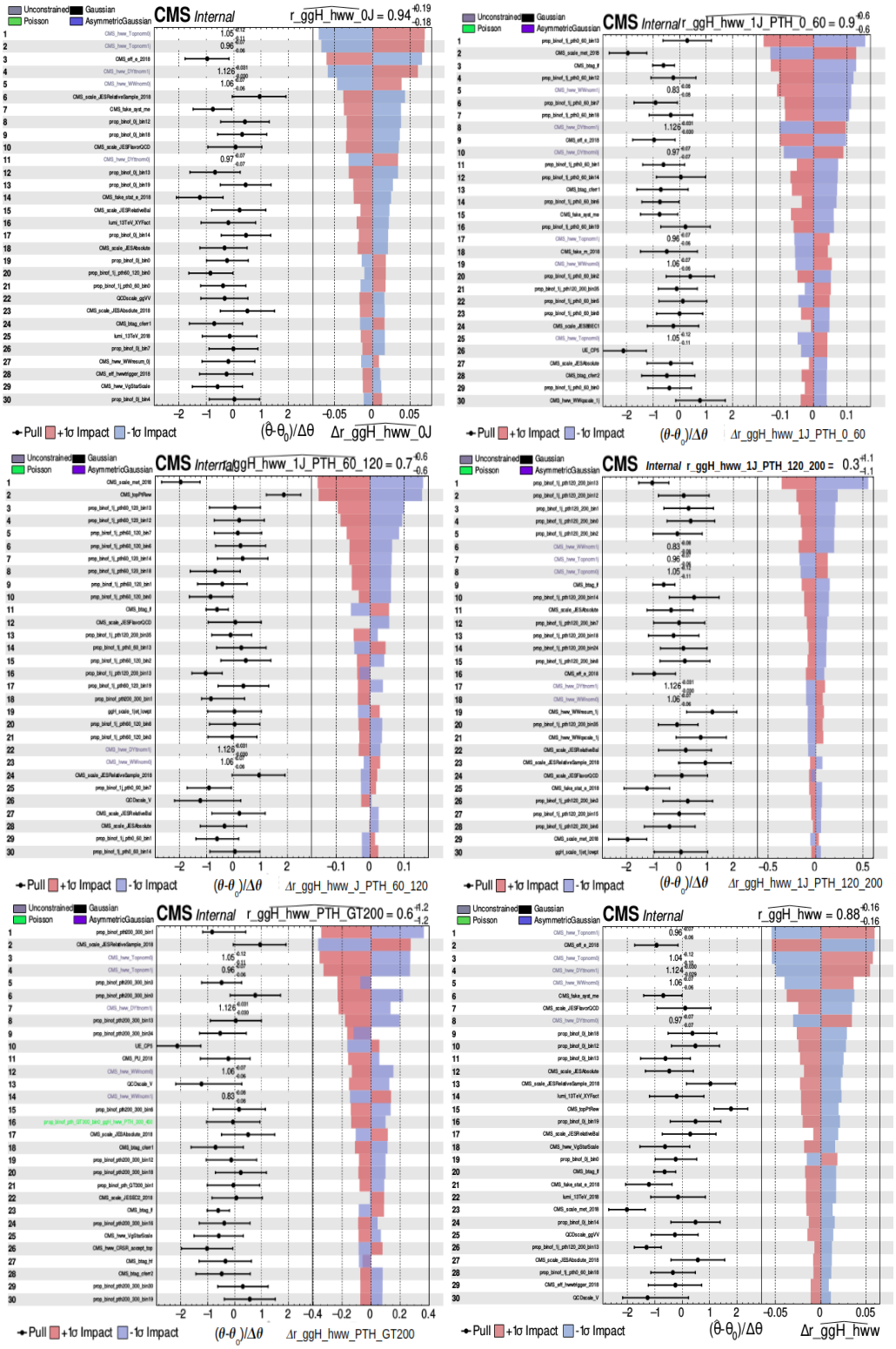


Figure 5.30: Impact of the first 30 nuisance parameters with highest ranking on the measured signal strengths for 0-jet (top left), 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV (top right), 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV (middle left), 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV (middle right) and $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV (bottom left) ggH STXS bins, obtained with the 2018 dataset. The combination of the five bins is shown as well (bottom right).

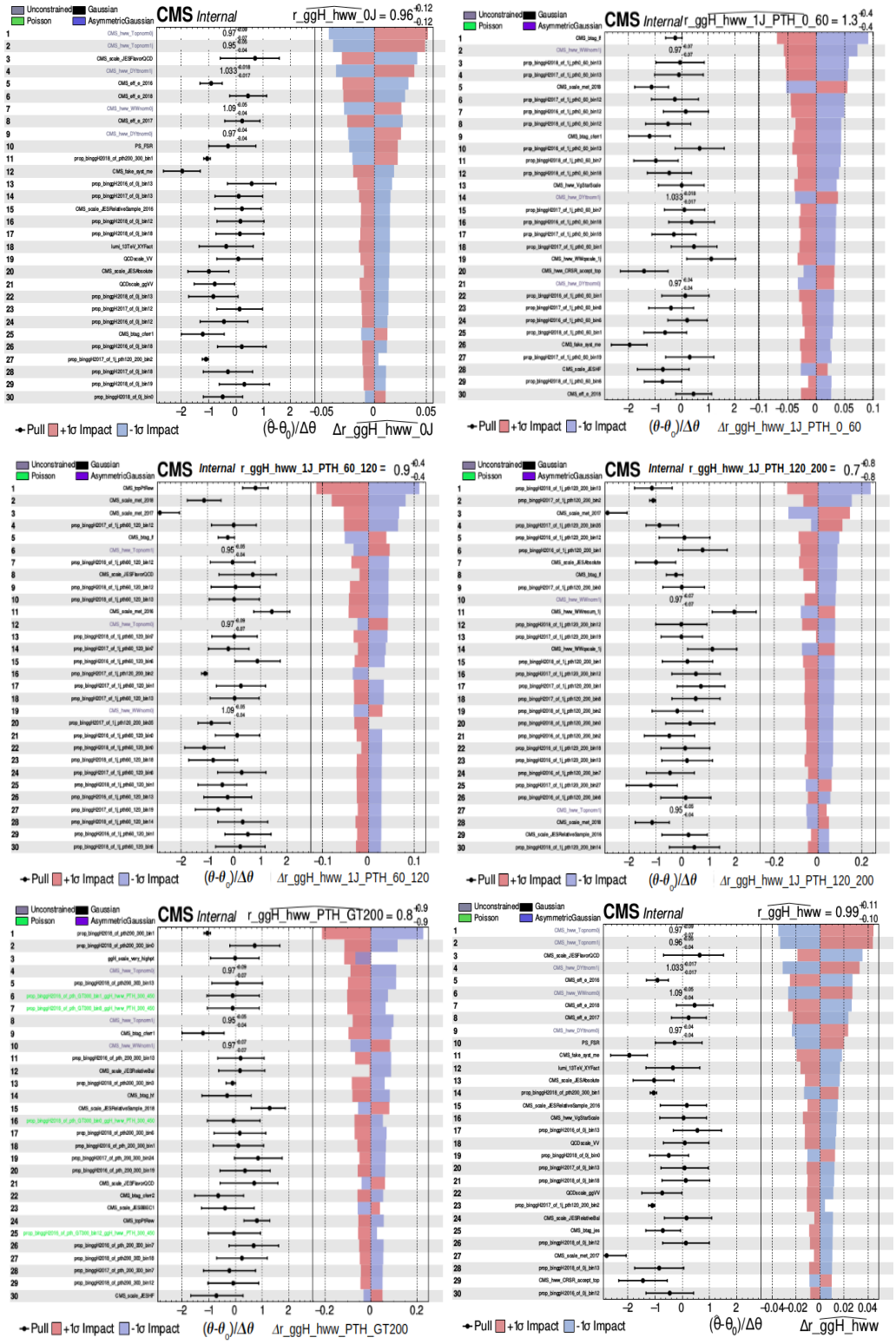


Figure 5.31: Impact of the first 30 nuisance parameters with highest ranking on the measured signal strengths for 0-jet (top left), 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV(top right), 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV(middle left), 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV(middle right) and high $p_{\mathrm{T}}^{\mathrm{H}}$ (bottom left) ggH STXS bins, obtained with the combination of the three datasets. The combination of the five bins is shown as well (bottom right).

Chapter 6

Summary and conclusion

This thesis presents measurements on the production of the SM Higgs boson in proton-proton collisions at a centre-of-mass energy of 13 TeV. It analyses the data sample collected by the CMS experiment at the LHC at CERN during the years 2016, 2017 and 2018, with an equivalent integrated luminosity of 137 fb^{-1}.

Within the SM, the Higgs boson decays into a pair of W bosons with one of the highest branching fractions, $\mathcal{B}(\text{H} \to \text{WW})$=0.2152±0.0033, at a mass of m_{H}=125.09 GeV. In addition, the leptonic decay of the W boson into an electron or a muon and a neutrino offers a relatively clean final state with a good signal to noise ratio. The CMS collaboration is currently involved in an in-depth analysis of the Higgs boson properties in the Higgs boson decay channel into a pair of W bosons, with the subsequent leptonic decay of the W bosons. The topology associated to all the main Higgs boson production mechanisms are explored: gluon fusion, vector boson fusion and vector boson associated production.

This thesis focuses on the study of the Higgs production properties in the WW decay channel with the subsequent decay of a W boson into an electron and a neutrino and the decay of the other W boson into a muon and a neutrino. It exploits the features from the gluon fusion mechanism, which has the largest cross section at LHC energies, and which is characterised by the absence of high energy jets. The final state analysed is thus characterised by a pair of opposite-charge leptons of different flavour (electron-muon), a significant imbalance in the reconstructed momentum in the transverse plane and few high energy jets.

The Higgs boson production rate is quantified in terms of a signal strength modifier, defined as the ratio of the observed cross section over the prediction from the SM. The measurement is performed inclusively and in exclusive phase space regions called Simplified Template Cross Section (STXS) bins. The so-called STXS framework aims to deliver measurements in finely-grained regions according to event kinematics and minimising theory dependencies. It also allows to isolate phase space regions where BSM signals may arise. The definition of STXS bins only depends on the production mode, being common to all Higgs boson decay modes, thus allowing combinations between measurements performed in different channels. The STXS framework has been developed in the context of the LHC Higgs Working group and has been adopted by the CMS and ATLAS collaborations with the final goal of providing a common LHC combination. The granularity of the STXS bins has been increased as the collected luminosity allowed for a further division of the available data samples. The most recent binning scheme proposed, the so-called stage 1.2 STXS definition, has been tested extensively in this work and applied for the first time to the H → WW analysis.

Starting from the analysis methodology followed in CMS with the analysis of the data taken at $\sqrt{s}$ = 7 and 8 TeV, the full analysis strategy has been revised in view of the upgrades performed in the CMS detector and the changes in the operating conditions of the LHC, which includes a higher pileup because of the higher instantaneous luminosity in the Run 2 period.

The signal strength modifier in both the inclusive and the STXS analysis are obtained from a maximum likelihood fit to the two dimensional distribution of the dilepton invariant mass and the transverse mass defined using the transverse momentum of the dilepton system and the missing transverse momentum. Control regions are defined to validate Monte Carlo predictions and to constrain the most important background contributions, those arising from DY and top quark production.

Instrumental background contribution from jets misidentified as leptons are not reliably reproduced by the simulation and have to be estimated with data control samples. Part of this thesis work has been dedicated to more precisely estimate the contribution from the misidentified leptons. Several studies were carried out in order to better understand this background, estimate its contribution to the signal region and to assess its uncertainty. From the results of these studies the lepton definition had to be optimised.

For the inclusive analysis, a global signal strength modifier is obtained using signal samples of different production modes as well as WW and $\tau\tau$ decay modes, but targeting the gluon fusion production mode and the WW decay. A categorisation of the events in the signal region is performed depending on the number of jets, the flavour and charge of the leading lepton, and the transverse momentum of the subleading lepton. Each of these categories has a different signal to noise ratio as they are differently affected by the several backgrounds, and using this categorisation results in an optimised overall sensitivity to the Higgs boson signal.

The main division depends on the number of jets in the final state. We split the sample into a 0-jet category (no high p_{T} jets in the final state), a 1-jet category (1 high p_{T} jet in the final state) and $\geq$2-jet category (events with two or more high p_{T} jets in the final state), where the 0-jet category carries most of the sensitivity of the analysis. Due to the overlap between the gluon fusion and the vector boson fusion production modes in the $\geq$2-jet STXS bins, no measurement is delivered for that region in this thesis, although the expected performance of this category is included in the studies. The following inclusive signal strength modifiers are obtained:

- 2016 0-jet category: $\mu = 0.88^{+0.19}_{-0.17}$
- 2016 1-jet category: $\mu = 1.36^{+0.31}_{-0.28}$
- 2017 0-jet category: $\mu = 0.85^{+0.17}_{-0.17}$
- 2017 1-jet category: $\mu = 1.09^{+0.15}_{-0.23}$
- 2018 0-jet category: $\mu = 0.82^{+0.15}_{-0.14}$
- 2018 1-jet category: $\mu = 0.72^{+0.20}_{-0.18}$

The results from each year are combined into:

- 2016 combination: $\mu = 1.01^{+0.17}_{-0.16}$
- 2017 combination: $\mu = 0.93^{+0.16}_{-0.16}$
- 2018 combination: $\mu = 0.79^{+0.13}_{-0.13}$

The results for each category from the three years are combined into:

- 0-jet category combination: $\mu = 0.86^{+0.11}_{-0.10}$
- 1-jet category combination: $\mu = 0.99^{+0.17}_{-0.15}$
- Full combination: $\mu = 0.90^{+0.10}_{-0.09}$

All the values of the measured signal strength modifiers are consistent with 1, meaning that the data is in agreement with the predictions from the SM. Comparing the results obtained with 2016 data to the previous CMS $H \to WW$ result with that dataset, there is a difference coming from the 0-jet category, the one that drives the sensitivity of the analysis. This can be explained by the changes in the analysis since then, such as redefinitions of the physics object identification criteria, another b-tagging algorithm being employed, slightly different samples being used and many other more subtle changes. The uncertainties have been reduced compared to the previous results.

The combination with the data from the three years for the global signal strength, $\mu_{global} = 0.90^{+0.10}_{-0.09}$ measures the signal strength modifier with a 10% uncertainty. It can be decomposed in its theoretical, experimental and statistical sources in this way:

$$\mu_{global} = 0.90^{+0.04}_{-0.05}\ (\text{stat.})^{+0.06}_{-0.05}\ (\text{exp.})^{+0.07}_{-0.06}\ (\text{theo.})\ . \tag{6.1}$$

Theoretical uncertainties have a larger contribution than experimental uncertainties, and the statistical uncertainties are the ones that contribute the least. The most important sources of systematic uncertainty in these measurements are the top quark normalisation and the ggH theoretical uncertainties, followed by the electron efficiency and DY normalisation in the 0-jet channel and top p_{T} reweighting and b-tagging in the 1-jet channel. The impact of the systematic uncertainties is quite similar in the three years.

Sensitivity studies have been carried on for the first time to test the proposed binning of the stage 1.2 STXS scheme for the $H \to WW$ analysis. Possible redefinition of the reconstructed phase space bins was studied and the merging of some of the bins in this analysis was proposed to avoid high statistical uncertainties due to low populated bins. Finally, signal strength modifiers were calculated from fits in the ggH STXS bins, as multiplying factors for the events of the gluon fusion $H \to WW$ samples only. The following signal strengths were obtained:

- 2016 0-jet bin: $\mu = 0.89^{+0.20}_{-0.20}$
- 2016 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin: $\mu = 2.0^{+0.8}_{-0.7}$
- 2016 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV bin: $\mu = 1.8^{+1.0}_{-0.9}$
- 2016 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin: $\mu = 1.4^{+1.6}_{-1.4}$
- 2016 $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV bin: $\mu = 0.7^{+1.8}_{-1.7}$
- 2017 0-jet bin: $\mu = 1.01^{+0.21}_{-0.26}$
- 2017 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin: $\mu = 1.1^{+0.7}_{-0.7}$
- 2017 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV bin: $\mu = 0.9^{+0.7}_{-0.7}$
- 2017 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin: $\mu = 1.9^{+2.0}_{-2.3}$
- 2017 $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV bin: $\mu = -0.8^{+2.1}_{-1.9}$
- 2018 0-jet bin: $\mu = 0.94^{+0.19}_{-0.18}$
- 2018 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin: $\mu = 0.9^{+0.6}_{-0.6}$
- 2018 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV bin: $\mu = 0.7^{+0.6}_{-0.6}$
- 2018 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin: $\mu = -0.3^{+1.1}_{-1.1}$
- 2018 $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV bin: $\mu = 0.6^{+1.2}_{-1.2}$

The sources of systematic uncertainties are similar to the ones in the inclusive analysis, with the main difference being the reduction of theoretical uncertainties and an increase in the statistical uncertainties, especially in the high $p_{\mathrm{T}}^{\mathrm{H}}$ and 1-jet bins with $p_{\mathrm{T}}^{\mathrm{H}} > 60$ GeV. The results for each category from the three years are combined into:

- 0-jet bin combination: $\mu = 0.96^{+0.12}_{-0.12}$
- 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin combination: $\mu = 1.3^{+0.4}_{-0.4}$
- 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV bin combination: $\mu = 0.9^{+0.4}_{-0.4}$
- 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin combination: $\mu = -0.7^{+0.8}_{-0.8}$
- $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV bin combination: $\mu = 0.8^{+0.9}_{-0.9}$

The results of the signal strengths are in all cases consistent with one, in agreement with the SM predictions. A ggH (with H $\rightarrow$ WW) signal strength of $\mu_{ggH}^{STXS} = 0.99^{+0.11}_{-0.10}$ is obtained from the combination of the five STXS bins, which can be decomposed in its theoretical, experimental and statistical sources in this way:

$$\mu_{ggH}^{STXS} = 0.99^{+0.11}_{-0.10} = 0.99^{+0.07}_{-0.07}\ (\text{stat.})^{+0.07}_{-0.07}\ (\text{exp.})^{+0.04}_{-0.04}\ (\text{theo.})\ . \tag{6.2}$$

The theoretical uncertainty is lower than the experimental, as is expected in the STXS measurements. The statistical uncertainty is lower than any of the systematic uncertainties, since all the data of the STXS bins is being used in this combination.

The difference between the inclusive analysis signal strength and this one comes from two factors. First, the STXS analysis is not measuring the same signal strength modifier as the inclusive analysis, since in the inclusive analysis a global signal strength is computed that multiplies other Higgs boson processes with different production and decay modes, whereas the STXS analysis measures the gluon fusion production mode with the WW decay only. However, our selection targets the ggH H $\rightarrow$ WW process and therefore it is the one that contributes the most in the inclusive analysis, which means the signals are not that different. Second, there is also another difference in the reconstructed phase space used in each case, since there is no splitting in 8 categories according to flavour, charge and subleading lepton p_{T} in the STXS analysis and there is a splitting in the bins instead. Because of these two reasons there is a difference in the uncertainties of the signal strengths as well.

The signal strength modifiers in the STXS bins can be translated into cross sections multiplying the signal strength by the SM prediction. The following cross sections are obtained:

- 0-jet bin combination: $\sigma = 606 \pm 76$ fb
- 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV bin combination: $\sigma = 202 \pm 64$ fb
- 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV bin combination: $\sigma = 80 \pm 36$ fb
- 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV bin combination: $\sigma = 11 \pm 13$ fb
- $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV bin combination: $\sigma = 12 \pm 14$ fb

Finally, the cross section obtained for gg$\rightarrow$H $\rightarrow$ WW using the combination of the STXS bins is $\sigma_{ggH}^{STXS} = 1.08^{+0.12}_{-0.11}$ pb.

These are the first cross section measurements in the stage 1.2 of the Simplified Template Cross Section scheme in the H $\rightarrow$ WW channnel. They will be combined with similar measurements performed in other Higgs boson decay channels to deliver a global CMS combination.

Appendix A

Additional plots using 2016 and 2017 datasets

A.1 Background control regions for 2016 and 2017

This appendix contains the DY, top quark and WW control region distributions that were shown for the 2018 dataset in section 4.3, this time obtained for the 2016 and 2017 datasets. The validation plots for the nonprompt lepton background discussed for 2018 in section 4.3.5 are also shown here for 2016 and 2017.

A.1.1 DY background control regions for 2016 and 2017

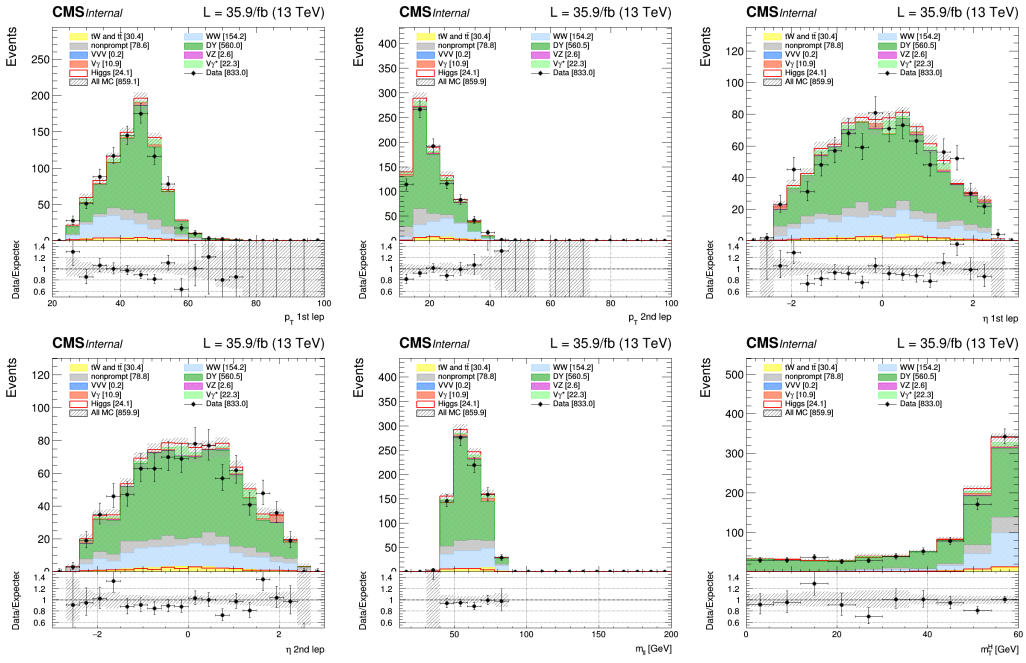


Figure A.1: Distributions of the 2016 0-jet DY control region.

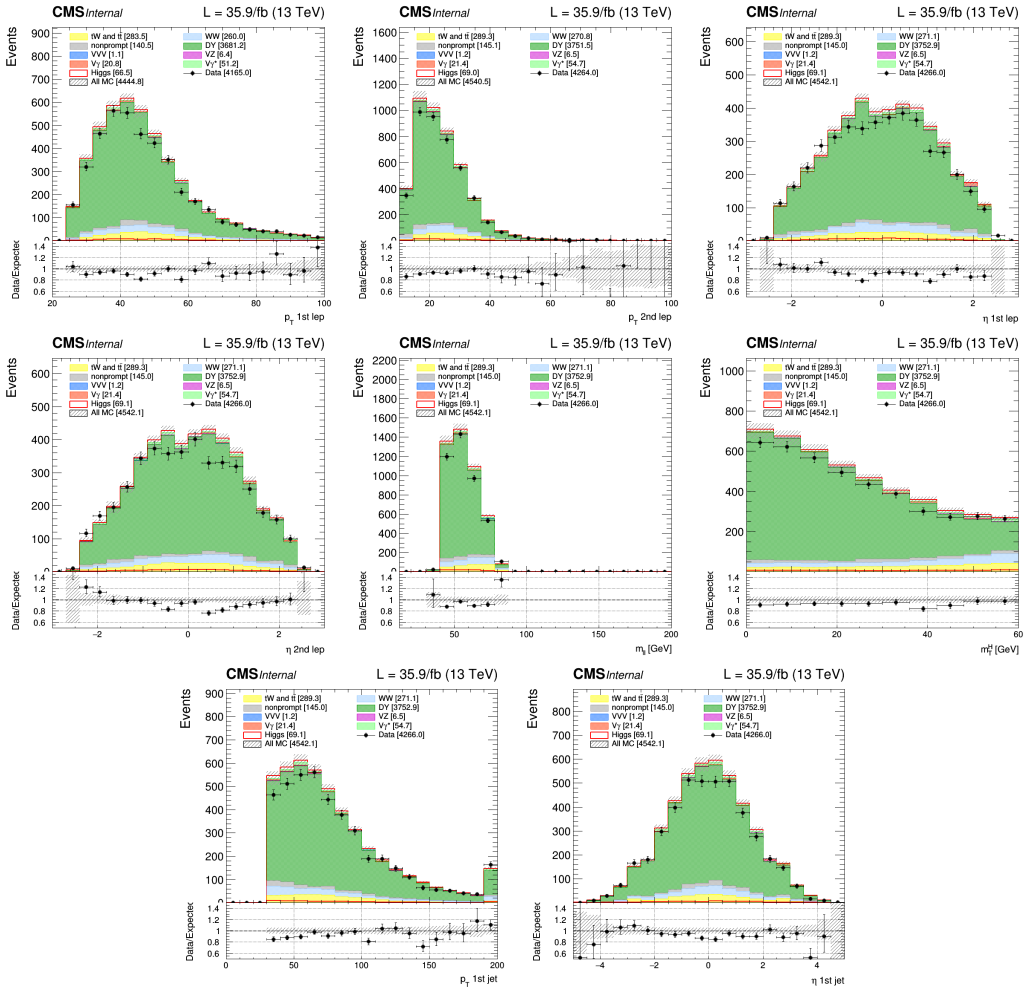


Figure A.2: Distributions of the 2016 1-jet DY control region.

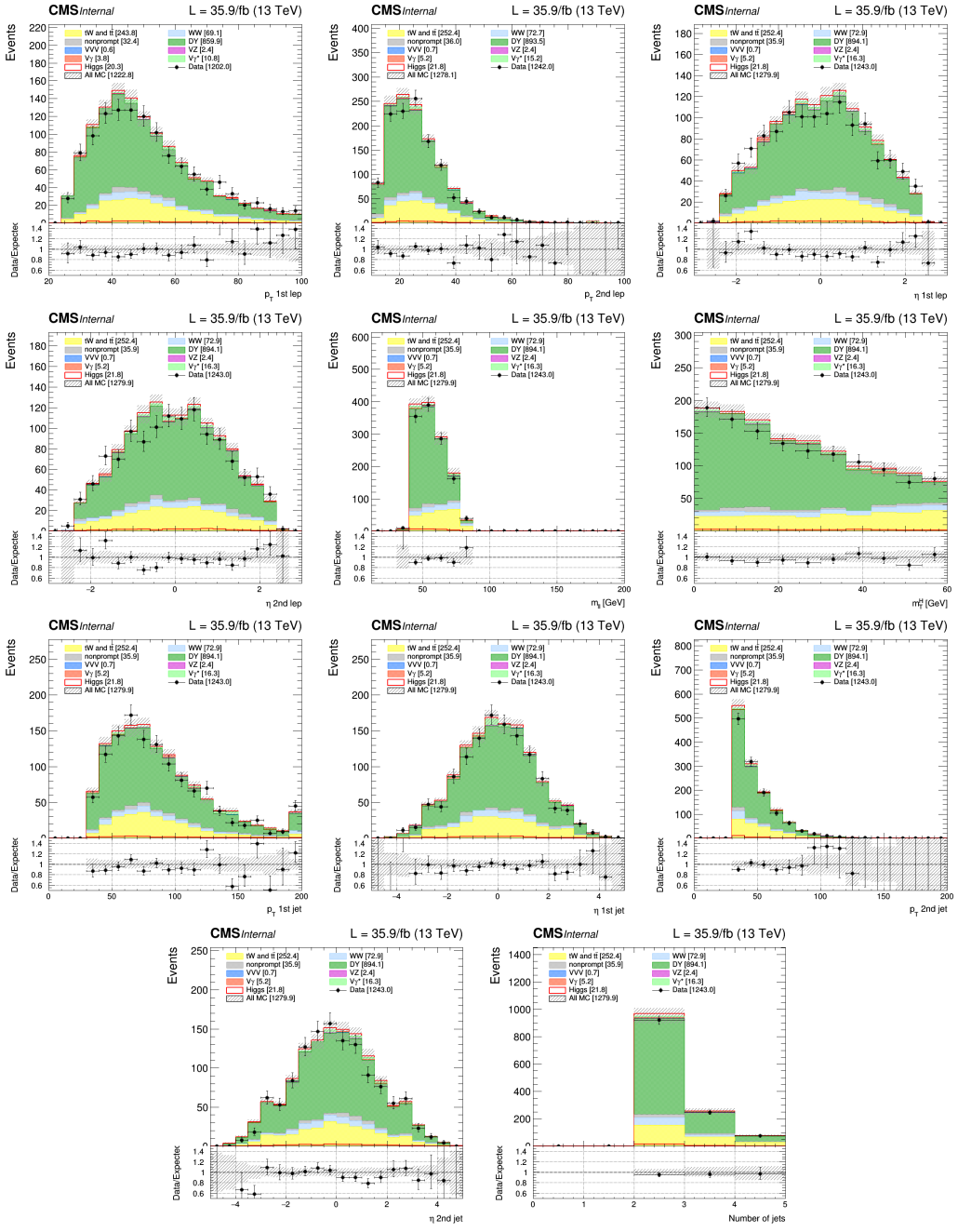


Figure A.3: Distributions of the 2016 ≥2-jet DY control region.

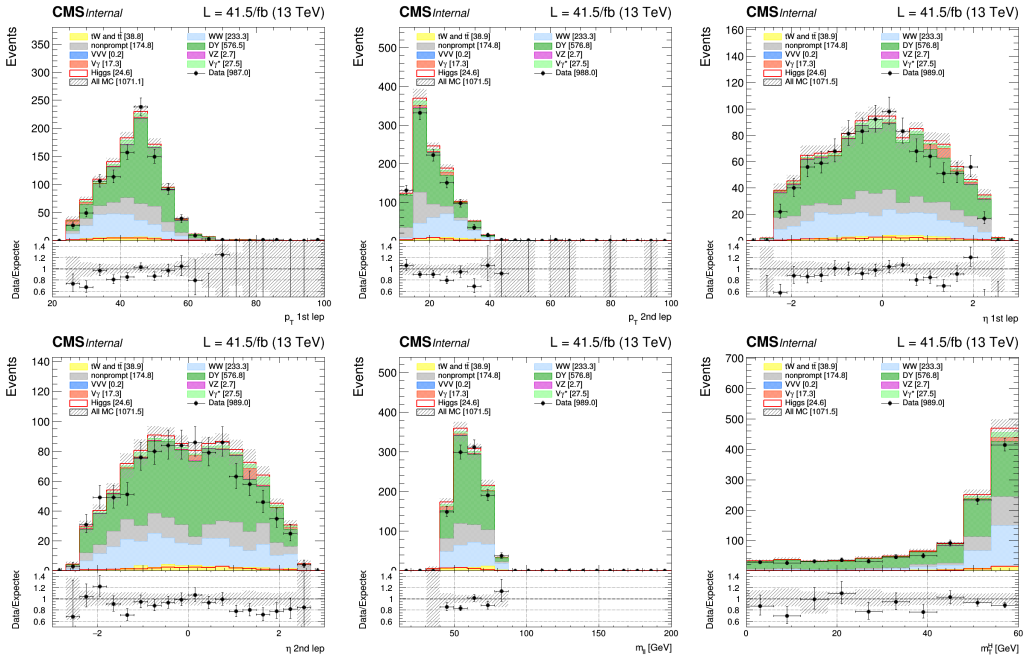


Figure A.4: Distributions of the 2017 0-jet DY control region.

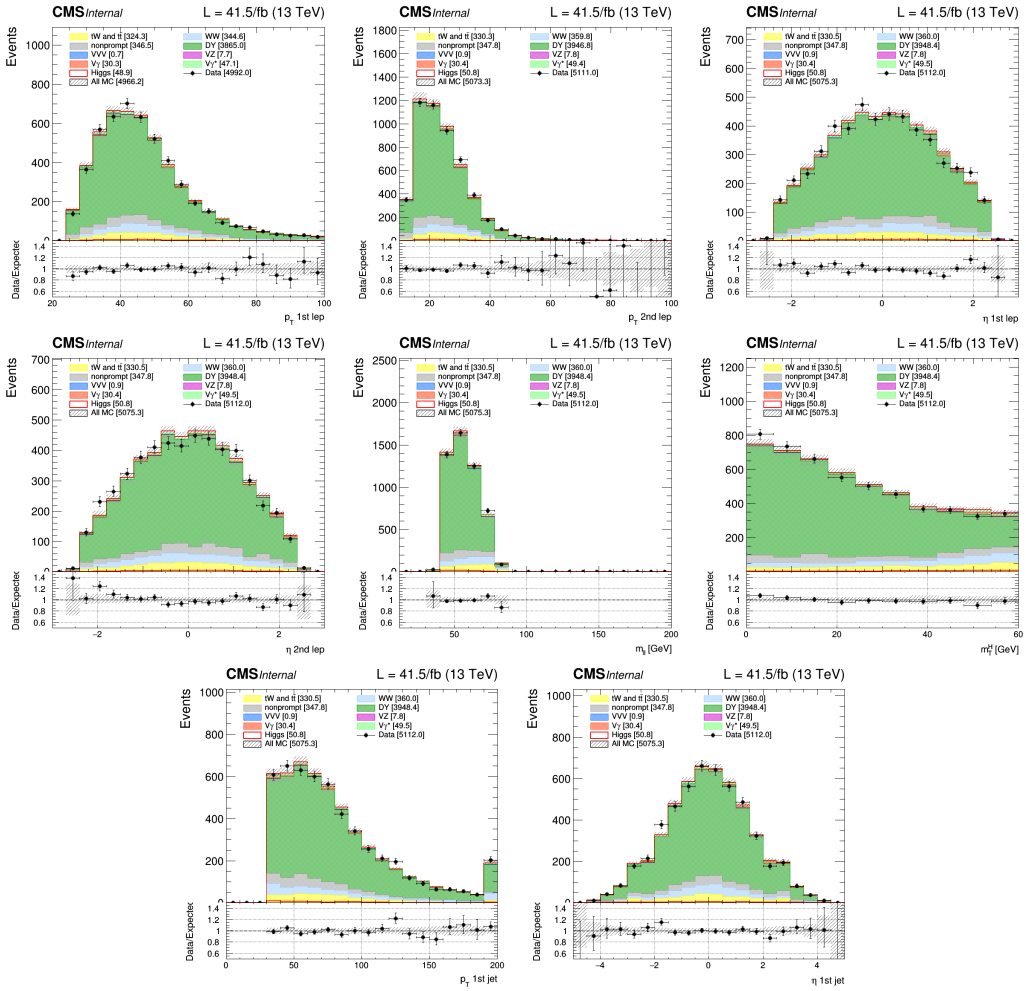


Figure A.5: Distributions of the 2017 1-jet DY control region.

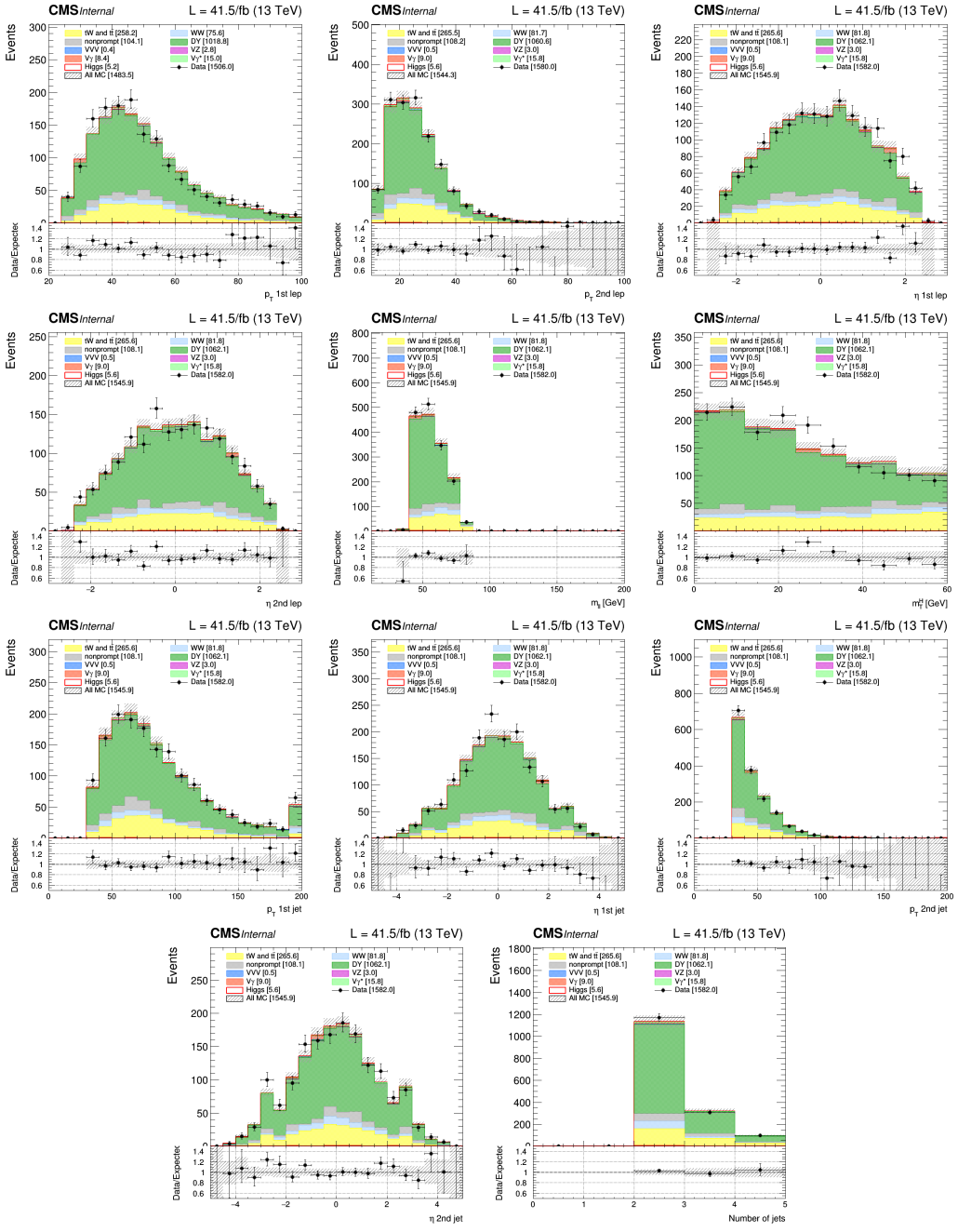


Figure A.6: Distributions of the 2017 ≥2-jet DY control region.

A.1.2 Top quark background control regions for 2016 and 2017

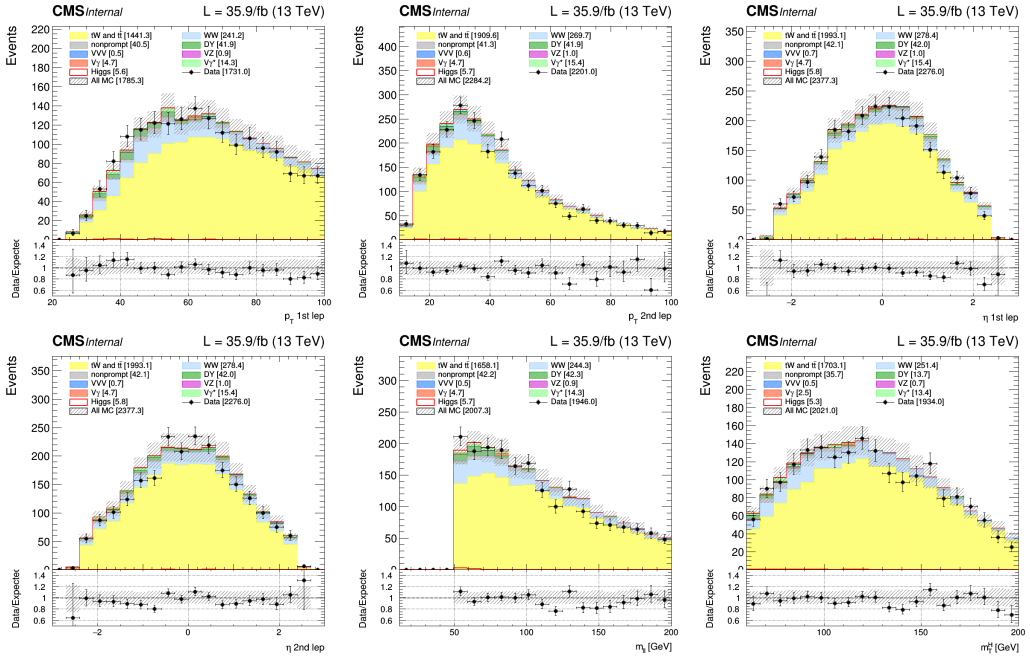


Figure A.7: Distributions of the 2016 0-jet top control region.

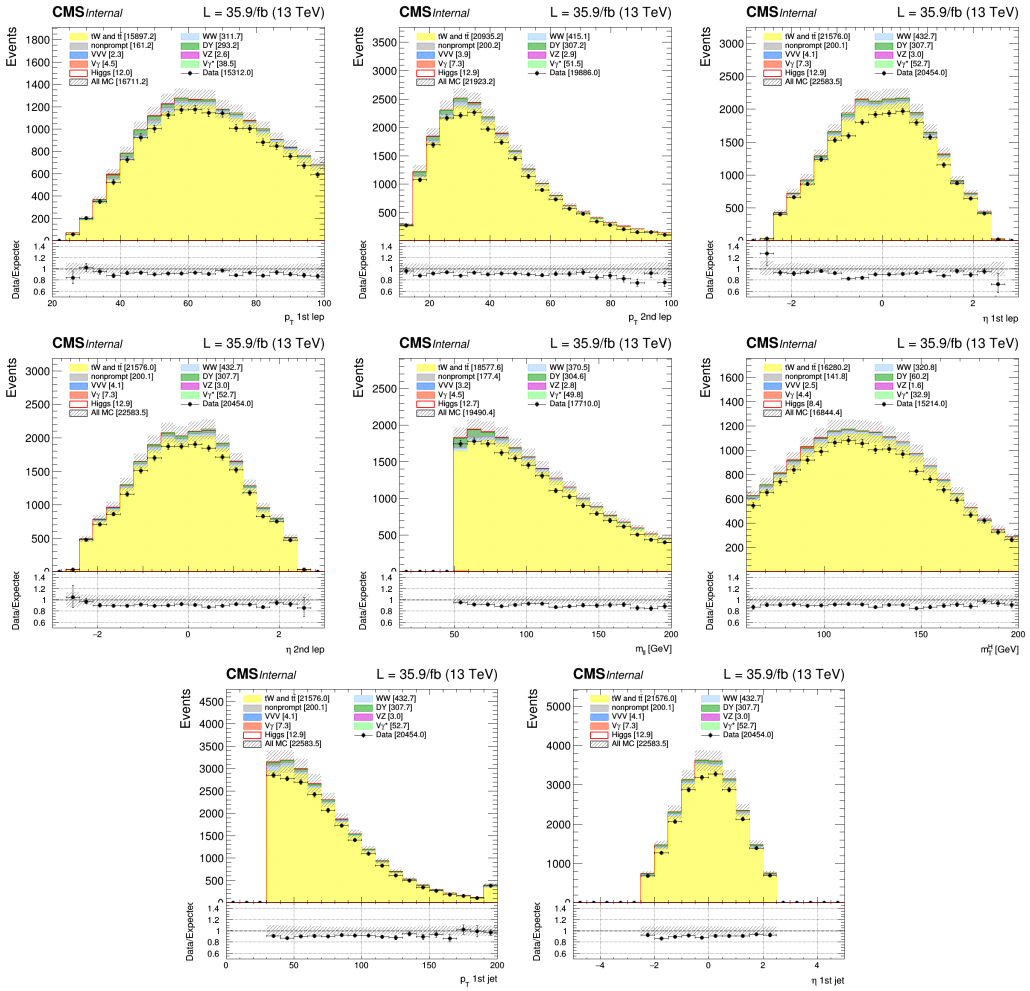


Figure A.8: Distributions of the 2016 1-jet top control region.

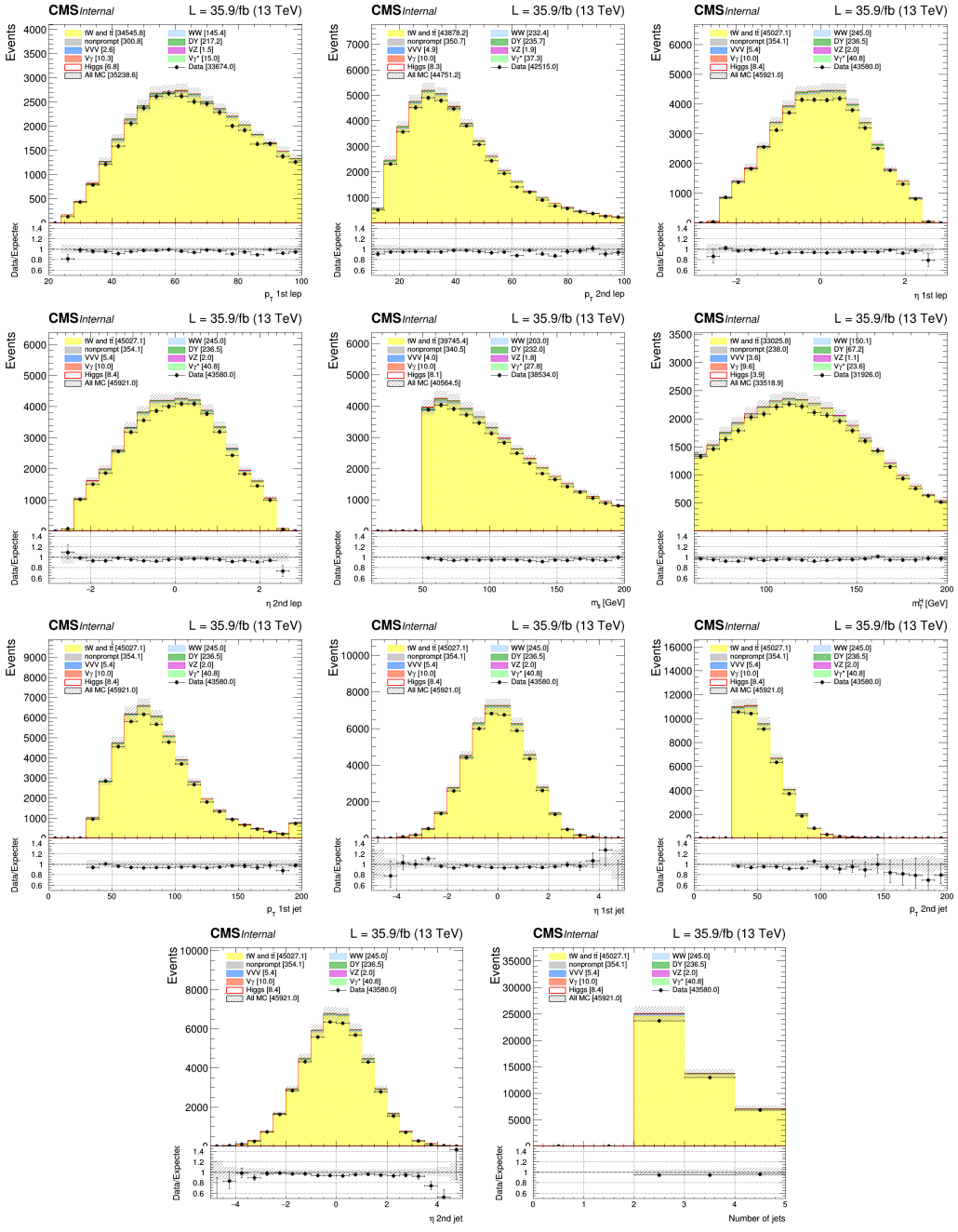


Figure A.9: Distributions of the 2016 ≥2-jet top control region.

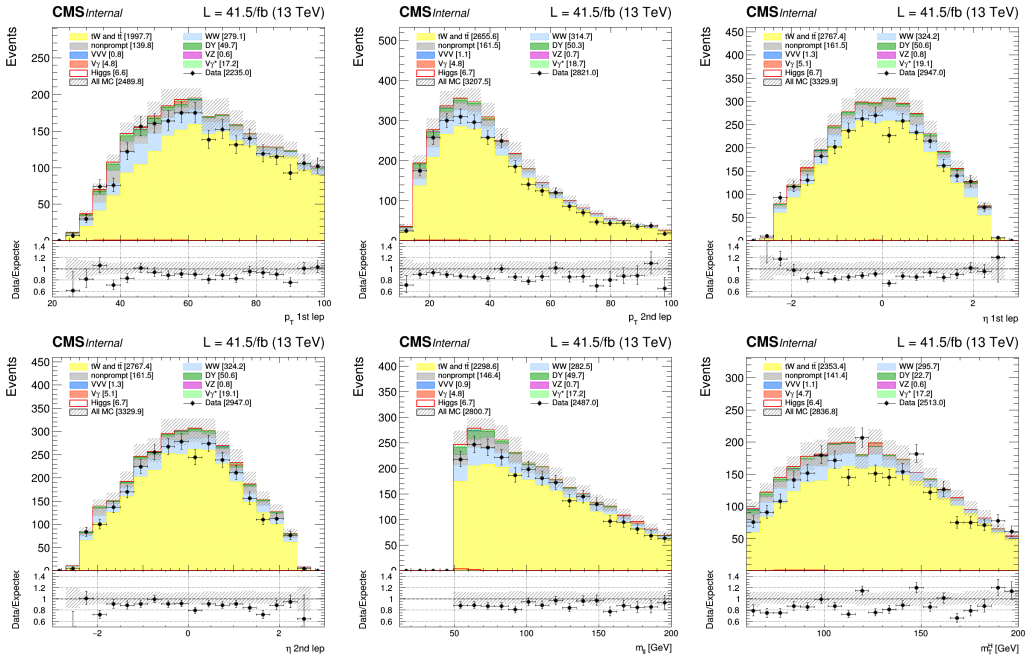


Figure A.10: Distributions of the 2017 0-jet top control region.

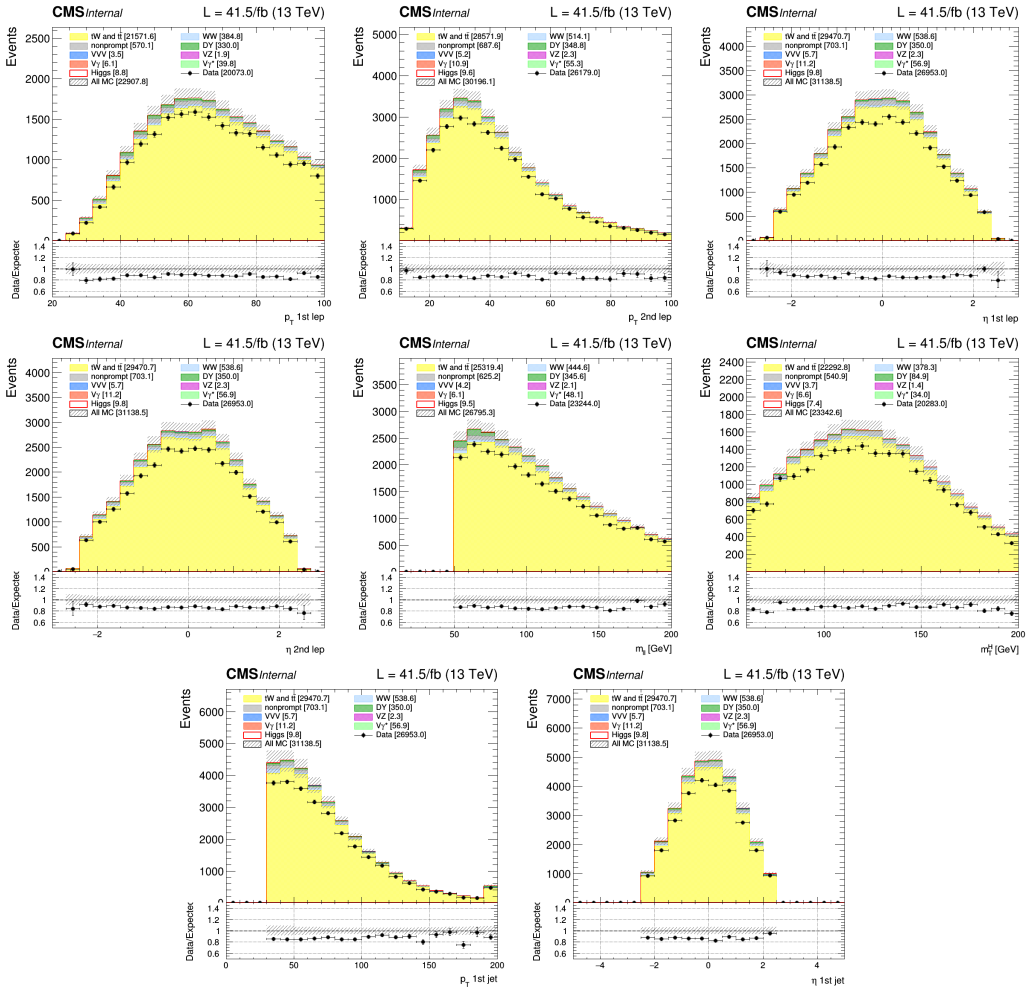


Figure A.11: Distributions of the 2017 1-jet top control region.

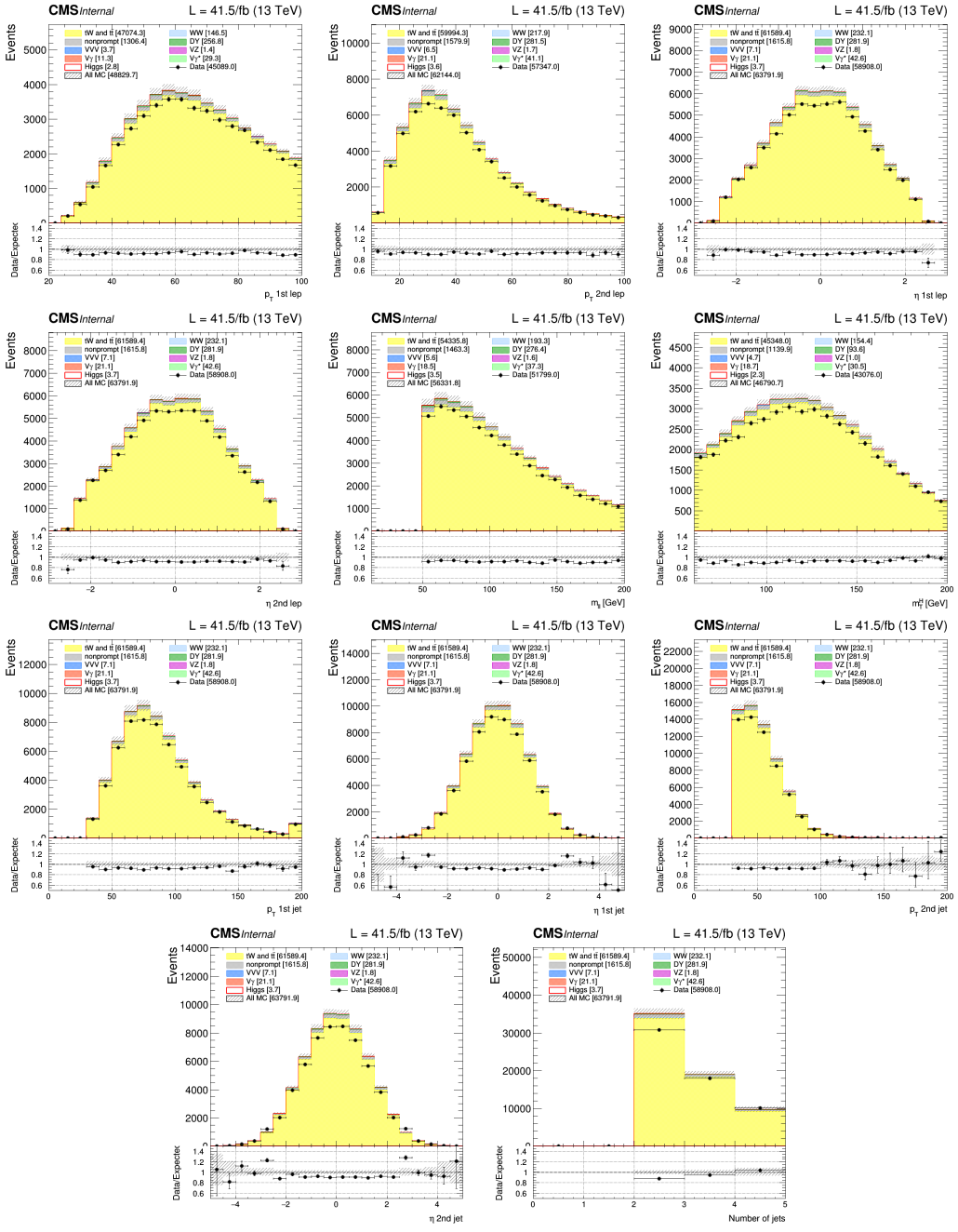


Figure A.12: Distributions of the 2017 ≥2-jet top control region.

A.1.3 WW background control region plots for 2016 and 2017

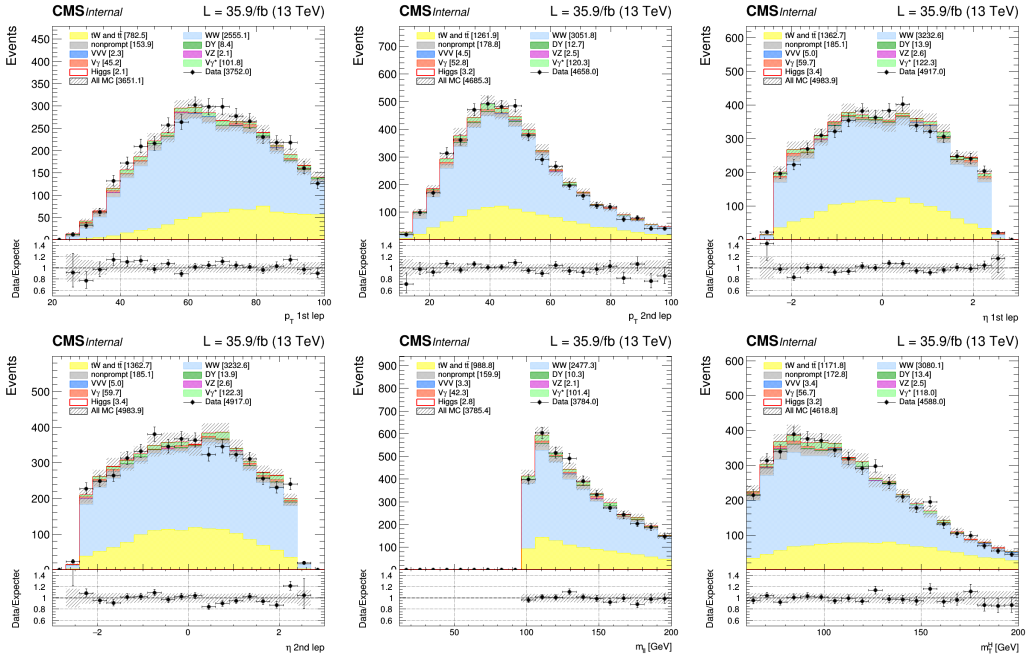


Figure A.13: Distributions of the 2016 0-jet WW control region.

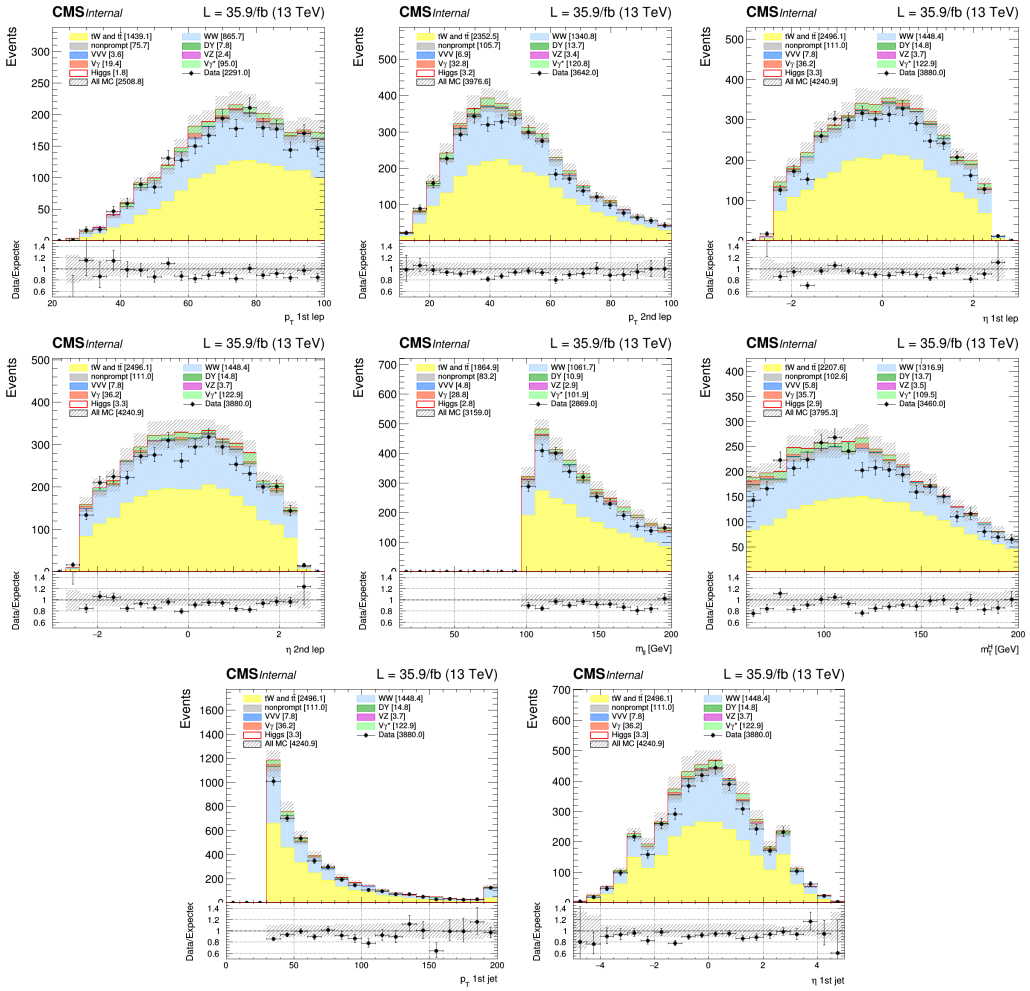


Figure A.14: Distributions of the 2016 1-jet WW control region.

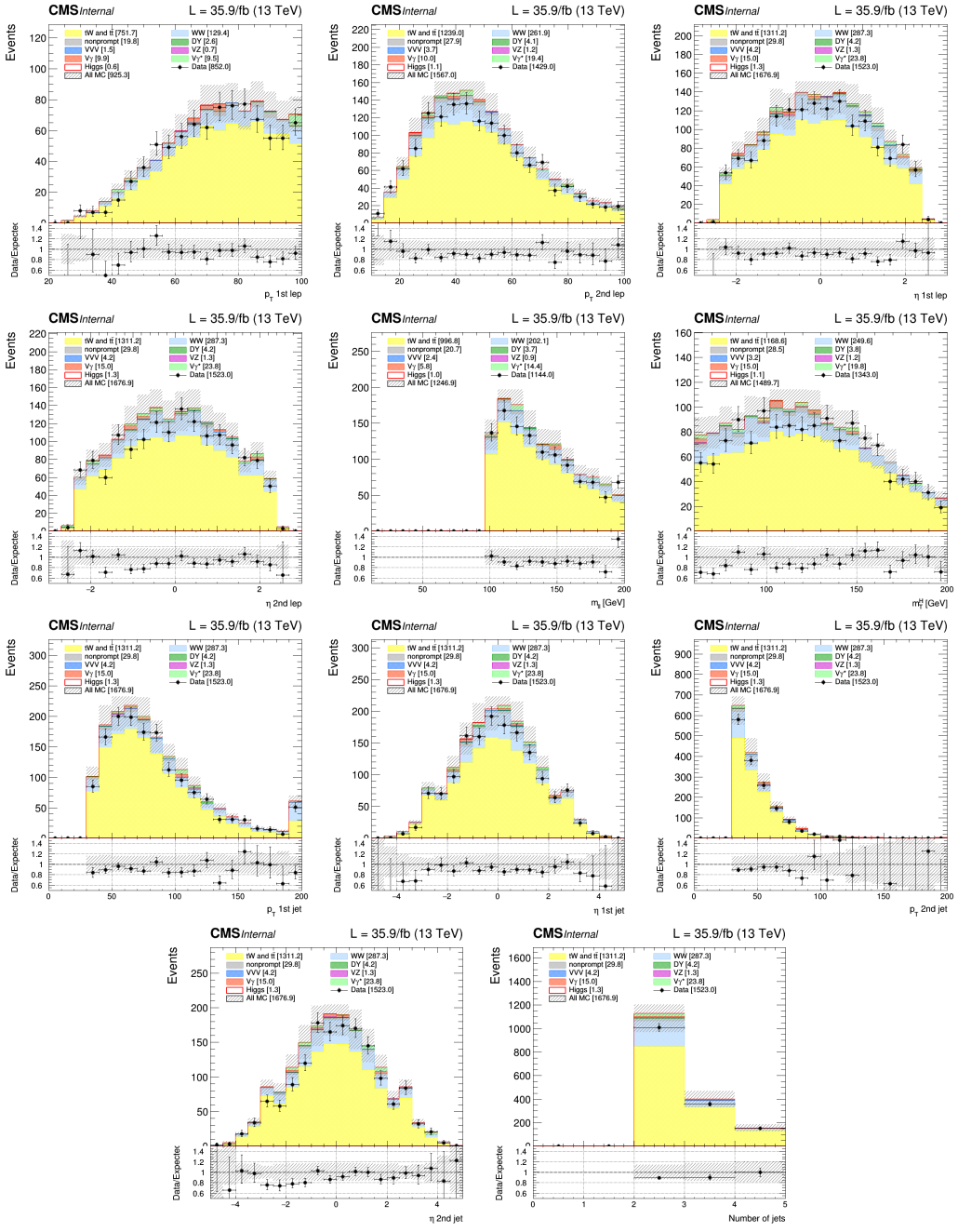


Figure A.15: Distributions of the 2016 ≥2-jet WW control region.

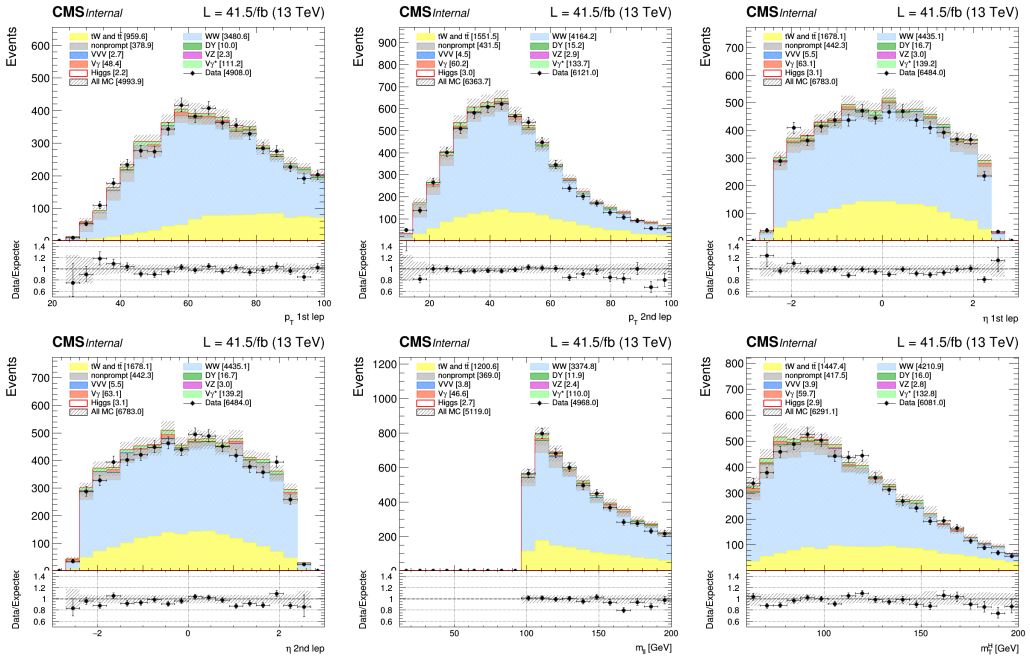


Figure A.16: Distributions of the 2017 0-jet WW control region.

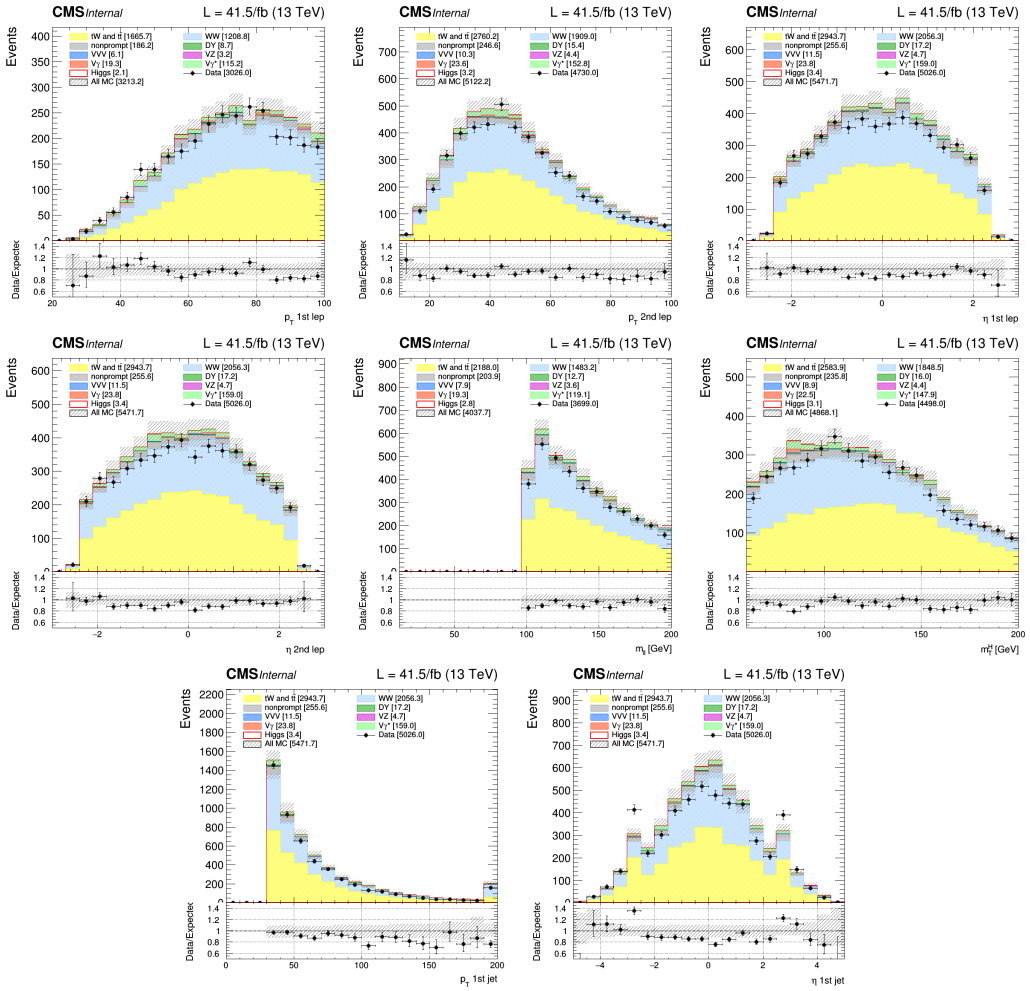


Figure A.17: Distributions of the 2017 1-jet WW control region.

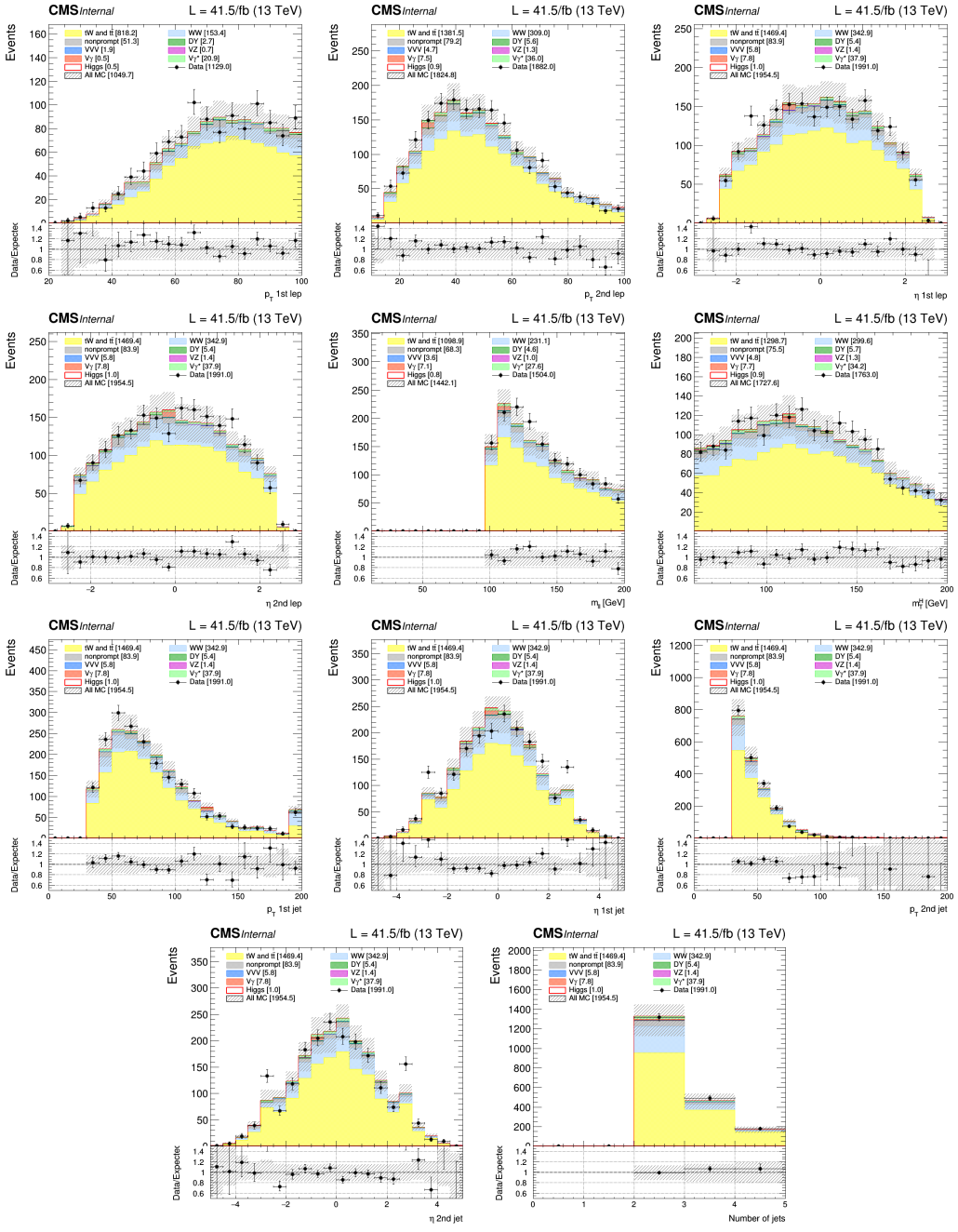


Figure A.18: Distributions of the 2017 ≥2-jet WW control region.

A.1.4 Nonprompt background validation for 2016 and 2017

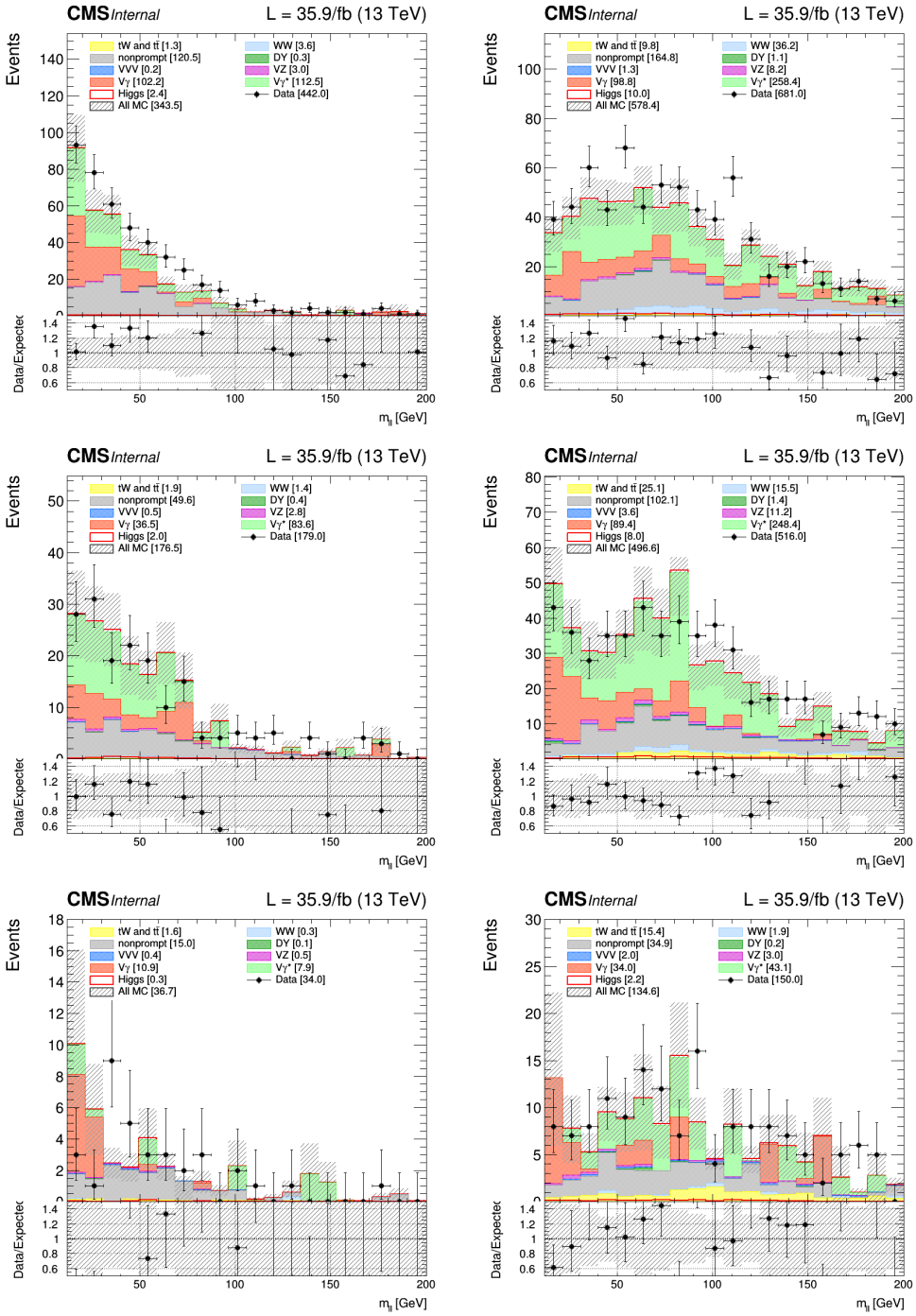


Figure A.19: Dilepton invariant mass distributions in the same-sign control region for eμ events for the year 2016 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

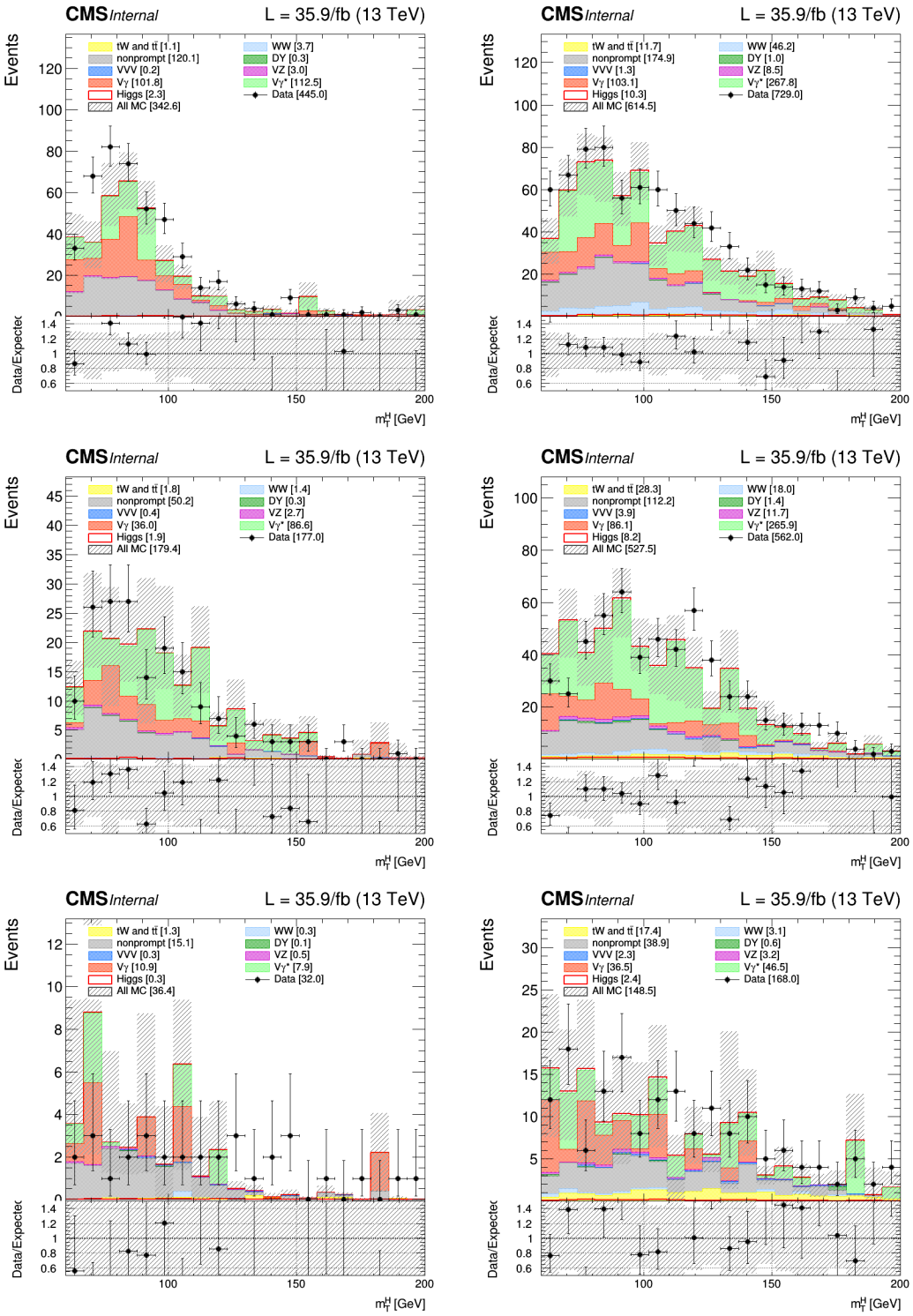


Figure A.20: Trasverse mass distributions in the same-sign control region for eμ events for the year 2016 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

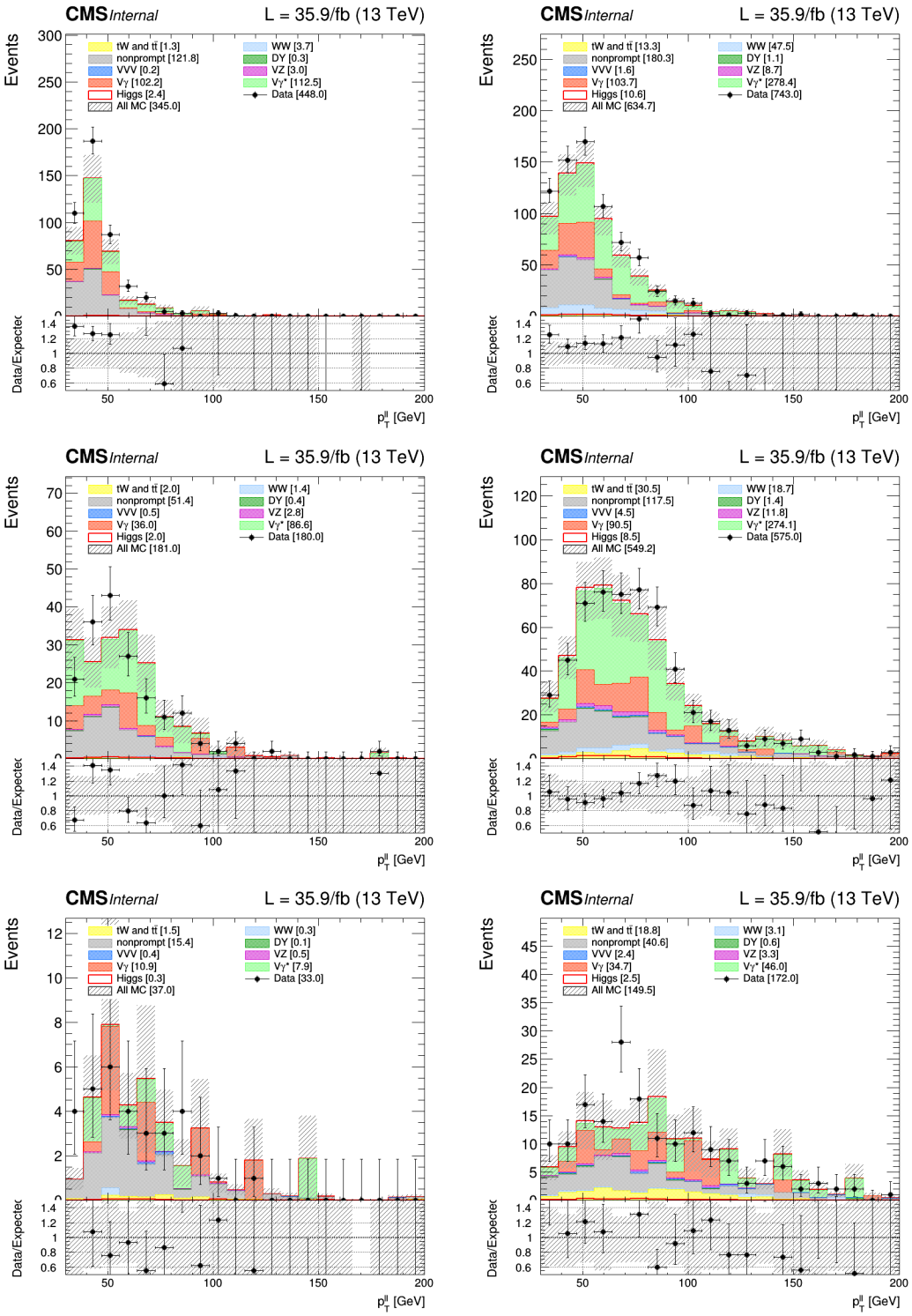


Figure A.21: Dilepton transverse momentum distributions in the same-sign control region for eμ events for the year 2016 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

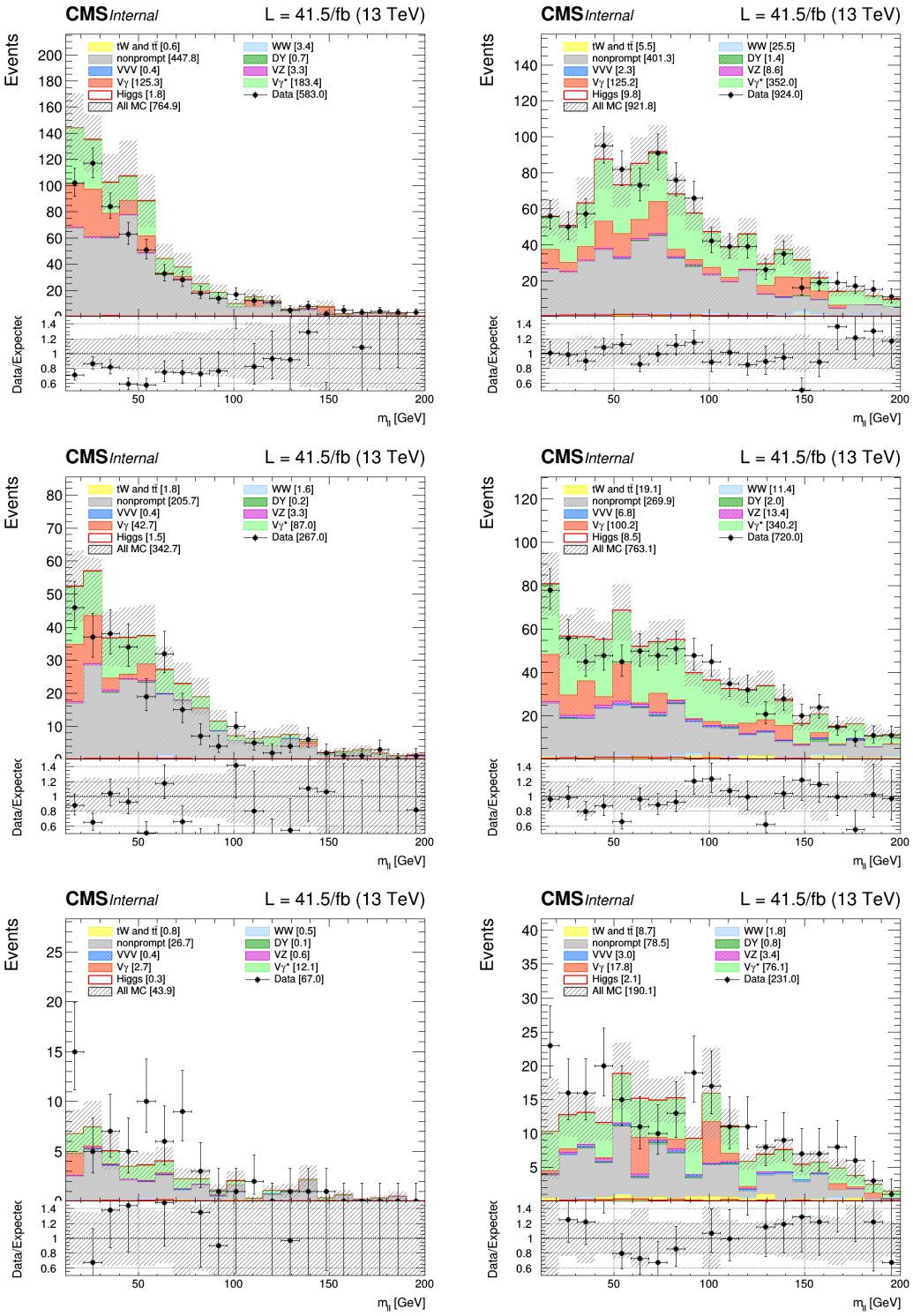


Figure A.22: Dilepton invariant mass distributions in the same-sign control region for eμ events for the year 2017 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

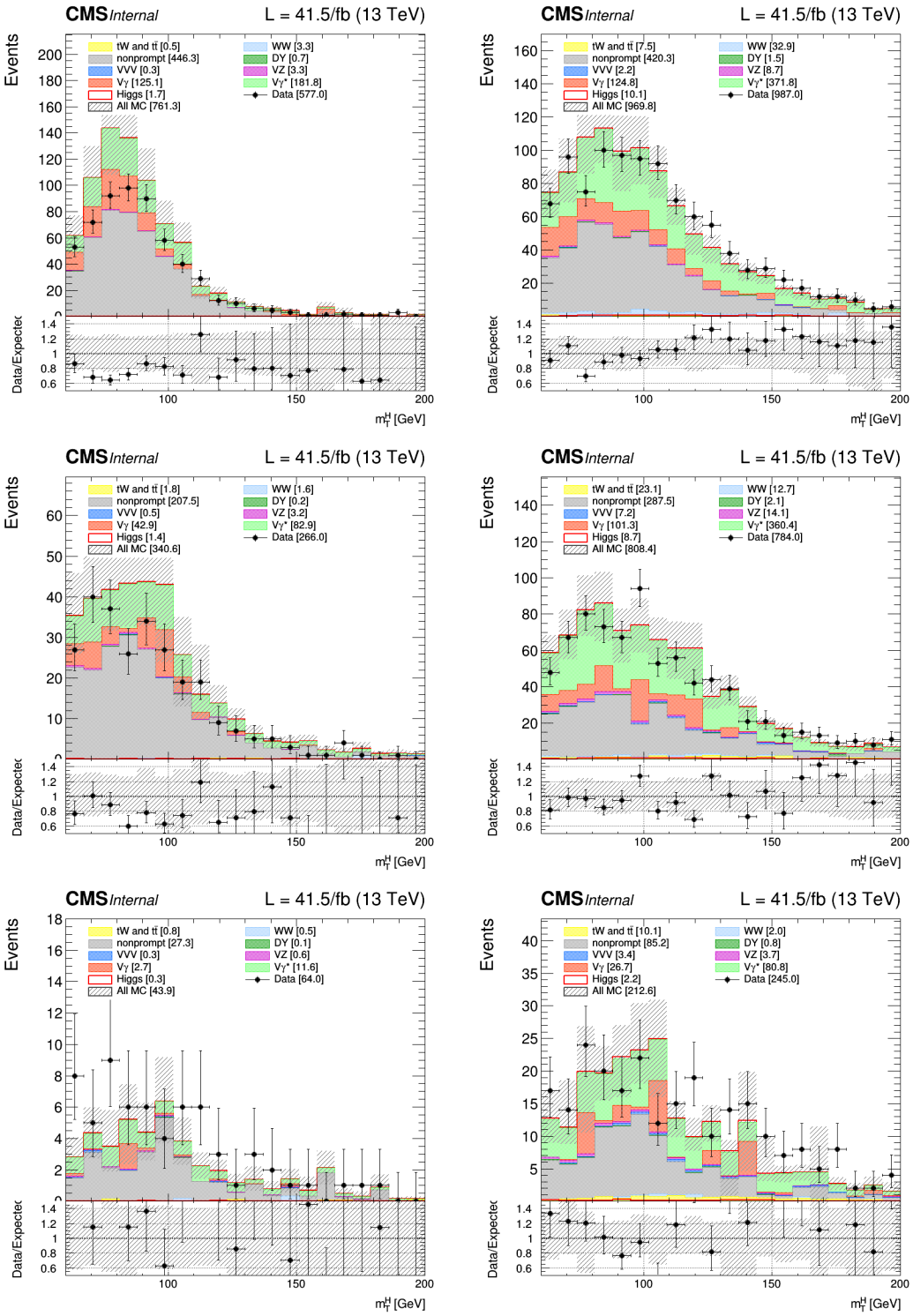


Figure A.23: Trasverse mass distributions in the same-sign control region for eμ events for the year 2017 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

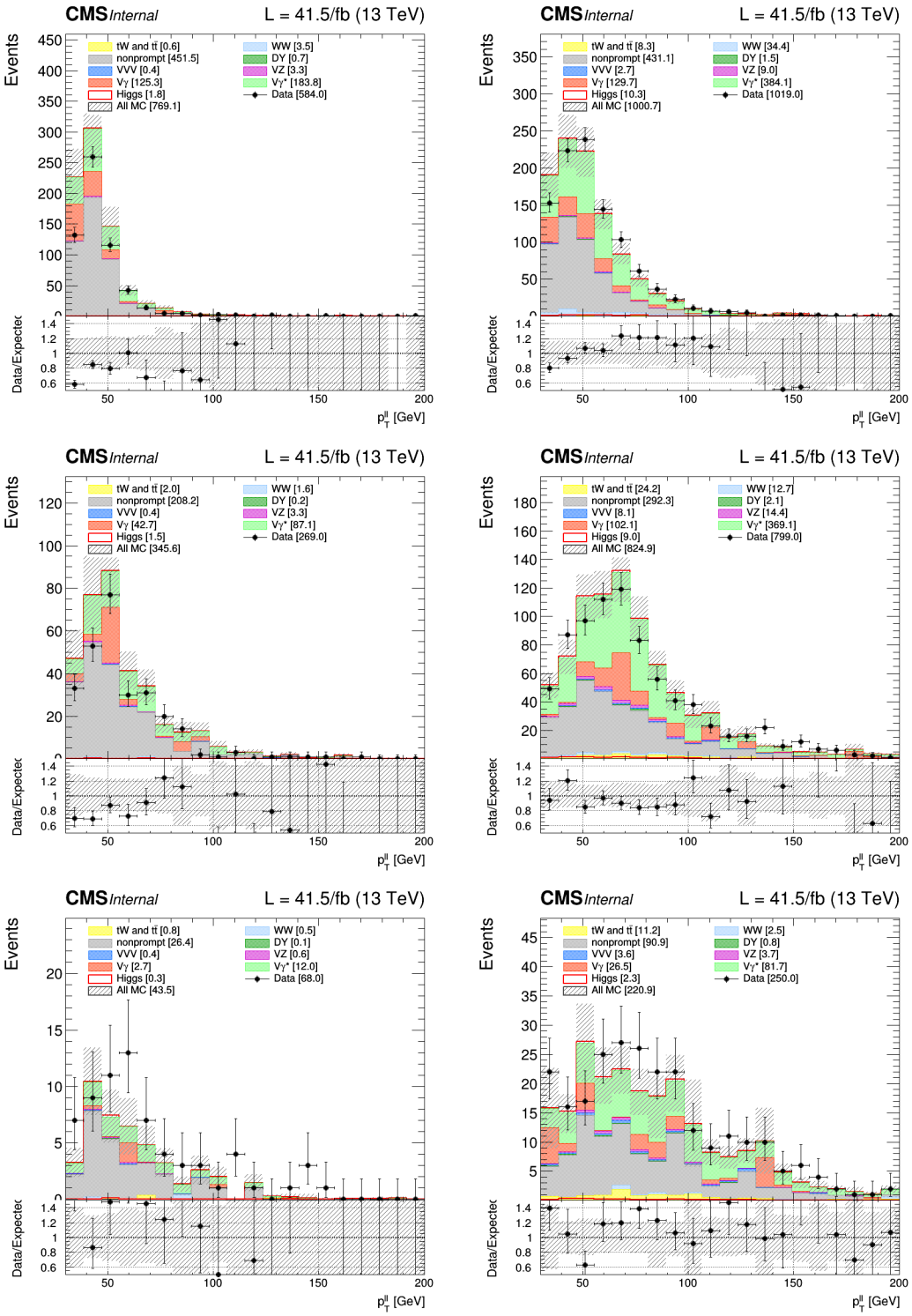
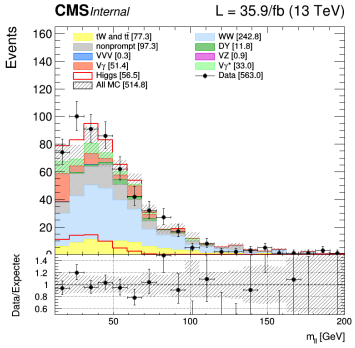


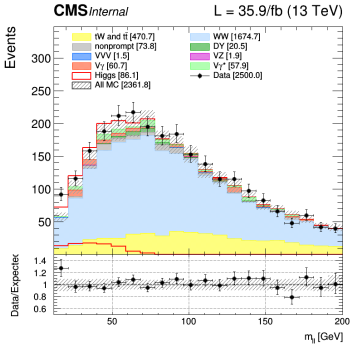
Figure A.24: Dilepton transverse momentum distributions in the same-sign control region for eμ events for the year 2017 and divided into low ($<$ 20 GeV, left) and high ($\geq$ 20 GeV, right) subleading lepton p_{T} and jet multiplicity (top: 0 jets, middle: 1 jet, bottom: $\geq$ 2jets). The shown uncertainties are statistical only.

A.2 Signal region distributions for 2016 and 2017

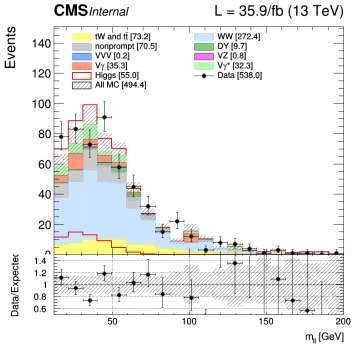
This appendix contains the ggH signal region $m_{\ell\ell}$, $m_{\mathrm{T}}^{\mathrm{H}}$ and the two-dimensional ($m_{\ell\ell}$, $m_{\mathrm{T}}^{\mathrm{H}}$) distributions that were shown for the 2018 dataset in section 5.2.3, this time obtained with the 2016 and 2017 datasets.



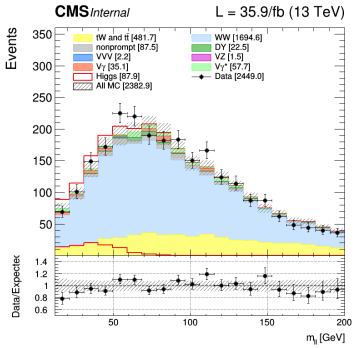
(a) $\mu^-e^+p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



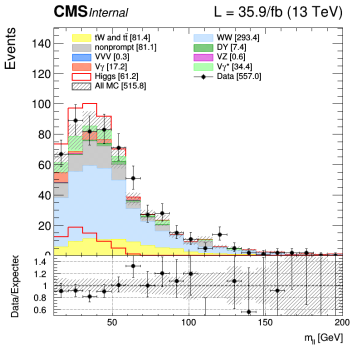
(b) $\mu^-e^+p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



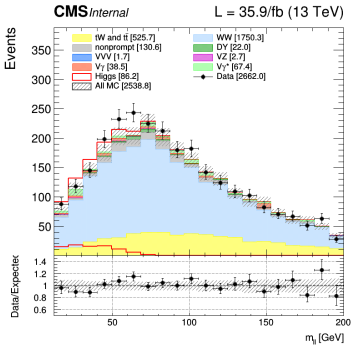
(c) $\mu^+e^-p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



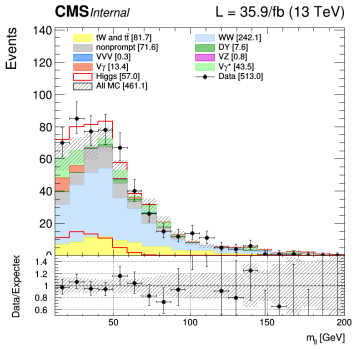
(d) $\mu^+e^-p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



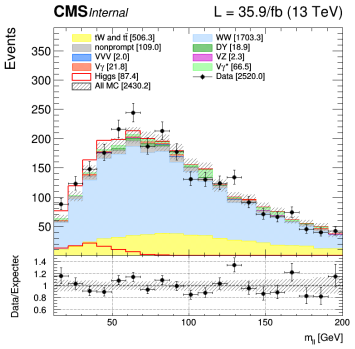
(e) $e^-\mu^+p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



(f) $e^-\mu^+p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

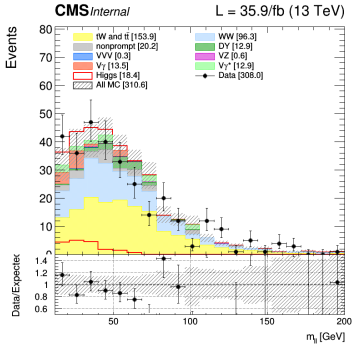


(g) $e^+\mu^-p_{\mathrm{T}2} < 20$ GeV 0-jet ggH

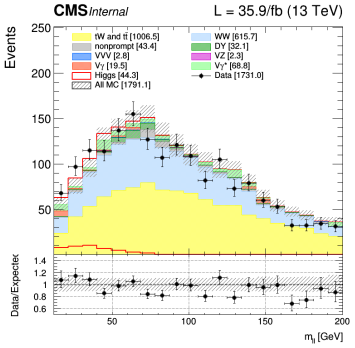


(h) $e^+\mu^-p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

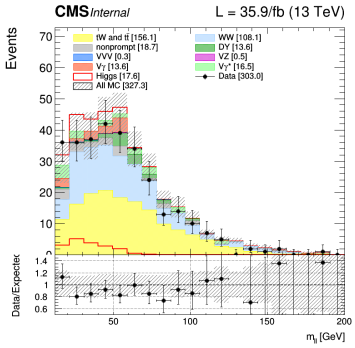
Figure A.25: Dilepton mass distributions in the 0-jet ggH subcategories for 2016



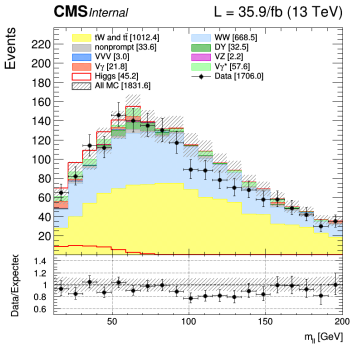
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



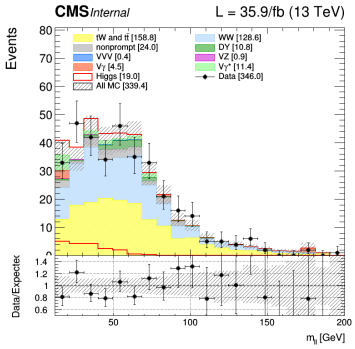
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



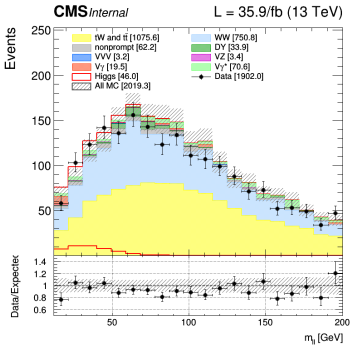
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



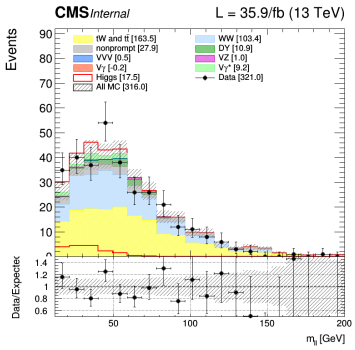
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



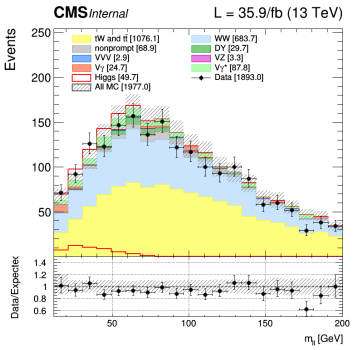
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

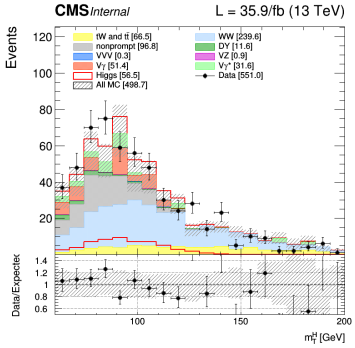


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

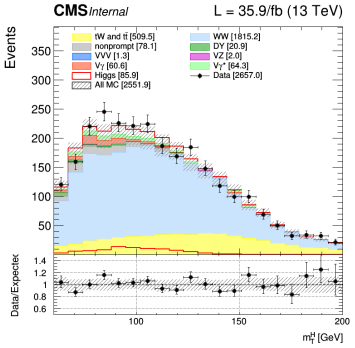


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

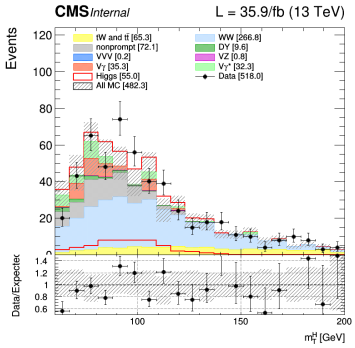
Figure A.26: Dilepton mass distributions in the 1-jet ggH subcategories for 2016



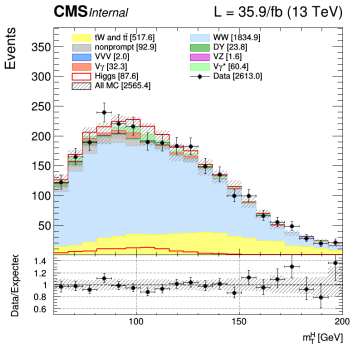
(a) $\mu^- e^+ p_{T2} < 20$ GeV 0-jet ggH



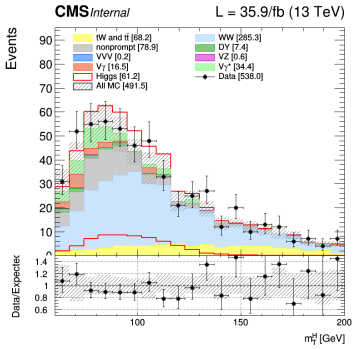
(b) $\mu^- e^+ p_{T2} > 20$ GeV 0-jet ggH



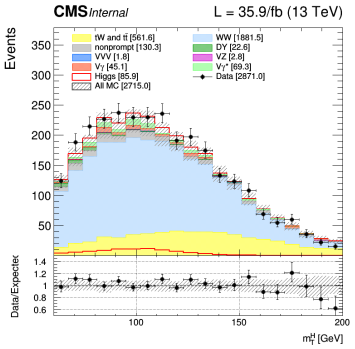
(c) $\mu^+ e^- p_{T2} < 20$ GeV 0-jet ggH



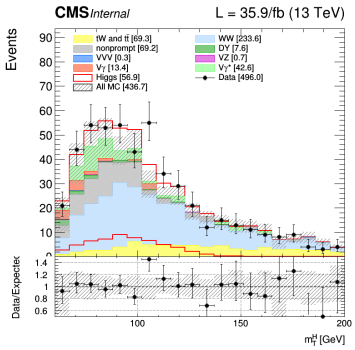
(d) $\mu^+ e^- p_{T2} > 20$ GeV 0-jet ggH



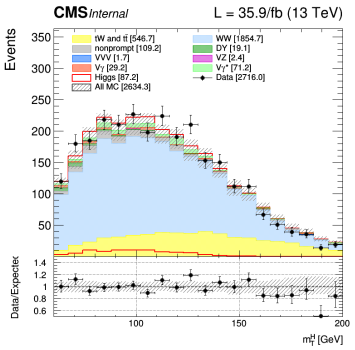
(e) $e^- \mu^+ p_{T2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{T2} > 20$ GeV 0-jet ggH

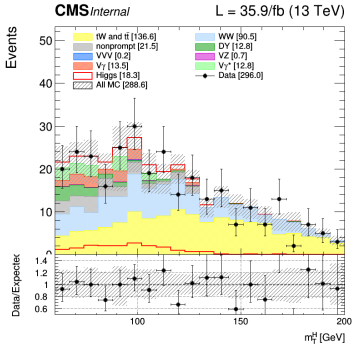


(g) $e^+ \mu^- p_{T2} < 20$ GeV 0-jet ggH

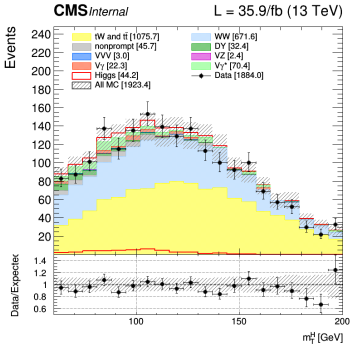


(h) $e^+ \mu^- p_{T2} > 20$ GeV 0-jet ggH

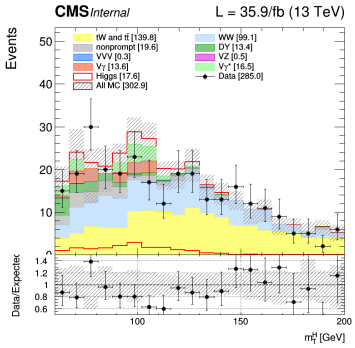
Figure A.27: Transverse mass distributions in the 0-jet ggH subcategories for 2016



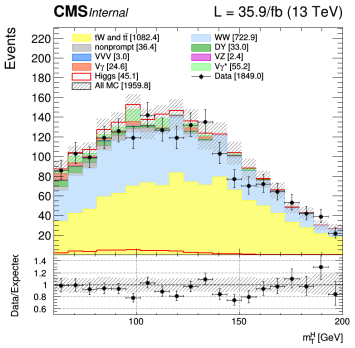
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



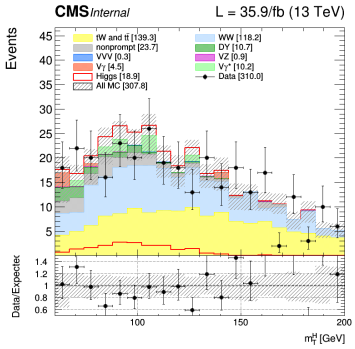
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



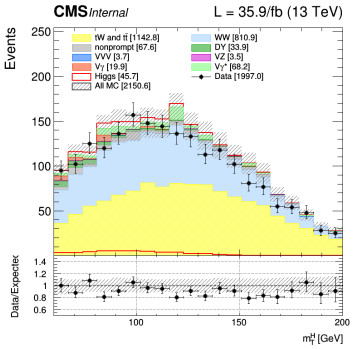
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



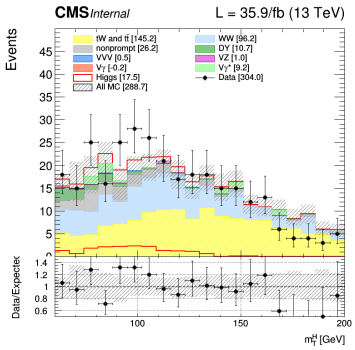
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



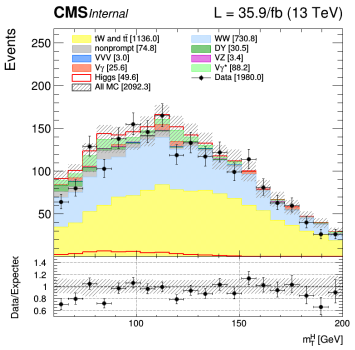
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

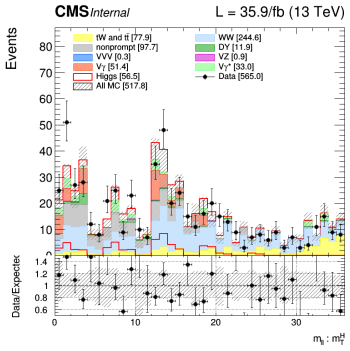


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

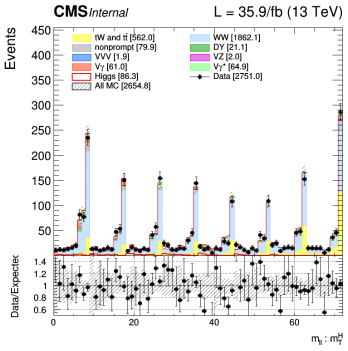


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

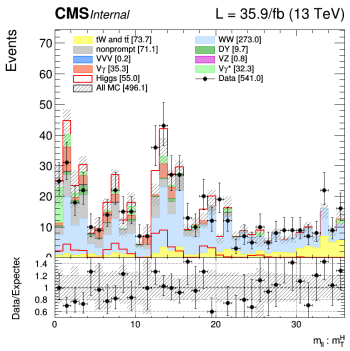
Figure A.28: Transverse mass distributions in the 1-jet ggH subcategories for 2016



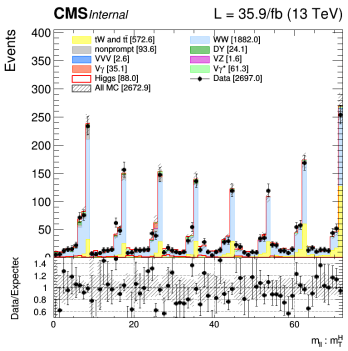
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



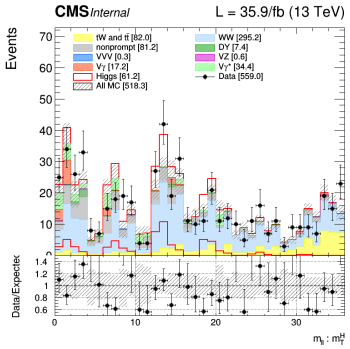
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



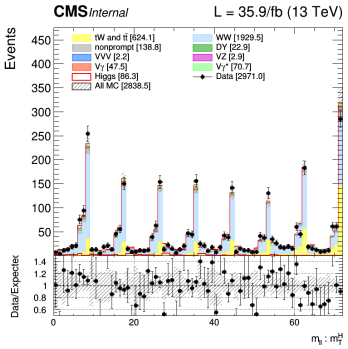
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



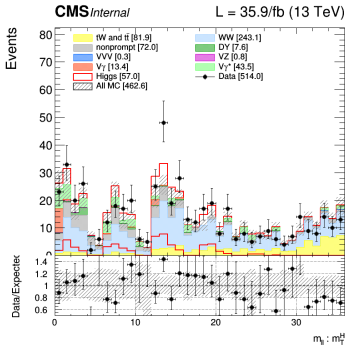
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



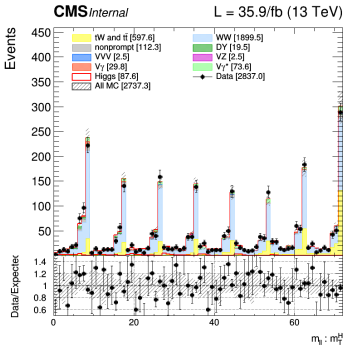
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

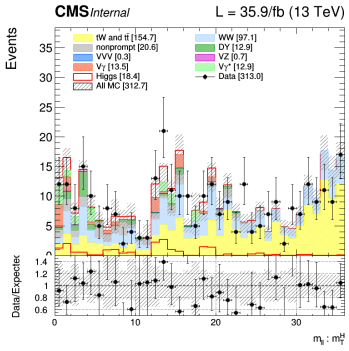


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH

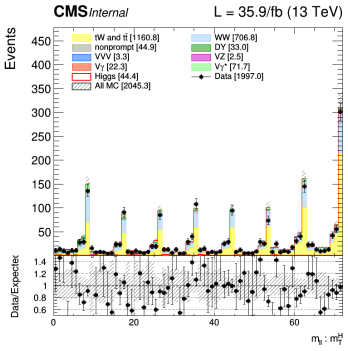


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

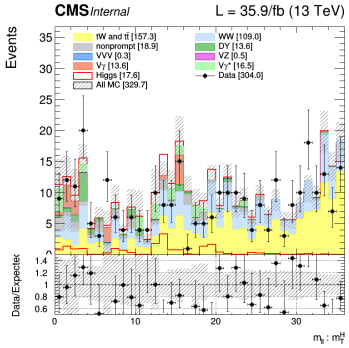
Figure A.29: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 0-jet ggH subcategories for 2016. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.



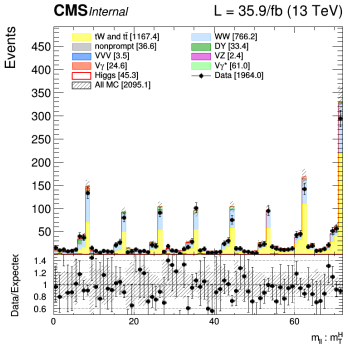
(a) $\mu^- e^+ p_{T2} < 20$ GeV 1 jet ggH



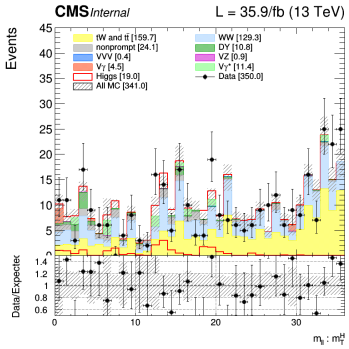
(b) $\mu^- e^+ p_{T2} > 20$ GeV 1 jet ggH



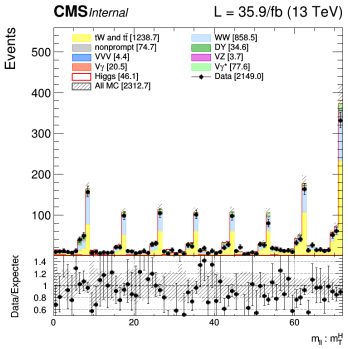
(c) $\mu^+ e^- p_{T2} < 20$ GeV 1 jet ggH



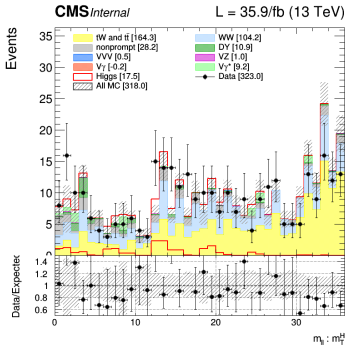
(d) $\mu^+ e^- p_{T2} > 20$ GeV 1 jet ggH



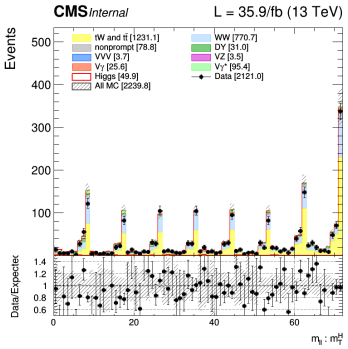
(e) $e^- \mu^+ p_{T2} < 20$ GeV 1 jet ggH



(f) $e^- \mu^+ p_{T2} > 20$ GeV 1 jet ggH

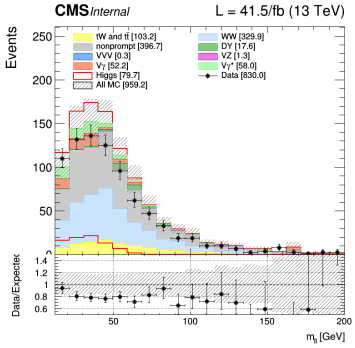


(g) $e^+ \mu^- p_{T2} < 20$ GeV 1 jet ggH

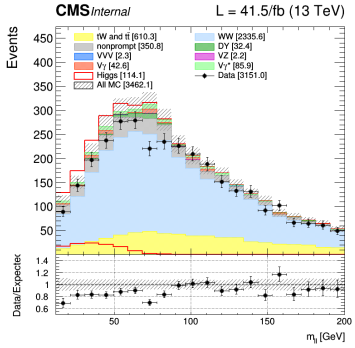


(h) $e^+ \mu^- p_{T2} > 20$ GeV 1 jet ggH

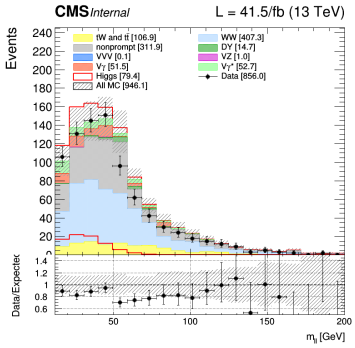
Figure A.30: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 1-jet ggH subcategories for 2016. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.



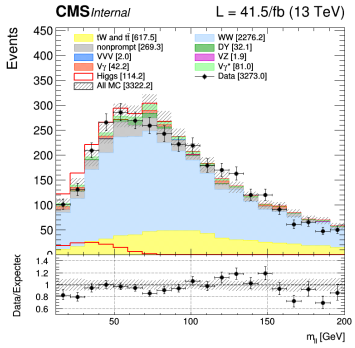
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



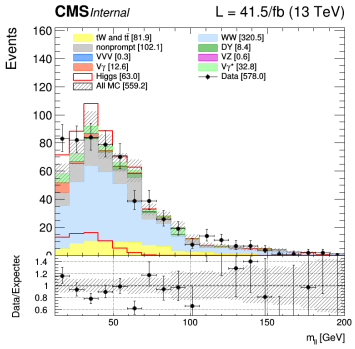
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



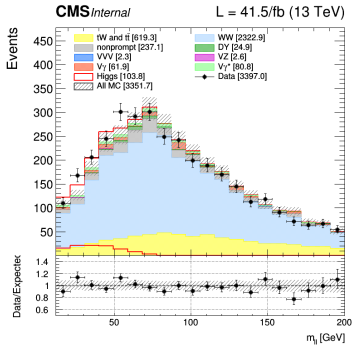
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



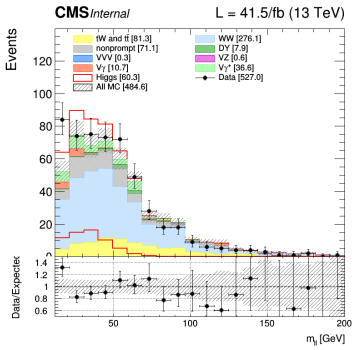
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH



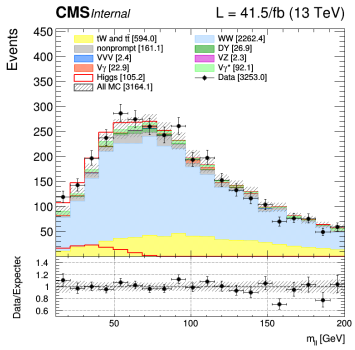
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

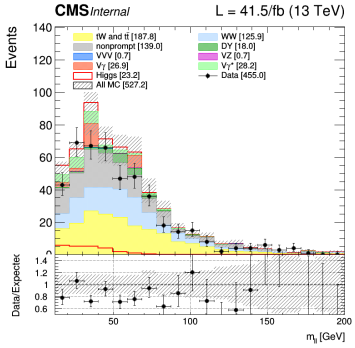


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 0-jet ggH

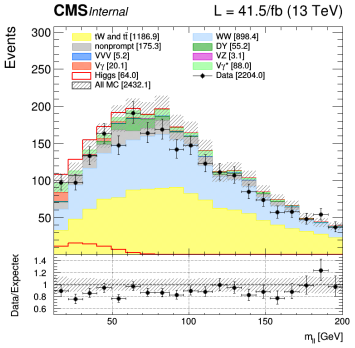


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 0-jet ggH

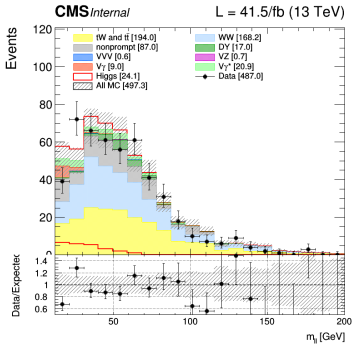
Figure A.31: Dilepton mass distributions in the 0-jet ggH subcategories for 2017



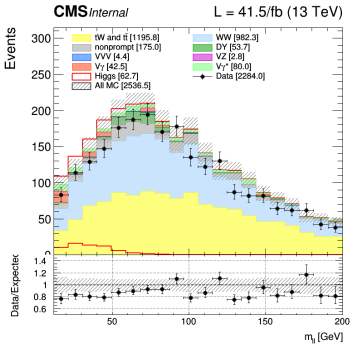
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



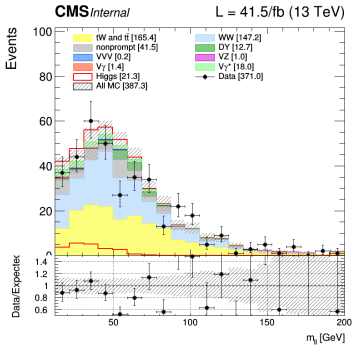
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



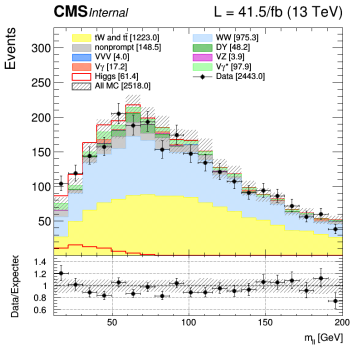
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



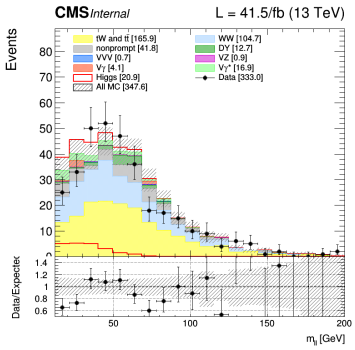
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



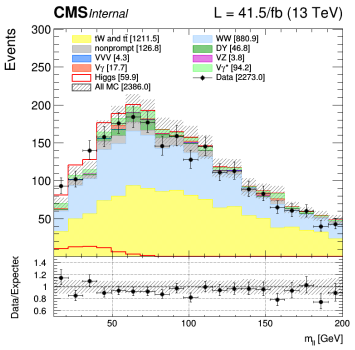
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

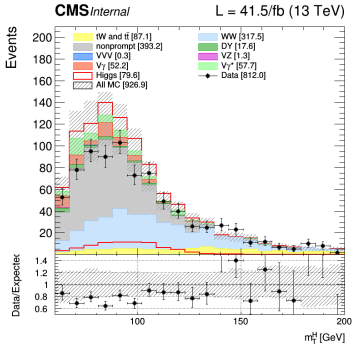


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

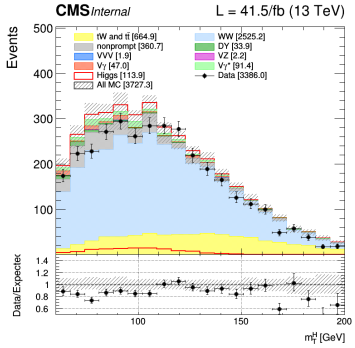


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

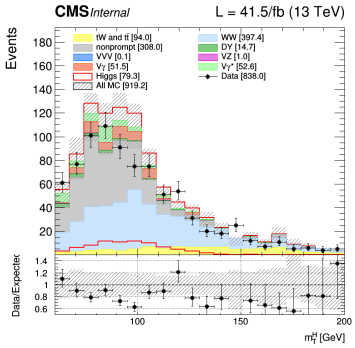
Figure A.32: Dilepton mass distributions in the 1-jet ggH subcategories for 2017



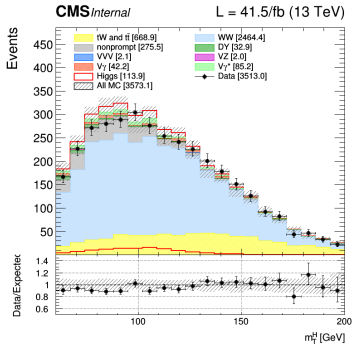
(a) $\mu^- e^+ p_{T2} < 20$ GeV 0-jet ggH



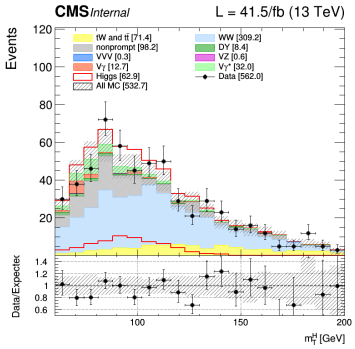
(b) $\mu^- e^+ p_{T2} > 20$ GeV 0-jet ggH



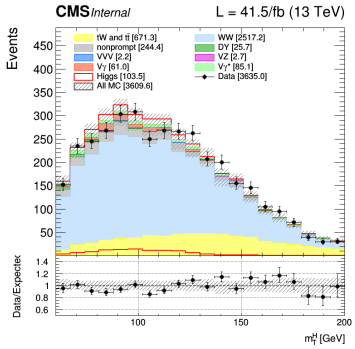
(c) $\mu^+ e^- p_{T2} < 20$ GeV 0-jet ggH



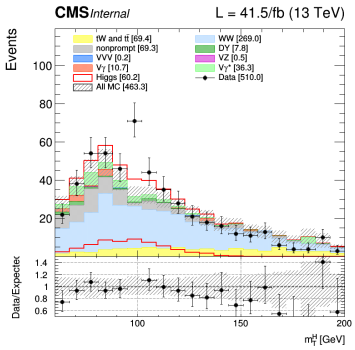
(d) $\mu^+ e^- p_{T2} > 20$ GeV 0-jet ggH



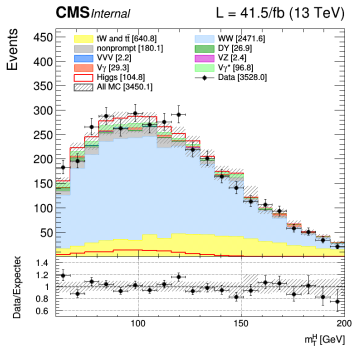
(e) $e^- \mu^+ p_{T2} < 20$ GeV 0-jet ggH



(f) $e^- \mu^+ p_{T2} > 20$ GeV 0-jet ggH

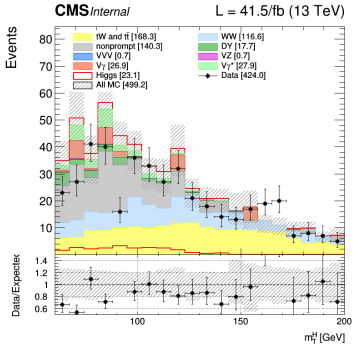


(g) $e^+ \mu^- p_{T2} < 20$ GeV 0-jet ggH

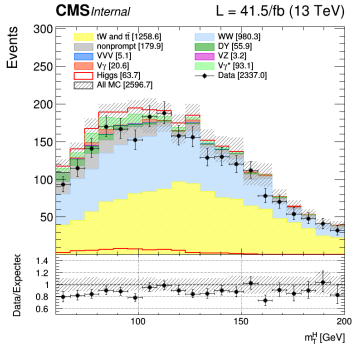


(h) $e^+ \mu^- p_{T2} > 20$ GeV 0-jet ggH

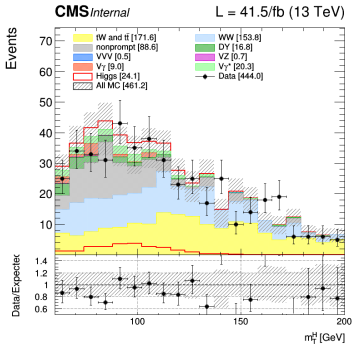
Figure A.33: Transverse mass distributions in the 0-jet ggH subcategories for 2017



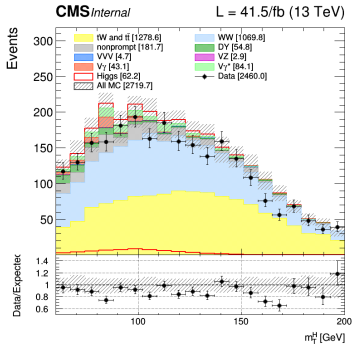
(a) $\mu^- e^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



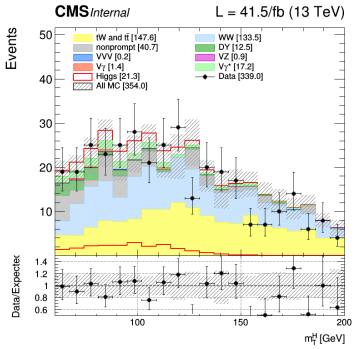
(b) $\mu^- e^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



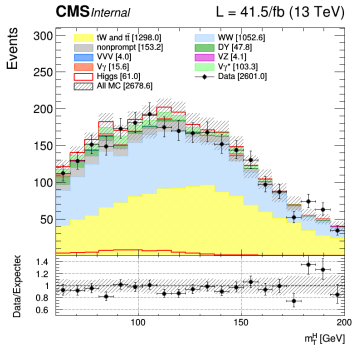
(c) $\mu^+ e^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



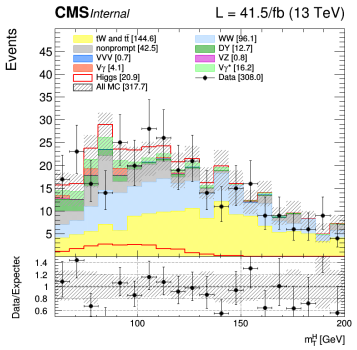
(d) $\mu^+ e^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH



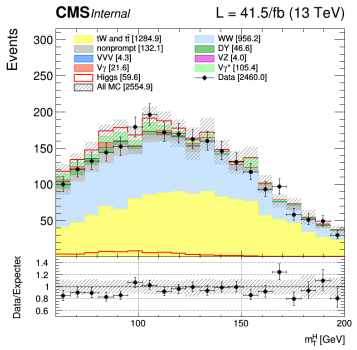
(e) $e^- \mu^+ p_{\mathrm{T}2} < 20$ GeV 1-jet ggH



(f) $e^- \mu^+ p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

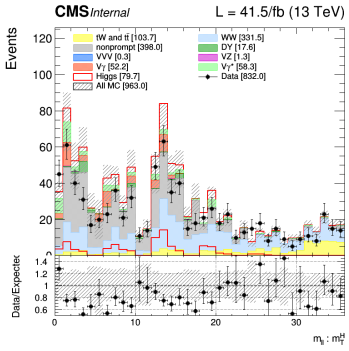


(g) $e^+ \mu^- p_{\mathrm{T}2} < 20$ GeV 1-jet ggH

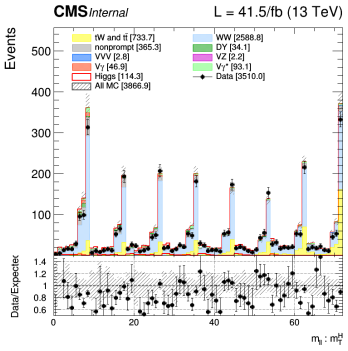


(h) $e^+ \mu^- p_{\mathrm{T}2} > 20$ GeV 1-jet ggH

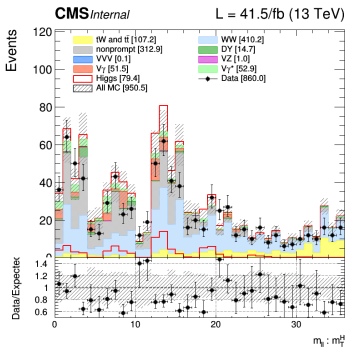
Figure A.34: Transverse mass distributions in the 1-jet ggH subcategories for 2017



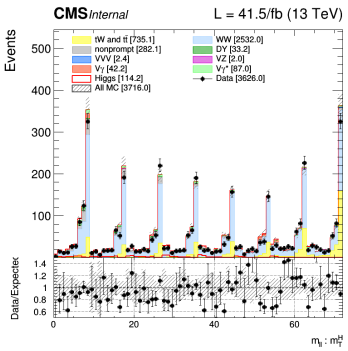
(a) $\mu^-e^+p_{\mathrm{T2}} < 20$ GeV 0-jet ggH



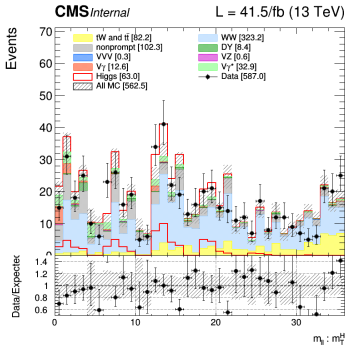
(b) $\mu^-e^+p_{\mathrm{T2}} > 20$ GeV 0-jet ggH



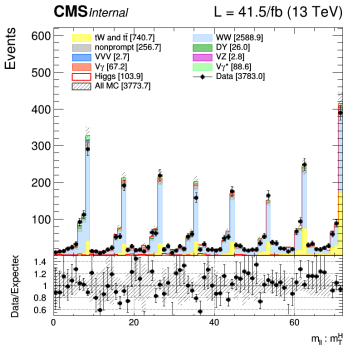
(c) $\mu^+e^-p_{\mathrm{T2}} < 20$ GeV 0-jet ggH



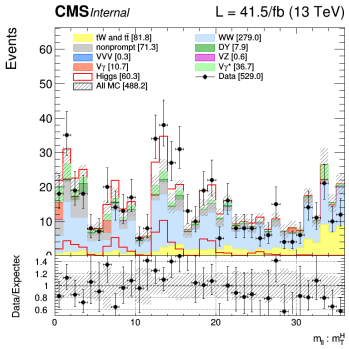
(d) $\mu^+e^-p_{\mathrm{T2}} > 20$ GeV 0-jet ggH



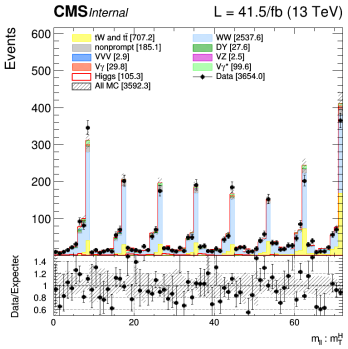
(e) $e^-\mu^+p_{\mathrm{T2}} < 20$ GeV 0-jet ggH



(f) $e^-\mu^+p_{\mathrm{T2}} > 20$ GeV 0-jet ggH

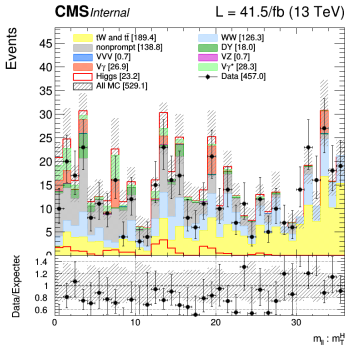


(g) $e^+\mu^-p_{\mathrm{T2}} < 20$ GeV 0-jet ggH

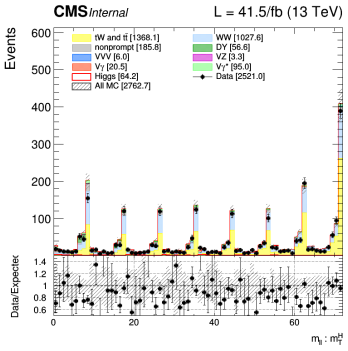


(h) $e^+\mu^-p_{\mathrm{T2}} > 20$ GeV 0-jet ggH

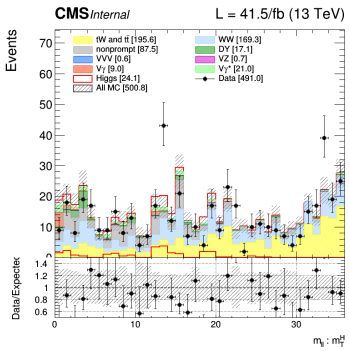
Figure A.35: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 0-jet ggH subcategories for 2017. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.



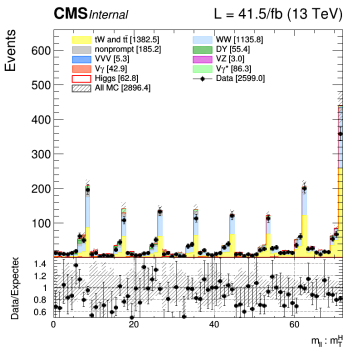
(a) $\mu^-e^+p_{\mathrm{T2}} < 20$ GeV 1 jet ggH



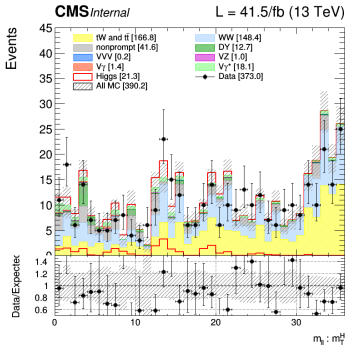
(b) $\mu^-e^+p_{\mathrm{T2}} > 20$ GeV 1 jet ggH



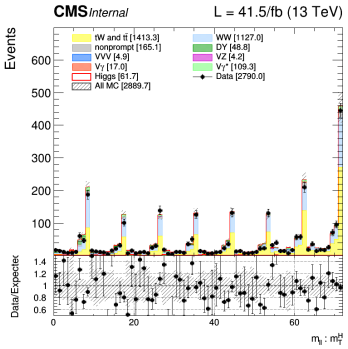
(c) $\mu^+e^-p_{\mathrm{T2}} < 20$ GeV 1 jet ggH



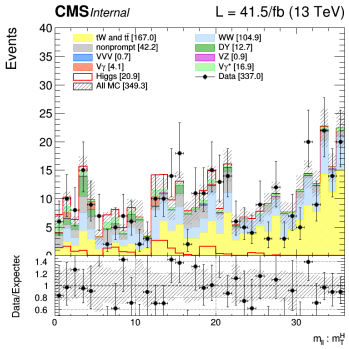
(d) $\mu^+e^-p_{\mathrm{T2}} > 20$ GeV 1 jet ggH



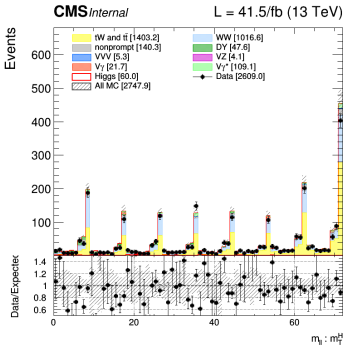
(e) $e^-\mu^+p_{\mathrm{T2}} < 20$ GeV 1 jet ggH



(f) $e^-\mu^+p_{\mathrm{T2}} > 20$ GeV 1 jet ggH



(g) $e^+\mu^-p_{\mathrm{T2}} < 20$ GeV 1 jet ggH



(h) $e^+\mu^-p_{\mathrm{T2}} > 20$ GeV 1 jet ggH

Figure A.36: Two-dimensional ($m_{\ell\ell}$,$m_{\mathrm{T}}^{\mathrm{H}}$) distributions in 1-jet ggH subcategories for 2017. The distributions are unrolled to one dimensional histograms such that identical values of $m_{\mathrm{T}}^{\mathrm{H}}$ are in adjacent bins.

Appendix B

Expected results in categories with two or more jets

In this appendix the expected results of the impacts from the systematic uncertainties in the ≥2-jet category using the three datasets are shown. Given the reduced number of events in this category, there are many statistical uncertainties among the 30 most relevant ones in each year. Of the experimental uncertainties, the first one in all years is the top p_{T} reweighting, the same thing that happened in the 1-jet categories. DY and top background normalisation and the theoretical uncertainties for ggH are the other common uncertainties in the three years that have an estimated effect of changing the signal strength up to 10%, approximately.

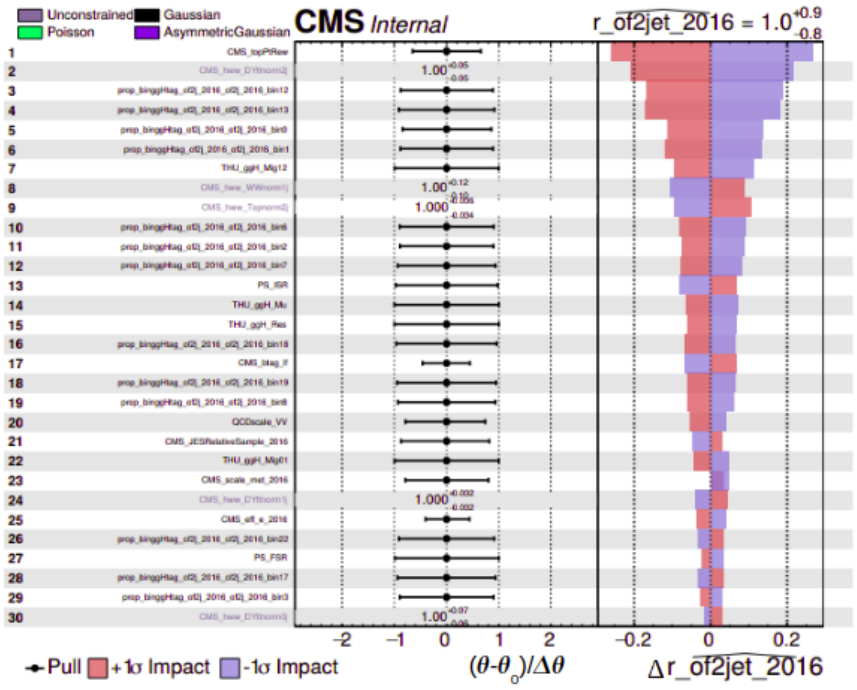


Figure B.1: Expected impacts of uncertainties in the ≥2-jet category in 2016.

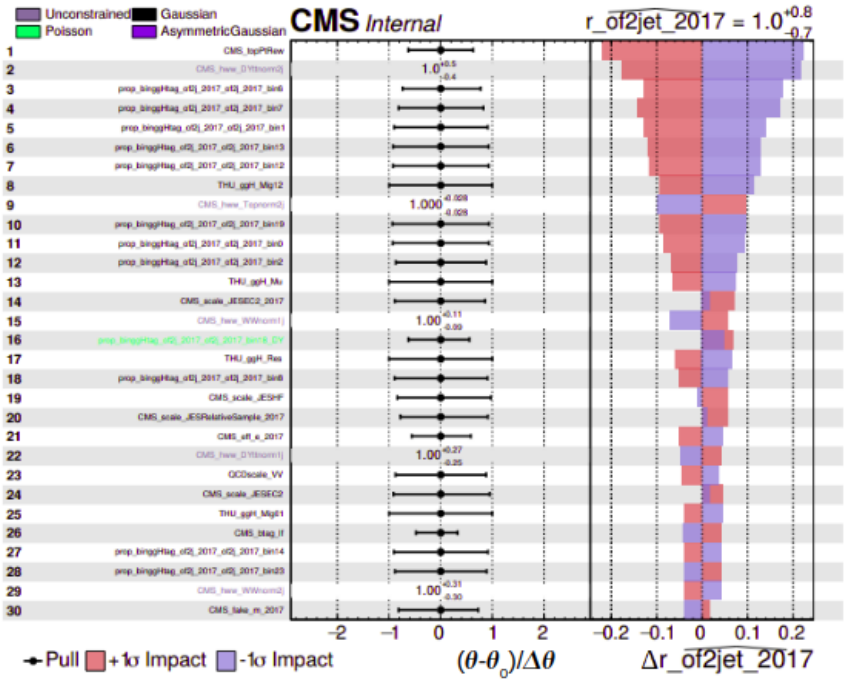


Figure B.2: Expected impacts of uncertainties in the ≥2-jet category in 2017.

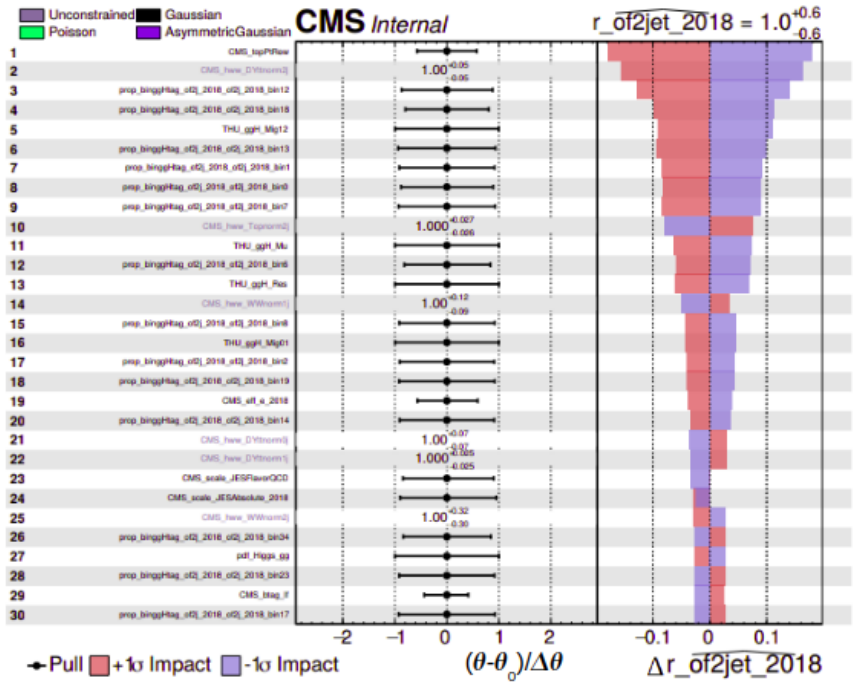


Figure B.3: Expected impacts of uncertainties in the ≥2-jet category in 2018.

Appendix C

Impacts from the uncertainties in 2016 and 2017 in the STXS analysis

In section 5.3.6 the uncertainties that most affect the result of the signal strengths of each STXS bin in 2018 are evaluated using impact plots. The thirty uncertainties that have the largest impacts in 2016 and 2017 in the 0- and 1-jet and high $p_{\mathrm{T}}^{\mathrm{H}}$ STXS bins are shown in Fig. C.1 and Fig. C.2. The uncertainties are classified according to their impact, with the highest ranked ones having the greatest effect in the signal strength.

In the 0-jet bin of 2016 there are no large pulls or constraints, and the high ranked non-statistical uncertainties are similar to the ones in 2018. The 2017 0-jet bin, however, has a large pull for the E_{T}^{miss} scale uncertainty and a very constrained statistical uncertainty due to the statistics in the second bin of the 1-jet high $p_{\mathrm{T}}^{\mathrm{H}}$ signal region. Considering the highest ranked uncertainties, 2017 0-jet has similar systematic uncertainties as the other years. One difference with respect to 2018 is the PS model uncertainty corresponding to the final state radiation, which is one of the most relevant ones in both 2016 and 2017 but not as much in 2018.

Moving on to the 1-jet bins, the systematic uncertainties with the highest impacts are similar to 2018 for both years, with b-tagging, top p_{T} reweighting, background normalisation and E_{T}^{miss} scale being the most important ones. It seems the E_{T}^{miss} scale might have something wrong which at this time is not understood, since in 2017 it causes large pulls in the 1-jet bins and in 2016 it has one-sided impacts.

Aside from the E_{T}^{miss} scale, the high $p_{\mathrm{T}}^{\mathrm{H}}$ bins have one-sided in the STXS acceptance uncertainty of the very high $p_{\mathrm{T}}^{\mathrm{H}}$ bins in 2016 and the DY QCD scale uncertainty in 2017. The combination of the bins, however, looks good in 2016, with no big asymmetries, pulls or constraints. The 2017 combination also looks good, with only a few very constrained statistical uncertainties.

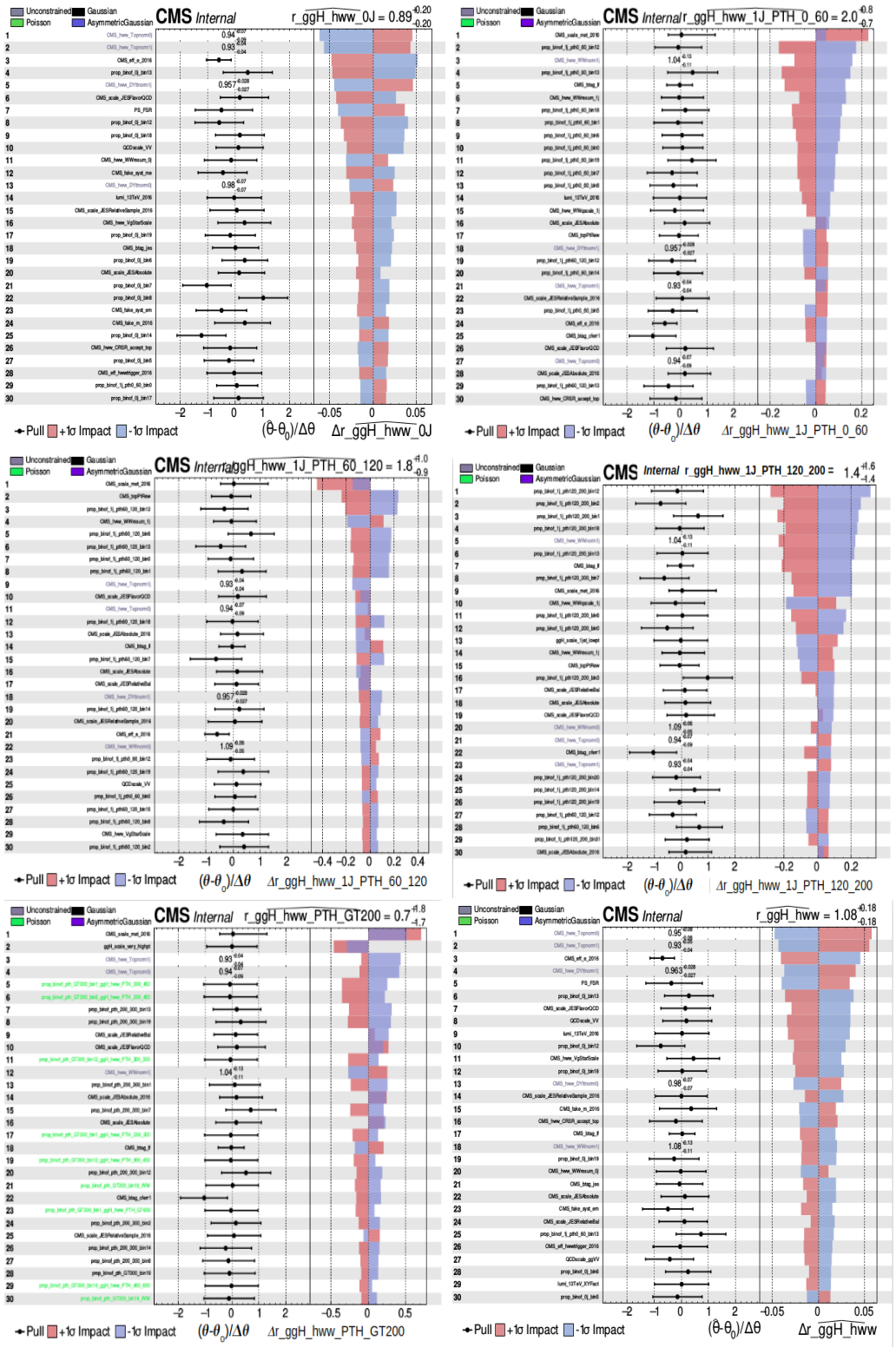


Figure C.1: Impact of the first 30 uncertainties with highest ranking on the observed signal strengths for 0-jet (top left), 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV (top right), 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV (middle left), 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV (middle right) and high $p_{\mathrm{T}}^{\mathrm{H}}$ (bottom left) ggH STXS bins, obtained with the 2016 dataset. The combination of the five bins is shown as well (bottom right).

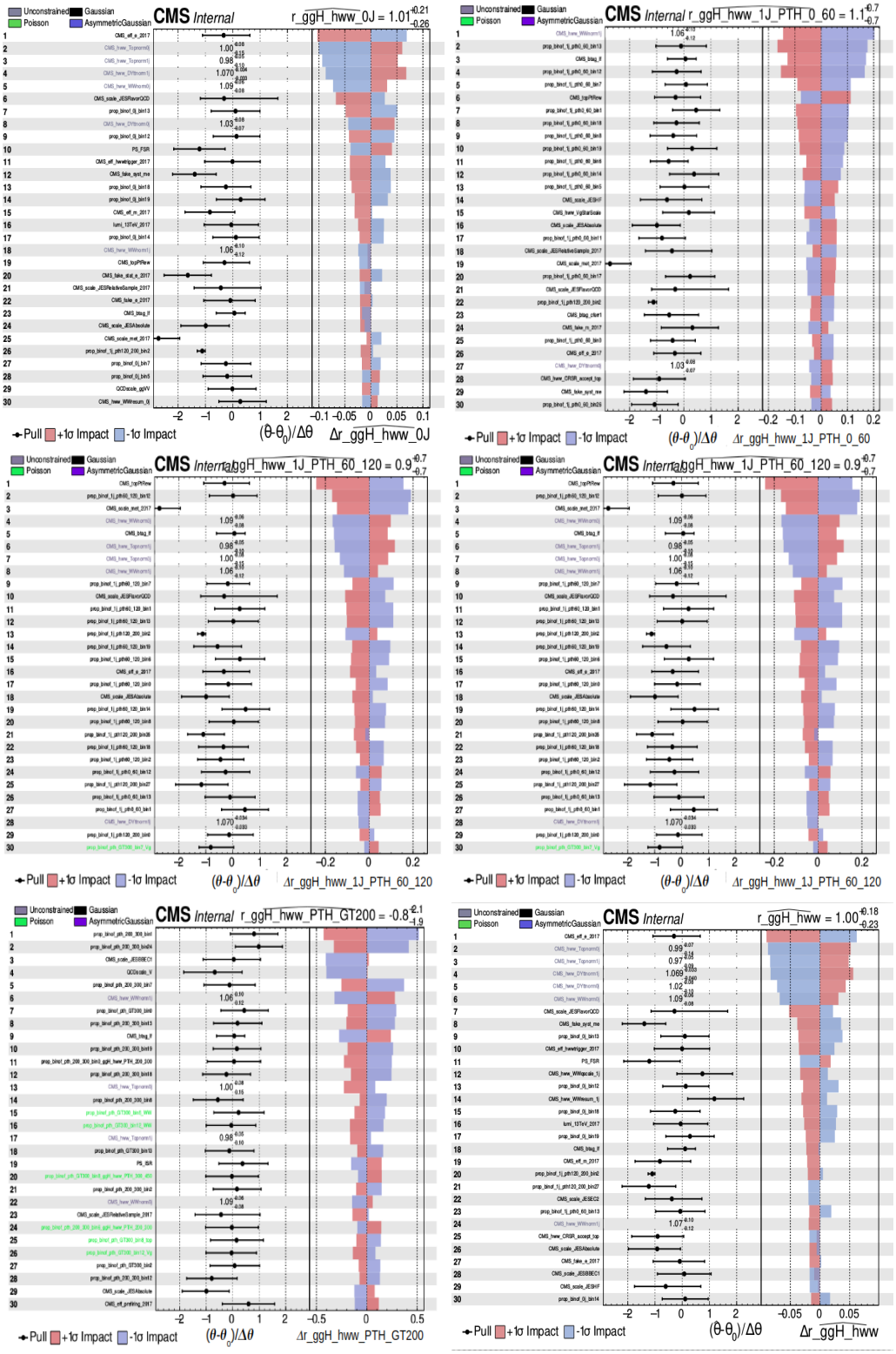


Figure C.2: Impact of the first 30 uncertainties with highest ranking on the observed signal strengths for 0-jet (top left), 1-jet $p_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV (top right), 1-jet $60 < p_{\mathrm{T}}^{\mathrm{H}} < 120$ GeV (middle left), 1-jet $120 < p_{\mathrm{T}}^{\mathrm{H}} < 200$ GeV (middle right) and high $p_{\mathrm{T}}^{\mathrm{H}}$ (bottom left) ggH STXS bins, obtained with the 2017 dataset. The combination of the five bins is shown as well (bottom right).

Bibliography

[1] General description of the Standard Model in wikipedia: URL: https://en.wikipedia.org/wiki/Standard_Model.

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

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

[4] N. Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Phys. Rev. Lett.* 10 (12 June 1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531. URL: https://link.aps.org/doi/10.1103/PhysRevLett.10.531.

[5] M. Kobayashi and T. Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652. eprint: https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-652.pdf. URL: https://doi.org/10.1143/PTP.49.652.

[6] S. Weinberg. *The Quantum theory of fields. Vol. 1: Foundations*. Cambridge University Press, June 2005.

[7] Website of the LHC Higgs Cross-section Working Group with public cross section results: URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG.

[8] 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.

[9] S. Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett.* B716.CMS-HIG-12-028, CERN-PH-EP-2012-220 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[10] S. Chatrchyan et al. "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV". In: *JHEP* 06.CMS-HIG-12-036, CERN-PH-EP-2013-035 (2013), p. 081. DOI: 10.1007/JHEP06(2013)081. arXiv: 1303.4571 [hep-ex].

[11] G. Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett.* B716.CERN-PH-EP-2012-218 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[12] A. M. Sirunyan et al. "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2017.11 (Nov. 2017). ISSN: 1029-8479. DOI: 10.1007/jhep11(2017)047. URL: http://dx.doi.org/10.1007/JHEP11(2017)047.

[13] CMS collaboration. *Measurements of properties of the Higgs boson in the four-lepton final state at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-HIG-18-001. CERN, 2018. URL: http://cds.cern.ch/record/2621419.

[14] CMS collaboration. *Measurements of Higgs boson production via gluon fusion and vector boson fusion in the diphoton decay channel at* $\sqrt{s} = 13$ *TeV*. Tech. rep. CMS-PAS-HIG-18-029. CERN, 2019. URL: http://cds.cern.ch/record/2667225.

[15] A. M. Sirunyan et al. "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2018.11 (Nov. 2018). ISSN: 1029-8479. DOI: 10.1007/jhep11(2018)185. URL: http://dx.doi.org/10.1007/JHEP11(2018)185.

[16] A. M. Sirunyan et al. "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector". In: *Physics Letters B* 779 (Apr. 2018), pp. 283–316. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.02.004. URL: http://dx.doi.org/10.1016/j.physletb.2018.02.004.

[17] CMS collaboration. *Measurement of Higgs boson production and decay to the $\tau\tau$ final state.* Tech. rep. CMS-PAS-HIG-18-032. CERN, 2019. URL: http://cds.cern.ch/record/2668685.

[18] CMS collaboration. *Search for the standard model Higgs boson decaying to charm quarks.* Tech. rep. CMS-PAS-HIG-18-031. CERN, 2019. URL: http://cds.cern.ch/record/2682638.

[19] A. M. Sirunyan et al. "Search for the decay of a Higgs boson in the $ll\gamma$ channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2018.11 (Sept. 2018). ISSN: 1029-8479. DOI: 10.1007/jhep11(2018)152. URL: http://dx.doi.org/10.1007/JHEP11(2018)152.

[20] A. M. Sirunyan et al. "Search for the Higgs Boson Decaying to Two Muons in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV". In: *Physical Review Letters* 122.2 (Jan. 2019). ISSN: 1079-7114. DOI: 10.1103/physrevlett.122.021801. URL: http://dx.doi.org/10.1103/PhysRevLett.122.021801.

[21] CMS collaboration. "Evidence for Higgs boson decay to a pair of muons". In: (2020). arXiv: 2009.04363 [hep-ex].

[22] M. Aaboud et al. "Measurements of gluon–gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s}$=13TeV with the ATLAS detector". In: *Physics Letters B* 789 (Feb. 2019), pp. 508–529. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.11.064. URL: http://dx.doi.org/10.1016/j.physletb.2018.11.064.

[23] S. Chatrchyan et al. "Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states". In: *Journal of High Energy Physics* 2014.1 (Jan. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep01(2014)096. URL: http://dx.doi.org/10.1007/JHEP01(2014)096.

[24] CMS collaboration. "Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s}$=13TeV". In: *Physics Letters B* 791 (2019), pp. 96–129. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2018.12.073. URL: http://www.sciencedirect.com/science/article/pii/S0370269319301169.

[25] CMS collaboration. "Measurement of the inclusive and differential Higgs boson production cross sections in the leptonic WW decay mode at $\sqrt{s} = 13$ TeV". In: (2020). arXiv: 2007.01984 [hep-ex].

[26] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev.* D98.3 (2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001.

[27] Website with public results of luminosity and pileup: URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[28] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001.

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

[30] ATLAS Collaboration et al. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (2008), S08003–S08003. DOI: 10.1088/1748-0221/3/08/s08003.

[31] ALICE collaboration. "ALICE: Technical proposal for a large ion collider experiment at the CERN LHC". In: CERN-LHCC-95-71, CERN-LHCC-P-3 (1995).

[32] LHCb collaboration. *LHCb reoptimized detector design and performance: Technical Design Report*. Technical Design Report LHCb CERN-LHCC-2003-030. CERN, 2003. URL: https://cds.cern.ch/record/630827.

[33] P. Lujan. "Van der Meer calibration of the CMS luminosity detectors in 2017". In: *PoS* ICHEP2018.CMS-CR-2018-339 (2019), p. 860. DOI: 10.22323/1.340.0860.

[34] A. M. Sirunyan et al. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13.CMS-MUO-16-001, CERN-EP-2018-058 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528.

[35] V. Karimäki et al. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS CERN-LHCC-98-006. CERN, 1997. URL: https://cds.cern.ch/record/368412.

[36] CMS collaboration. *The CMS tracker: addendum to the Technical Design Report*. Technical Design Report CMS CERN-LHCC-2000-016. CERN, 2000. URL: https://cds.cern.ch/record/490194.

[37] A. Dominguez et al. *CMS Technical Design Report for the Pixel Detector Upgrade*. Tech. rep. CERN-LHCC-2012-016. CMS-TDR-11. 2012. URL: https://cds.cern.ch/record/1481838.

[38] CMS collaboration. "2017 tracking performance plots". In: CMS-DP-2017-015 (Apr. 2017). URL: https://cds.cern.ch/record/2290524.

[39] G. Apollinari et al. "High Luminosity Large Hadron Collider HL-LHC". In: (2017). arXiv: 1705.08830 [physics.acc-ph].

[40] CMS collaboration. "The CMS Outer Tracker Upgrade for the HL-LHC". In: *Nucl. Instrum. Methods Phys. Res., A* 924 (2019), 256–261. 6 p. DOI: 10.1016/j.nima.2018.09.144. URL: http://cds.cern.ch/record/2717424.

[41] CMS collaboration. *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS CERN-LHCC-97-033. CERN, 1997. URL: https://cds.cern.ch/record/349375.

[42] S. Chatrchyan et al. "Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in *pp* Collisions at $\sqrt{s} = 7$ TeV". In: *JINST* 8.CMS-EGM-11-001, CERN-PH-EP-2013-097 (2013), p. 9009. DOI: 10.1088/1748-0221/8/09/P09009. arXiv: 1306.2016 [hep-ex].

[43] B. Clerbaux et al. "Saturation and Energy Corrections for TeV Electrons and Photons". In: (Dec. 2006).

[44] Z. Antunovic et al. "Radiation hard avalanche photodiodes for the CMS detector". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 537.1 (2005). Proceedings of the 7th International Conference on Inorganic Scintillators and their Use in Scientific adn Industrial Applications, pp. 379–382. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2004.08.047. URL: http://www.sciencedirect.com/science/article/pii/S016890020401856X.

[45] S. Baccaro et al. "Radiation damage effect on avalanche photodiodes". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 426.1 (1999), pp. 206–211. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(98)01493-4. URL: http://www.sciencedirect.com/science/article/pii/S0168900298014934.

[46] K. W. Bell et al. "Vacuum phototriodes for the CMS electromagnetic calorimeter endcap". In: *IEEE Transactions on Nuclear Science* 51.5 (2004), pp. 2284–2287.

[47] CMS collaboration. *The CMS hadron calorimeter project: Technical Design Report*. Technical Design Report CMS CERN-LHCC-97-031. CERN, 1997. URL: http://cds.cern.ch/record/357153.

[48] S. Abdullin et al. "Design, performance, and calibration of CMS hadron-barrel calorimeter wedges". In: *Eur. Phys. J.* C55.FERMILAB-PUB-08-246-CMS, CERN-CMS-NOTE-2006-138, CMS-NOTE-2006-138 (2008), pp. 159–171. DOI: 10.1140/epjc/s10052-008-0573-y.

[49] S. Chatrchyan et al. "Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data". In: *JINST* 5.CMS-CFT-09-009 (2010), T03012. DOI: 10.1088/1748-0221/5/03/T03012. arXiv: 0911.4991.

[50] G. Baiatian et al. "Design, Performance and Calibration of the CMS Forward Calorimeter Wedges". In: (Jan. 2008).

[51] J. G. Layter. *The CMS muon project: Technical Design Report*. Technical Design Report CMS CERN-LHCC-97-032. CERN, 1997. URL: https://cds.cern.ch/record/343814.

[52] L. Sonnenschein. "Drift velocity and pressure monitoring of the CMS muon drift chambers". In: (2011). arXiv: 1107.0358.

[53] V. Khachatryan et al. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020–P01020. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/01/p01020. URL: http://dx.doi.org/10.1088/1748-0221/12/01/P01020.

[54] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-k_T jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ex].

[55] CMS collaboration. "Fine synchronization of the CMS muon drift-tube local trigger using cosmic rays". In: *Journal of Instrumentation* 5.03 (Mar. 2010), T03004–T03004. DOI: 10.1088/1748-0221/5/03/t03004.

[56] M. Aldaya et al. "Fine synchronization of the CMS muon drift tubes local trigger". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 564.1 (2006), pp. 169–177. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2006.04.046. URL: http://www.sciencedirect.com/science/article/pii/S0168900206006632.

[57] CMS collaboration. "Calibration of the CMS drift tube chambers and measurement of the drift velocity with cosmic rays". In: *Journal of Instrumentation* 5.03 (Mar. 2010), T03016–T03016. DOI: 10.1088/1748-0221/5/03/t03016.

[58] CMS Collaboration et al. "CMS The Computing Project, Technical Design Report". In: 23 (July 2005), pp. 1–137.

[59] Website explaining the CMS computing model: URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel.

[60] S. Agostinelli et al. "GEANTfour—a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), p. 250. DOI: 10.1016/S0168-9002(03)01368-8.

[61] R. Brun and F. Rademakers. "ROOT — An object oriented data analysis framework". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.1 (1997). New Computing Techniques in Physics Research V, pp. 81–86. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(97)00048-X. URL: http://www.sciencedirect.com/science/article/pii/S016890029700048X.

[62] CMS collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.CMS-PRF-14-001, CERN-EP-2017-110 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965.

[63] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *Journal of Instrumentation* 9.10 (2014), P10009–P10009. ISSN: 1748-0221. DOI: 10.1088/1748-0221/9/10/p10009. URL: http://dx.doi.org/10.1088/1748-0221/9/10/P10009.

[64] R. Frühwirth. "Application of Kalman filtering to track and vertex fitting". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.2 (1987), pp. 444–450. ISSN: 0168-9002. DOI: https://doi.org/10.1016/0168-9002(87)90887-4. URL: http://www.sciencedirect.com/science/article/pii/0168900287908874.

[65] CMS collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s}$ = 7TeV". In: *Journal of Instrumentation* 7.10 (2012), P10002–P10002. ISSN: 1748-0221. DOI: 10.1088/1748-0221/7/10/p10002. URL: http://dx.doi.org/10.1088/1748-0221/7/10/P10002.

[66] G. L. Bayatian et al. "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software". In: CERN-LHCC-2006-001, CMS-TDR-8-1 (2006).

[67] CMS collaboration. "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$= 8 TeV". In: *Journal of Instrumentation* 10.06 (2015), P06005–P06005. ISSN: 1748-0221. DOI: 10.1088/1748-0221/10/06/p06005. URL: http://dx.doi.org/10.1088/1748-0221/10/06/P06005.

[68] E. Meschi et al. "Electron Reconstruction in the CMS Electromagnetic Calorimeter". In: CMS-NOTE-2001-034, CERN-CMS-NOTE-2001-034 (June 2001).

[69] W. Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". In: *Journal of Physics G: Nuclear and Particle Physics* 31.9 (2005), N9–N20. DOI: 10.1088/0954-3899/31/9/n01.

[70] CMS collaboration. *Jet Performance in pp Collisions at 7 TeV*. Tech. rep. CMS-PAS-JME-10-003. CERN, 2010. URL: https://cds.cern.ch/record/1279362.

[71] CMS collaboration. *Pileup Removal Algorithms*. Tech. rep. CMS-PAS-JME-14-001. CERN, 2014. URL: https://cds.cern.ch/record/1751454.

[72] CMS collaboration. "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS". In: *JINST* 6.arXiv:1107.4277. CMS-JME-10-011. CERN-PH-EP-2011-102 (July 2011), P11002. 67 p. DOI: 10.1088/1748-0221/6/11/P11002. URL: http://cds.cern.ch/record/1369486.

[73] CMS collaboration. "Plans for Jet Energy Corrections at CMS". In: CMS-PAS-JME-07-002 (2008).

[74] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.CMS-BTV-16-002, CERN-EP-2017-326 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158.

[75] S. S. Ghosh. "Performance of MET reconstruction and pileup mitigation techniques in CMS". In: *Nuclear and Particle Physics Proceedings* 273-275 (2016), pp. 2512–2514. ISSN: 2405-6014. DOI: 10.1016/j.nuclphysbps.2015.09.442. URL: http://dx.doi.org/10.1016/j.nuclphysbps.2015.09.442.

[76] D. Bertolini et al. "Pileup Per Particle Identification". In: *JHEP* 10.FERMILAB-PUB-14-238-PPD, EFI-14-18, MIT-CTP-4558 (2014), p. 059. DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013 [hep-ph].

[77] A. M. Sirunyan et al. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s}$ = 13 TeV using the CMS detector". In: *JINST* 14 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004. arXiv: 1903.06078 [hep-ex].

[78] CMS collaboration. *CMS Luminosity Measurements for the 2016 Data Taking Period*. Tech. rep. CMS-PAS-LUM-17-001. 2017. URL: https://cds.cern.ch/record/2257069.

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

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

[81] T. Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191.LU-TP-14-36, MCNET-14-22, CERN-PH-TH-2014-190, FERMILAB-PUB-14-316-CD, DESY-14-178, SLAC-PUB-16122 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[82] R. D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04.EDINBURGH-2014-15, IFUM-1034-FT, CERN-PH-TH-2013-253, OUTP-14-11P, CAVENDISH-HEP-14-11 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[83] R. D. Ball et al. "Parton distributions with QED corrections". In: *Nucl. Phys.* B877.EDINBURGH-2013-20, IFUM-1014-FT, FR-PHENO-2013-008, CERN-PH-TH-2013-075 (2013), pp. 290–320. DOI: 10.1016/j.nuclphysb.2013.10.010. arXiv: 1308.0598 [hep-ph].

[84] R. D. Ball et al. "Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO". In: *Nucl. Phys.* B855.EDINBURGH-2011-14, IFUM-979-FT, FR-PHENO-2011-010, RWTH-TTK-11-24 (2012), pp. 153–221. DOI: 10.1016/j.nuclphysb.2011.09.024. arXiv: 1107.2652 [hep-ph].

[85] V. Khachatryan et al. "Event generator tunes obtained from underlying event and multiparton scattering measurements". In: *Eur. Phys. J. C* 76.CMS-GEN-14-001, CERN-PH-EP-2015-291 (2016), p. 155. DOI: 10.1140/epjc/s10052-016-3988-x. arXiv: 1512.00815 [hep-ex].

[86] A. M. Sirunyan et al. "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements". In: *The European Physical Journal C* 80.1 (Jan. 2020). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-7499-4. URL: http://dx.doi.org/10.1140/epjc/s10052-019-7499-4.

[87] S. Frixione, P. Nason, and C. Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". In: *JHEP* 11.BICOCCA-FT-07-9, GEF-TH-21-2007 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[88] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06.DESY-10-018, SFB-CPP-10-22, IPPP-10-11, DCPT-10-22 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[89] K. Hamilton, P. Nason, and G. Zanderighi. "Finite quark-mass effects in the NNLOPS POWHEG+MiNLO Higgs generator". In: *JHEP* 05.CERN-PH-TH-2015-006 (2015), p. 140. DOI: 10.1007/JHEP05(2015)140. arXiv: 1501.04637 [hep-ph].

[90] 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: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: http://dx.doi.org/10.1007/JHEP07(2014)079.

[91] CMS and ATLAS collaborations. "Combined Measurement of the Higgs Boson Mass in pp Collisions at sqrt[s]=7 and 8 TeV with the ATLAS and CMS Experiments". In: *Phys Rev Lett* 114 (May 2015), p. 191803.

[92] CMS collaboration. "A measurement of the Higgs boson mass in the diphoton decay channel". In: *Phys. Lett. B* 805.arXiv:2002.06398. CMS-HIG-19-004-003 (Feb. 2020), 135425. 37 p. DOI: 10.1016/j.physletb.2020.135425. URL: https://cds.cern.ch/record/2709657.

[93] K. Hamilton, P. Nason, and G. Zanderighi. "Finite quark-mass effects in the NNLOPS POWHEG+MiNLO Higgs generator". In: *Journal of High Energy Physics* 2015 (Jan. 2015). DOI: 10.1007/JHEP05(2015)140.

[94] K. Hamilton et al. "NNLOPS simulation of Higgs boson production". In: *JHEP* 10.MCNET-13-11, CERN-PH-TH-2013-205, OUTP-13-18P (2013), p. 222. DOI: 10.1007/JHEP10(2013)222. arXiv: 1309.0017 [hep-ph].

[95] S. Bolognesi et al. "On the spin and parity of a single-produced resonance at the LHC". In: *Phys. Rev. D* 86.ANL-HEP-PR-12-62, FERMILAB-PUB-12-475-PPD (2012), p. 095031. DOI: 10.1103/PhysRevD.86.095031. arXiv: 1208.4018 [hep-ph].

[96] The LHC Higgs Cross Section Working Group. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: FERMILAB-FN-1025-T, CERN-2017-002-M (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[97] The LHC Higgs Cross Section Working Group. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties". In: (2013). arXiv: 1307.1347 [hep-ph].

[98] J. M. Campbell and R. K. Ellis. "An Update on vector boson pair production at hadron colliders". In: *Phys. Rev. D* 60.FERMILAB-PUB-99-146-T (1999), p. 113006. DOI: 10.1103/PhysRevD.60.113006. arXiv: hep-ph/9905386 [hep-ph].

[99] J. M. Campbell, R. K. Ellis, and C. Williams. "Vector boson pair production at the LHC". In: *JHEP* 07.FERMILAB-PUB-11-182-T (2011), p. 018. DOI: 10.1007/JHEP07(2011)018. arXiv: 1105.0020 [hep-ph].

[100] J. M. Campbell, R. K. Ellis, and W. T. Giele. "A Multi-Threaded Version of MCFM". In: *Eur. Phys. J. C* 75.FERMILAB-PUB-15-043-T (2015), p. 246. DOI: 10.1140/epjc/s10052-015-3461-2. arXiv: 1503.06182.

[101] P. Jaiswal and T. Okui. "Explanation of the WW excess at the LHC by jet-veto resummation". In: *Phys. Rev. D* 90 (7 2014), p. 073009. DOI: 10.1103/PhysRevD.90.073009. URL: https://link.aps.org/doi/10.1103/PhysRevD.90.073009.

[102] P. Meade, H. Ramani, and M. Zeng. "Transverse momentum resummation effects in W^+W^- measurements". In: *Phys. Rev. D* 90 (11 2014), p. 114006. DOI: 10.1103/PhysRevD.90.114006. URL: https://link.aps.org/doi/10.1103/PhysRevD.90.114006.

[103] F. Caola et al. "QCD corrections to W+W- production through gluon fusion". In: *Physics Letters B* 754 (2016), pp. 275–280. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2016.01.046. URL: http://www.sciencedirect.com/science/article/pii/S0370269316000630.

[104] A. Hoecker et al. "TMVA - Toolkit for Multivariate Data Analysis". In: (2007). arXiv: physics/0703039 [physics.data-an].

[105] CMS collaboration. "Measurement of the associated production of a Higgs boson with a top quark pair in final states with electrons, muons and hadronically decaying τ leptons in data recorded in 2017 at $\sqrt{s} = 13$ TeV". In: CMS-PAS-HIG-18-019 (Nov. 2018).

[106] A. Bodek et al. "Extracting Muon Momentum Scale Corrections for Hadron Collider Experiments". In: *Eur. Phys. J. C* 72 (2012), p. 2194. DOI: 10.1140/epjc/s10052-012-2194-8. arXiv: 1208.3710 [hep-ex].

[107] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *Journal of Instrumentation* 13.05 (2018), P05011–P05011. ISSN: 1748-0221. DOI: 10.1088/1748-0221/13/05/p05011. URL: http://dx.doi.org/10.1088/1748-0221/13/05/P05011.

[108] A. M. Sirunyan et al. "Measurements of differential Z boson production cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2019.12 (2019). ISSN: 1029-8479. DOI: 10.1007/jhep12(2019)061. URL: http://dx.doi.org/10.1007/JHEP12(2019)061.

[109] A. M. Sirunyan et al. "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data". In: *JINST* 14.CMS-TAU-18-001, CERN-EP-2019-012 (2019), P06032. DOI: 10.1088/1748-0221/14/06/P06032. arXiv: 1903.01216 [hep-ex].

[110] CMS collaboration. "Measurement of differential cross sections for top quark pair production using the lepton + jets final state in proton-proton collisions at 13 TeV". In: *Phys. Rev. D* 95 (9 2017), p. 092001. DOI: 10.1103/PhysRevD.95.092001. URL: https://link.aps.org/doi/10.1103/PhysRevD.95.092001.

[111] J. Brochero et al. "Data Driven Estimates of Non-Prompt Leptons Background for the Higgs Boson Decaying to WW and WW Cross Section in the Leptonic Final State measurements". In: *CMS Physics Analysis Note* CMS-AN-13-184 (2013).

[112] R. J. Barlow and C. Beeston. "Fitting using finite Monte Carlo samples". In: *Comput. Phys. Commun.* 77 (1993), pp. 219–228. DOI: 10.1016/0010-4655(93)90005-W.

[113] ATLAS collaboration, CMS collaboration, and LHC Higgs Combination Group. *Procedure for the LHC Higgs boson search combination in Summer 2011.* Tech. rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11. CERN, 2011. URL: http://cds.cern.ch/record/1379837.

[114] G. Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0.

[115] J. Han. "W and Z Boson Cross Section and W Asymmetry at CMS". In: (2011). arXiv: 1110.0152 [hep-ex].

[116] CMS collaboration. *Measurements of properties of the Higgs boson in the four-lepton final state in proton-proton collisions at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-HIG-19-001. CERN, 2019. URL: http://cds.cern.ch/record/2668684.

[117] CMS collaboration. *Measurements of Higgs boson properties in the diphoton decay channel at* $\sqrt{s} = 13$ *TeV*. Tech. rep. CMS-PAS-HIG-19-015. CERN, 2020. URL: http://cds.cern.ch/record/2725142.

[118] ATLAS collaboration. "Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment". In: *Phys. Rev. D* 101 (1 Jan. 2020), p. 012002. DOI: 10.1103/PhysRevD.101.012002. URL: https://link.aps.org/doi/10.1103/PhysRevD.101.012002.

[119] Website of the Higgs Cross-section Working Group with a description of the STXS bins: URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGFiducialAndSTXS.