NNT: 2016SACLS254

THESE de DOCTORAT
de l'Université Paris-Saclay

préparée au Service de Physique des Particules du CEA Saclay

Ecole Doctorale 576: PHENIICS

Particules, Hadrons, Energie, Noyau, Instrumentation, Imagerie, Cosmos et Simulation

Spécialité: Physique des particules

par

Martina MACHET

Higgs boson production in the diphoton decay channel with CMS at the LHC: first measurement of the inclusive cross section in 13 TeV pp collisions, and study of the Higgs coupling to electroweak vector bosons

Soutenue publiquement à l'Orme des Merisiers le 26 septembre 2016 devant le jury:

Rapporteur et Président	M. Josè OCARIZ	Professeur (*CNRS/LPNHE Paris*)
Rapporteur	Mme Lucia DI CIACCIO	Professeur (*CNRS/LAPP Annecy-le-Vieux*)
Directeur de thèse	M. Patrick JARRY	Docteur d'état (*CEA Saclay*)
Encadrant	M. Fabrice COUDERC	Ingénieur de recherche (*CEA Saclay*)
Examinateur	M. Christophe GROJEAN	Professeur (*DESY Hambourg*)
Examinateur	M. Gautier HAMEL DE MONCHENAULT	Directeur de recherche (*CEA Saclay*)
Examinateur	M. Pascal PRALAVORIO	Directeur de recherche (*CNRS/CPPM Marseille*)

Titre: Production du boson de Higgs dans le canal de désintégration en 2 photons au LHC dans l'expérience CMS: première mesure de la section efficace inclusive dans des collisions proton-proton à 13 TeV, et étude du couplage de Higgs aux bosons vecteurs

Résumé: Dans ce document deux analyses des propriétés du boson de Higgs se désintégrant en 2 photons dans l'expérience CMS située auprès du LHC (Large Hadron Collider) sont présentées. Le document commence par une introduction théorique sur le Modèle Standard et sur la physique du boson de Higgs, suivie par une description détaillée de l'expérience CMS. En deuxième lieu, les algorithmes de réconstruction et identification des photons sont présentés, avec une attention particulière aux différences entre le premier et le deuxième run du LHC, le premier run (Run 1) ayant été pris entre 2010 et 2012 avec une énergie dans le centre de masse de 7 puis 8 TeV, le deuxième (Run 2) ayant commencé en 2015 avec une énergie dans le centre de masse de 13 TeV. Les performances des reconstructions du Run 1 et du Run 2 en ce qui concerne l'identification des photons sont comparées. Ensuite l'algorithme d'identification des photons pour l'analyse $H \to \gamma\gamma$ et optimisé pour le Run 2 est présenté. Pour ce faire une méthode d'analyse multivariée est utilisée. Les performances de l'identification des photons à 13 TeV sont enfin étudiées et une validation données-simulation est effectuée. Ensuite l'analyse $H \to \gamma\gamma$ avec les premières données du Run 2 est présentée. Les données utilisées correspondent à une luminosité intégrée de 12.9 fb^{-1}. Une catégorisation des événements est faite, afin de rendre maximale la signification statistique du signal et d'étudier les différents modes de production du boson de Higgs. La signification statistique observée pour le boson de Higgs du Modèle Standard à $m_H = 125.09$ GeV est 5.6 σ, pour une signification attendue de 6.2 σ, et la signification maximale de 6.1 σ est observée à $m_H = 126.0$ GeV. Enfin une étude de faisabilité ayant pour but de contraindre les couplages anomaux du boson de Higgs aux bosons de jauges est présentée. Pour cette analyse les données à 8 TeV collectées pendant le Run 1 du LHC, correspondant à une luminosité intégrée de 19.7 fb^{-1} sont utilisées. Cette analyse exploite la production du boson de Higgs par fusion de bosons-vecteurs (VBF), avec le Higgs se désintégrant ensuite en 2 photons. Les distributions cinématiques des jets et du Higgs, qui dépendent de l'hypothèse de spin-parité, sont utilisées pour construire des discriminants capables de séparer les différentes hypothèses de spin-parité. Ces discriminants permettent de définir différentes régions de l'espace des phases enrichies en signaux de spin-parité différent. Les différents nombres d'événements de signal sont extraits dans chaque région par un ajustement de la masse invariante diphoton, permettant de déterminer les contributions respectives des différents signaux et permettant ainsi de contraindre la production de boson de Higgs pseudo-scalaire (spin-parité 0^-).

Mots-clés: LHC, CMS, boson de Higgs, identification de photons, couplage de Higgs, fusion de bosons-vecteurs

Title: Higgs boson production in the diphoton decay channel with CMS at the LHC: first measurement of the inclusive cross section in 13 TeV pp collisions, and study of the Higgs coupling to electroweak vector bosons

Abstract: In this document two analyses of the properties of the Higgs boson in the diphoton decay channel with the CMS experiment at the LHC (Large Hadron Collider) are presented. The document starts with a theoretical introduction of the Standard Model and the Higgs boson physics, followed by a detailed description of the CMS detector. Then, photon reconstruction and identification algorithms are presented, with a particular focus on the differences between the first and the second run of the LHC, where the first run (Run 1) took place from 2010 to 2012 with a centre-of-mass energy of 7 and then 8 TeV, while the second run (Run 2) started in 2015 with a centre-of-mass energy of 13 TeV. Performances of Run 1 and Run 2 reconstructions from the photon identification point of view are compared. Then the photon identification algorithm for the $H \to \gamma\gamma$ analysis optimised for Run 2 is presented. To do that a multivariate analysis method is used. Performances of the photon identification at 13 TeV are finally studied and a data-simulation validation is performed. Afterwards, the $H \to \gamma\gamma$ analysis using the first Run 2 data is presented. The analysis is performed with a dataset corresponding to an integrated luminosity of 12.9 fb^{-1}. An event classification is performed to maximize signal significance and to study specific Higgs boson production modes. The observed significance for the Standard Model Higgs boson at $m_H = 125.09$ GeV is 5.6 σ, while 6.2 σ was expected, and the maximum significance of 6.1 σ is observed at $m_H = 126.0$ GeV.
Finally a feasibility study, having the aim of constraining the anomalous couplings of the Higgs boson to the vector bosons, is presented. This analysis is performed using the data collected at 8 TeV during Run 1 at the LHC, corresponding to an integrated luminosity of 19.7 fb^{-1}. This analysis exploits the production of the Higgs boson through vector boson fusion (VBF), with the Higgs decaying to 2 photons. The kinematic distributions of the di-jet system and the Higgs, which depend from the spin-parity hypothesis, are used to build some discriminants able to discriminate between different spin-parity hypotheses. These discriminants allow to define different regions of the phase-space enriched with a certain spin-parity process. The Higgs boson signal yield is extracted in each region from a fit to the diphoton mass, allowing to determine the contributions of the different processes and then constrain the production of a pseudo-scalar (spin-parity 0^-) Higgs boson.

Keywords: LHC, CMS, Higgs boson, photon identification, Higgs coupling, vector boson fusion

Contents

Introduction

The Standard Model is, up to now, the most successful theory of subatomic elementary particles. Developed in the early 1970s, it provides an elegant mathematical framework which describes how the fundamental constituents of the matter interact between each other, through the electromagnetic, weak and strong forces. Furthermore, it has successfully explained several experimental results and precisely predicted a wide variety of phenomena. The Standard Model of particle physics predicts the existence of a unique physical Higgs scalar boson associated to the spontaneous electroweak symmetry breaking, whose mass is a free parameter of the theory, and which is regarded as the responsible for the masses of all known elementary particles. This particle, whose search is one of the main goals of the LHC collider and its experiments installed at the CERN laboratory in Geneva, was discovered by both ATLAS and CMS experiments in 2012, with a measured mass of $\sim$ 125 GeV. After the Higgs boson discovery the main objectives are the measurement of its properties and the tests of consistency with the Standard Model.
The inclusive production of SM Higgs boson followed by the decay to two photons is one of the most sensitive channels for the Higgs search at $\sim$ 125 GeV. In fact, despite its very small rate, it has a clear experimental signature thanks to an excellent diphoton mass resolution.
The general context of this thesis is therefore the measurement of the Higgs boson properties in the $H \to \gamma\gamma$ decay channel. The analysed datasets were collected in proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 8$ TeV and $\sqrt{s} = 13$ TeV, recorded with the CMS detector, comprising a total integrated luminosity of 19.7 fb^{-1} and 12.9 fb^{-1} respectively.

In Chapter 1, a theoretical introduction of the Standard Model and the Higgs boson physics is given, along with the presentation of the main Higgs results achieved at the LHC during Run 1.
In Chapter 2, after an introduction to the LHC performances, the CMS detector is described, with particular focus on the electromagnetic calorimeter, the sub-detector used to identify and reconstruct photons.
In Chapter 3 photon reconstruction and identification algorithms are presented, concen-

trating in particular on the differences between the first and the second run of the LHC. The first run (Run 1) took place from 2010 to 2012 with a centre-of-mass energy of 7 and then 8 TeV, while the second run (Run 2) started in 2015 with a centre-of-mass energy of 13 TeV. Performances of Run 1 and Run 2 reconstructions from the photon identification point of view are compared. Then the photon identification algorithm for the $H \to \gamma\gamma$ analysis optimised for Run 2 is presented. To do that a multivariate analysis method is used. Performances of the photon identification at 13 TeV are finally studied and a data-simulation validation is performed.

Chapter 4 presents the $H \to \gamma\gamma$ analysis using the first Run 2 data. The analysis is performed with a dataset corresponding to an integrated luminosity of 12.9 fb^{-1}. An event classification is performed to maximize signal significance and to study specific Higgs boson production modes.

Finally a feasibility study, having the aim of constraining the anomalous couplings of the Higgs boson to the vector bosons, is presented in Chapter 5. This analysis is performed using the data collected at 8 TeV during Run 1 at the LHC, corresponding to an integrated luminosity of 19.5 fb^{-1}. This analysis exploits the production of the Higgs boson through vector boson fusion (VBF), with the Higgs decaying to 2 photons. The kinematic distributions of the dijet and diphoton systems, which depend from the spin-parity hypothesis, are used to build some discriminants able to discriminate between different spin-parity hypotheses. These discriminants allow to define different regions of the phase-space enriched with a certain spin-parity process. The Higgs boson signal yield is extracted in each region from a fit to the diphoton mass, allowing to determine the contributions of the different processes and then constrain the production of a pseudo-scalar Higgs boson.

Chapter 1

The Standard Model Higgs boson at LHC

The fundamental components of matter and their interactions are nowadays best described by the Standard Model of Particle Physics (SM) [1, 2, 3], a quantum field theory which describes the electroweak interaction (Glashow-Weinberg-Salam model or GWS) and the strong interaction (Quantum Chromo-Dynamics or QCD). The SM predicts the existence of a single physical scalar boson, the Higgs boson, associated to the spontaneous electroweak symmetry breaking via the Brout-Englert-Higgs (BEH) mechanism [4, 5]. The mass m_H of this boson is a free parameter of the theory. The BEH mechanism gives origin to the mass of both fermions and gauge bosons, in agreement with experimental results. This occurs without explicitly breaking the gauge invariance, thus preserving the renormalizability of the theory.

After a brief introduction of the theoretical framework, in the following the BEH mechanism, the Higgs boson phenomenology at the LHC collider and the main Higgs results of the Run 1 at the LHC (data recorded from 2010 to 2012) are described.

1.1 The Standard Model of elementary particles

The SM describes the matter as composed by twelve elementary particles, the *fermions*, all having half-integer spin. Fermions can be divided into two main groups, *leptons* and *quarks*, whose classification is given in Table 1.1. Leptons can just interact via electroweak bosons, while quarks are subject to both strong and electroweak interactions. Moreover, quarks do not exist as free states, but only as elementary constituents of a wide class of particles, the *hadrons*, such as protons and neutrons.
In the SM the interactions between elementary particles are mediated by *bosons*, integer-spin particles. The main characteristics of bosons and of the corresponding interactions

Table 1.1: Classification of the three families of fundamental fermions.

Fermions	1^{st} fam.	2^{nd} fam.	3^{rd} fam.	Charge	Interactions
Quarks	u	c	t	$+\frac{2}{3}$	All
	d	s	b	$-\frac{1}{3}$	
Leptons	e	μ	τ	-1	Weak, Electromagnetic
	ν_e	ν_μ	ν_τ	0	Weak

Table 1.2: Properties of the three fundamental interactions (gravitational interaction is not taken into account).

	Electromagnetic	Weak	Strong
Quantum mediator	Photon (γ)	$W^{\pm}$, Z	Gluons
Mass [GeV/c^2]	0	80, 90	0
Coupling constant	$\alpha(Q^2 = 0) \approx \frac{1}{137}$	$\frac{G_F}{(\hbar c)^3} \approx 1.2 \cdot 10^{-5}\ GeV^{-2}$	$\alpha_s(m_Z) \approx 0.1$
Range [cm]	∞	10^{-16}	10^{-13}

are summarised in Table 1.2.
The gravitational interaction is not taken into account, as it is not relevant at the typical energy scales of particle physics.

This complex phenomenology arises from a mathematical formalism according to which the SM is a perturbatively renormalizable quantum field theory (QFT) based on the *local gauge symmetries* of its Lagrangian. According to Noether's theorem, a conservation law corresponds to each of these local invariances, explaining why they are so important. The SM is therefore a local gauge quantum field theory describing three of the four fundamental interactions: electromagnetic, weak and strong interaction. It is based on the symmetry group

$$SU(3)_C \otimes SU(2)_I \otimes U(1)_Y,$$

the direct product of $SU(3)_C$, the color symmetry group upon which Quantum Chromo Dynamics (QCD) is built, the gauge groups of weak isospin, $SU(2)_I$, and hypercharge, $U(1)_Y$. Electromagnetic and weak interactions are unified in the *electroweak* gauge group $SU(2)_I \otimes U(1)_Y$, upon which the Glashow-Weinberg-Salam Model is built.
Despite this symmetry predicts with precision and accuracy the phenomenology of particle interactions, it is broken by both fermion mass term and gauge boson mass term of the Lagrangian. A necessary ingredient of the SM is therefore that the electroweak symmetry is broken, which allows to introduce mass terms to the Lagrangian.

1.2 The electroweak theory

From a historical point of view, the starting point of the study of electroweak interactions is the Fermi's theory of muon decay [6], which is based on an effective four-fermion Lagrangian:

$$\mathcal{L} = -\frac{4G_F}{\sqrt{2}}\bar{\nu}_\mu\gamma^\alpha\frac{1-\gamma_5}{2}\mu\bar{e}\gamma_\alpha\frac{1-\gamma_5}{2}\nu_e, \tag{1.1}$$

where G_F is the Fermi coupling constant reported in Table 1.2, e, μ, ν_e and ν_μ are the fermionic fields of the electron, muon and electron and muon neutrinos respectively, while γ_α and γ_5 are Dirac matrices.
Equation 1.1 represents a "point-like" interaction, with only one vertex and without any intermediate boson exchanged. It is usually referred to as $V-A$ interaction, being formed by a vectorial and an axial component. The term $\frac{1}{2}(1-\gamma_5)$ that appears in it is the left-handed projector. Only the left-handed component of fermions takes part to this interaction.

Fermi's Lagrangian is not renormalizable and it results in a non-unitary scattering matrix. Both problems of renormalizability and unitarity are overcome, as already said, requiring the weak interaction Lagrangian to be invariant under local transformations generated by the elements of a Lie group (*gauge transformations*). The resulting Lagrangian must reduce to Equation 1.1 in the low energy limit.
A gauge theory for weak interactions is conceived as an extension of the theory of electromagnetic interaction, the Quantum Electro-Dynamics (QED), which is based on the gauge group $U(1)_{EM}$, associated to the conserved quantum number Q (electric charge). In this case, the condition of local invariance under the $U(1)_{EM}$ group leads to the existence of a massless vector boson, the *photon*.
A theory reproducing both the electromagnetic and weak interaction phenomenology is achieved by extending the gauge symmetry to the group $SU(2)_I \otimes U(1)_Y$. In this sense, the weak and electromagnetic interactions are said to be unified. The generator of $SU(2)_I$ is the weak isospin operator and the generator of $U(1)_Y$ is the weak hypercharge Y operator. The corresponding quantum numbers satisfy the Gell-Mann-Nishijima formula

$$Q = I_3 + \frac{Y}{2},$$

where I_3 is the third component of the weak isospin. Fermions can be divided in doublets of left-handed particles and singlets of right-handed particles, as follows:

$$L_L = \begin{pmatrix} \nu_{\ell,L} \\ \ell_L \end{pmatrix},\ \ell_R,\ Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix},\ u_R,\ d_R, \tag{1.2}$$

where $\ell = e, \mu, \tau,\ u = u, c, t$ and $d = d, s, b$. In Table 1.3, I_3, Y and Q quantum numbers of all fermions are reported. As well as for QED, the requirement of local gauge invariance

Table 1.3: Isospin (I_3), hypercharge (Y) and electric charge (Q) of all fermions.

	I_3	Y	Q
$\begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} \frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}$	$\begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$
u_R, d_R	0, 0	$\frac{4}{3}$, $-\frac{2}{3}$	$+\frac{2}{3}$, $-\frac{1}{3}$
$\begin{pmatrix} \nu_{\ell,L} \\ \ell_L \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$
ℓ_R	0	-2	-1

with respect to the $SU(2)_I \otimes U(1)_Y$ group introduces now four massless vector fields (gauge fields), $W_\mu^{1,2,3}$ and B_μ, which couple to fermions with two different coupling constants, g and g'. Note that B_μ does not represent the photon field. The gauge-invariant Lagrangian for fermion fields can be written as follows:

$$\mathcal{L} = \bar{\Psi}_L \gamma^\mu \Big(i\partial_\mu + g t_a W_\mu^a - \frac{1}{2} g' Y B_\mu \Big) \Psi_L + \bar{\psi}_R \gamma^\mu \Big(i\partial_\mu - \frac{1}{2} g' Y B_\mu \Big) \psi_R, \tag{1.3}$$

where

$$\Psi_L = \begin{pmatrix} \Psi_L^1 \\ \Psi_L^2 \end{pmatrix}$$

and where Ψ_L and Ψ_R are summed over all the possibilities in Equation 1.2. As already stated, $W_\mu^{1,2,3}$ and B_μ do not represent physical fields, which are given instead by linear combinations of the four mentioned fields: the charged bosons W^+ and W^- correspond to:

$$W_\mu^\pm = \frac{1}{\sqrt{2}} \left(W_\mu^1 \mp i W_\mu^2 \right), \tag{1.4}$$

while the neutral bosons γ and Z correspond to

$$A_\mu = B_\mu cos\theta_W + W_\mu^3 sin\theta_W \tag{1.5}$$

$$Z_\mu = -B_\mu sin\theta_W + W_\mu^3 cos\theta_W \tag{1.6}$$

obtained by mixing the neutral fields W_μ^3 and B_μ with a rotation angle being named the *Weinberg angle* θ_W. In terms of the fields in Equations 1.4 to 1.6, the interaction term between gauge fields and fermions, taken from the Lagrangian in Equation 1.3, becomes

$$\mathcal{L}_{int} = \frac{1}{2\sqrt{2}} g (J_\alpha^+ W^{(+)\alpha} + J_\alpha^- W^{(-)\alpha}) + \frac{1}{2} \sqrt{g'^2 + g^2} J_\alpha^Z Z^\alpha - e J_\alpha^{EM} A^\alpha, \tag{1.7}$$

where J^{EM} is the electromagnetic current connected to the photon field, while J^+, J^- and J^Z are the three weak isospin currents. A^α can then be identified with the photon field, yielding

$$g \sin\theta_W = g' \cos\theta_W = e. \tag{1.8}$$

The GWS model thus predicts the existence of two charged gauge fields, which only couple to left-handed fermions, and two neutral gauge fields, which interact with both left- and right-handed components. Note that the Z boson interacts differently with right and left part, while the photon does not.

1.3 The Higgs mechanism

In order to correctly reproduce the phenomenology of weak interactions, both fermion and gauge boson fields must acquire mass, in agreement with experimental results. Up to this point, however, all particles are considered massless: in the electroweak Lagrangian, in fact, a mass term for the gauge bosons would violate gauge invariance, which is needed to ensure the renormalizability of the theory. Masses are thus introduced with the BEH mechanism [4, 5, 7], which allows fermions and $W^{\pm}$, Z bosons to be massive, while keeping the photon massless. Such mechanism is accomplished by means of a doublet of complex scalar fields,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix}, \tag{1.9}$$

which is introduced in the electroweak Lagrangian within the term

$$\mathcal{L}_{BEH} = (D^\mu \phi)^\dagger (D_\mu \phi) + V(\phi^\dagger \phi), \tag{1.10}$$

where $D_\mu = \partial_\mu - igt_a W_\mu^a + \frac{i}{2} g' Y B_\mu$ is the covariant derivative. The Lagrangian in Equation 1.10 is invariant under $SU(2)_I \otimes U(1)_Y$ transformations, since the kinetic part is written in terms of covariant derivatives and the potential V only depends on the product $\phi^\dagger \phi$. The ϕ field is characterised by the following quantum numbers:

	I_3	Y	Q
$\begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Writing the potential term as follows (see also Figure 1.1 for a graphical representation)

$$V(\phi^\dagger \phi) = -\mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2, \tag{1.11}$$

with $\mu^2 < 0$ and $\lambda > 0$, it results to have a minimum for

$$\phi^\dagger \phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}. \tag{1.12}$$

This minimum is not found for a single value of ϕ, but for a manifold of non-zero values. The choice of (ϕ^+, ϕ^0) corresponding to the ground state, i.e. the lowest energy state or vacuum, is arbitrary, and the chosen point is not invariant under rotations in the (ϕ^+, ϕ^0)

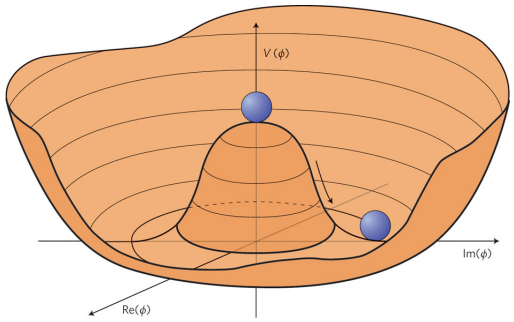


Figure 1.1: Shape of the Higgs potential of Equation 1.11.

plane: this is referred to as *spontaneous symmetry breaking*. If one chooses to fix the ground state on the ϕ^0 axis (if not the vacuum would be charged), the vacuum expectation value of the ϕ field is

$$< \phi > = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad v^2 = -\frac{\mu^2}{\lambda}. \tag{1.13}$$

The ϕ field can thus be rewritten in a generic gauge, in terms of its vacuum expectation value:

$$\phi = \frac{1}{\sqrt{2}} e^{\frac{i}{v}\phi^a t_a} \begin{pmatrix} 0 \\ H + v \end{pmatrix}, \quad a = 1, 2, 3, \tag{1.14}$$

where the three fields ϕ^a are called *Goldstone fields* and H is the Higgs boson scalar field. These 3 Goldstone bosons are massless and in the SM can actually be eliminated by choosing an ad hoc gauge named the *unitary gauge*, given by the transformation

$$\phi \to \phi' = e^{-\frac{i}{v}\phi^a t_a} \phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ H + v \end{pmatrix}. \tag{1.15}$$

The remaining field, the *Higgs field*, has now a zero expectation value.
Rewriting the Lagrangian in Equation 1.10 with the ϕ field in the unitary gauge, $\mathcal{L}_{BEH}$ can be written as:

$$\mathcal{L}_{BEH} = \frac{1}{2}\partial_\mu H \partial^\mu H - \frac{1}{2} m_H^2 - \sqrt{\frac{\lambda}{2}} m_H H^3 - \frac{\lambda}{4} H^4 + \left[m_W^2 W^{+\mu} W_\mu^- + \frac{m_Z^2}{2} Z^\mu Z_\mu \right] \left(1 + \frac{H}{v} \right)^2, \tag{1.16}$$

where

$$m_H = \sqrt{2\mu^2} = \sqrt{2\lambda} v. \tag{1.17}$$

Equation 1.16 now contains mass terms for fields $W^\pm$ and Z: each of the three gauge bosons has acquired mass and an additional degree of freedom, corresponding to the longitudinal polarization. At the same time, the three Goldstone bosons have disappeared from the Lagrangian $\mathcal{L}_{BEH}$, thus preserving the total number of degrees of freedom: the degrees

related to the missing Goldstone bosons have become the longitudinal degrees of the vector bosons. Only the H scalar field is still present and has acquired mass itself: it is the Higgs boson.
Summarizing, the BEH mechanism is used to explain weak boson masses without explicitly breaking the gauge invariance and thus preserving the renormalizability of the theory. When a symmetry is "spontaneously" broken, it is only "hidden" by the choice of the ground state. It can be shown that the minimum of the Higgs field is still invariant under the $U(1)_{EM}$ group. The electromagnetic symmetry is therefore unbroken and photons do not couple to the Higgs boson at tree level and remain massless.

1.3.1 Vector boson masses and couplings

The masses of vector bosons $W^{\pm}$ and Z are related to the parameter v, characteristic of the BEH mechanism, and to the electroweak coupling constants:

$$\begin{cases} m_W = \frac{1}{2}vg \\ m_Z = \frac{1}{2}v\sqrt{g^2+g'^2} \end{cases} \rightarrow \frac{m_W}{m_Z} = \frac{g}{\sqrt{g^2+g'^2}} = cos\theta_W. \tag{1.18}$$

The couplings of vector bosons to the Higgs are found to depend on the square of m_W and m_Z:

$$g_{HW} = \frac{1}{2}vg^2 = \frac{2}{v}m_W^2 \tag{1.19}$$

$$g_{HZ} = \frac{1}{2}v(g^2+g'^2) = \frac{2}{v}m_Z^2. \tag{1.20}$$

A relation between the decay ratios of the Higgs boson to a W pair and to a Z pair can be derived from Equations 1.19 and 1.20:

$$\frac{BR(H \rightarrow W^+W^-)}{BR(H \rightarrow ZZ)} = \left(\frac{g_{HW}}{\frac{1}{2}g_{HZ}}\right)^2 = 4\left(\frac{m_W^2}{m_Z^2}\right)^2 \simeq 2.4.$$

Finally, the electroweak symmetry breaking energy scale can be determined from the relation between the v parameter and the Fermi constant G_F:

$$v = \left(\frac{1}{\sqrt{2}G_F}\right)^{\frac{1}{2}} \simeq 246\ GeV. \tag{1.21}$$

1.3.2 Fermion masses and couplings

Fermion fields can be split into a left-handed and right-handed part, that are chirality eigenstates, according to:

$$\psi = \psi_L + \psi_R\ , \psi_{L,R} = \frac{\left(1 \mp \gamma^5\right)}{2}\psi \tag{1.22}$$

Left and right part of fermion fields fill different multiplets of electroweak gauge group, to account for parity violation of weak interactions. For the first generation one can write:

$$q_L \equiv \begin{pmatrix} u_L \\ d_L \end{pmatrix}, u_R, d_R, l_L \equiv \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}, e_R \tag{1.23}$$

It is not possible to introduce in the fermion lagrangian an explicit mass term like $m\bar{\psi}\psi$ because it is not gauge invariant: from Equation (1.23) it is clear that ψ_L and ψ_R behave differently under $SU(2)$ transformations. However, with a Higgs doublet as introduced in the GWS model there is a gauge invariant interaction that looks like a fermion mass term when the Higgs gets its vacuum expectation value, that is called *Yukawa coupling*. For the electron:

$$\mathcal{L}_{Yuk} = -Y_e \bar{l_L} \phi e_R + h.c., \tag{1.24}$$

where *h.c.* is the hermitian conjugate. Using unitary gauge (Equation 1.15) one gets, for the terms proportional to v:

$$-\frac{Y_e v}{\sqrt{2}} (\bar{e_L} e_R + \bar{e_R} e_L) = \frac{Y_e v}{\sqrt{2}} \bar{e}e \Rightarrow m_e = \frac{Y_e v}{\sqrt{2}} \tag{1.25}$$

From Equation 1.25 one can see that electron has acquired a mass, proportional to v. Like for vector bosons W and Z, also fermions get mass from spontaneous symmetry breaking: the mass is proportional to Y_f, which is the strength of the coupling of the fermion f to the Higgs.
For the down quarks (*d, s, b*), the Yukawa coupling term is the same as for electron:

$$\mathcal{L}_{Yuk} = -Y_d \bar{q_L} \phi d_R + h.c. \tag{1.26}$$

but for the up quarks (*u, c, t*) something different is needed, because the vacuum expectation value of the Higgs field is placed in the down part of the doublet (see Equation 1.13). Defining:

$$\phi^C = \frac{1}{\sqrt{2}} \begin{pmatrix} H + v \\ 0 \end{pmatrix} \tag{1.27}$$

one can get a new Yukawa invariant interaction for the up quarks:

$$\mathcal{L}_{Yuk} = -Y_u \bar{q_L} \phi^C u_R + h.c. \tag{1.28}$$

Finally the mass terms also for up and down quarks can be derived:

$$m_u = \frac{Y_u v}{\sqrt{2}}\ ,\ m_d = \frac{Y_d v}{\sqrt{2}}.$$

In this brief description we have omitted all the cross terms which give couplings and which are eliminated by means of the CKM matrix giving rise to CP violation. Note that this important mechanism is the only known source of CP violation in the electroweak sector of the SM.

1.4 The Higgs boson phenomenology at the LHC

The electroweak theory has been extensively tested in the last thirty years of the 20th century, proving that the Standard Model offers a valid explanation of the nature of particle interactions. The Higgs boson has been the only missing piece for more than three decades and has been searched by experiments at LEP, Tevatron and LHC until it was discovered by both ATLAS and CMS experiments at the LHC in 2012, with a measured mass of about 125 GeV [8, 9].

1.4.1 Higgs boson production

While the Higgs boson mass cannot be predicted by the theory, the Higgs couplings to the fermions and bosons are predicted to be proportional to the corresponding particle mass. For this reason the Higgs production and decay processes are dominated by channels involving the coupling of Higgs boson to heavy particles, especially to $W^{\pm}$ and Z bosons and to the third generation of fermions. About the other gauge bosons, the Higgs does not couple to photons and gluons at tree level, but only by one-loop diagram, where the main contribution is given by t loops, for the $gg \to H$ channel, and by W^+W^- and t loops for the $\gamma\gamma \to H$ channel.
In proton-proton collisions at $\sqrt{s} = 7 - 14$ TeV, like those at the Large Hadron Collider, the main processes contributing to the Higgs boson production are represented by the Feynman diagrams in Figure 1.2. In Figure 1.3 (left) the Higgs cross section for the different production mechanisms is shown as a function of the centre-of-mass energy for a Higgs mass of 125 GeV. Figure 1.3 (right) shows how the total Higgs production cross section increases going from 7-8 TeV to 14 TeV.
Below the main Higgs production mechanisms are described.

Gluon-gluon fusion

The gluon gluon fusion (ggF) mechanism,

$$pp \to gg \to H, \tag{1.29}$$

is dominant at the LHC in the whole Higgs mass range. The coupling of the gluons to the Higgs boson is mediated through a triangular loop of virtual quarks where the t contribution plays the dominant role because of the large top mass. The theoretical cross section has been computed including the QCD corrections up to next-to-next-to-next-to-leading-order (N3LO) and next-to-next-to-leading-log (NNLL), whereas the electroweak corrections are known at next-to-leading order (NLO).

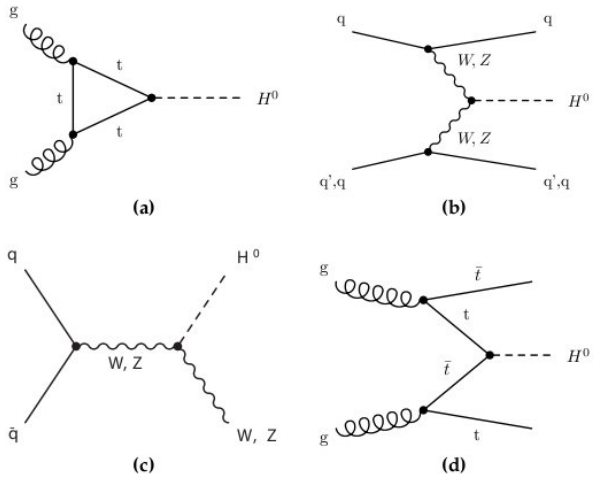


Figure 1.2: Higgs boson production mechanisms at tree level in proton-proton collisions: (a) gluon-gluon fusion; (b) VV fusion; (c) W and Z associated production (or Higgsstrahlung); (d) $t\bar{t}$ associated production.

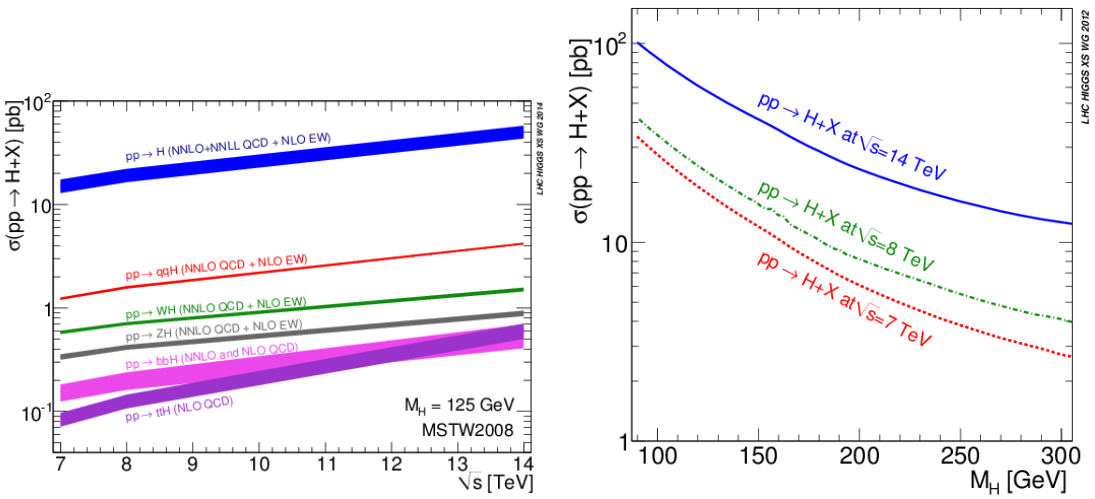


Figure 1.3: On the left, cross section for the different Higgs boson production mechanisms, as a function of the centre-of-mass energy for a Higgs mass of 125 GeV. On the right, total cross section for $pp \to H + X$ for $\sqrt{s} = 7$ TeV, $\sqrt{s} = 8$ TeV and $\sqrt{s} = 14$ TeV, in the Higgs mass range $100 < M_H < 300$ GeV.

Vector boson fusion

The vector boson fusion mechanism (VBF),

$$pp \to qq \to qqH, \tag{1.30}$$

is the second-most important mechanism at the LHC. For the Higgs boson at $m_H = 125$ GeV, its contribution to the total cross section is of the order of 8%. However, this channel is very interesting because of its clear experimental signature: the presence of two spectator jets with high invariant mass in the forward region provides a powerful tool to tag the signal events and discriminate the backgrounds, improving in this way the signal to background ratio, despite the low cross section. The VBF cross section has been calculated up to NNLO.

Associated production

The associated production mechanism (VH), described by the process

$$pp \to q\bar{q} \to VH, \tag{1.31}$$

is smaller due to the presence of an antiquark that, in a pp machine, does not come from the valence but from the sea. In this case, the selection of an event can be performed by reconstructing the original boson that radiated the Higgs. The VH cross section has been calculated including the QCD corrections up to NNLO, whereas the electroweak corrections are known at NLO.

Associated production with a $t\bar{t}$ pair

The production mechanism of the Higgs boson associated with a $t\bar{t}$ pair,

$$pp \to q\bar{q} \to t\bar{t}H, \tag{1.32}$$

is the production mechanism with the smallest yield, but the presence of a $t\bar{t}$ pair in the final state provides a clean experimental signature. Furthermore, going from 8 to 13 TeV its cross section increases by a factor of 4, and this production mode, giving a direct access to top quark coupling, plays an important role in testing the SM. This process has been computed with a precision up to NLO.

1.4.2 Higgs boson decay

The branching ratios of the different Higgs decay channels are shown in Figure 1.4 as a function of the Higgs mass. Fermion decay modes dominate in the low mass region (up to about 150 GeV). In particular, the channel $H \to b\bar{b}$ gives the largest contribution. When the decay channels into vector boson pairs open up, they quickly dominate. A peak in

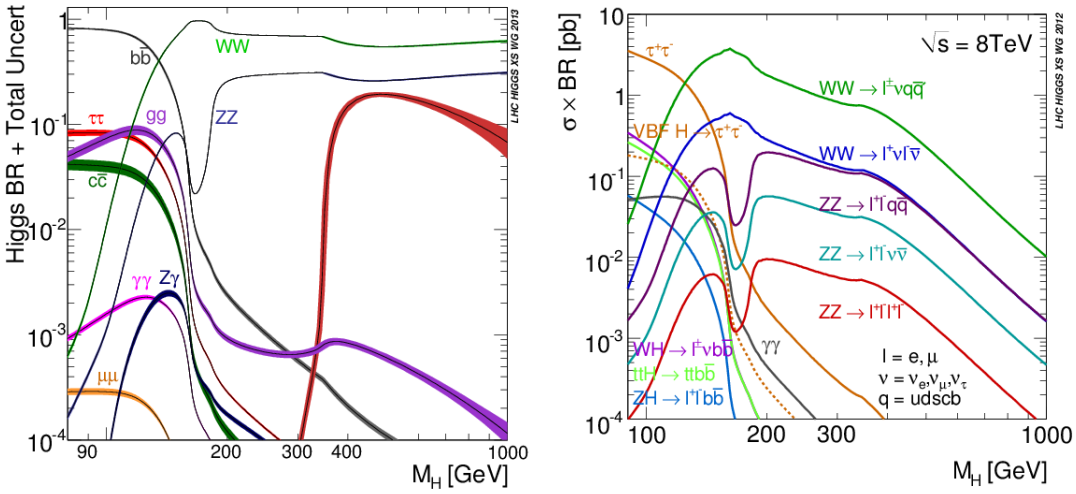


Figure 1.4: On the left, branching ratio of different Higgs boson decay channels as a function of the Higgs boson mass. On the right, Standard Model Higgs boson production cross section times branching ratio at $\sqrt{s} = 8$ TeV in the whole mass range up to 1 TeV [10, 11].

the $H \to W^+W^-$ channel is visible around 160 GeV, when the production of two on-shell W bosons becomes possible and the production of a real ZZ pair is still not allowed. At high masses (about 350 GeV) the $t\bar{t}$ channel opens. As shown in Figure 1.4, the branching ratios change dramatically across the possible mass range, requiring different strategies for the Higgs identification depending on its mass. The most promising decay channels for the Higgs discovery do not only depend on the corresponding branching ratios, but also on the capability to experimentally distinguish signal from background. Fully hadronic events are the most copious final states from Higgs boson decays, but they cannot be easily distinguished from QCD background. For this reason topologies with leptons or photons are preferred, even if they have smaller branching ratios.
Such channels are illustrated in the following, depending on the Higgs mass range:

- **High mass region** (180 GeV $< m_H <$ 1 TeV). This range is characterised by the fact that $m_H > 2m_W, 2m_Z$. The Higgs boson decays into two massive bosons, W^+, W^- or ZZ pairs, with a branching ratio of $\sim 70\%$ in the WW and $\sim 30\%$ in the ZZ final state. In the WW case, its $\sigma \times BR$ is high; however, the presence of neutrinos does not allow reconstruction of the final state: the measurement of the rate is important but difficult since several backgrounds can mimic the signals. On the other hand, the ZZ decay channel into 4 leptons, electrons or muons, possesses an excellent mass resolution as well as a negligible background rate.

- **Intermediate mass region** (130 GeV $< m_H <$ 180 GeV). The Higgs boson still decays into 4 fermions, through a pair of massive gauge bosons, one of them being

virtual. As m_H approaches 130 GeV, the decay channel into $b\bar{b}$ pairs is present as well, with a branching ratio $\sim 50\%$. When the WW threshold ($m_H \sim 2m_W$) is reached, the $b\bar{b}$ decay channel contribution gets less important. Until the production threshold of two real Z bosons ($m_H \sim 2m_Z$), the WW decay completely dominates the branching ratio.

- **Low mass region** (110 GeV $< m_H <$ 130 GeV). In the low mass range, the main decay channel is $b\bar{b}$, then $\tau^+\tau^-$, $c\bar{c}$ and gg (even if not accessible experimentally). In spite of its high value of $\sigma \times BR$, the $b\bar{b}$ mode is not easily accessible because of the overwhelming QCD background. The decay channels that are characterised by a virtual top or bottom loop, namely H $\to \gamma\gamma$ and $H \to Z\gamma$, are much rarer than $b\bar{b}$ or $\tau^+\tau^-$ although they are experimentally precious because of their very distinctive signature due to two high energetic photons forming a narrow invariant mass peak. Furthermore, H $\to \gamma\gamma$ decay channel only suffers from the $q\bar{q} \to \gamma\gamma$ and $Z \to e^+e^-$ backgrounds or jets faking photons. The expected signal rate is at least one order of magnitude smaller than the SM background rate.

Higgs boson to two photons decay channel

The photon is massless while the Higgs boson only couples to massive particles. Nevertheless the Higgs boson decays to two photons through a loop of massive charged particles, mainly W bosons and top quarks. The leading order Feynman diagrams, where the W loop and top quark loop interfere destructively, are shown in Figure 1.5.

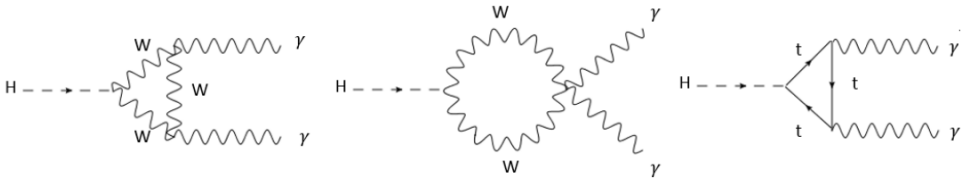


Figure 1.5: Leading order Feynman diagrams for Higgs boson decaying to two photons.

For $m_H = 125$ GeV the decay rate of the diphoton channel is smaller than those of the other four main channels (about 0.2%). Despite its very small rate, this channel is one of the most sensitive channels for the Higgs search in the low m_H region thanks to the two isolated high energy photons in the final state. In fact the photons, each carrying an energy of 62.5 GeV for a Higgs boson with $m_H = 125$ GeV decaying at the rest, are easily detected and identified. Their energies and momenta are well measured, from which the diphoton mass, $m_{\gamma\gamma}$, is reconstructed using the kinematic formula:

$$m_{\gamma\gamma} = \sqrt{2E^{\gamma 1}E^{\gamma 2}(1 - \cos(\theta_{\gamma\gamma}))}, \tag{1.33}$$

where $E^{\gamma 1}$ and $E^{\gamma 2}$ are the measured single photon energies and $\theta_{\gamma\gamma}$ is the measured angle between the momenta of the two photons.
A good resolution of both the measured photon energy and the measured open angle therefore leads to a narrow peak of diphoton mass spectrum associated to the Higgs resonance. Since the distribution of the background events is expected to be continuously falling, this narrow peak provides an eloquent evidence of the existence of the Higgs boson.
From the amplitude and the location of the peak, the relative total Higgs production cross section times the branching ratio to two photons with respect to the Standard Model Higgs expectation, the signal strength, and the Higgs mass are measured precisely. The rate of the decay, mediated through a loop of particles involving W boson and top quark, is sensitive to the magnitudes of both k_V and k_f as well as their relative sign, same sign for destructive interference between W loop and top quark loop as expected by the SM while opposite sign for constructive interference (see Section 1.5.4 for more details and for the definition of k_V and k_f). It is also sensitive to any possible new heavy charged particles in the loop, whose existence is quantified through measuring the effective Higgs coupling strength to photon, k_γ [12].
As explained in detail in Chapter 3, the dominant background consists of "irreducible" and "reducible" components. The "irreducible" component is real (prompt) diphoton events. The "reducible" component includes dijet and γ + jet events, in which jets are misidentified as photons. A jet typically fakes a photon when it results in a narrow concentration of photonic energy in the detector due to the decay of high energy neutral mesons, especially π^0's. The π^0 decays into two photons with small opening angle, which may appear as a single photon.

1.4.3 Higgs boson total width

The total width of the Higgs boson resonance is shown in Figure 1.6 as a function of m_H. Below the $2m_W$ threshold, the Higgs width is of the order of a few MeVs, then it rapidly increases, but it remains lower than 1 GeV up to $m_H \simeq 200$ GeV. The region at low mass is therefore the most challenging one, because the Higgs boson width is dominated by the experimental resolution. At 125 GeV, the Higgs boson width is expected to be 4 MeV.

1.5 Main Higgs results from LHC Run 1

As mentioned above, the Higgs boson was discovered by both ATLAS and CMS experiments at the LHC in 2012, with a measured mass of about 125 GeV [8, 9]. In this section the main results on the Higgs boson properties from LHC Run 1 are presented.

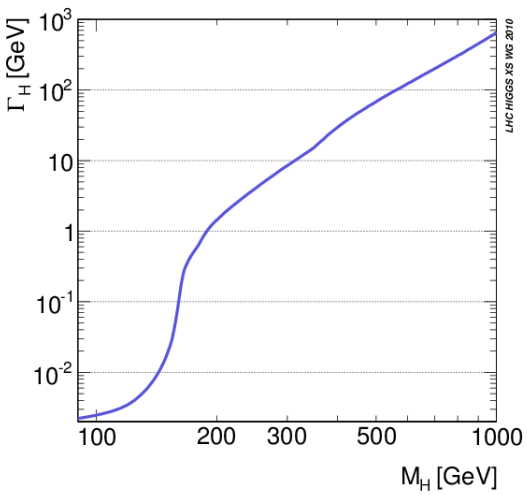


Figure 1.6: Total decay width of the Higgs boson as a function of its mass [10].

1.5.1 Mass measurement

The H $\to \gamma\gamma$ and $H \to ZZ \to 4\ell$ channels are used to determine the Higgs boson mass, thanks to their very good mass resolution (of about 1%) [13, 14]. Figure 1.7 shows the mass measurement for the two channels indipendently: the H $\to \gamma\gamma$ analysis, on the left, measures a narrow signal mass peak on top of a smoothly falling background mainly due to events from prompt nonresonant diphoton production. The plot on the right shows the four-lepton mass peak (in red) over a small continuum background mainly due to nonresonant ZZ production.
The mass measurements with H $\to \gamma\gamma$ and $H \to ZZ \to 4\ell$ data are combined assuming a single state [15]. Figure 1.8 on the left shows the scan of the test statistic as a function of the mass m_H separately for the H $\to \gamma\gamma$ and $H \to ZZ \to 4\ell$ channels, and for their combination. The production rates and decay ratios are left free in the m_H fit. The intersections of the likelihood curves with the 2 horizontal lines define the 68% and 95% CL confidence intervals for the mass of the observed particle. The mass is measured to be $m_H = 125.02 \pm 0.27(\text{stat.}) \pm 0.15(\text{syst.})$ GeV. The ATLAS+CMS combined result is: $125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})$ GeV [16].

Metastability of the EW vacuum

The discovery of the Higgs boson [8, 9] was expected to be the herald of new physics soon to be found at the TeV scale, but so far no signal of new physics nor any clear deviation from the SM Higgs properties have been detected at the LHC. Furthermore, the Higgs mass did not provide clear indications for new physics. The measured value (see previous section) is a bit high for supersymmetry and a bit low for composite models, making theoretical

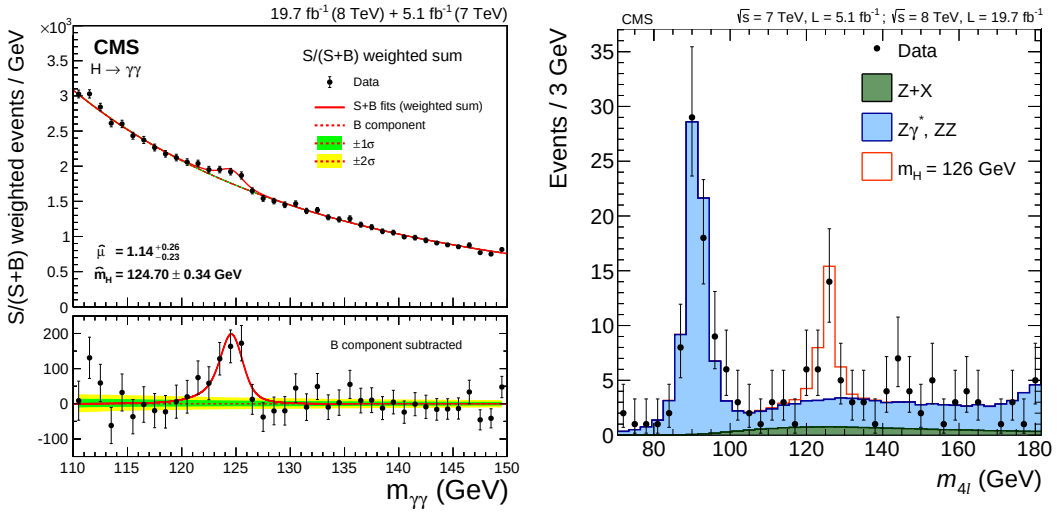


Figure 1.7: On the left, diphoton mass spectrum together with the background subtracted mass spectrum. On the right, distribution of the four-lepton reconstructed mass for the sum of the $4e$, $2e2\mu$ and 4μ channels. Shaded histograms represent the backgrounds, and the unshaded histogram represents the signal expectation for a mass hypothesis of $m_H = 125$ GeV.

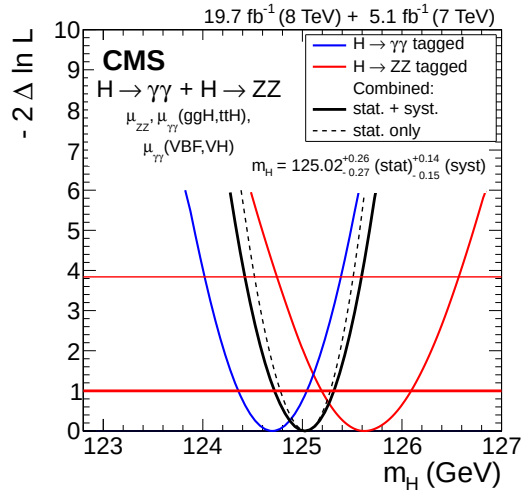


Figure 1.8: Scan of the test statistic $q(m_H)$= -2Δln$\mathcal{L}$ versus the mass of the boson m_H for the H $\to \gamma\gamma$ and $H \to ZZ \to 4\ell$ final states separately and for their combination.

interpretations more difficult. On the other hand the Higgs boson mass lies well within the parameter window in which the SM can be extrapolated up to the Planck mass, with no problem of consistency except remaining in the dark about naturalness. In the context of the SM, the measured Higgs boson mass is special because it corresponds to a near-critical situation in which the Higgs vacuum does not reside in the configuration of minimal energy, but in a metastable state close to a phase transition.
For decades, it has been understood that our universe might be in a metastable vacuum, i.e. a local minimum of the vacuum expectation value. If the universe was indeed in such a false vacuum state, a catastrophic bubble of more stable "true vacuum" could theoretically occur at any time and anywhere expanding outward at the speed of light.
The Standard Model of particle physics opens the possibility of calculating, from the masses of the Higgs boson and the top quark, whether the universe's present electroweak vacuum state is likely to be stable or merely long-lived. From the last calculations done using the measured value of m_H [17], it has been found that the Higgs boson mass lies very close to the boundary between stability and metastability regions. This result is presented in Figure 1.9, where the SM phase diagram in terms of Higgs and top pole masses is shown. The regions of stability, metastability, and instability of the EW vacuum are represented. The measured values of m_H and m_t appear to be rather special, in the sense that they place the SM vacuum in a near-critical condition, at the border between stability and metastability. The conclusion is that vacuum stability of the SM up to the Planck scale is excluded at 2.8σ (99.8% C.L. one-sided). The main source of uncertainty in this calculation comes from m_t, so any refinement in the measurement of the top mass is of great importance in view of the EW vacuum stability.

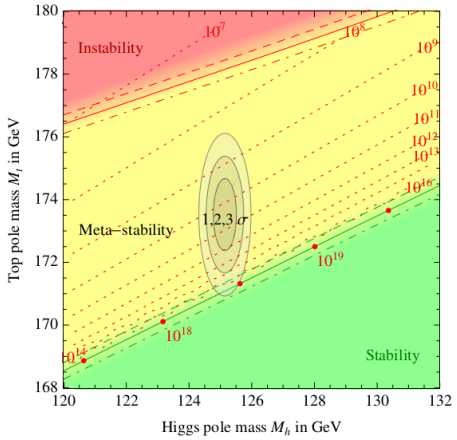


Figure 1.9: SM phase diagram in terms of Higgs and top pole masses in the region of the preferred experimental range of m_H and m_t. The plane is divided into regions of absolute stability, meta-stability and instability of the SM vacuum. The grey areas denote the allowed region at 1, 2, and 3σ.

1.5.2 Higgs boson width

The width of the SM Higgs boson is expected to be ∼4 MeV. Some upper limits on the width can be obtained through direct measurement from the mass peak, but they are limited by the experimental resolution (∼1 GeV). Another possibility is to use the off-shell/on-shell production and decay to 2 Z bosons ratio [18]. From this equation

$$\frac{d\sigma_{gg\to H\to ZZ}}{dm^2_{ZZ}} \sim \frac{g^2_{ggH} g^2_{HZZ}}{(m^2_{ZZ} - m^2_H)^2 + m^2_H \Gamma^2_H} \tag{1.34}$$

one can see that the gluon fusion production cross section depends on Γ_H through the Higgs boson propagator, where g_{ggH} and g_{HZZ} are the couplings of the Higgs boson to gluons and Z bosons, respectively. It was found that a measurement of the relative off-shell and on-shell production in the $H \to ZZ$ channel provides direct information on Γ_H, as long as the evolution of ggH and HZZ couplings as a function of the mass is the SM one. In Figure 1.10, the indirect measurement of the Higgs width, using the off-shell/on-shell ratio, is shown. CMS was able to put an upper limit of 22 MeV on the Higgs width, 5.4 times the expected value in the SM.

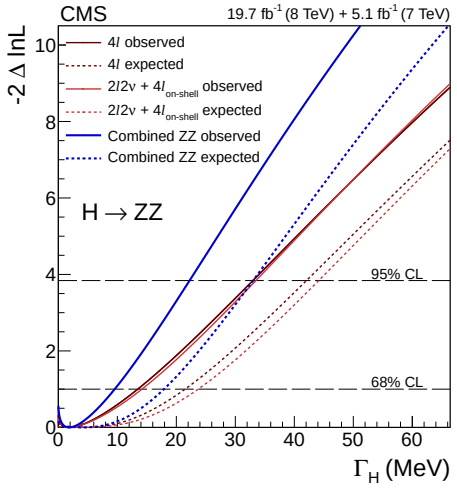


Figure 1.10: Scan of the negative log-likelihood, -2Δln$\mathcal{L}$, as a function of Γ_H for the combined fit of the 4ℓ and $2\ell 2\nu$ channels (blue thick lines), for the 4ℓ channel alone in the off-shell and on-shell regions (dark red lines), and for the $2\ell 2\nu$ channel in the off-shell region and 4ℓ channel in the on-shell region (light red lines). The solid lines represent the observed values, the dotted lines the expected values.

1.5.3 Spin-parity measurement

Concerning the spin-parity measurement, CMS was able to test some reasonable benchmark models for alternative J^{CP} hypotheses [19]. To do that, the angular distributions

were used, and the bosonic channels are the most sensitive: for example the $ZZ \to 4\ell$ system is fully reconstructed, and thanks to a Matrix Element Likelihood Approach that uses the kinematics of the decay products it is possible to distinguish between different spin-parity hypotheses. The WW final state is not fully reconstructed, but some kinematic variables, like $M_{\ell\ell}$ and M_T, are sensitive to the Higgs spin/CP. The $\gamma\gamma$ system can be used for spin determination through p_T and $cos\theta^*$ and it allows to exclude the spin-1 hypothesis.

The most general expression that one can write for HVV scattering amplitude is this one

$$A(\mathrm{HVV}) \sim \left[a_1^{\mathrm{VV}} + \frac{\kappa_1^{\mathrm{VV}} q^2_{\mathrm{V1}} + \kappa_2^{\mathrm{VV}} q^2_{\mathrm{V2}}}{\left(\Lambda_1^{\mathrm{VV}}\right)^2} \right] m^2_{\mathrm{V1}} \epsilon^*_{\mathrm{V1}} \epsilon^*_{\mathrm{V2}} + a_2^{\mathrm{VV}} f^{*(1)}_{\mu\nu} f^{*(2),\mu\nu} + a_3^{\mathrm{VV}} f^{*(1)}_{\mu\nu} \tilde{f}^{*(2),\mu\nu} \quad (1.35)$$

where a_1 is the scalar SM term, a_2 indicates a scalar anomalous, while a_3 corresponds to the pseudo-scalar component. In the SM $a_1 = 1$ and $a_2 = a_3 = 0$. Beyond the SM a_2 and a_3 can be different from zero.
The SM hypothesis 0^+ was tested against many alternative ones. Figure 1.11 shows the distributions of the test-statistic for the spin 2 J^P models tested against the SM Higgs boson hypothesis in the $X \to ZZ$ analyses. The expected median and the 68%, 95% and 99.7% CL regions are shown for the SM Higgs boson in orange and for the alternative J^P hypotheses in blue. The observed q values are indicated by the black circles.
Finally the 0^+ hypothesis is preferred to all the alternate tested models at more than the 3σ level.

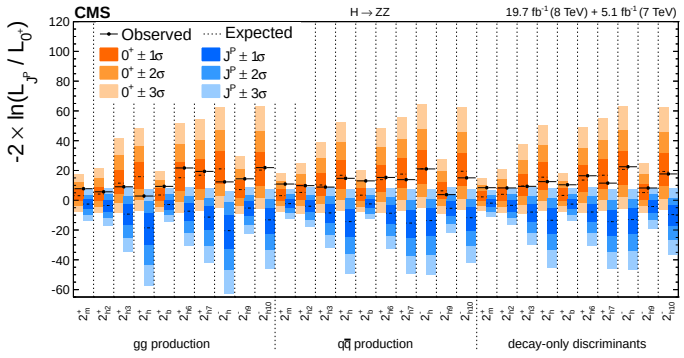


Figure 1.11: Distributions of the test statistic $q = -2ln(\mathcal{L}_{J^P}/\mathcal{L}_{0^+})$ for the spin-two J^P models tested against the SM Higgs boson hypothesis in the $X \to ZZ$ analyses. The expected median and the 68.3%, 95.4%, and 99.7% CL regions for the SM Higgs boson (orange, the left for each model) and for the alternative J^P hypotheses (blue, right) are shown. The observed q values are indicated by the black dots.

1.5.4 Higgs couplings

The couplings of the Higgs boson were also probed for deviations in magnitude from the SM predictions, and no significant deviations were found [15]. To test the observed data for possible deviations from the rates expected for the SM Higgs boson in the different channels, coupling modifiers were introduced, denoted by the scale factors k_i. The scale factors are defined for production processes by $k_i^2 = \sigma_i/\sigma_i^{SM}$, for decay processes by $k_i^2 = \Gamma_{ii}/\Gamma_{ii}^{SM}$, and for the total width by $k_H^2 = \Gamma_{tot}/\Gamma_{SM}$. It is possible to group the Higgs boson couplings into vectorial and fermionic sets fitting 2 parameters, k_V and k_f, which indicate the Higgs couplings with vector bosons and with fermions respectively. The $H \to \gamma\gamma$ is the only channel sensitive to the relative sign of k_V and k_f, because the partial width $\Gamma_{\gamma\gamma}$ is induced via loops with virtual W bosons or top quarks and scales as a function of both k_V and k_f. Figure 1.12 (left) shows the results of this combined analysis in the different decay channels and for the combination. One can see that the $H \to \gamma\gamma$ region is the only not symmetric in the 2 quadrants. Figure 1.12 (right) shows the scaling of the coupling of the Higgs as a function of the particle mass. Both plots indicate that the observation is compatible with the SM expectation.

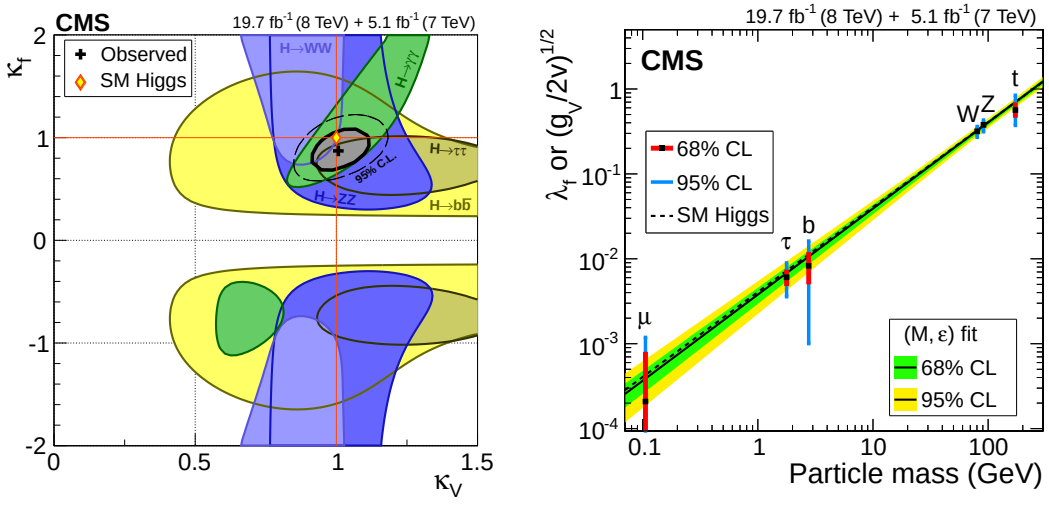


Figure 1.12: On the left the 68% CL confidence regions for individual channels (coloured swaths) and for the overall combination (thick curve) for the k_V and k_f parameters are shown. The cross indicates the global best-fit values. The dashed contour bounds the 95% CL confidence region for the combination. The diamond represents the SM expectation, (k_V,k_f) = (1,1). On the right the scaling of the coupling of the Higgs is shown as a function of the particle mass.

1.6 Beyond the SM Higgs

The Higgs mechanism as in the SM, conveniently solved several problems, like the existence of massive gauge bosons and the non-unitarity of longitudinal weak boson scattering. Nevertheless, despite its success at describing experiments, the SM fails to explain a number of phenomena observed in the universe.
It is thought that more than 95% of the known universe consists of dark matter ($\sim$ 27%) and dark energy ($\sim$ 68%) [20]. Since there is currently no way to explain either dark matter or dark energy within the SM, the latter can only attempt to explain about 5% of the energy of the universe.
Furthermore, the overabundance of matter, as opposed to anti-matter, in the universe, is a phenomenon known as the baryon asymmetry. It was shown by Sakharov [21] that a model of baryogenesis must satisfy three necessary conditions: baryon-number violation, charge-symmetry and charge-parity-symmetry violation (CP-violation), and interactions which are out of thermal equilibrium at early stages of the universe. Even if it has been shown that the SM contains the three necessary conditions for baryogenesis, it is believed to be insufficient to explain the degree of baryonic asymmetry in the visible universe [22, 23]. Additional sources of CP-violation in the SM would therefore provide a promising solution to the baryon-asymmetry problem.
The hierarchy problem, linked to the expected naturalness of electroweak symmetry breaking, is also often cited as a big reason to expect physics beyond the SM. The question is why the Higgs boson is so much lighter than the Planck mass: one would expect that the large quantum contributions to the square of the Higgs boson mass would inevitably make the mass huge, comparable to the scale at which new physics appears, unless there is an incredible fine-tuning cancellation between the quadratic radiative corrections and the bare mass [24]. Since the Higgs measured mass is $\sim$ 125 GeV, it should exist some kind of physics able to explain this 10^{17} fine-tuning of our SM parameters. For this reason, we expect there to be some kind of new physics accessible at TeV energies to explain why the Higgs should be right around that scale rather than being at the Planck mass.
It exists a number of proposed solutions to the fine-tuning problem, some of which could also provide solutions to some of the problems noted above, for example, Supersymmetry (SUSY) [24]. Since SUSY predicts that all fermions have a symmetry with a corresponding boson, all Feynman diagrams which provide quantum corrections to the Higgs boson mass have a canceling partner which removes the large quantum corrections. SUSY is also thought to provide a natural dark matter candidate and is a prerequisite for string theory, which naturally incorporates gravity. Finally, it is possible for SUSY to allow for additional CP-violation in the Higgs sector. Recently this idea was studied in the more generic framework of type-II 2 Higgs doublet models (2HDM) and found that the amount of additional CP-violation possible in the Higgs sector could provide a reasonable model for baryogenesis [25].

Other explanations of fine tuning include composite Higgs models or Randall-Sundrum models of gravity. Composite Higgs models interpret the Higgs mechanism as only an effective theory and introduce a new strongly interacting QCD-like force above the electroweak scale. It was shown by Randall and Sundrum [26] that higher-dimensional models with warped space-time metrics can provide a natural explanation of the hierarchy problem and thus fine-tuning.

1.7 Summary

In this chapter a theoretical introduction of the Standard Model and the electroweak symmetry breaking is given. The Higgs boson phenomenology at the LHC is then presented, with particular attention to the Higgs to two photons decay channel, main subject of this thesis. Finally the main Higgs results achieved at the LHC during Run 1 are summarised. After the discovery of the Higgs boson and the first measurements of its properties, it is important to perform more precise measurements of this particle, in order to detect possible small deviations from the Standard Model, hints of new physics.

Chapter 2

LHC and the CMS detector

2.1 The Large Hadron Collider restart

The Large Hadron Collider (LHC) [27] is the world's biggest particle accelerator. Proposed and realized by the European Organization for Nuclear Research (CERN), it was designed to collide protons, as well as lead ions, at an unprecedented energy and rate, in order to answer some of the most fundamental questions of physics.
On 23 november 2009, the accelerator produced the first proton-proton collisions. After few pilot runs at energies of 450 GeV and 1.18 TeV per beam, the energy was ramped up to 3.5 TeV and, on 30 march 2010, the first collisions at a centre-of-mass energy of 7 TeV, the highest ever reached at a particle collider, were recorded by the experiments ATLAS, CMS and LHCb. During these years the LHC has performed beyond expectations and the total integrated luminosity delivered in 2011 was 6.13 fb^{-1} (see Figure 2.1). In 2012 the centre-of-mass energy has been incremented up to 8 TeV, and an integrated luminosity of 23.30 fb^{-1} was reached (see Figure 2.1). The Run 1, concluded in february 2013, has been rich of scientific results, such as the discovery of the Higgs Boson and of several hadrons (like the χ_b $(3P)$ bottomonium state) and the first observation of the very rare decay of the B_s meson into two muons, which severely constrains supersymmetry models for instance.
After these three very fruitful years of collisions at the LHC started a two-year break, the LHC's first long shutdown (LS1). This period without beam aimed at improving the accelerator as well as the detectors towards the restart for Run II at a centre-of-mass energy of 13 TeV. During the technical stop about 10.000 electrical interconnections between the magnets were consolidated, magnet protection systems were added, while cryogenic, vacuum and electronics were improved and strengthened. Furthermore, the beams have been set up in such a way that they will produce more collisions by bunching protons closer together, with the time separating bunches being reduced from 50 ns to 25 ns.
In this second period of operation at an unprecedented energy of 13 TeV, after the discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations, physicists will be

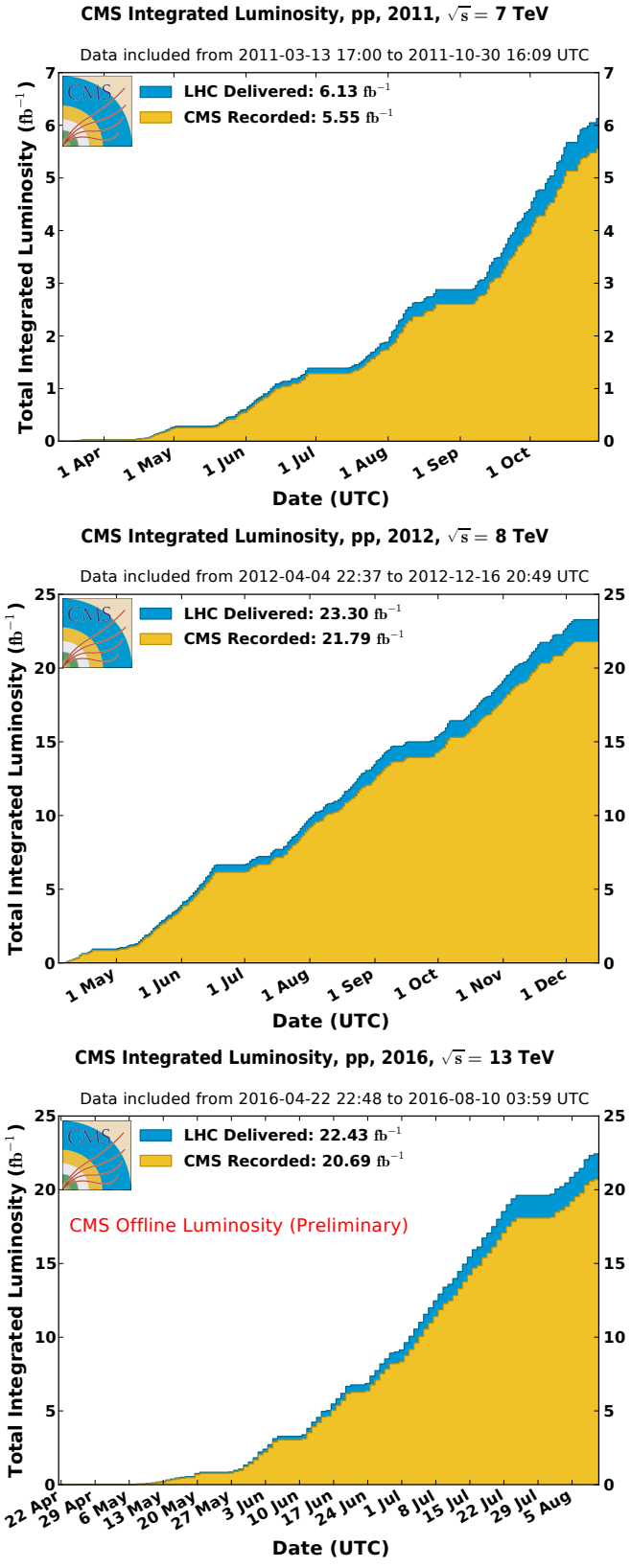


Figure 2.1: Cumulative luminosity versus day delivered (blue) to, and recorded (orange) by CMS during stable beams and for p-p collisions at 7 TeV centre-of-mass energy in 2011 (top), 8 TeV centre-of-mass energy in 2012 (middle) and 13 TeV centre-of-mass energy in 2016 (bottom).

putting the Standard Model of particle physics to its most stringent test yet, searching for new physics beyond this well-established theory describing particles and their interactions. Run 2 will be also the occasion to perform more precise studies of the physics processes already known, and in particular it will be possible to analyze more in detail the Higgs boson properties.
The first beams started circulating in april 2015 at an energy of 450 GeV, to reach gradually the expected energy of 13 TeV. The first part of Run 2 finished in november 2015 and it was followed by a brief technical stop during winter. The total integrated luminosity delivered was 4.22 fb^{-1}, out of which CMS recorded 3.81 fb^{-1}.
After the technical stop, beams started circulating again in 2016, and the LHC performed very well, delivering about 22 fb^{-1} so far, as shown in Figure 2.1. The dataset used for the $H \to \gamma\gamma$ analysis presented at ICHEP amounts to 12.9 fb^{-1}, recorded in 2016.

2.2 The CMS detector

The Compact Muon Solenoid (CMS) [28] is one of the two general-purpose detectors which operate at LHC. Its main physics goals are the search for the Higgs boson, the search for new physics beyond the SM and precision measurements of already known physics processes. For these reasons an excellent lepton reconstruction and particle identification are required.
The main characteristics of the CMS detector are a compact design with a strong magnetic field, which is obtained using a 3.8 T superconducting solenoid, a robust and redundant muon system, a good electromagnetic calorimeter and a high quality central tracking system.
CMS is composed by a cylindrical barrel, with several layers coaxial to the beam axis, closed at both ends by endcap disks orthogonal to the beam direction. Its full length is 28.7 m, the diameter 15 m and the total weight about 14.000 t.
As already mentioned, the core of the apparatus is the magnet, which contains, from inside out, the following detectors:

- the *tracker*, made of a silicon *pixel detector* in the inner region, closest to the beam, and of silicon *microstrip detectors* in the outer region, used to reconstruct charged particle tracks and primary and secondary interaction vertices,
- the *electromagnetic calorimeter* (ECAL), which allows for precise measurement of electron and photon energies; it is made of lead tungstate ($PbWO_4$) scintillating crystals, both in the barrel and in the endcaps, and extended by a forward *preshower detector* in the endcap region,
- the *hadron calorimeter* (HCAL), used for jet direction and transverse energy measurements, extended in the forward region with the "*very forward calorimeter*".

Outside the magnet coil, the iron return yoke of the magnet hosts the *muon spectrometer*, used for reconstruction of muon tracks: *drift tubes* (DT) in the barrel and *cathode strip chambers* (CSC) in the endcaps, complemented overall by *resistive plate chambers* (RPC), to ensure redundancy and robustness to the muon trigger. Schematic views of the CMS detector are shown in Figures 2.3 and 2.4.
In the following we will give a brief description of each subdetector, developing in particular the electromagnetic calorimeter section, since this thesis work is focused on photons.

2.2.1 Coordinate conventions

The CMS coordinate system (see Figure 2.2) used to describe the detector geometry is a right-handed Cartesian frame, with the x axis pointing to the centre of the LHC ring, the z axis parallel to the beam and the y axis directed upwards.
Because of the cylindrical symmetry of the CMS design, the reconstruction algorithms use a cylindrical coordinate system: the azimuthal angle Φ is measured in the x - y plane from the x axis, while the polar angle θ is measured from the z axis. Instead of θ, the pseudorapidity η is used, which is defined as:

$$\eta = -\ln\tan\frac{\theta}{2}.$$

The transverse momentum p_T is defined from x,y components of the momentum as:

$$p_T = \sqrt{p_x^2 + p_y^2}.$$

The transverse energy is defined as $E_T = Esin\theta$ and the missing transverse energy is denoted with E_T^{miss}.
Finally, the ΔR parameter is defined as:

$$\Delta R = \sqrt{\Delta\Phi^2 + \Delta\eta^2}.$$

2.2.2 The tracker

The tracker [29] is the innermost subdetector and the closest to the interaction point. Its goal is to reconstruct charged tracks with high efficiency and momentum resolution, to measure their impact parameter and to reconstruct primary and secondary vertices. The tracker is solely based on several layers of silicon detectors and its dimensions are $|z| < 270\ cm$ and $r < 120\ cm$. Close to the interaction point, the first layers, composed by finely segmented pixel detectors, are fundamental for the measurement of the impact parameters and have to cope with a very high particle flux. The rest of the tracker is made

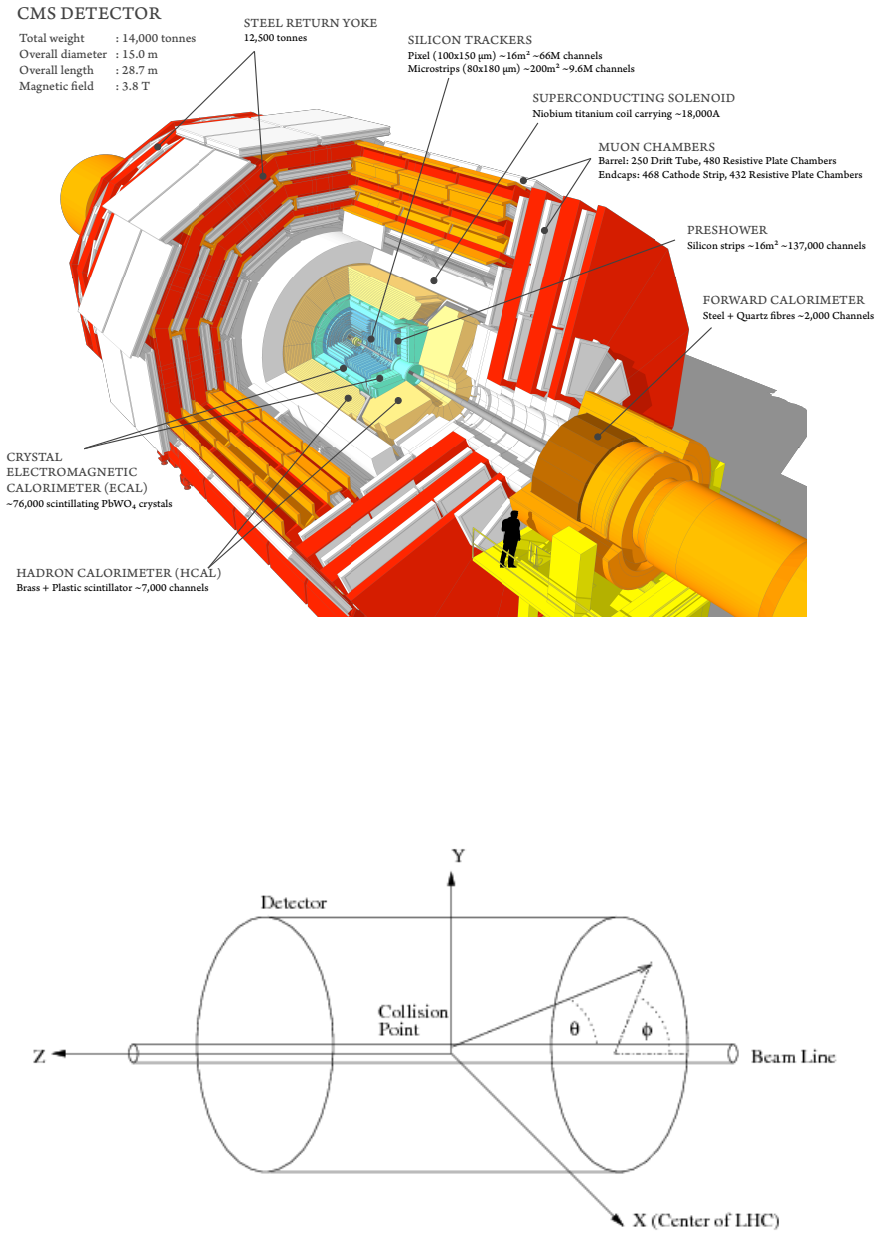


Figure 2.2: A three dimensional view of the CMS detector with the conventional coordinate system.

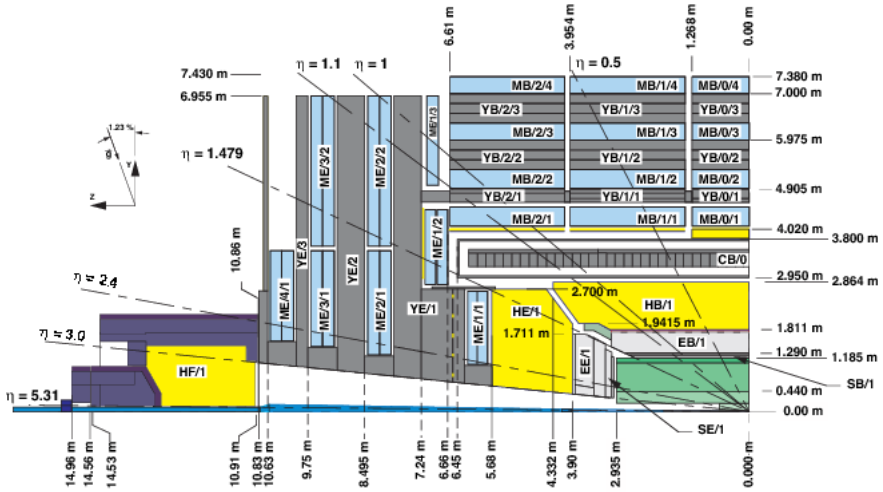


Figure 2.3: Longitudinal view of one quarter of the CMS detector.

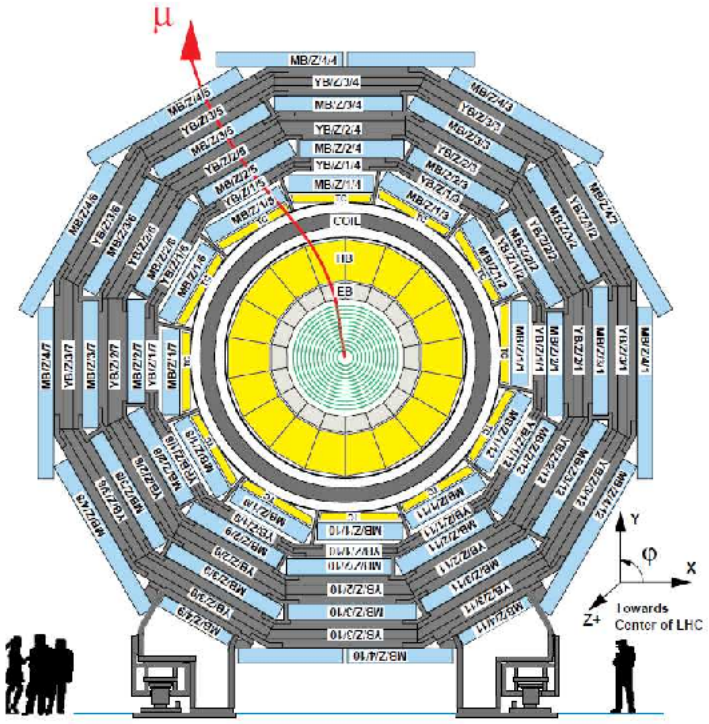


Figure 2.4: Transverse view of the CMS barrel region.

up of single-sided and double-sided silicon strip detectors.
In order to limit the radiation damage to silicon sensors due to the high particle flux, both pixel and microstrip detectors have to be kept at a working temperature of $-10\ ^oC$.

Pixel detector

The pixel detector (see Figure 2.5) provide high-resolution three-dimensional measurements, that are used for charged track reconstruction. Its excellent resolution allows the measurement of track impact parameters, the identification of b- and τ-jets and the reconstruction of vertices in three dimensions. This detector consists of three barrel layers and two endcap disks for each side. The barrel layers, extending from z = -26.5 cm to z = +26.5 cm, are placed at mean radii of 4.4 cm, 7.3 cm and 10.2 cm. The two disks of the endcaps, placed on each side at z = 34.5 cm and 46.5 cm, have the inner radius of 6 cm and the outer of 15 cm.
The pixel detector consists of 66 million pixel elements, each 100 μm $\times$ 150 μm in dimension, spread across 1440 modules. Each pixel consists of a p-n semiconductor junction. When a charged particle crosses the junction, it excites electron-hole pairs, and the charge is collected by the readout electronics connected to the junction. In order to keep the data volume reasonable given the very large number of channels, zero suppression is performed by electronics on the sensor modules, in which only pixels with signal above a set threshold are read out. A charged particle crossing the module will generally deposit charge in at least two adjacent pixels, with the amount of charge deposited in each pixel inversely related to the distance between the particle position and the pixel. A measurement of the charge sharing between adjacent pixels therefore allows a single hit position resolution substantially smaller than the dimensions of a single pixel. In order to exploit the sharing of charge among adjacent pixels, the signal amplitude is digitized with 5 to 8 bits of information, allowing a single hit position resolution of $15 - 20\ \mu m$.

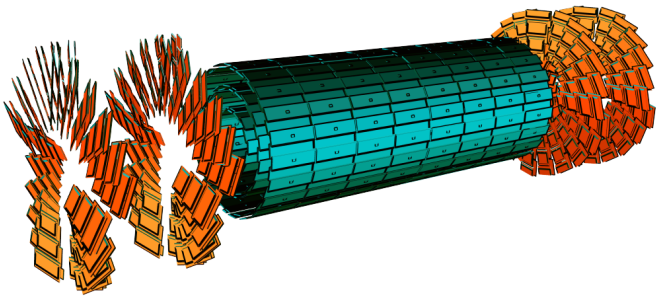


Figure 2.5: The pixel detector: the barrel section and the two disks of the endcaps are visible.

Microstrip detector

The outer part of the tracking detectors, the silicon microstrip detector, provides measurements precisely localized in only two-dimensions, with most strips oriented perpendicular to the ϕ direction. In the barrel, this consists of the Tracker Inner Barrel (TIB) region, comprised of four layers between 20 and 55 cm in radius, as well as the Tracker Outer Barrel (TOB) region, consisting of an additional 6 layers between 50 and 116 cm in radius. In the endcap region, the strip tracker consists of the Tracker Inner Disk (TID) region of three layers between $|z|$ of 80 and 90 cm, plus a Tracker EndCap (TEC) region of nine layers located between $|z|$ of 124 and 280 cm. A fraction of the layers includes double layered modules, with a second set of strips oriented at an angle of 100 mrad with respect to the first. The combination with these stereo measurements can give a position measurement in the third dimension with a precision ranging from 230 to 530 μm. A schematic view of the tracking detectors, labeled by region, is shown in Figure 2.6.

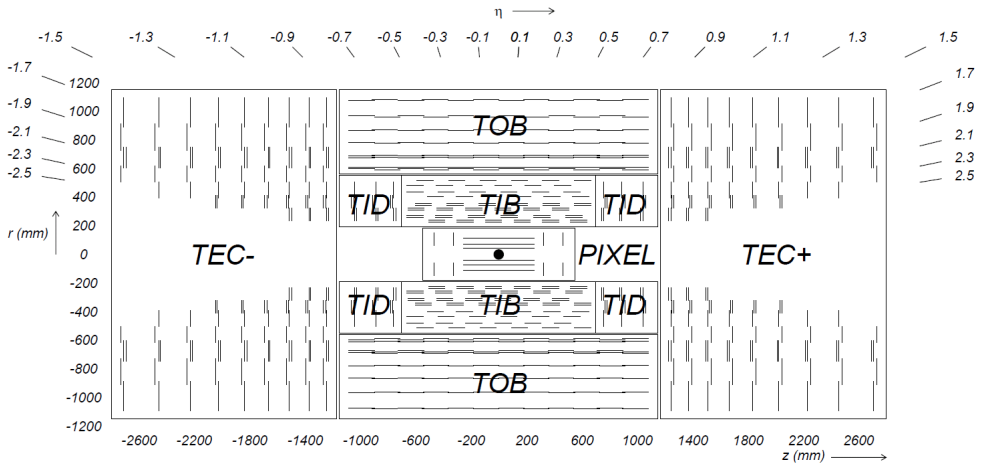


Figure 2.6: A diagram of the CMS inner tracking detectors, showing the layers of the silicon pixel and strip tracking detectors.

The combined tracking detector system provides coverage up to $|\eta| = 2.5$, with an average of 13-17 measurements per charged particle, depending on the pseudorapidity region.
The silicon strip detector consists of about 9.3 million strips across 15148 modules, with strips as well consisting of p-n junctions across which charge carriers are ionized by charged particles as they cross the strip. Depending on the region of the detector, the strip pitch varies between 80 and 184 μm. By exploiting charge sharing between strips, analogous to charge sharing between adjacent pixels, the single hit resolution along the ϕ direction ranges from 23 to 53 μm.

The large amount of silicon in the inner tracking detectors, combined with the sophis-

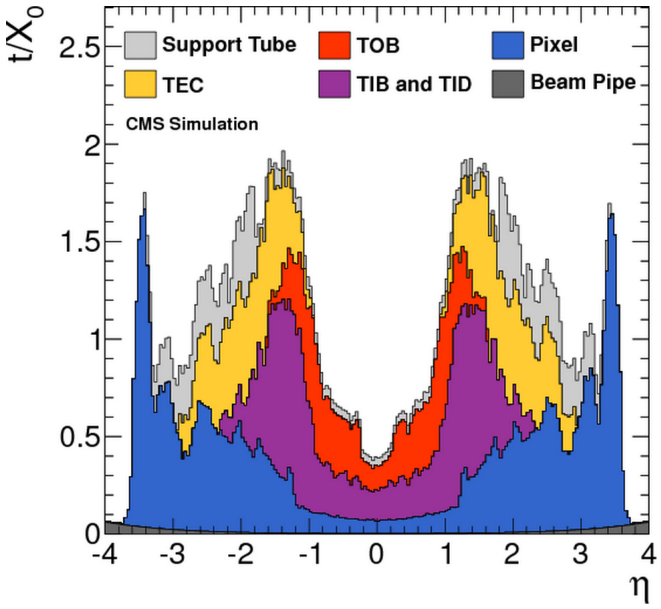


Figure 2.7: The amount of material in the inner tracking detectors, measured in units of radiation lengths, and broken down by detector regions.

ticated electronics leads to a substantial requirement for cabling and cooling services. For this reason the amount of material in the detector is relatively large. The estimated material budget, as a function of pseudorapidity, is shown in Figure 2.7. The estimated total material budget ranges from about 0.4 radiation lengths in the very central barrel, to a peak of about 1.8 radiation lengths in the vicinity of $|\eta| = 1.5$, near the barrel-endcap transition region.

2.2.3 The electromagnetic calorimeter

The goal of the Electromagnetic Calorimeter (ECAL) [30] [31] is the accurate measurement of the position and energy of electrons and photons. In particular, one of the main objectives of this subdetector is the search for the Higgs boson in the channel $H \to \gamma\gamma$, considered as the golden channel for low Higgs masses. Thus the electromagnetic calorimeter performance has to deal with the diphoton mass resolution, which depends both on energy and angular resolution:

$$\frac{\sigma_M}{M} = \frac{1}{2}\left(\frac{\sigma_{E_1}}{E_1} \oplus \frac{\sigma_{E_2}}{E_2} \oplus \frac{\sigma_\theta}{\tan\frac{\theta}{2}}\right),$$

where $E_{1,2}$ are the energies of the two photons, θ is the photon angular separation and $\oplus$ indicates a quadratic sum.

The energy resolution $\frac{\sigma_E}{E}$ can be parametrized as

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c,$$

where a, b, c are respectively the stochastic, noise and constant term and will be discussed in details later in this section.
In order to achieve the best possible energy resolution, the CMS calorimeter is homogeneous, nearly hermetic and made of lead tungstate ($PbWO_4$) crystals, coupled to photodetectors. There are about 76000 crystals in total, divided between the ECAL barrel detector (EB) and the ECAL endcap detector (EE). Incident electrons and photons initiate showers inside the crystals, and the showering particles produce scintillation light as they interact with the crystal. This scintillation light is then measured by photodetectors, and the amount of scintillation light is used to determine the energy deposited in each crystal. In the endcap region, in the η range $1.653 < |\eta| < 2.6$, a sampling preshower detector made of lead and silicon active layers is installed in front of the ECAL in order to improve angular resolution for photon/π^0 separation.
Lead tungstate has been chosen for the crystal material because of its high density, corresponding short radiation length ($X_0 = 0.89\ cm$) and small Molière radius ($2.2\ cm$), thus having a high compactness, allowing a very fine granularity. Moreover, the very short scintillation decay time of these crystals allows to collect about 80% of the light within 25 ns, so that they can be used at the crossing rate of 40 MHz.
The use of high density crystals has allowed the design of a calorimeter which is fast, has fine granularity and is radiation resistant, all important characteristics in the LHC environment.

Lead tungstate crystals

The main parameters of the $PbWO_4$ crystals compared to other crystals typically used for electromagnetic calorimetry are summarized in Table 2.1. The crystals emit blue-green

Table 2.1: $PbWO_4$ compared to other crystals.

	$PbWO_4$	$NaI(Tl)$	BGO
density (g/cm^3)	8.28	3.67	7.13
radiation length (cm)	0.89	2.59	1.12
Molière radius (cm)	2.2	4.5	2.4
maximum emission (nm)	440	410	480
emission time (ns)	5-15	250	300

scintillation light with a Gaussian-shaped distribution peaking at about 440 nm with a

range from 360 nm to 570 nm at 10% of the maximum.
The scintillation decay time of these crystals is of the same order of magnitude as the LHC bunch crossing time: about 80% of the light is emitted in 25 ns. The light output is relatively low and varies with temperature (variations of -2%/°C at room temperature). The temperature dependence of the light yield is represented in Figure 2.8. For this reason the detector cooling system, which much stabilize the crystal temperature to 0.05 °C, is fundamental.
To exploit the total internal reflection for optimum light collection on the photodetector, the crystals are polished after machining, on all but one side for EB crystals. Since the truncated pyramidal shape makes the light collection non-uniform along the crystal length, the needed uniformity is achieved by depolishing one lateral face. On the contrary, for endcap crystals the light collection is naturally more uniform because the crystal geometry is nearly parallelepipedic. Pictures of barrel and endcap crystals with their photodetectors attached are shown in figure 2.9.

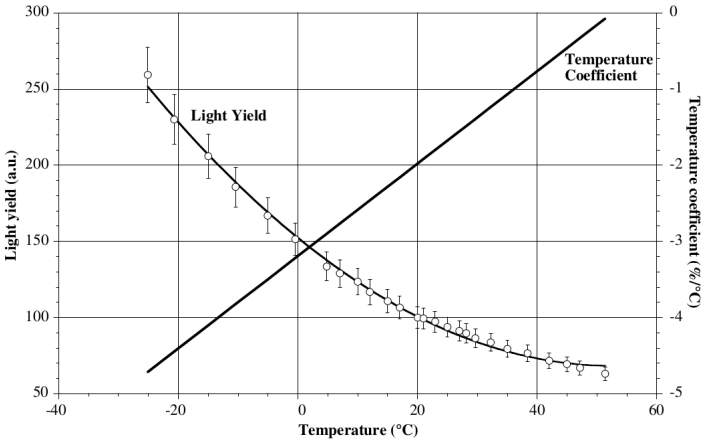


Figure 2.8: Temperature dependence of the $PbWO_4$ light yield.

The crystals have to withstand the radiation levels and particle fluxes present throughout the duration of the experiment. For this reason lead tungstate is intrinsically radiation hard. Nevertheless crystals suffer from radiation damage: ionizing radiation produces absorption bands through the formation of colour centres due to oxygen vacancies and impurities in the lattice. As a consequence the lead tungstate transparency is altered within a few percent, while the scintillation mechanism is not affected. The loss in the transmission efficiency can be corrected for by monitoring the optical transparency with injected laser light, as briefly described later. In this way most of the radiation damages are recovered.

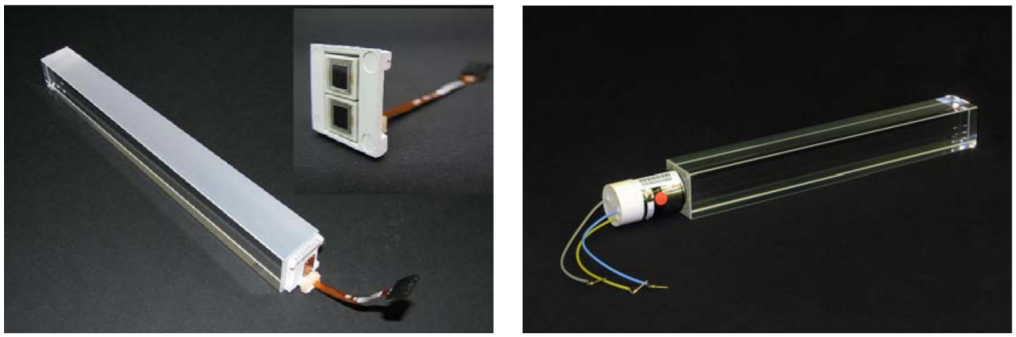


Figure 2.9: $PbWO_4$ crystals with photodetectors attached. Left figure: a barrel crystal with the upper face depolished and the APD capsule. In the insert, a capsule with the two APDs. Right figure: an endcap crystal and VPT.

The ECAL layout and mechanics

The CMS electromagnetic calorimeter consists of a barrel part and of two endcaps; a three dimensional view of the calorimeter is given in Figure 2.10. The main design characteristics are strongly prescribed by the need to have accurate measurements of electrons, photons and missing energy. The mechanical design should in particular minimize the amount of material in front of the calorimeter, optimize the interface with the tracking system and with the Hadron Calorimeter, ensure the best possible hermeticity by minimizing the gaps between crystals and the barrel/endcaps transition region, stabilize the crystal temperature within a tenth of a degree.

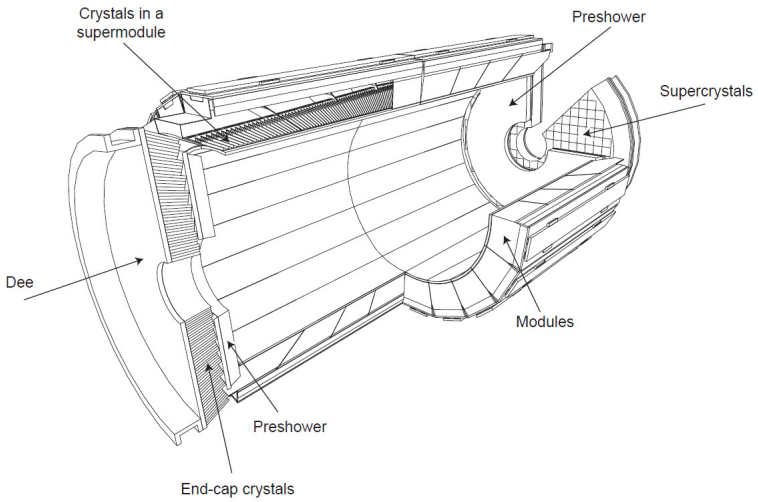


Figure 2.10: The layout of the barrel and endcap ECAL subdetectors, along with the preshower detector.

The EB detector consists of 61200 crystals arranged in a 90×360 $\eta - \phi$ grid, with coverage up to $|\eta| = 1.479$. In order to have an hermetic coverage of the detector, the crystals must be tightly packed. Moreover, to minimize the leakage of the electromagnetic shower in the small gaps between crystals, the EB geometry is such that crystals are tilted at an angle of 3° with respect to the trajectory of particles incident from the nominal interaction point. In order to obtain this tilt, together with the tight packing requirement, EB crystals are shaped as truncated pyramids, with a number of different particular variations needed depending on the precise location in EB. The crystals surface is 22×22 mm^2 at their front face, and 26×26 mm^2 at their back face, comparable to the Molière radius, such that a large fraction of the energy from an electromagnetic shower is expected to be contained within a radius of a few crystals with respect to the crystal on which the electron or photon was incident. Each crystal is 23 cm long, corresponding to 25.8 radiation lengths of material, such that longitudinal leakage of the electromagnetic showers is negligible. The back of each crystal is attached to an Avalanche Photodiode (APD), that detects the scintillation light from the crystal. The EB crystals are arranged into 36 supermodules, 18 for each of the $+z$ and -z sides of the detector, such that each supermodule subtends 20° in ϕ. Each supermodule is composed of 4 modules, ranging from $\eta = 0$ to $\eta = \pm 1.479$. These modules are enumerated 1-4 from the center of the detector outwards. Module 1 in each supermodule consists of 25×20 crystals in the $\eta \times \phi$ direction, whereas modules 2, 3 and 4 in each supermodule consist of 20×20 crystals. The barrel granularity is of $\Delta\eta \times \Delta\Phi = 0.0175 \times 0.0175$.
All supermodules are provided with a cooling system supplying a stability of the crystal array and readout devices within a tight spread of 0.05 °C.

The EE detectors consist of 15000 crystals. The geometry in the endcap is different from that in the barrel, with crystals arranged in an x-y grid in groups of 5×5 crystals, such that all EE crystals share the same geometry. In this way the angle between the crystal axes and the trajectory of particles from the interaction point is between 2 and 8 degrees. EE crystals are 25×25 mm^2 at the front and 30×30 mm^2 at the back, with a length of 22 cm, corresponding to 24.7 radiation lengths. Because of the higher radiation dose in the endcap region, Vacuum Photo Triodes (VPT's) are used as photodetectors for the EE crystals instead of APD's. The EE coverage extends from $|\eta| = 1.479$ to $|\eta| = 3.0$.

The preshower detector is located in front of the EE in the region $1.653 < |\eta| < 2.6$. It consists of two alternating layers of passive lead and active silicon, acting as a sampling calorimeter. The first lead layer corresponds to two radiation lengths of material, whereas the second layer corresponds to one additional radiation length. The silicon layers consist of active silicon strips with a pitch of 1.9 mm. The additional spatial resolution provided by the preshower is designed in principle to improve the separation between prompt photons and neutral mesons in the endcap region.

A longitudinal section of the electromagnetic calorimeter is shown in Figure 2.11.

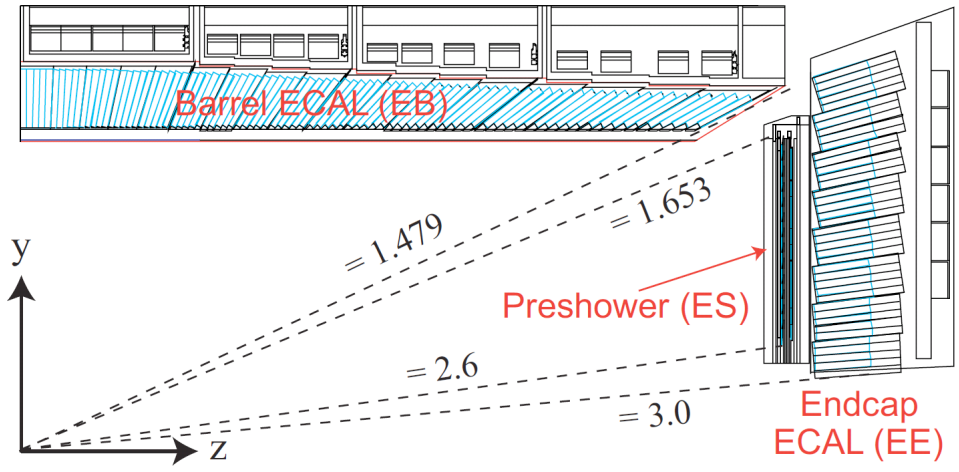


Figure 2.11: Pseudorapidity coverage of a quarter of the CMS electromagnetic calorimeter.

Photodetectors

Due to the crystals low light yield, the choice of the readout devices used to extract the crystal signal is very important. Photodetectors with an internal gain are needed, in order to give a first amplification stage for the signal before the injection in the electronic readout chain. Furthermore the photodetectors need to be fast, radiation tolerant, and be able to operate in the longitudinal 4-T magnetic field. The requirement of radiation hardness and the presence of a strong magnetic field lead to the choice of Avalanche PhotoDiodes (APDs) for the barrel region and of Vacuum PhotoTriodes (VPTs) in the endcaps. Two different devices are needed in order to face up the different radiation level and magnetic field conditions.

Barrel: avalanche photodiodes

The use of APDs presents several advantages: they are fast detectors ($\approx 2\ ns$ of rise time), they have a very good quantum efficiency of 70%-80% around $\lambda = 420\ nm$ and they are higly insensitive to the magnetic field. They are compact devices (overall thickness of $2\ mm$) with a high radiation resistance, and can be manufactured in large quantities with a small spread in the performance parameters. Each APD has an active area of $5 \times 5\ mm^2$. Two of them are glued to the back of each crystal. The APD basic structure is shown in Figure 2.12.
The light enters via the p^+ layer and is absorbed in the p layer behind, where electron-hole pairs are generated if the photon energy is higher than the gap energy. A drift in the

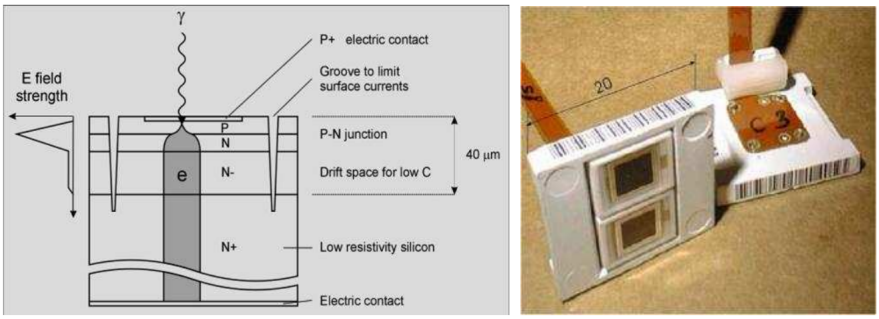


Figure 2.12: On the left, structure of a barrel APD; on the right, pair of APDs to be installed on a crystal rear face.

p-n transition region is followed by an amplification stage in the n volume (with a gain tunable between 50 and more than 10^3) and by an intrinsic drift region before the charge is collected by the cathode.
The APD is supplied by a reverse voltage: by changing its value, the charge multiplication gain can vary from 0 to 200, but the optimum gain to operate with the CMS front-end electronics sits between 50 and 100 and it has been decided to operate at gain 50.
Not only the $PbWO_4$ light yield, but also the APD gain is temperature dependent: for this reason, one tenth of the APD pairs glued to the crystals has a sensor for the temperature measurement.

Endcap: vacuum phototriodes

The APDs used in the barrel are insufficiently radiation-hard to be used also in the endcap region, where VPTs are employed. The VPT basic structure is represented in Figure 2.13.

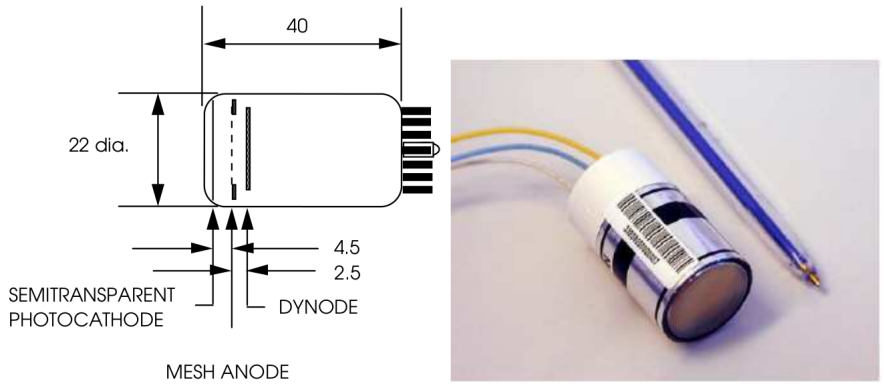


Figure 2.13: On the left, structure of an endcap VPT; on the right, picture of a VPT detector.

The photocathode is semitransparent and made of a radiation-hard glass: the photoelectrons produced are accelerated by an ultra fine mesh (100 wires/mm) placed $4-5\ mm$ far from the photocathode, and impact on a dynode producing secondary electrons (emission factor of about 20). The secondary electrons are attracted back to the anode mesh where a substantial fraction is captured, leading a total effective gain for the VPT greater than 8 in a magnetic field of 4 T. The VPTs lower quantum efficiency with respect to the APDs is compensated by a larger active area of about 280 mm^2, so that the total detector response is almost the same for barrel and endcap regions.

Trigger tower and readout electronics

The CMS Electromagnetic Calorimeter needs a very fast electronics readout, in order to match the bunch crossing time of 25 ns and to provide very precise energy measurements over a wide range. The additional requirement of radiation hardness and large amount of channels are two other aspects to take into account.
The signal produced by photodetectors is amplified and then digitized, passing through 3 different boards as shown in Figure 2.14.

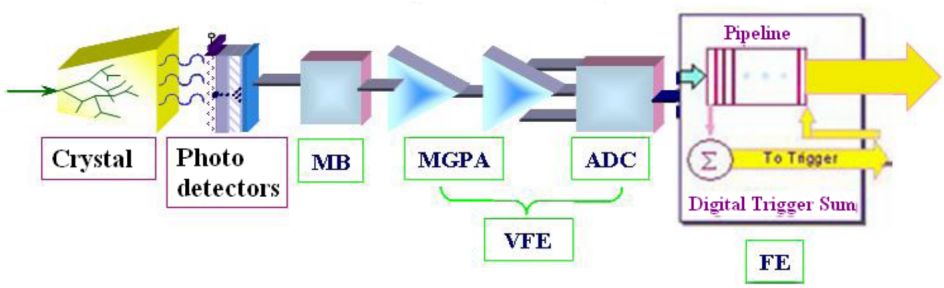


Figure 2.14: ECAL readout chain, from the crystal light emission to the digitized signal.

The basic building block of the readout electronics is a 5x5 matrix of crystals, called trigger tower in the barrel, or supercrystal in the endcap, and is made up of 1 Mother Board (MB), a Low Voltage Regulator Board (LVRB), 5 Very Front End (VFE) boards and 1 Front End (FE) card. A trigger tower covers a region of $\Delta\eta \times \Delta\Phi = 0.087 \times 0.087$.
The Mother Board is a totally passive board located beneath the cooling system for the electronics; it is necessary to route the signals from the photodetectors to the VFE cards, to distribute high voltage to the photodetectors and low voltage to the VFE cards.
The LVRBs are connected directly to the external Low Voltage power supplies which sit in the CMS racks attached to the outside of the CMS iron yoke, approximately 20 metres from the supermodule. Each LVRB contains radiation-hard voltage regulators which provide the 2.5 V needed by the front end electronics. This regulated 2.5 V is distributed to the FE card by a small connector on the LVRB, and to the 5 VFE cards in a trigger tower via the MB.

The VFE board is needed to amplify the signal and convert it from analog to digital, using very fast components (40 MHz) compatible with the 25 ns time that separates two LHC bunch crossings. Each VFE contains amplification and digitization for the signals from 5 crystals adjacent in Φ.
The VFE board feeds the digitized outputs to a FE board, which stores and processes the data during the Level-1 trigger latency of $\sim$ 3 μs. The trigger data are then transmitted to the off-detector electronics through an optical link operating at 800 Mbyte/s. After a L1-trigger reception, the data stored on the FE card corresponding to the triggered event are transmitted through a second optical link to the off-detector electronics for further trigger analysis (High-Level Trigger).

CMS trigger and data acquisition

At the LHC expected energy and instantaneous luminosity, the interaction rate ($\approx 40MHz$) leads to $\approx 10^9$ interactions/sec, that is orders of magnitude larger than what can be reasonably processed by the readout system and archived for later off-line analysis. In fact, data from only about 100 crossings/sec can be written to archival media. For this reason a very good online selection is needed and a rejection of nearly 10^7 with respect to the active bunch crossings at the LHC has to be achieved. In CMS this selection is performed in two physical steps: the *Level-1 Trigger* and the *High Level Trigger* (HLT). The L1 Trigger is based on processors which perform fast selections ($\approx 3\ \mu s$) for each 25-ns bunch crossing; it is built of mostly custom-made hardware and it performs detector analysis in a coarse way. On the contrary, the HLT can operate on longer timescales, it is basically a processor farm which inspects the events that have already passed the L1 trigger and executes software algorithms. In Figure 2.15 the data flow in the CMS trigger and data acquisition system is shown. In this section a brief description of L1 trigger and HLT is provided; then, a particular attention is paid to the calorimeter trigger.

Level 1 Trigger versus High Level Trigger

The total time allocated for the L1 trigger decision to keep or discard data from a particular beam crossing is 3.2 μs. This time includes the transit time for signals from the front-end electronics to reach the services cavern housing the Level-1 trigger logic and return back to the detector front-end electronics. During this time, the detector data must be held in buffers, while trigger data are collected from the front-end electronics and decisions are performed. The Level-1 triggers involve the calorimetry and muon systems, as well as some correlation of information between them. The Level-1 decision is based on the presence of "trigger primitive" objects such as photons, electrons, muons, and jets above E_T or p_T thresholds. It also employs global sums of E_T and E_T^{miss}. The L1 trigger reduces event rates from 40 MHz to 100 kHz (design value).

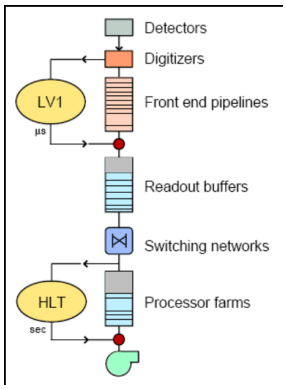


Figure 2.15: Data flow in the CMS Trigger/DAQ system. The software-based High-Level Trigger filters via the Data Acquisition system (DAQ) the events passing hardware-based Level-1 trigger. Time axis goes from upside down.

The HLT reduces event rates furthermore, down to 100 Hz, running the reconstruction algorithms and performing more sophisticated selections. The HLT code is developed starting from the idea of partial reconstruction: rather than reconstruct all possible objects in an event, whenever possible only those objects and regions of the detector that are actually needed are reconstructed. Moreover, since events have to be discarded as soon as possible, many virtual trigger levels are provided: calorimeter and muon information are used, followed by the tracker pixel data and finally the use of the full event information (including full tracking). HLT selection can then be seen as a sequence of filters of increasing complexity, using the information of calorimeters, pixel association and track reconstruction.

The Electromagnetic Calorimeter Trigger

The ECAL front end electronics is in charge of:

- amplifying and shaping the signal from the sensors,
- digitizing the signal at 40 MHz,
- using the digitized data to calculate trigger primitives for the Level-1 Trigger decision,
- buffering the data until reception of the Level-1 trigger decision,
- transmitting the data to the off-detector electronics for insertion in the CMS data stream.

The building block of the front end electronics is the Trigger Tower, previously described. At each bunch crossing, trigger primitive generation is first started in the Front End boards, and then finalized and synchronized in an electronic board (called Trigger Concentration

Card, TCC) before the transmission to the regional calorimeter trigger. Each trigger primitive refers to a single trigger tower and consists of the summed transverse energy deposited in the tower (8 bits)[1], plus a compactness bit which characterizes the lateral extension of the electromagnetic shower ("fine grain veto"). The encoded trigger primitives are time aligned and stored in the TCC during the Level-1 latency for subsequent reading: each TCC collects trigger data from 68 FE boards in the barrel, corresponding to a supermodule, and from 48 FE boards in the endcaps, corresponding to the inner or outer part of a 20° sector. Finally, trigger primitives are sent to the Level-1 regional calorimeter trigger, where together with HCAL trigger primitives the electron/photon and jets candidates are computed as well as the total transverse energy.

Calorimeter towers, ECAL plus HCAL, and Global Calorimeter Trigger

Readout cells in HCAL are arranged in a tower pattern in $\eta - \Phi$ space. The cells in the barrel region have a segmentation of $\Delta\eta \times \Delta\Phi = 0.087 \times 0.087$, becoming progressively larger in the endcap and forward regions. Since the ECAL granularity is much finer than HCAL, calorimeter towers (ECAL plus HCAL) are formed by addition of signals in $\eta - \Phi$ bins corresponding to individual HCAL cells.
Local calorimeter trigger information refers to energy depositions in the trigger towers of the electromagnetic and hadronic calorimeters. The Regional Calorimeter Trigger uses the trigger primitives to find candidate electrons or photons, jets and isolated hadrons from the decay of τs and to calculate transverse energy sums in different detector regions. All calorimeter trigger objects described so far are forwarded to the Global Calorimeter Trigger, which sorts the electrons or photons, τs and jets according to energy and quality, and sends the four objects with the highest rank in each category to the Global Trigger. The input for the physics trigger algorithm calculations are the trigger objects ordered by rank: an algorithm is defined as a logic combination of the trigger objects together with a set of energy or momentum thresholds, windows in η and/or Φ and topological conditions. All thresholds and space parameters, except for some exceptions, are only applied at the Global Trigger stage.

Energy resolution

As already mentioned at the beginning of this section, the energy resolution of an homogeneous electromagnetic calorimer can be parametrized as:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c,$$

[1] in the barrel the trigger tower is divided into 5 Φ-oriented strips, whose energy deposits are summed by the FE board trigger pipeline to give the total transverse energy of the tower, called the main trigger primitive

where a, b, c are respectively the stochastic, noise and constant term. The individual contributions are discussed below.

The stochastic term

The stochastic term a is a consequence of the statistics associated with the electromagnetic shower development in the calorimeter and the following scintillation light collection. This term represents the intrinsic resolution of an ideal calorimeter, that is a calorimeter of infinite size and without response deterioration caused by instrumental effects. The initial energy E_0 of a particle incident in the calorimeter is directly proportional to the total track length T_0, defined as the sum of all ionization tracks due to the charged particles in the electromagnetic shower. Since T_0 is proportional to the number of track segments in the cascade and the shower development is a stochastic process, the intrinsic resolution, from purely statistical arguments, can be written as:

$$\frac{\sigma_E}{E} \propto \frac{\sqrt{T_0}}{T_0} \propto \frac{1}{\sqrt{E_0}}.$$

In the case of a real calorimeter, this term also absorbs the effects related to statistical fluctuations in the scintillation light collection due to geometry effects, quantum efficiency and electron multiplication processes inside the photodetectors.
For the CMS ECAL calorimeter, the contribution due to fluctuations on the lateral containment of the shower is of 1.5% considering the energy deposited in a cluster of 5×5 crystals. The contributions due to the photostatistics are kept below 2.3% if the photodetectors produce more than 4000 photoelectrons per GeV. In the endcap regions, where a preshower is present in front of the calorimeter, an additional contribution of $\sim 5\%$, related to the fluctuations on the energy deposited in the absorber, needs to be taken into account.

The noise term

The noise term is strongly related to the features of the readout circuit (detector capacitance, devices, cables etc.). In ECAL photodetectors contribute because of their intrinsic capacitance and leakage currents. Furthermore there are the noise introduced by the preamplifier stage of the electronic readout and the one introduced in the digitization step: the first one is $\sim 30-40$ MeV in the barrel and 150 MeV in the endcaps, while the second one is negligible. A final contribution to the noise term comes from pileup: in a cluster of 5×5 crystals the contribution is of ~ 30 MeV in the barrel and of ~ 175 MeV in the endcaps, thus compatible with the total electronic noise.

The constant term

The constant term c is very important since it is the asymptotic value of the energy resolution at high energies. All the defects connected to the detector construction and assembly, as well as the instability of temperature, voltage etc. during its operation, contribute to this term. The main contributions are described below:

- Non-uniformity of the longitudinal light collection: as a consequence of the truncated pyramid-shape of the crystal and the high refractive index ($n = 2.16$), a strong focusing effect on the scintillation light causes non-uniformity in the light-yield. One of the lateral faces of the crystals is depolished during the production process to avoid this effect. In this way the contribution is kept below 0.3%,
- Longitudinal shower containment and uncorrected geometrical effects: thanks to an accurate simulation and to test beam studies has been shown that the constant term contribution due to these effects is lower than 0.2%,
- Inter-calibration errors: inter-calibration of the channels is a crucial issue for physics performance. The main source of channel-to-channel response variation is the crystal-to-crystal variation of scintillation light yield, together with readout variations,
- Temperature stability: as already discussed, both the emission of scintillation light and the APD gain are temperature dependent. A temperature stability within 0.05 $^{\circ}C$ is needed over the full detector volume in order to keep the contribution to the constant term below 0.1%,
- High voltage stability: the APD gain strongly depends on the bias voltage. In order to keep this contribution below 0.1% the stability on the high voltage has to be better than 30 mV.

Calibration

In order to achieve a constant term contribution of 0.5% in the energy resolution, a big effort has to be made to obtain the best possible calibration of the calorimeter [32]. ECAL calibration is naturally seen as composed of a global component, giving the absolute energy scale, and a channel-to-channel relative component, which is referred to as intercalibration. The essential issues are uniformity over the whole ECAL and stability, so that showers in different locations in the ECAL in data recorded at different times are accurately related to each other.
The intercalibration is principally done using π^0 events in the barrel and $W \to e\nu$ events in the endcap. Concerning the absolute energy scale, physics events in which a particle (namely a Z boson) decays into an electron-positron couple are used. The kinematical

constraint given by the invariant mass of the particle gives the absolute calibration of the calorimeter.

Monitoring

The crystal transparency is expected to decrease with the amount of radiation absorbed, and then recover after a certain time. For this reason it is fundamental to continuously monitor the light transmission of each crystal during the LHC operation. For this purpose a laser-based monitoring system is designed to inject pulses into each single crystal to measure the light transmission near the scintillation spectrum peak ($\lambda \approx 440\ nm$) and, as a crosscheck, at a longer wavelength ($\lambda \approx 800\ nm$). The loss in transparency due to irradiation for the laser light (R) and for the scintillation light (S) are related by

$$\frac{S}{S_0} = \left(\frac{R}{R_0}\right)^{\alpha},$$

where R_0 and S_0 are the signal intensity, respectively for laser light and scintillation light, before irradiation. Specific test beam studies have shown that the coefficient α for the different crystals has the same value within 5%. The irradiation damage being small ($< 6\%$) in Run 1 for crystals in the barrel, it is possible to use one single value of α for all the crystals in order to correct the crystal response for the transparency loss. This allows to keep the contribution to the constant term in the resolution $< 0.3\%$, within the design specification.

2.2.4 The hadronic calorimeter

The hadronic calorimeter (HCAL, see Figure 2.16) [33] plays an essential role measuring the direction and energy of jets, the total transverse energy and the imbalance in the transverse energy (missing E_T). To achieve this goal a high hermeticity is required. In particular, the HCAL angular coverage must include the very forward region, since the identification of forward jets is very important for the rejection of many backgrounds.
The Hadronic Calorimeter can be divided in four regions, which provide a good segmentation, a moderate energy resolution and a full angular coverage up to $|\eta| = 5$. The *barrel hadronic calorimeter* (HB) surrounds the electromagnetic calorimeter and covers the central pseudorapidity region up to $|\eta| = 1.3$. The endcap regions are covered up to $|\eta| = 3$ by the two *endcap hadron calorimeters* (HE). The HB and HE are located inside the solenoid magnet. To satisfy the hermeticity requirements, two *forward hadronic calorimeters* (HF) surround the beam pipe at $|z| = 11\ m$, extending the pseudorapidity coverage up to $|\eta| = 5$. The magnet and an additional layer of scintillation detectors, which is referred to as the *outer hadronic calorimeter* (HO), installed outside of the coil increase the material thickness in the barrel pseudorapidity region, such that the hadronic showers are fully absorbed before reaching the muon system.

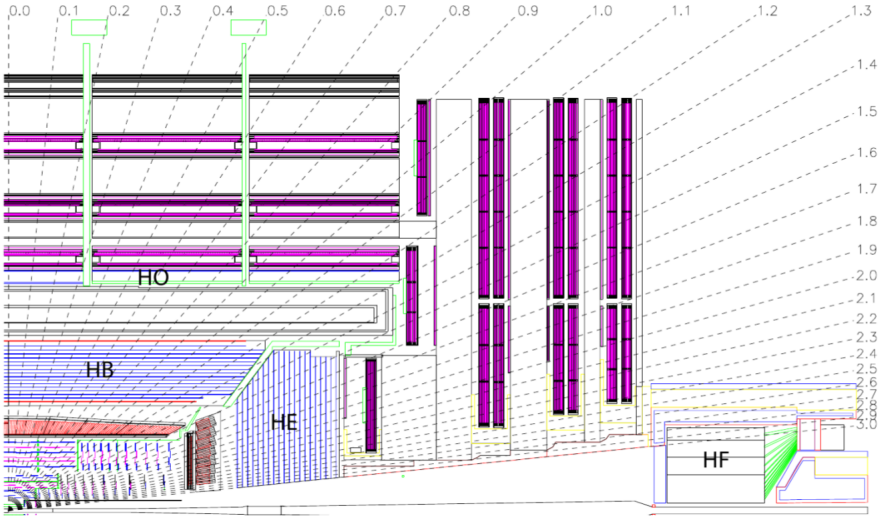


Figure 2.16: The layout of the various HCAL subdetectors showing their respective coverage in pseudorapidity.

The HB and HE are sampling calorimeters with active plastic scintillators interleaved with brass plates. This absorber material has been chosen because of its reasonably short interaction length and because it is non-magnetic. The read-out system is constituted by wavelength-shifting fibres. The first layer is read out separately, while all others are read out together in towers of $\Delta\eta \times \Delta\Phi = 0.087 \times 0.087$ rad.
The energy resolution (expressed in GeV) is:

- $\frac{\sigma_E}{E} \sim 65\%\sqrt{E} \oplus 5\%$ in the barrel,
- $\frac{\sigma_E}{E} \sim 85\%\sqrt{E} \oplus 5\%$ in the endcaps,
- $\frac{\sigma_E}{E} \sim 100\%\sqrt{E} \oplus 5\%$ in the very forward calorimeter.

2.2.5 The CMS solenoid

In order to achieve a good momentum resolution for momenta up to 1 TeV/c, CMS needs a high magnetic field. The CMS solenoid is the central device around which the experiment is built and its dimensions limited the size of the total apparatus. Its purpose is to bend the paths of particles emerging from high-energy collisions. The higher the particle momentum is, the less its trajectory is curved by the magnetic field. A higher strength field, combined with high-precision position measurement in the tracker and muon system, gives accurate measurement of momentum.
The CMS magnet [34] is a 13 m long superconducting solenoid, the largest ever built. It is

able to generate a uniform magnetic field of 4 T in the inner region, storing about 2.5 GJ of energy (Figure 2.17). It operates at a temperature of 4 K, ensured by a sophisticated

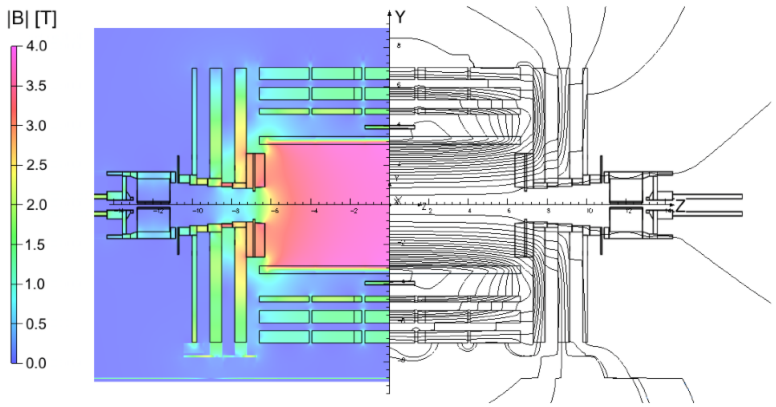


Figure 2.17: Layout of the magnetic field of CMS.

helium cooling system. At such temperature, the flat NiTb cable becomes superconducting, allowing a 20 kA current to flow without appreciable loss. The whole magnet is then contained in an enormous vacuum cylinder, which isolates it from the external environment. Outside, an iron structure composed by five barrel layers and three disks for each endcap constitutes the iron yoke, needed to return the magnetic flux.

2.2.6 The muon system

The aim of the muon spectrometer [35] is to identify muons and measure, in combination with the inner tracker, their transverse momentum accurately. As a matter of fact, since high-p_T muons provide a clean signature for many processes, the muon system plays an important role in the trigger. The muon spectrometer, placed outside the magnet, is embedded in the iron return yoke, so that the magnetic field bends the tracks and allows muon p_T measurements. The muon system consists of 3 types of gaseous particle detectors (see Figure 2.18):

- Drift Tube (DT) Chambers in the barrel, covering the region ($|\eta| < 1.2$),
- Cathode Strip Chambers (CSC) in the endcaps, covering the region ($0.9 < |\eta| < 2.4$),
- Resistive Plate Chambers (RPC) in both the barrel and the endcaps, covering the region ($|\eta| < 1.6$).

These different technologies are used because of the different particle rates and occupancies, both higher in the endcaps, and the intensity of the residual magnetic field, which is lower in the barrel.

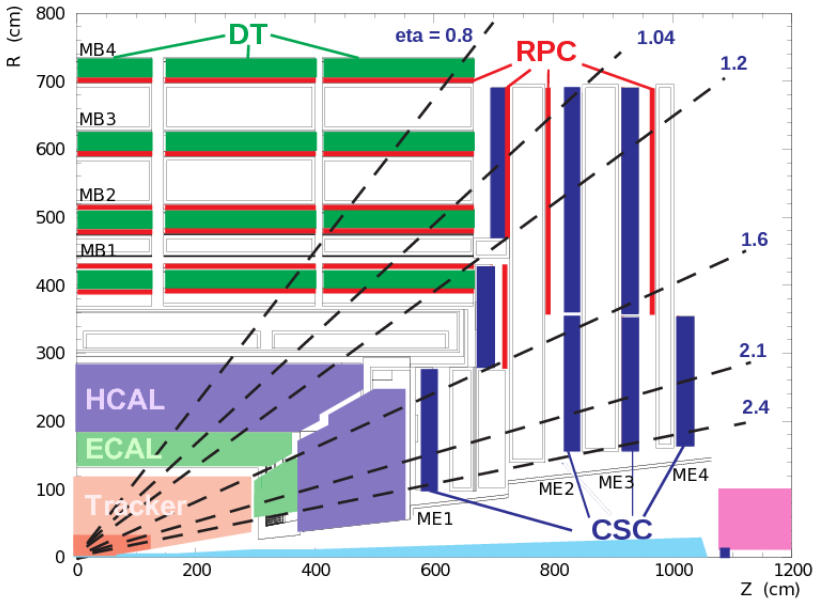


Figure 2.18: Longitudinal view of one quarter of the muon system.

Drift tube chambers

Since in the barrel region the expected occupancy is low ($< 10\ Hz/cm^2$) drift tubes were chosen. The DT segmentation follows that of the iron plates of the yoke, which consists of 5 wheels along the z-axis, each one divided in 12 sectors. Chambers are arranged in 4 *stations* named MB1,...,MB4 as shown in Figure 2.18. Each station consists of 12 chambers, except for MB4 which has 14 chambers.
The basic detector element is a drift tube cell, whose section is shown in Figure 2.19. A

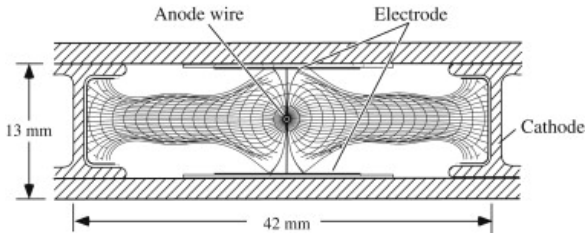


Figure 2.19: Section of a drift tube cell.

layer of cells is made of parallel aluminium plates, with cells obtained with perpendicular "I"-shaped aluminium cathodes. The anodes are 50 μm diameter steel wires placed between the cathodes. The distance of the track from the wire is measured by the drift time of electrons; to improve the distance-time linearity, additional field shaping is obtained with two positively-biased insulated strips, glued on the planes in correspondence to the wire.

The internal volume is filled with a gas mixture of Ar (85%) and CO_2 (15%) at atmospheric pressure, which provides good quenching properties and a saturated drift velocity of about 5.6 $cm/\mu s$. A single cell has an efficiency close to 100% and a resolution of about 180 μm.

Cathode strip chambers

Because of the large occupancy of the endcap regions, from few Hz/cm^2 to more than 100 Hz/cm^2, and the intense and inhomogeneous magnetic field, cathode strip chambers were chosen in this region.
CSC chambers are arranged in four disks (*stations*) placed between the iron disks of the yoke and named ME1,...,ME4 (see Figure 2.18). The CSCs, multiwire proportional chambers with good spatial and time resolution, are composed of one cathode plane segmented in strips orthogonal to the wires. An avalanche developed on a wire induces a distributed charge on the cathode plane. The orthogonal orientation of the cathode strips with respect to the wires allows the determination of two coordinates from a single detector plane, as shown in Figure 2.20. Each chamber is formed by 6 trapezoidal layers, with strips in the radial direction for a precise measurement of the azimuthal coordinate Φ. The wires

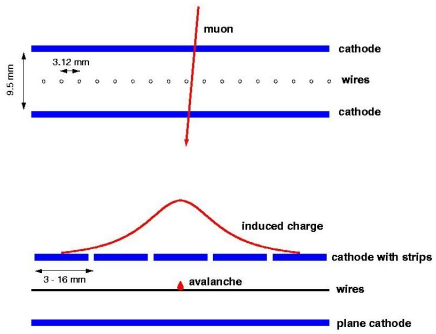


Figure 2.20: Orthogonal sections of a cathode strip chamber.

resolution is of the order of about 0.5 cm, while for the strips is of about 50 μm.

Resistive plate chambers

Resistive plate chambers are installed both in the barrel and in the endcap regions, in order to add robustness and redundancy to the muon trigger. They have a limited spatial resolution, but an excellent time resolution, of the order of few nanoseconds.
The RPCs used in CMS are composed of 4 bakelite planes forming two coupled gaps 2 mm thick, as shown in Figure 2.21. The gaps are filled with a mixture of 90% $C_2H_2F_4$ (freon) and 5% $i-C_4H_{10}$ (isobutane).

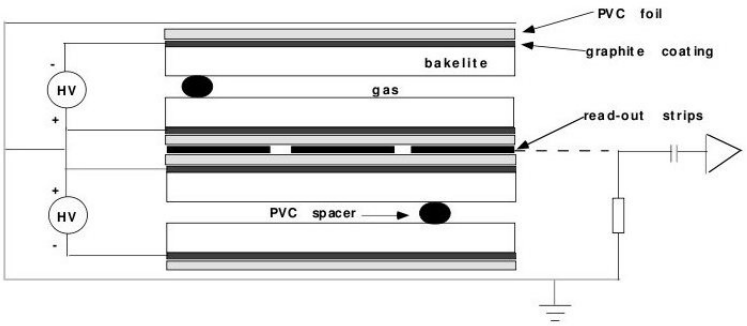


Figure 2.21: Section of a double gap resistive plate chamber.

They operate in avalanche mode rather than in the more common streamer mode. This is obtained with a moderate electric field across the gap which allows to sustain higher rate. However the gas multiplication is reduced, and improved electronic amplification is needed.

2.3 Summary

This chapter gave an overview of the LHC performances since its start and the structure of the CMS experiment. The electromagnetic calorimeter is presented in more detail with respect to the other subdetectors, since it plays a fundamental role in the detection of photons. The LHC data taking is going on very smoothly, so for the end of the year more than 40 fb^{-1} of data should be collected.

Chapter 3

Photon reconstruction and identification

One of my main contribution to the H $\rightarrow \gamma\gamma$ analysis concerns the study of photon reconstruction and photon identification, which is presented in this chapter. As a first step, I performed a comparison study between different kinds of photon reconstruction. As explained in Section 3.1, the photon reconstruction in Run 2 is very different from the Run 1 one. The aim of my study was to compare these different reconstructions, in particular the performances obtained by their respective photon identifications, and optimise the variables in the new reconstruction. The details and the results of this first study are presented in Section 3.3.1.
In a second step, I performed the training of the photon identification algorithm specific for the H $\rightarrow \gamma\gamma$ analysis using Monte Carlo samples at 13 TeV. As illustrated in Section 3.3.2, I demonstrated that a dedicated training was needed at 13 TeV, because more performant than the Run 1 training applied on 13 TeV samples. I performed this training testing different configurations: different values of the MVA parameters, different kinds of MVA algorithms, new variables added as input. The final training, used in the public H $\rightarrow \gamma\gamma$ analysis, was performed through the boosted decision tree technique. Furthermore, I studied the photon identification efficiency as a function of various kinematic variables and number of vertices in the event, both for signal and background.
Once 13 TeV data became available, I performed a comparison between data and simulation, in order to see to which extent the MVA inputs and output in data are well modeled by Monte Carlo. This validation was performed mainly using electrons reconstructed as photons in $Z \rightarrow ee$ events. Finally the photon identification systematics were assessed. The data-simulation comparison and the treatment of systematic uncertainties are presented in Section 3.3.3.

3.1 Photon reconstruction

A photon, produced at the interaction point, first passes through the tracker, then enters ECAL and loses all its energy through electromagnetic shower, which can be spread over several neighbouring crystals. The photon can have two different behaviours. In about 75% of the cases the photon is unconverted, it goes through the tracker without interacting and deposits about 94% (97%) of its energy into 3×3 (5×5) crystal matrix in the ECAL. In the remaining cases the photon converts to an electron-positron pair before entering the ECAL, the electron and positron bend under the magnetic field and deposit their energies in a larger region in ϕ. To include all the photon energy deposits, photons are thus reconstructed by clustering the energy deposits in the ECAL crystals into so-called superclusters, which are a collection of EM clusters close together [36, 37, 38].
The clustering algorithm used in Run 2 is different from that used in Run 1, allowing for a finer reconstruction of the photon energy. In the following section a more detailed description of the clustering algorithm is given, with a particular attention to the differences between Run 1 and Run 2.

3.1.1 Clustering

Clustering in Run 1

Photons are reconstructed in the detector pseudorapidity range $|\eta| < 1.479$ in the barrel and $1.479 < |\eta| < 2.5$ in the endcap. Clustering of ECAL shower energy is performed on intercalibrated, reconstructed signal amplitudes. The clustering algorithm collects the energy from radiating electrons and converted photons that get spread in the ϕ direction by the magnetic field. This algorithm evolves from fixed matrices of 5×5 crystals, which provide the best reconstruction of unconverted photons, by allowing extension of the energy collection in the ϕ direction, to form "superclusters".
Clusters are built starting from a "seed crystal", which is the crystal with highest E_T above a certain threshold among the crystals not included in any other cluster yet. In the barrel, where the crystals are arranged in an (η, ϕ) grid, the clusters have a fixed width of five crystals centered on the seed crystal, in the η direction. Then, 5×1 matrices of crystals (bars) each centered at the same η of the seed crystal are built, within the range $\pm$ 17 crystals in ϕ from the seed crystal. The bars with total energy above a certain threshold connected in ϕ are further grouped into clusters called basic clusters. The basic clusters with the highest bar energy above a certain threshold are finally grouped to form a supercluster. In Figure 3.1 a schematic view of the $\eta - \phi$ window opened to build the supercluster in the barrel is shown.
Clustering in the endcaps uses fixed matrices of 5×5 crystals. After a seed cluster has been defined, further 5×5 matrices are added if their centroid lies within a small η window and within a ϕ distance roughly equivalent to the 17 crystals span used in the barrel. The

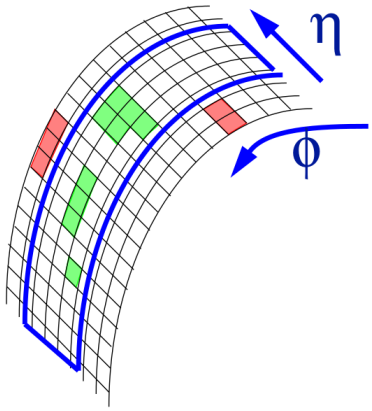


Figure 3.1: A supercluster algorithm collects all calorimetric clusters, satisfying a given geometric condition (e.g. lying in a certain region around the "main" cluster), into a collection of clusters.

5×5 matrices are allowed to partially overlap one another. For unconverted photons, the superclusters resulting from both the barrel and endcap algorithms are usually simply 5×5 matrices.
The raw photon energy E_{RAW} is obtained by summing the calibrated energy deposits in the crystals of the supercluster, and the energy deposited in the preshower detector is added for photons in the endcap. The photon position in $\eta - \phi$ is obtained from the mean position of basic clusters weighted by energy, and the position of basic clusters is calculated from mean positions of crystals corresponding to the shower depth weighted by the logarithm of the crystal energy [38].

Clustering in Run 2

The algorithm for the photon reconstruction in Run 2 is quite different, allowing for a finer reconstruction of the photon energy. The main changes applied to the reconstruction are listed below:

- Particle Flow clustering: the particle flow technique, which consists in a combination of the information from all sub-detectors, allowing the improvement of the global event description, was just partially used during Run 1. In Run 2 the potential of this technique is fully exploited, with the so-called particle flow (PF) clusters grown from the seed by aggregating crystals with at least one side in common with a cell already included in the cluster, and with an energy above a given threshold (zero suppression method). This threshold represents about two standard deviations of the electronic noise in the ECAL, that is 80 MeV in the barrel and up to 300 MeV in the endcaps. Furthermore, an energy-sharing algorithm was developed. It consists

in sharing the energy of each cell among contiguous clusters according to the cell-cluster distance with an iterative determination of the cluster energies and positions assuming a gaussian shower profile.
A more detailed description of the particle flow algorithm is given in Section 3.1.2, where it is explained how this technique plays an important role not just in the cluster growing, but also in the supercluster building.

- Dynamic Superclustering: PF clusters are dynamically merged into superclusters. Dynamic superclustering allows good energy containment, robustness against pileup and takes into account the detector geometrical variations with η (e.g. endcap crystal size). Clusters lying in the area between two parabolas, function of η and centered around the most energetic cluster, are dynamically gathered giving to the supercluster a mustache-like shape. This is important especially moving to higher $|\eta|$ regions and for low energy clusters, as the shape of the shower starts to extend also in η. A fixed supercluster size, as was used in Run 1, would suffer from pileup in low-pseudorapidity region and miss some bremsstrahlung electron at high-η. In Figure 3.2 the mustache-effect is shown. The $\Delta\phi$ and $\Delta\eta$ distances between each cluster and the max-energy cluster are shown for $|\eta|$ going from 0 to 3, in slices of $|\eta| = 0.25$. It is evident that moving to higher $|\eta|$ the shower shape becomes different and starts to extend not only in ϕ but also in η. The mustache shape depends from the cluster energy and it is more pronounced for clusters with low energy, of about 1 GeV. The clusters recovered with the mustache clustering are originating from electrons coming from converted photons, the latter probably being low energy bremsstrahlung.

In Run 2 variables describing shower shapes can be computed in two manners:

- Particle Flow: only crystals from PF clusters are used,
- "full5×5": all crystals in a 5×5 matrix around the seed are used and both the energy sharing fraction and the zero suppression are ignored (zero suppression excludes crystals with an energy lower than a certain threshold). This is very similar to Run 1.

I performed some comparison studies between the Run 1 reconstruction (called "RECO" in the following), the standard particle flow reconstruction (called "GED") and the "full5×5" one (called "GED5×5"). These studies are presented in Section 3.3.1 of this chapter.

3.1.2 Particle flow

The particle flow algorithm [39] allows to exploit the versatility of the CMS apparatus in an attempt to identify and reconstruct individually each particle arising from LHC proton-proton collisions with a combination of the information from all sub-detectors. This can

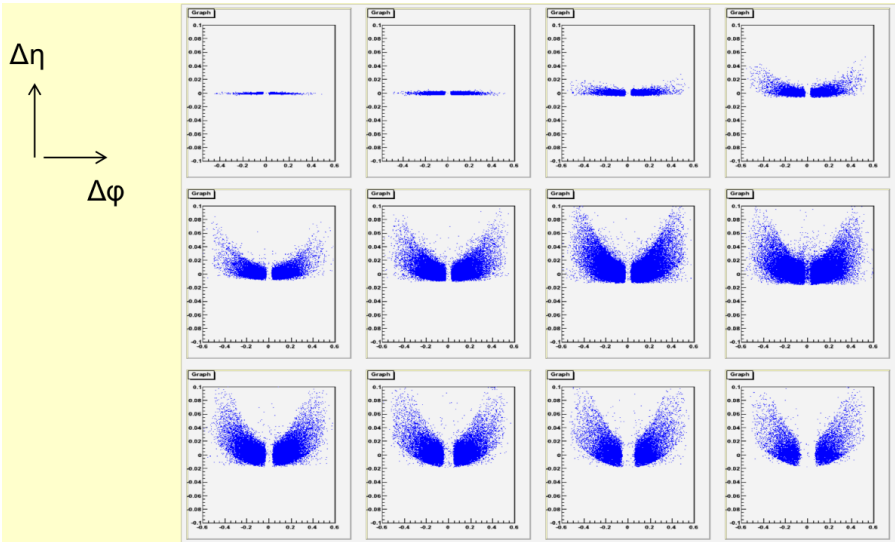


Figure 3.2: Representation of $\Delta\phi$ and $\Delta\eta$ distances between each cluster and the max-energy cluster for $|\eta|$ going from 0 to 3, in slices of $|\eta| = 0.25$. The simulation is done using a photon gun with p_T between 10 and 100 GeV. Taken from an internal presentation.

lead to a better global event description. The reconstruction and the identification of each particle are optimised, allowing to avoid redundancy while keeping a good level of precision.
An important ingredient for the success of the particle flow technique is the fine spatial granularity of the sub-detectors. Indeed, if the sub-detectors are coarse-grained the signals from different particles could merge, reducing the identification and reconstruction capability for the different types of particles. If the granularity is instead sufficient to provide a good separation between individual particles, a complete event description becomes possible.
The particle flow algorithm proceeds in steps, associating clusters and tracks with a newly reconstructed particle at each progressive step. In the following the different steps of the particle flow algorithm are briefly described:

- first, clusters and tracks are gathered in a list called "unassociated objects";
- tracks and clusters identified as being associated with hits and segments in the muon chambers are tagged as muons and removed from the list of unassociated objects;
- then, tracks and clusters identified as being associated with electrons (and all the possible individual bremsstrahlung photons) are tagged and removed from further processing;
- next, in the case of HCAL clusters linked to a track, the calibrated HCAL cluster energy is compared with the track momentum. If the cluster energy is compatible

with the track momentum, a charged hadron is created with energy determined from a weighted average of the track momentum and the cluster energy. If the difference between the cluster energy and the track momentum is significant (with respect to the expected calorimeter energy resolution and the measured track momentum uncertainty), a neutral hadron is created from the cluster energy in excess;

- similarly, if an ECAL cluster and an HCAL cluster are linked together with a track, the calibrated combined energy of the ECAL cluster and HCAL cluster is compared with the track momentum. If the combined ECAL and HCAL calorimeter energy is compatible with the track momentum, a charged hadron is created, otherwise either a neutral hadron or a photon is created from the excess of calorimeter energy. This is done with a multivariate analysis that takes into account the track momentum, the relative energy deposits in ECAL and HCAL, the cluster-track link quality and the transverse cluster shapes;

- once these tracks and clusters have been removed from the list of unassociated objects, only clusters not linked to any track remain uncleaned from the event. In this case any ECAL clusters are assumed to be photons and any HCAL clusters (or HCAL clusters linked with ECAL clusters) are assumed to be neutral hadrons.

The particle flow algorithm allows to improve in particular the jets and taus reconstruction, along with the missing transverse energy and isolation computation. For instance, jets are commonly reconstructed by clustering the energy deposits observed in the calorimeters. The particle flow algorithm, on the other hand, makes it possible to build jets of reconstructed particles, thereby resolving the jet constituents and providing unique insight on the jet substructure and the parton fragmentation process.
An example of the benefit brought by the particle flow approach is shown in Figure 3.3, where it can be seen that the tau jets PF reconstruction is characterised by an improved energy and angular resolution with respect to the calorimeter-based algorithm. In fact, the limited energy and angular resolution of the calorimeter-based jets is dominated by the hadron calorimeter resolution and granularity. Since the tau decays mostly in photons and charged pions, the particle flow-based jets benefit from the tracker and electromagnetic calorimeter better resolutions.

3.2 Photon identification

3.2.1 Principles of photon identification

The great challenge of the Higgs to diphoton decay channel, as shown in Figure 3.4, is to identify a small peak in the diphoton mass distribution over a background that is several orders of magnitude larger. Diphoton events include potential Higgs signal events but mostly

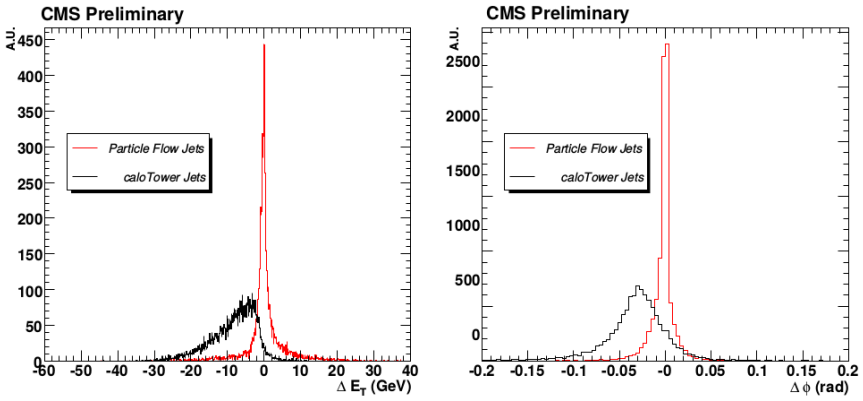


Figure 3.3: Comparison between particle flow reconstruction (red) and calorimeter-based (black) reconstruction of single taus with $p_T = 50$ GeV. In the left plot the difference, in GeV, between the reconstructed and the true visible transverse momentum is shown. The right plot shows the difference, in radian, between the reconstructed and the true azimuthal angle. Taken from Reference [39].

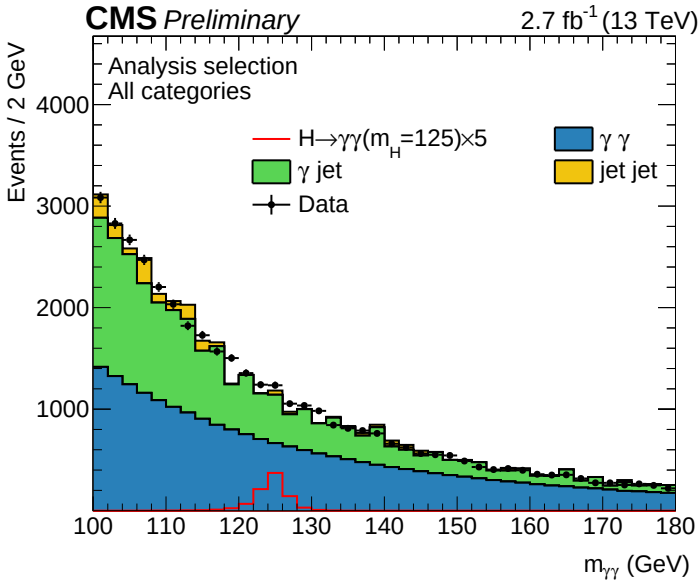


Figure 3.4: Diphoton invariant mass distribution for events passing the selection of the analysis, for data and simulation. Backgrounds are represented by the filled histograms, while signal for a Higgs boson with a mass of 125 GeV (scaled by a factor of 5) is shown by the red line.

a mixture of background events. The backgrounds can be separated in two components, "irreducible" and "reducible". The "irreducible" component consists of prompt diphoton events, where a prompt photon originates from the hard scatter vertex. The "reducible" component includes dijet and γ + jet events, in which jets are misidentified as photons (fake). The primary mechanism for a jet to fake a photon involves the fragmentation of the majority of the jet energy into a leading π^0 or η meson, which subsequently decay to two photons. For the energy range of interest, the π^0 or η are significantly boosted, such that the two photons from the decay are nearly collinear and are difficult to distinguish from a prompt photon.
The goal of photon identification is to reduce these backgrounds, in general applying isolation requirements against additional energy from jet fragmentation, as well as exploiting differences in shower profile in the calorimeter to distinguish between a prompt incident photon and a pair of incident photons from a neutral meson decay.
In Section 3.2.2 the variables able to discriminate between prompt and fake photons are presented. The discriminating power of these variables can be used through several methods, as explained in Section 3.2.3.

3.2.2 Discriminating variables

In general two groups of variables can be used for discrimination between prompt and fake photons, because of their connection with the two intrinsic differences between a prompt photon and a fake photon: the electromagnetic shower shape variables and the isolation variables. As discussed, the most photon-like jets result from fragmentation into a leading π^0 or η, which subsequently decay into a pair of photons with small opening angle. Even if the two photons cannot be cleanly distinguished, such objects have wider shower profiles on average than a single photon in the calorimeter. This is true in particular along the η axis of the cluster, since the discriminating power resulting from the Φ profile of the shower is partially washed out by the effect of the magnetic field, which can expand the electromagnetic cluster in the Φ direction for both converted photons. In addition, because π^0 or η result from jet fragmentation, there are in general additional charged and neutral particles produced. Since jets tend to be collimated objects, these additional particles are likely to be produced close to the reconstructed photon in the detector. This leads to a class of discriminating variables called isolation, which generally consist of the sum of some particular type (EM, charged...) of energy in a cone around the reconstructed object. To do this, it is convenient to define cones around the direction of the photon, with cone radius defined by $R = \sqrt{\Delta\eta^2 + \Delta\Phi^2}$, where $\Delta\eta$ and $\Delta\Phi$ are the distances in η and Φ, respectively, between the photon direction and the direction of selected objects in the cone. In order to ensure that the energy from the photon itself is not included in this sum, we define a smaller veto region inside the cone, which is excluded from the isolation cone. Two isolations are included:

- one with respect to the best estimate of the primary vertex,
- one which is the worst isolation of a photon among all isolations computed with respect to all reconstructed vertices. This is due to the fact that a prompt photon is generally isolated with respect to all vertices, while fake photons are not.

Event-level variables are also included such that the distributions of shower shape and isolation variables are used differentially as function of pileup contamination measured by ρ, and photon kinematics measured by η_{SC} and E_{RAW}.
In order to reduce the photon kinematic dependence of the photon identification BDT and the associated mass dependence in the diphoton BDT, explicit use of kinematic differences between prompt photons and fake photons in the training sample is avoided by reweighting the 2D p_T-η_{SC} distribution of the signal to that of the background.

The input variables for the photon identification BDT are given below, with their definition:

- Shower shape cluster variables:
  - R_9: $E_{3\times3}/E_{SC}$, where $E_{3\times3}$ is the energy sum of the 3×3 crystals surrounding the supercluster seed crystal and E_{SC} is the energy sum of the supercluster.
  - S_4: $E_{2\times2}/E_{5\times5}$, the ratio of the maximum energy 2×2 crystal grid and the energy in a 5×5 crystal grid centered on the seed crystal.
  - $\sigma_{i\eta i\eta}$: the log-energy weighted standard deviation of single crystal η in crystal index within the 5×5 crystals centered at the seed crystal. The weight per-crystal is 4.7 plus the logarithm of the ratio between the energy in the crystal and the energy in the 5×5 crystals. If the weight is negative then 0 is used instead.
  - $\text{cov}_{i\eta i\Phi}$: the log-energy weighted covariance of single crystal $\eta - \Phi$ in crystal index within the 5×5 crystals centered at the seed crystal.
- Shower shape supercluster variables:
  - SC η-width (σ_η): the energy-weighted standard deviation of single crystal η in detector coordinate within supercluster. The weight per-crystal is the ratio of the single crystal energy to the supercluster energy.
  - SC Φ-width (σ_Φ): the energy-weighted standard deviation of single crystal Φ in detector coordinate within supercluster. The weight per-crystal is the ratio of the single crystal energy to the supercluster energy.
  - Preshower σ_{RR} (endcap only): the sum in quadrature of the energy-weighted standard deviation of the strip index in the x and y planes of the preshower detector.

These variables have also the full 5×5 version, which uses only 5×5 matrix around the seed crystal, and does not use PF energy sharing neither zero suppression (this means that all crystals are included).

- Particle flow-based isolation variables:
  - *PF Photon ISO*: transverse energy sum associated with all particles identified as photons by the particle flow algorithm falling inside a cone of size $R = 0.3$ around the photon candidate direction. For photons in the barrel, an inner veto strip of $|\Delta\eta| < 0.015$ is excluded from the isolation sum, and for photons in the endcap instead an inner veto cone of $\Delta R < 0.07$ is excluded.
  - *PF Charged ISO (selected vertex)*: transverse energy sum associated with all particles identified as charged hadrons by the particle flow algorithm falling inside a cone of size $R = 0.3$ around the photon candidate direction, excluding an inner veto cone of $R < 0.02$. It is measured with respect to the selected primary vertex.
  - *PF Charged ISO (worst vertex)*: transverse energy sum associated with all particles identified as charged hadrons by the particle flow algorithm falling inside a cone of size $R = 0.3$ around the photon candidate direction, excluding an inner veto cone of $R < 0.02$. It is measured with respect to the worst vertex, that is the one which yields the largest isolation sum.
- Auxiliary variables:
  - ρ: the estimate of transverse energy per unit area in the $\eta - \Phi$ plane contributed by the pileup interactions and underlying-event effects in the event. This is constructed from the median transverse energy density of all anti-k_T-reconstructed jets in the event, where the anti-k_T algorithm [40] tends to produce a large number of soft jets such that the median of this distribution is relatively insensitive to the hard interaction.
  - SC η: the η of the supercluster corresponding to the reconstructed photon, computed from the pseudorapidity of the vector joining the point (0,0,0) to the reconstructed supercluster position in the electromagnetic calorimeter.
  - SC E_{RAW}: the sum of crystal energy in the supercluster corresponding to the reconstructed photon.

Other variables are interesting because used in the preselection of the H $\rightarrow \gamma\gamma$ analysis (see Section 4.4):

- H/E: the energy collection by the HCAL towers within a cone of R=0.15 centered on the supercluster position, divided by the supercluster energy.

- *Tracker isolation in a hollow cone* (TrackerIso in Table 4.2): transverse momentum sum associated with all tracks falling in a cone size of R=0.3 around the photon candidate direction. Tracks falling in an inner cone of size R=0.04 are not included in the p_T sum.

3.2.3 Methods for photon identification

Cut-based technique

The goal of photon identification is to discriminate as well as possible prompt from fake photons. The first ones are thus considered as signal, the second ones as background. A simple approach to this problem is to encode information about the photons present in the event into a set of one-dimensional variables, and then apply simple cuts on those variables. In the analysis photons which pass the cuts are kept, those which fail are discarded. This approach, called "cut-based identification", is easy to understand and to describe, but has a few important drawbacks:

- events are either retained or discarded. In principle events can be used in a more fine-grained way, according to how signal-like or background-like they are,
- correlations between variables are neglected. Generally it is difficult to encode all of the relevant information into fully uncorrelated variables, and therefore neglecting correlations between variables leads to a loss of discriminating power,
- the selection of cut values which optimize the performance of an analysis is a complex problem, especially with a large number of variables. Most cut-based analyses therefore rely on a relatively small number of variables and/or use suboptimal cut values, again neglecting potentially useful information.

Boosted Decision Trees (BDT)

The optimal separation between two classes of events given a set of variables $\bar{x}$, is given by the likelihood ratio [41]

$$L_R = \frac{\mathcal{L}_s(\bar{x})}{\mathcal{L}_s(\bar{x}) + \mathcal{L}_b(\bar{x})}$$

where $\mathcal{L}_s(\bar{x})$ and $\mathcal{L}_b(\bar{x})$ represent the full multidimensional likelihood functions for signal and background events respectively. If the set of variables $\bar{x}$ encodes all of the relevant information, then this likelihood ratio contains all of the relevant information for distinguishing the two classes of events. This ratio quantifies the probability that a given event with features $\bar{x}$ is a signal event as opposed to a background event.
In high energy physics, the input variables $\bar{x}$ are the output of a complicated detector response, and therefore there is no known analytic form for the multidimensional likelihood

functions above. Therefore this likelihood is estimated using a finite sample of events representing each class, either from Monte Carlo simulation or from carefully defined control regions in data. There are many techniques designed to estimate the multidimensional likelihood ratio from a finite set of training events in this case. One such technique is the Boosted Decision Tree, able to address some of the limitations listed for the cut-based technique.

Boosted Decision Trees [42, 43, 44] is one of the MVA (multivariate analysis) techniques employed in experimental particle physics to estimate a multidimensional function. If the function has discrete values, like signal or background, this is called classification. If the output is continuous, like energy corrections, this is called regression. We use a BDT in this analysis for its ability to handle large number of input variables and their correlations, as well as its simple mechanism and robustness against overtraining. BDT are used to combine all the relevant information in an event into a single variable which discriminates signal from background for classification (for example for the photon identification) or precisely estimates a particular target property (for the regression, see Section 5.11). Whereas a cut-based technique is able to select only one hypercube of the phase space, the decision tree is able to split the phase space into a large number of hypercubes, each of which is identified as either "signal-like" or "background-like".
To train a BDT we provide two simulated samples with known identity, one for the signal and one for the background, and a set of input variables with discriminating power $\vec{x} = \{x_1, x_2, ..., x_n\}$. A single decision tree, which is a binary tree structured classifier like the one sketched in Figure 3.5, is first trained.

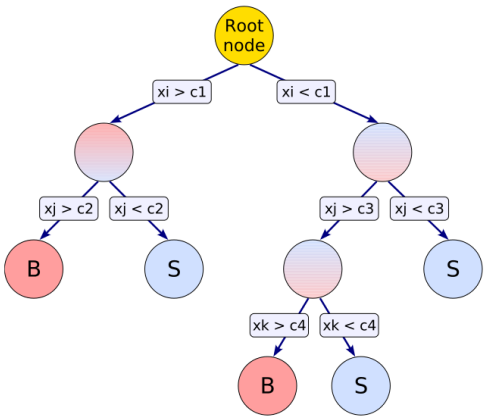


Figure 3.5: Schematic view of a decision tree.

Starting from the root node representing the entire variable phase space, a sequence of binary splits using the discriminating variables x_i is performed. Every split uses the variable that at this node gives the best discrimination between signal and background when

being cut on. The same variable may be used at several nodes, while others might not be used at all. The terminal nodes at the end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes. In more detail, the tree building starts from the root node, with the number of signal and background events reweighted such that both total weights for signal and for background are equal. The node is then split by selecting a single variable and a cut value on it. The splitting continues iteratively until a predetermined limit is reached, such as a maximum depth of the tree or a minimum number of events in a node. The limit is chosen to reduce the bias due to statistical fluctuation of the training samples, the overtraining.
A shortcoming of the single decision trees is their instability with respect to statistical fluctuations in the training sample from which the tree structure is derived. This instability can cause an overtraining and a misclassification for some of the events in the final nodes. This problem is overcome by constructing a forest of decision trees, all derived from the same training sample, with the events undergoing subsequently so-called "boosting", a procedure which modifies their weights in the sample. With this method we train a set of trees and we assign a score to an event as the weighted average of the scores of all the trees.
For the photon identification studies we use the boosting procedure called Gradient Boost, whose expression is as following:

$$F(\vec{x};P) = \sum_{m=1}^{M} \beta_m f(\vec{x};\alpha_m),$$

where $F(\vec{x};P)$ represents the function with the set of parameters P corresponding to the BDT made up of M trees, $f(\vec{x};\alpha_m)$ is the function corresponding to the m_{th} tree, α_m represents the parameters of the m_{th} tree including the splitting variables and cut values at each node, and β_m is the weight of the m_{th} tree. The parameter set P is determined by minimizing the deviation between the estimates provided by $F(\vec{x};P)$ and the true identities of the training events, given by the loss function:

$$L(F(\vec{x};P),y) = \sum_{n=1}^{N} ln(1+e^{-2F_n(\vec{x};P)y_n}),$$

where $F_n(\vec{x};P)$ is the estimated value for the n_{th} event, and y_n represents the true value $+1$ or -1 of the n_{th} event, and N is the total number of events.
The trained BDT function is then used to assign a score to each event, given its values of the input variables. The score is a continuous variable varying from -1 to $+1$, the more signal-like an event is, the higher value it gets.
A classifier based on a Boosted Decision Tree trained on a set of variables containing all of the relevant information itself contains all of the information needed to discriminate between signal and background. The simplest possible usage of such a BDT is to place a

cut on the output. Cutting on the BDT output rather than on the input variables directly addresses two of the three limitations listed for cut-based technique, namely that it properly exploits correlations among the input variables and is relatively simple to optimize.
This cut-based usage still retains the drawback of completely accepting or discarding events. In the case of the H $\rightarrow \gamma\gamma$ analysis, the photon identification multivariate discriminator is fed forward to the per-event multivariate discriminator described in Section 4.8, after a very loose cut on it.
For the photon identification studies we use a Toolkit for Multivariate Data Analysis (TMVA) [42] within CERN's ROOT framework [45] to train the BDT.

3.2.4 Training samples

The BDT technique is thus used in the H $\rightarrow \gamma\gamma$ analysis to perform the photon identification. As mentioned in Section 3.2.3, to train a BDT we provide two simulated samples, one signal and one background, each sample containing the set of input variables with discriminating power $\vec{x} = \{x_1, x_2, ..., x_n\}$. The prompt and fake photons for training, used as signal and background respectively, are both taken from the γ + jet Monte Carlo samples listed in Section 4.1.6. The events in these samples generally contain one prompt photon and one electromagnetically enriched object originating from a jet (more details on the EM-enrichment applied to some background MC samples can be found in Section 4.1.6). In order to properly cover the phase-space, there are two γ + jet samples used for the training. They are weighted according to the cross-section since they cover different p_T ranges. Prompt photons and non-prompt photons are required to pass the H $\rightarrow \gamma\gamma$ analysis preselection as defined in Section 4.4. Just the p_T requirement on the leading (subleading) photon is loosen from 30 (20) GeV to 18 GeV with respect to the preselection, which gives a p_T threshold of 18 GeV for both photons.
The signal sample consists of reconstructed photons which match prompt photons at the generator level, which means they originate from a quark. The fake photons belonging to the background sample are all the remaining photons.
50% of the γ + jet samples is used for the training, while the remaining is used for testing, that is to recompute the photon identification output variable using the trained BDT.

3.3 Results

In the next two sections my studies concerning the different photon identification algorithms and the relevant results on the photon identification are presented. Monte Carlo samples might differ from one study to another, yielding some slight differences in the variable shapes or in the performances achieved.

3.3.1 Results of photon reconstruction study

The first part of my study was dedicated to the comparison between Run 1 and Run 2 photon reconstruction from the identification point of view. These two reconstructions are called "RECO" (Run 1) and "GED" (Run 2) in the following. In addition, as explained in Section 3.1.1, a third identification was studied for Run 2 reconstruction, and it is called "GED5×5" in the following. We will refer to "RECO", "GED" and "GED5×5" as different reconstructions, though strictly speaking "GED" and "GED5×5" have the same clustering. There are three main steps: first I analysed the shower shape and isolation variables for the different reconstructions. Then I compared the performances of the different reconstructions using, for all of them, the BDT trained for Run 1 conditions. Finally I compared the performances of the reconstructions, this time after having performed a dedicated training at 13 TeV for each of them.

In Figures 3.6, 3.7 and 3.8 the distributions of shower shape and isolation variables are shown, for prompt photons (signal) and fake photons (background) belonging to γ + jet simulated sample and for the different reconstructions. These distributions are obtained applying the preselection cuts of Run 1 analysis. These cuts are in general quite similar to those presented in Section 4.4, and are responsible for discontinuity in the distribution of *PF Charged ISO (worst vertex)* and R_9.
In general the variable distributions of the three reconstructions are quite similar. The variables that show the biggest discrepancies between "RECO" and particle flow-based reconstructions are SC η-width and SC Φ-width. This is expected because these are supercluster level variables, whereas most of the others are related only to the seed cluster. In fact the superclusters are different between Run 1 and Run 2, because of the difference in the geometry of the clustering (mustache profile versus simple rectangular $\eta-\Phi$ region).

Looking at the differences in the variables is not sufficient to compare the three reconstructions. It is indeed necessary to compare the photon identification performances in the three cases. As a first step, the photon identification (in the following "ID MVA") output variable for the three reconstructions was computed using the Run 1 training. This is clearly a rough approach, since this training was obtained with the RECO reconstruction at 8 TeV, so it is non-optimized for the GED and GED5×5 reconstructions. Nevertheless in this way it is possible to have a first idea of the performances of the different reconstructions without doing a new training.
Once the ID MVA output variable, like the one in Figure 3.18, is obtained, it can be used to calculate the so-called "ROC curve", which is the background efficiency as a function of the signal efficiency obtained for several cuts of the ID MVA output variable. The more the curve is on the bottom right part of the plot, the better is the ID MVA performance, because in this region the signal efficiency is high while the background efficiency is low.

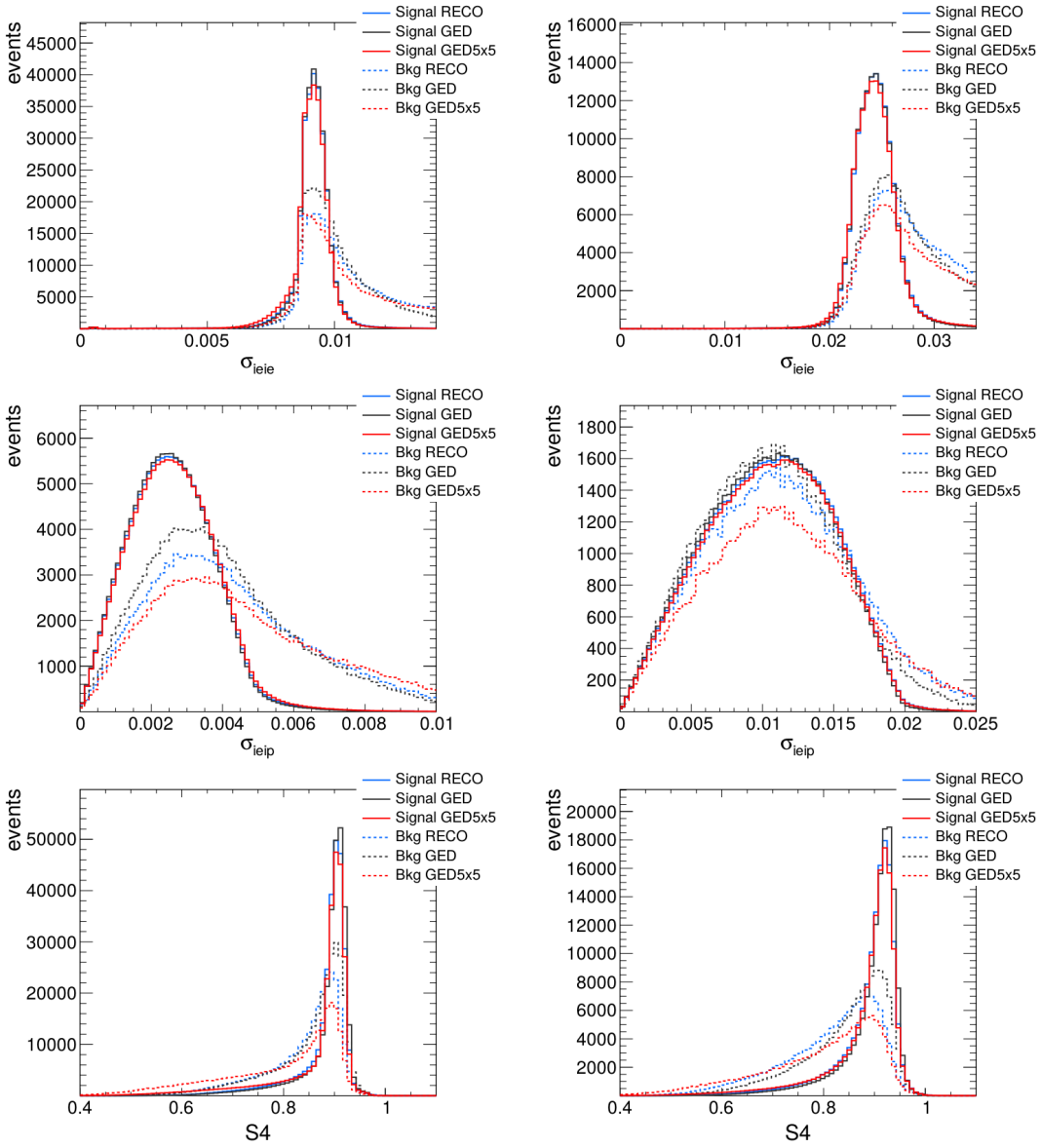


Figure 3.6: Distributions of shower shape variables $\sigma_{i\eta i\eta}$ (top row), $\sigma_{i\eta i\Phi}$ (middle row) and S_4 (bottom row) for signal prompt photons (solid line) and background fake photons (dashed line) in the barrel (left) and in the endcap (right). The distributions are shown for the three reconstruction algorithms, RECO (blue), GED (grey) and GED5×5 (red). Photons are from γ + jet samples passing the Run 1 preselection.

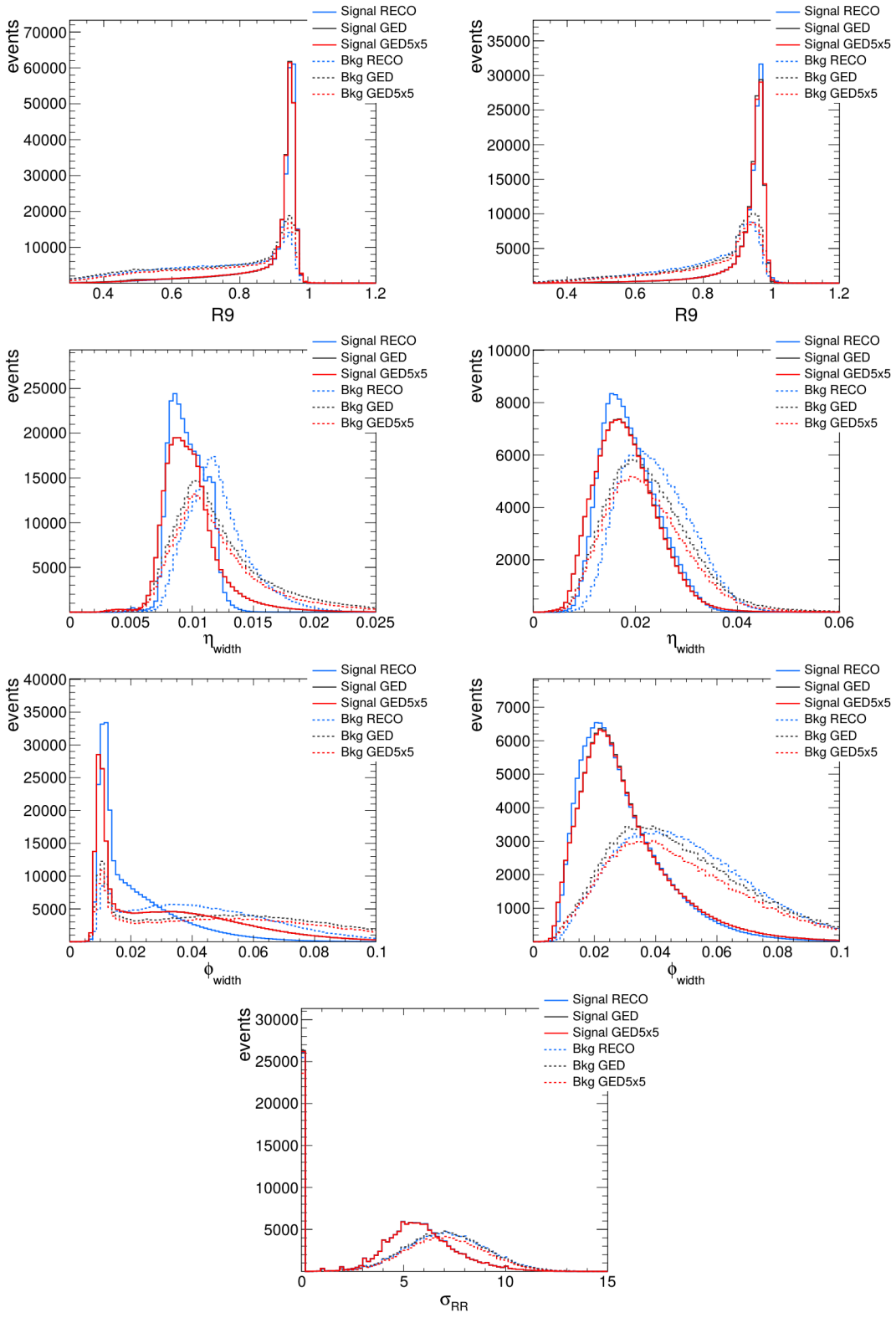


Figure 3.7: Distributions of shower shape variables R_9 (first row), SC η-width (second row) and SC Φ-width (third row) for signal prompt photons (solid line) and background fake photons (dashed line) in the barrel (left) and in the endcap (right), along with the distribution of preshower σ_{RR} (fourth row) for photons in the endcap only. The distributions are shown for the three reconstruction algorithms, RECO (blue), GED (grey) and GED5×5 (red). Photons are from γ + jet samples passing the Run 1 preselection.

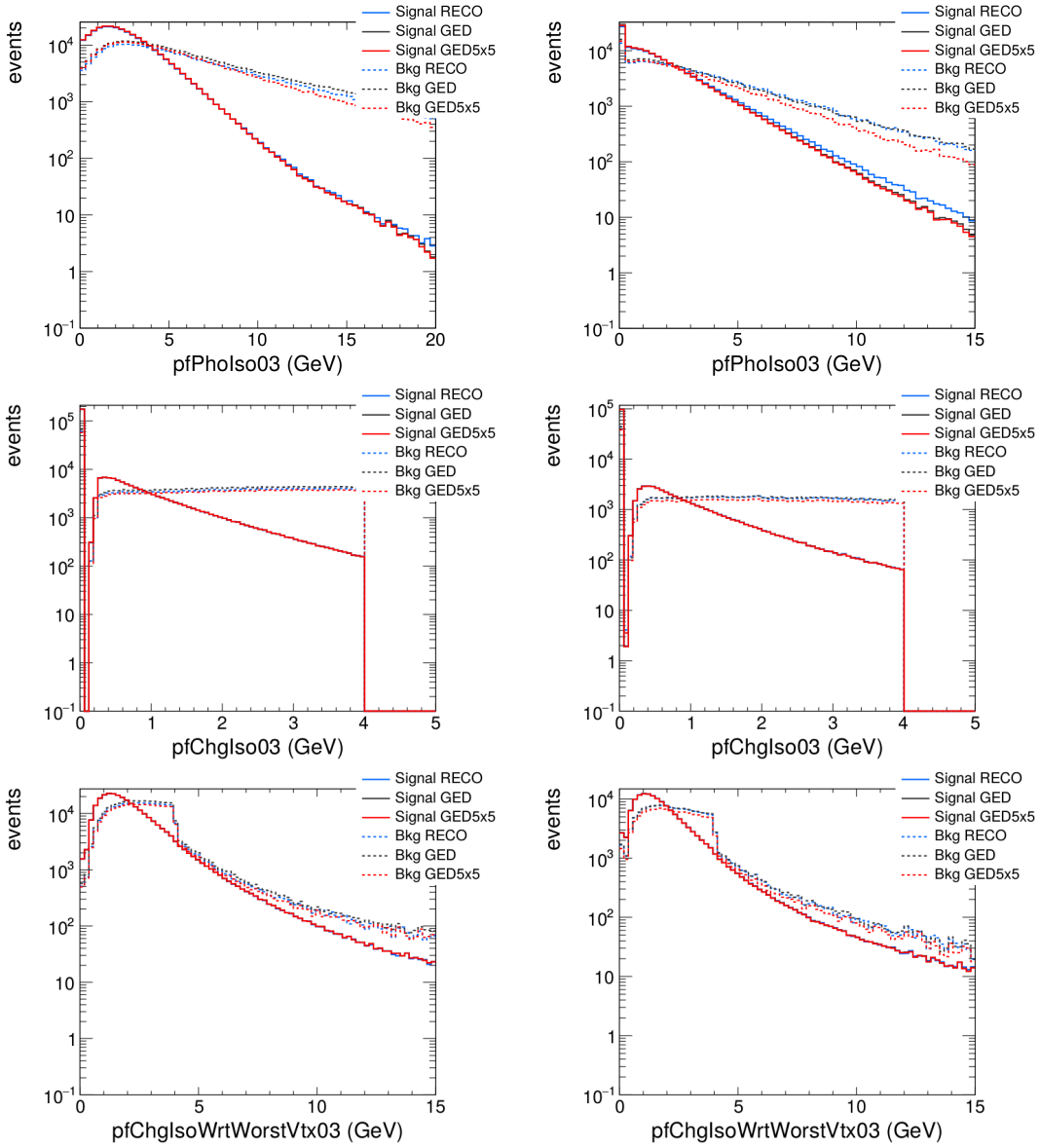


Figure 3.8: Distributions of isolation variables *PF Photon ISO* (top row), *PF Charged ISO (selected vertex)* (middle row) and *PF Charged ISO (worst vertex)* (bottom row) for signal prompt photons (solid line) and background fake photons (dashed line) in the barrel (left) and in the endcap (right). The distributions are shown for the three reconstruction algorithms, RECO (blue), GED (grey) and GED5×5 (red). Photons are from γ + jet samples passing the Run 1 preselection.

The ROC curve is a good figure of merit for the photon identification discriminating power. Therefore the ROC curves for each reconstruction are obtained, starting from the ID MVA output obtained applying the Run 1 training. Figure 3.9 shows the ROC curves obtained, in blue for RECO, in grey for GED and in red for GED5×5 reconstruction, for both barrel and endcap. From these plots it is evident that the Run 1 reconstruction has a better performance than the GED one, while the use of GED5×5 variables allows to have performances similar to RECO.
Nevertheless one has to remember that the training used in this case is optimised for RECO, so it is possible that the loss in performance of the particle flow-based reconstructions is just an artifact; furthermore, GED5×5 has a better performance than GED because it is more similar to RECO algorithm.
For this reason a dedicated training at 13 TeV for each reconstruction was performed, and the resulting ROC curves were compared. The result is quite different with respect to the previous one; as shown in Figure 3.10, with three dedicated training the different reconstructions have similar performances (solid lines represent the ROC curves coming from the dedicated trainings, while dashed lines represent the ROC curves obtained in the previous study).

It has thus been shown that the photon identification has the same performances for the Run 1 reconstruction and for the particle flow-based ones. The little differences between the three reconstructions present in the variable distributions seem to have no impact on the final performances of the photon identification and the MVA training is able to compensate by the correlations between variables.
It was eventually decided to use the GED5×5 reconstruction in the final $H \to \gamma\gamma$ analysis, which is slightly better than GED (see Figure 3.10).

3.3.2 Results of photon identification study

Discriminating variables and training tests

Figure 3.11 shows the p_T and η_{SC} distributions, for both barrel and endcap, for fake photons and for prompt photons before and after the 2D p_T-η_{SC} reweighting mentioned in Section 3.2.2. The distributions of the input variables for the signal and background training samples after the 2D reweighting are shown in Figures 3.12, 3.13, 3.14 and 3.15. The discontinuities visible in some variables are due to the preselection cuts. As expected, fake photons have a shower profile wider than prompt photons, and the isolation values are in general larger for the fake photons.

Before doing the final training, various trainings with different BDT parameters, different MVA techniques and input variables are performed. In the "Boosting" technique the final response is determined by a possibly large number of trees, so it is not necessary for

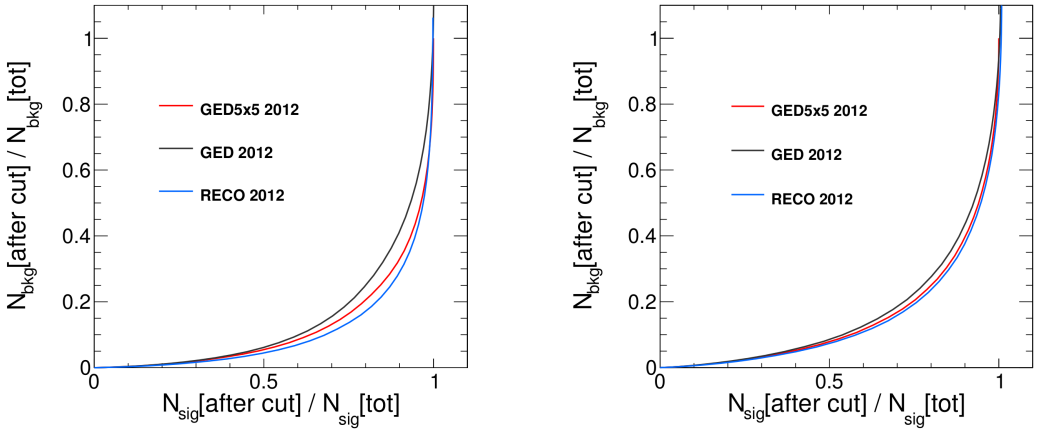


Figure 3.9: ROC curves obtained applying the Run 1 weights in the computation of ID MVA output variable, in blue for RECO, in grey for GED and in red for GED5×5 reconstruction, for both barrel (left) and endcap (right).

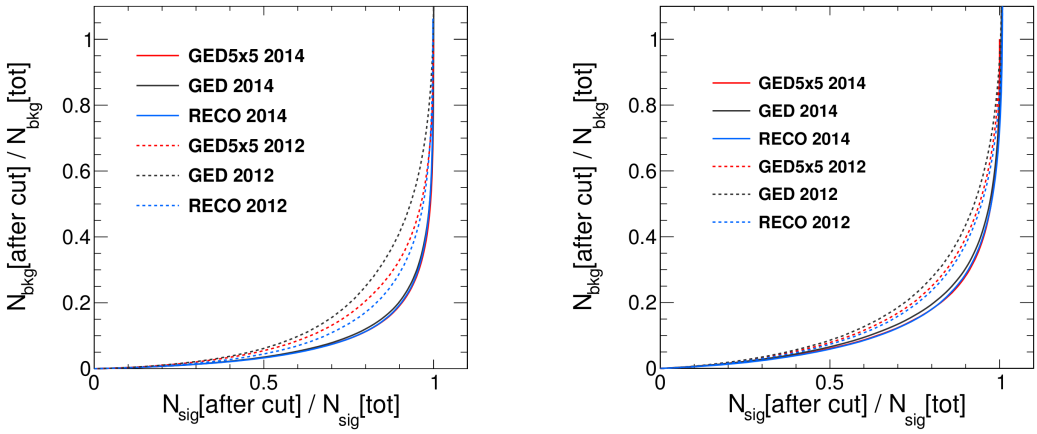


Figure 3.10: ROC curves obtained applying the Run 1 weights in the computation of ID MVA output variable (dashed lines) and ROC curves obtained with a dedicated training at 13 TeV for each reconstruction (solid lines), in blue for RECO, in grey for GED and in red for GED5×5, for both barrel (left) and endcap (right).

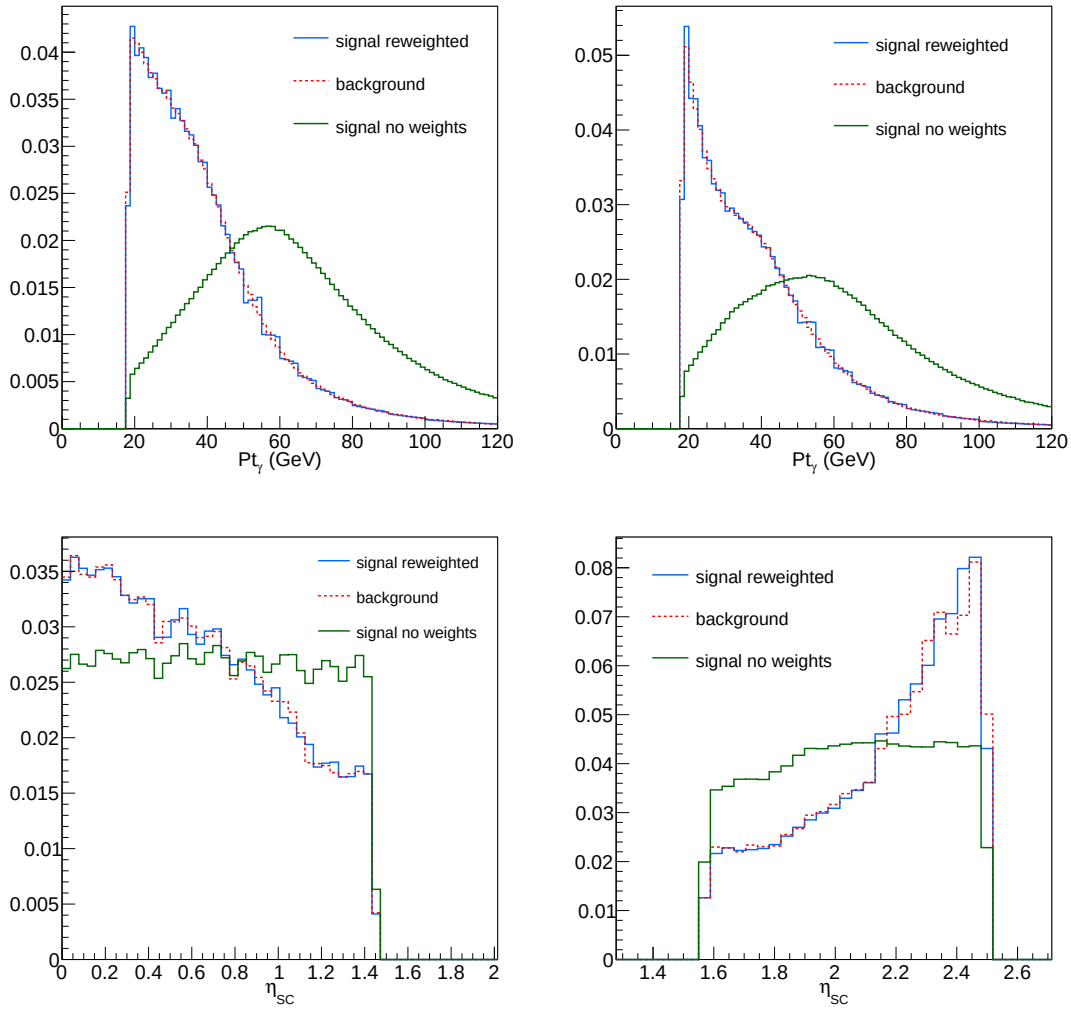


Figure 3.11: p_T and η_{SC} distributions, for both barrel (left) and endcap (right), for fake photons (dashed red histogram) and for prompt photons before (green) and after (blue) the 2D p_T-η_{SC} reweighting, where the 2D p_T-η_{SC} distribution of the signal is reweighted to that of the background.

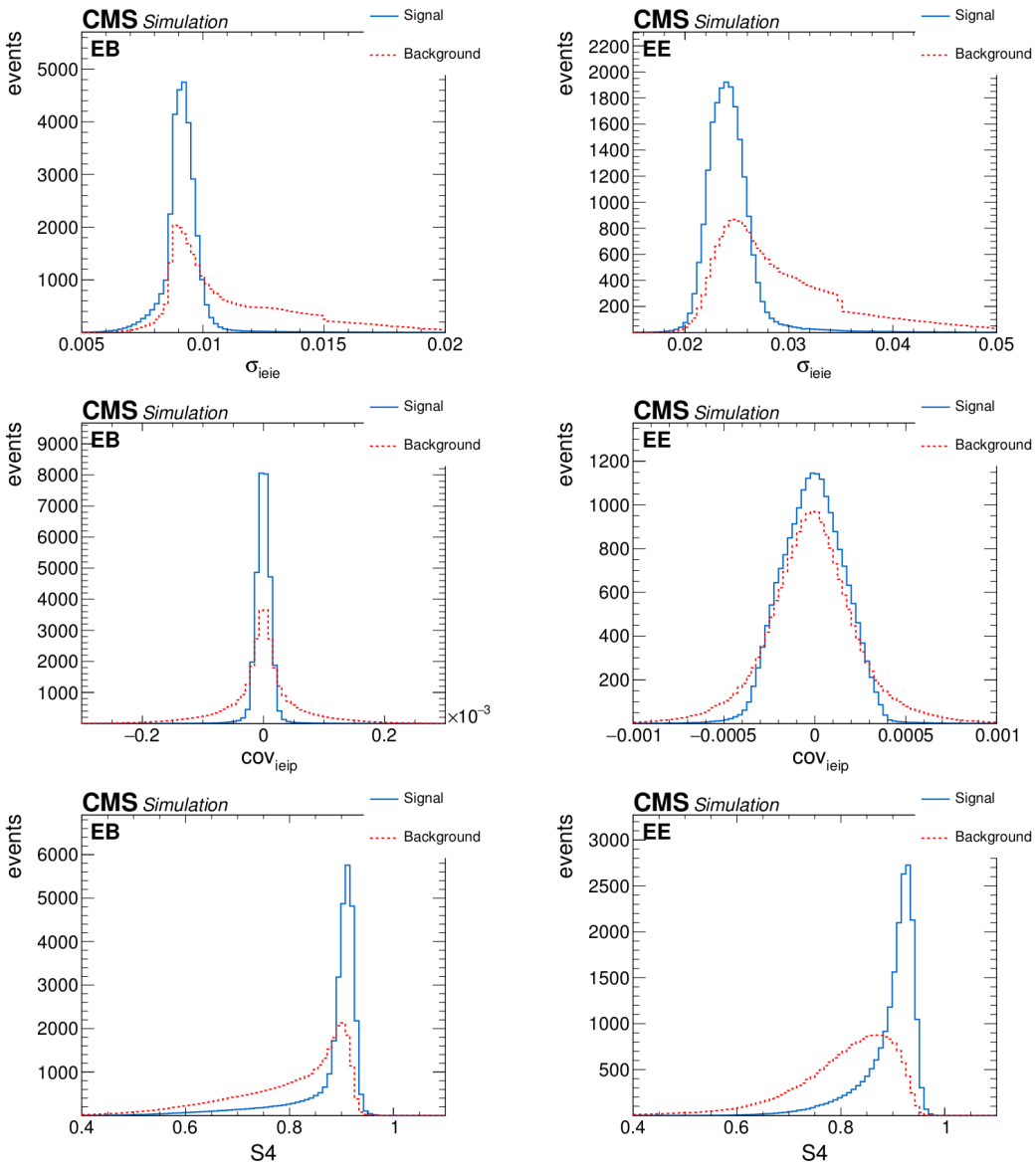


Figure 3.12: Distributions of photon identification BDT input variables $\sigma_{i\eta i\eta}$ (top row), $\text{cov}_{i\eta i\Phi}$ (middle row) and S_4 (bottom row) for signal prompt photons (blue) and background fake photons (red) in the barrel (left) and in the endcap (right) from pp collisions at 13 TeV. Photons are from the training samples passing the preselection with $p_T > 18$ GeV and after p_T-η_{SC} reweighting.

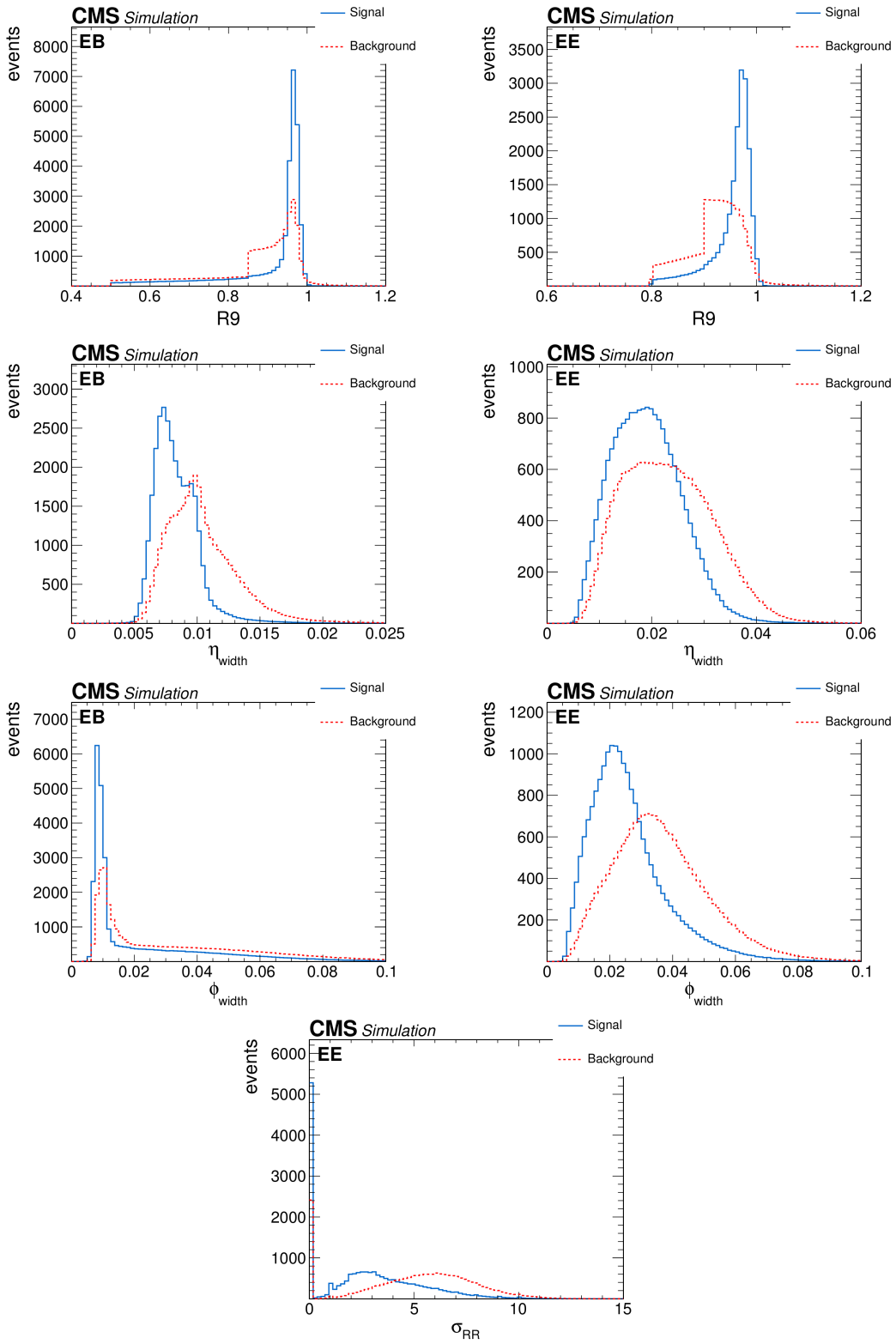


Figure 3.13: Distributions of photon identification BDT input variables R_9 (first row), SC η-width (second row) and SC Φ-width (third row) for signal prompt photons (blue) and background fake photons (red) in the barrel (left) and in the endcap (right), along with the distribution of preshower σ_{RR} (fourth row) for photons in the endcap only, from pp collisions at 13 TeV. Photons are from the training samples passing the preselection with $p_T > 18$ GeV and after p_T-η_{SC} reweighting.

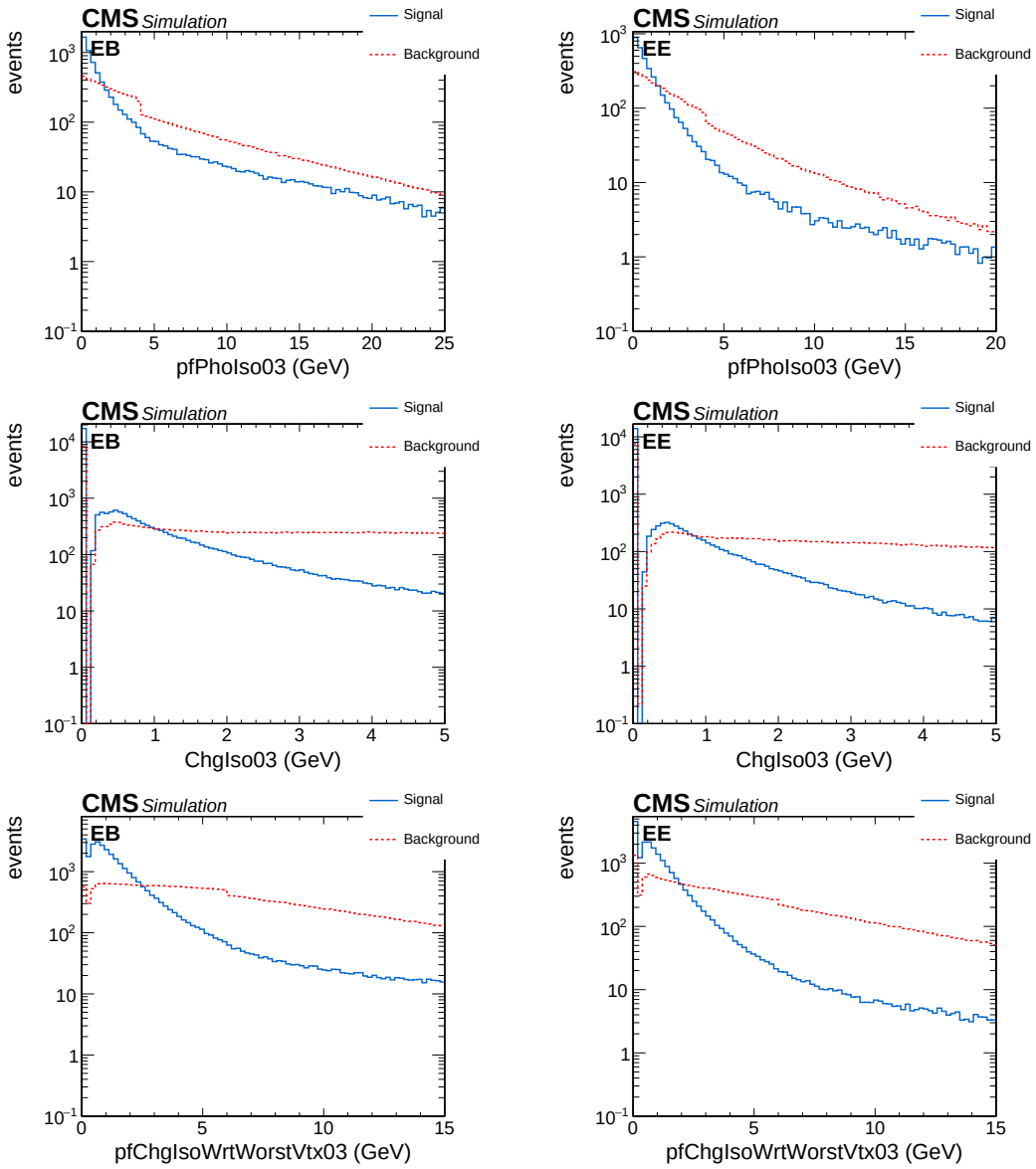


Figure 3.14: Distributions of photon identification BDT input variables *PF Photon ISO* (top row), *PF Charged ISO (selected vertex)* (middle row) and *PF Charged ISO (worst vertex)* (bottom row) for signal prompt photons (blue) and background fake photons (red) in the barrel (left) and in the endcap (right) from pp collisions at 13 TeV. Photons are from the training samples passing the preselection with $p_T > 18$ GeV and after p_T-η_{SC} reweighting.

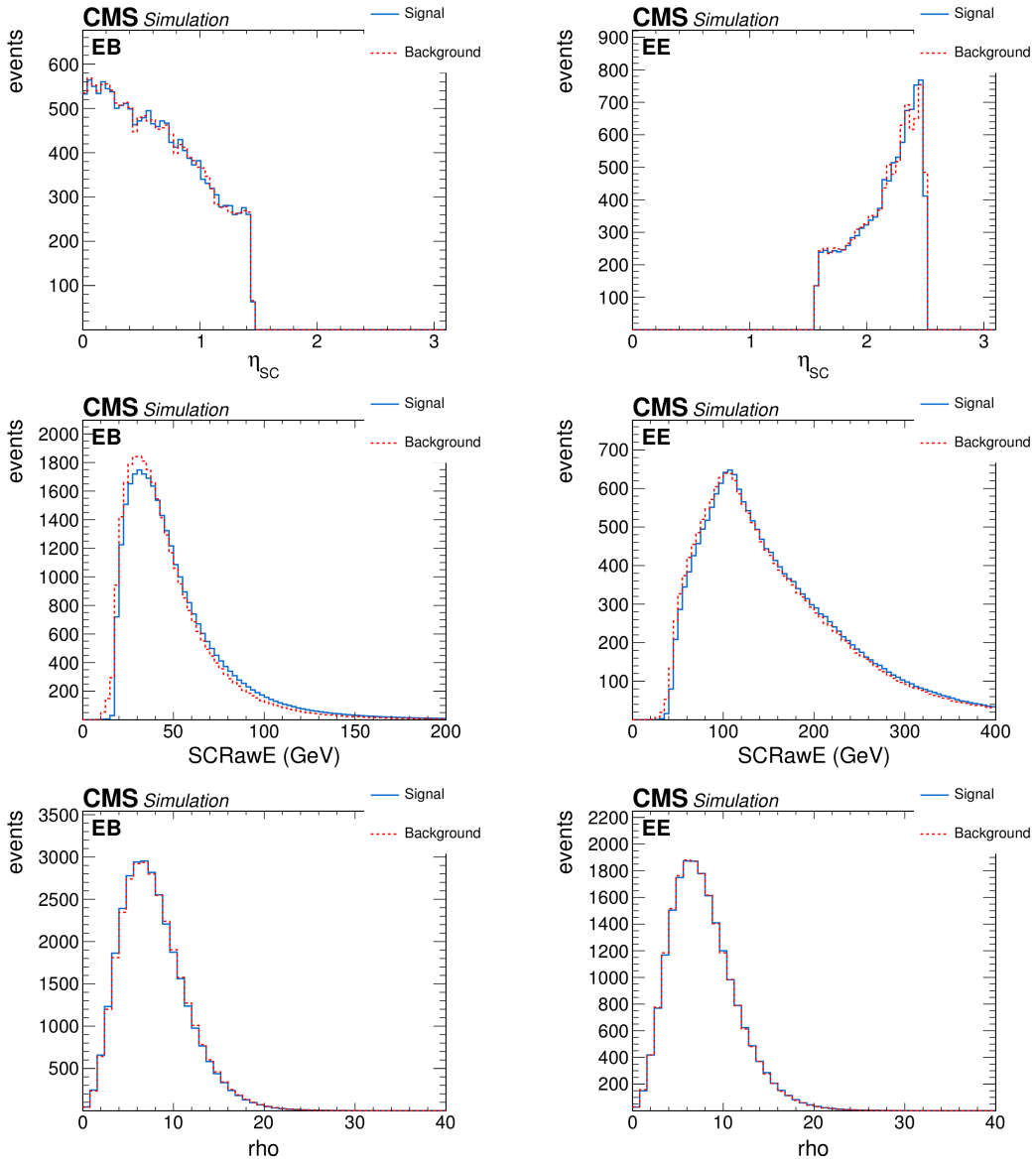


Figure 3.15: Distributions of photon identification BDT input variables η_{SC} (top row), SC E_{RAW} (middle row) and ρ (bottom row) for signal prompt photons (blue) and background fake photons (red) in the barrel (left) and in the endcap (right) from pp collisions at 13 TeV. Photons are from the training samples passing the preselection with $p_T > 18$ GeV and after p_T-η_{SC} reweighting.

each individual tree to fully explore the multidimensional phase space. The depth (number of terminal nodes) of each of the individual trees in the boosting series, referred to as the forest, can thus be limited. An additional handle for reducing overtraining is the addition of a shrinkage factor, which represents the learning rate of the gradient boost algorithm. Empirically for classification problems it has been found that optimal performance is obtained with a large number of small trees, and using a small shrinkage factor, of about 0.1. Boosted decision trees constructed in this way are extremely robust to the addition of variables which are either redundant, or only contribute additional information in a limited region of the input variable space. In general this allows the selection of variables based on the physical information which they contain, with the final performance being relatively insensitive to the precise definition or particular combination of variables.
Taking into account all these prescriptions, a shrinkage factor of 0.1 was used. Several trainings with different number of trees and depth were tested. I started with 700 trees, each one with a depth of 3; then I varied the number of trees for 500, 1000, and finally doubled the depth of the trees. These changes in the parameters have small effect on the training performances, as one can see in Figure 3.16, where the ROC curve with the initial parameters is shown along with the ROC curves obtained increasing the number of trees to 1000 and the depth of the trees to 6. Figure 3.16 shows also the ROC curve obtained adding in the training the variable H/E, which was tried since it has a good discriminating power between prompt and fake photons. But it does not seem to improve the training performances.

Several MVA techniques different from BDT were also tried. The more interesting is the Artificial Neural Network (ANN), that is any simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. By applying an external signal to some (input) neurons the network is put into a defined state that can be measured from the response of one or several (output) neurons. The neural network can therefore be seen as a mapping from a space of input variables $x_1, ..., x_{n_{var}}$ onto a, in case of a signal-versus-background discrimination problem, one-dimensional space of output variables y. The mapping is nonlinear if at least one neuron has a nonlinear response to its input.
The neural network MLP (Multilayer Perceptron) was used. While in principle a neural network with n neurons can have n^2 directional connections, the complexity can be reduced by organizing the neurons in layers and only allowing directional connections from one layer to the immediate next one. This kind of neural network is called *multilayer perceptron*. The first layer of a multilayer perceptron is the input layer, the last one the output layer, and all others are hidden layers. For a classification problem with n_{var} input variables and 2 output classes the input layer consists of n_{var} neurons that hold the input values, $x_1, ..., x_{n_{var}}$, and one neuron in the output layer that holds the output variable, the neural network estimator y_{MLP}. Each directional connection between the output of

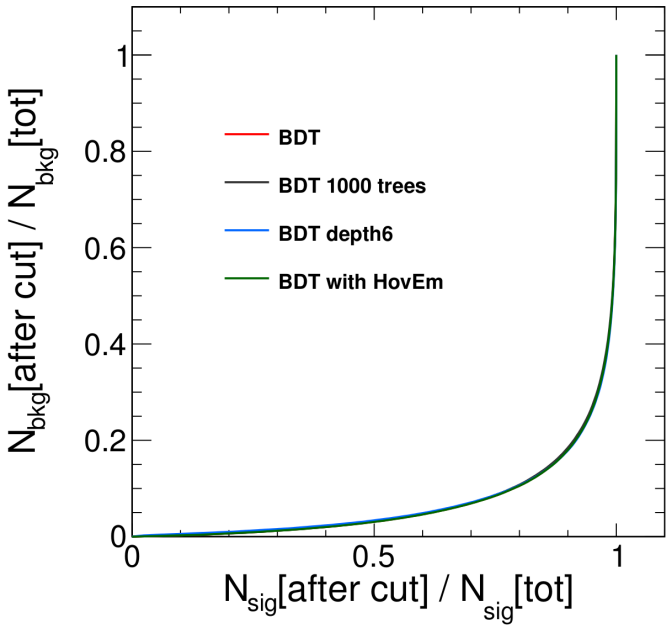


Figure 3.16: ROC curves obtained doing the training with different parameters/variables. In red the default training with Ntrees = 700 and depth = 3, in grey the training with Ntrees = 1000, in blue the training with depth = 6 and in green the training with the addition of the variable H/E.

one neuron and the input of another has an associated weight. The value of the output neuron is multiplied with the weight to be used as input value for the next neuron. The number of hidden layers can be tuned, keeping in mind that for a multilayer perceptron a single hidden layer is theoretically sufficient to approximate a given continuous correlation function to any precision, given an arbitrary large number of neurons in the hidden layer. If the available computing power and the size of the training data sample are sufficient, one can thus raise the number of neurons in the hidden layer until the optimal performance is reached.
Another parameter useful to tune is the number of training cycles, that has to be not too low, to exploit all the information available, but neither too high, to avoid overtraining. Keeping in mind all these aspects I decided to use one hidden layer with several neurons, $n_{var} + 5$, and to perform trainings with a different number of cycles. The more adequate number of cycles was found to be 3000, and Figure 3.17 shows the performance comparison between the training done with the neural network and the training done with the BDT. The performances are very similar, so the BDT was chosen for the final training, since neural network is very computing and time consuming.

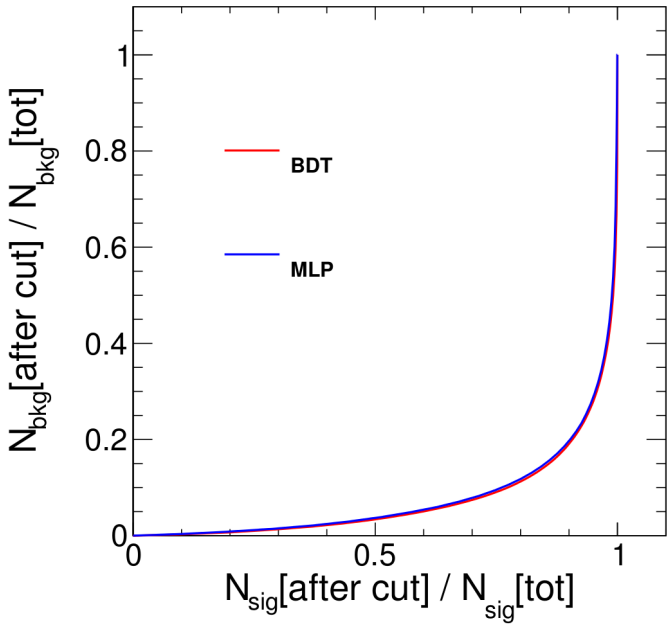


Figure 3.17: ROC curves obtained doing the training with different MVA techniques, with a BDT in red and with a neural network MLP in blue.

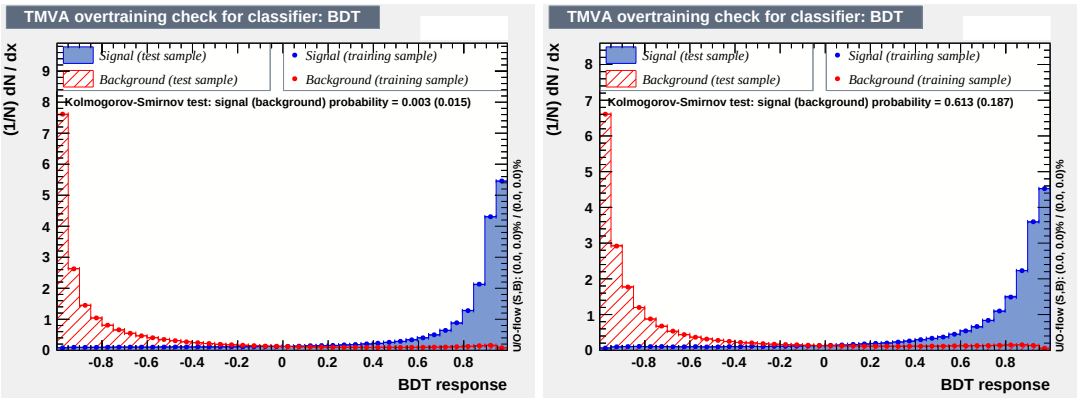


Figure 3.18: Output distribution of the BDT: barrel (left) and endcap (right). The blue histograms represent the prompt photons, the red histograms the fake photons. The points correspond to training samples, the histograms to test samples.

Final training and relevant results

The final training, used in the public H $\rightarrow \gamma\gamma$ analysis, along with the relevant results is presented in this section. The BDT technique is used, because quicker and easier to handle than neural network. The photon identification BDT output is a score assigned to each photon which ranges from -1 to 1. The higher the score assigned to a photon, the more likely the photon is a prompt photon rather than a fake photon. The BDT output for the double EM-Enriched photon plus jet simulated samples is shown in Figure 3.18, for barrel and endcap respectively. The signal (blue) and background (red) training samples (solid circles), and the corresponding testing samples (histograms) are shown. From this figure one can see that the discriminating power between prompt and fake photons is very good. Furthermore, good agreement between the distributions of the testing samples and those of the training samples is found, which proves the statistical stability of the BDT output.

The BDT output has been checked also for signal samples and for the others H $\rightarrow \gamma\gamma$ backgrounds (photon plus jet, jet-jet and diphoton). Figure 3.19 shows the photon identification BDT score of the lower-scoring photon in diphoton pairs with an invariant mass, $m_{\gamma\gamma}$, in the range $100 < m_{\gamma\gamma} < 180$ GeV, for events passing the preselection, in data and simulated background events. The sum of all the backgrounds is consistent with data, even if some discrepancies are visible in the high-score region. These discrepancies are taken into account in the treatment of the systematic uncertainties, presented in the next section.

Figure 3.20 shows background efficiency versus signal efficiency of the identification performances for the new ID MVA (in blue) and for identification used in the 8 TeV analysis (in red) applied in the 13 TeV environment. Efficiencies are relative to the preselection

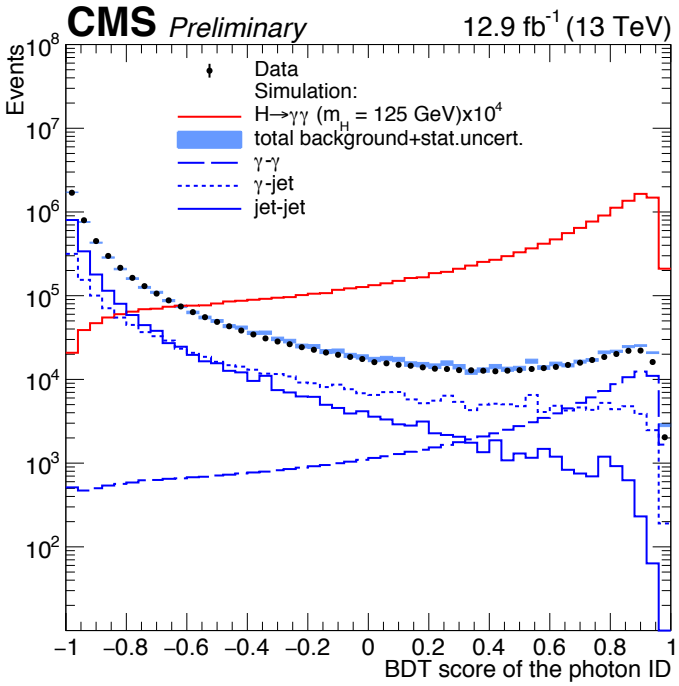


Figure 3.19: Photon identification BDT score of the lower-scoring photon of diphoton pairs with an invariant mass in the range $100 < m_{\gamma\gamma} < 180$ GeV, for events passing the preselection in the 13 TeV dataset (points), and for simulated background events (cyan histogram). Histograms are also shown for different components of the simulated background, in which there are either two, one, or zero prompt candidate photons. The distribution of the sum of all the simulated background events is scaled to data preserving the relative ratio of the single components, generated at leading order. The red histogram corresponds to simulated Higgs boson signal events.

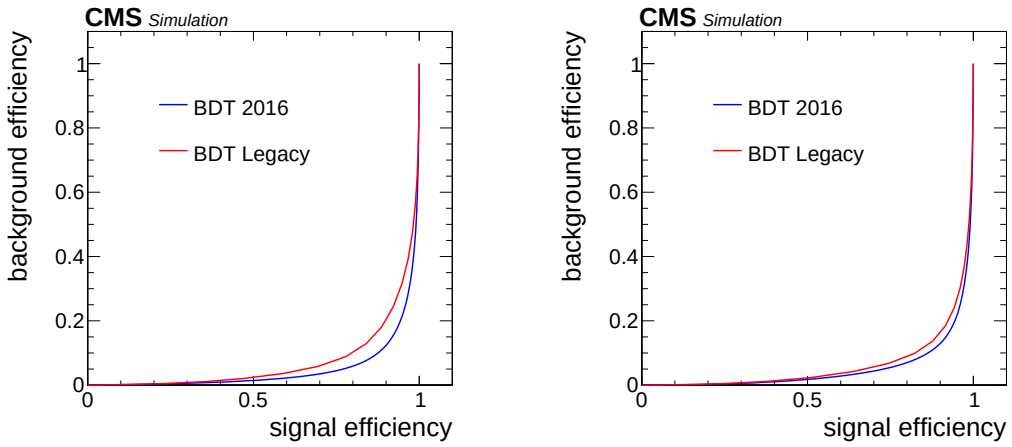


Figure 3.20: Curve of background efficiency as a function of signal efficiency of the training, both for the new 13 TeV training (in blue) and for the 8 TeV training applied to the 13 TeV samples (in red). Left plot refers to the barrel, right plot to the endcap.

described in Section 4.4. This figure shows that the new identification performs better than the 8 TeV one, and clearly demonstrates the benefit of a 13 TeV dedicated photon identification.

A working point, with signal efficiency 95% for the barrel and 90% for the endcap, is used to illustrate the performance of the ID MVA. The signal efficiency and the background efficiency are shown as functions of p_T, supercluster η and number of vertices, for the chosen working point, in Figure 3.21. It can be seen that in general the efficiency is quite flat, indicating that the photon identification performance is the same for different phase spaces. In particular, the flatness of the efficiency versus p_T and η_{SC} is a desirable feature due to the inclusion of η_{SC} and E_{RAW} into the input variables, and the 2D p_T-η_{SC} reweighting in the training. The flatness of the efficiency as a function of number of vertices is expected as a result of using ρ as an input variable.

In the analysis a selection at -0.9 is applied on the BDT output as a further preselection which guarantees 99% efficiency on signal photons. The BDT output values for each photon are used as an input to a diphoton event-level multivariate classifier, described in Section 4.8.

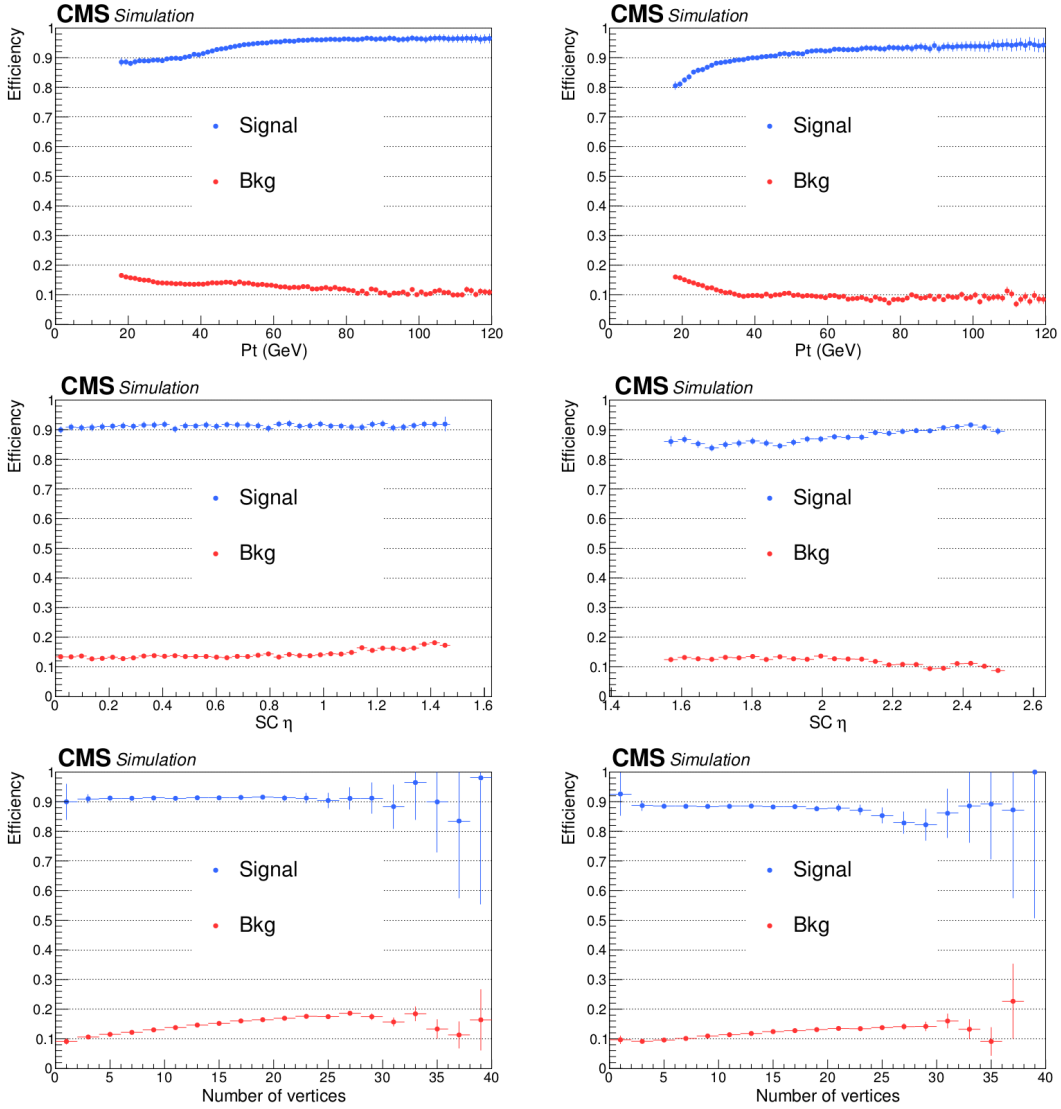


Figure 3.21: Signal and background efficiency versus p_T (top), supercluster η (middle) and number of vertices (bottom). Left plots refer to the barrel, right plots to the endcap.

3.3.3 Data-simulation comparison and systematic uncertainties

Some quantities, like the photon shower shapes and the correlation with other BDT input observables, are sensitive to the accuracy of the simulation of the detector response.
A validation of the extent to which the MVA inputs and output in data are well modeled by simulation can be performed using electrons reconstructed as photons in $Z \to ee$ events. Even if the Z boson differs from a Higgs signal in many aspects, a comparison for $Z \to ee$ events in simulation and data constitutes an important check to be sure that the modeling of the BDT input variables and their correlations in the simulation is sufficiently accurate.

Contrary to the standard analysis, in the validation events are selected with the inverted electron veto applied as part of the preselection, in order to keep both electrons decaying from the Z boson. In addition to the sliding $p_T/m_{\gamma\gamma}$ criteria applied in the preselection (see Section 4.4), the leading and subleading photons are also required to satisfy $p_T >$ 35 and 25 GeV, respectively, in order to prevent transverse momentum thresholds falling below trigger thresholds for low diphoton invariant masses. Finally, to have a purer di-electron sample, each pair is required to have an invariant mass between 86 and 94 GeV.

Data-simulation comparison showed a reasonable agreement for the photon identification input variables, except for three of them: important discrepancies arose for some shower shape variables, in particular for R_9, S_4 and SC η-width. These discrepancies are present especially for photons in the barrel, and their origin is still under investigation. In order to improve the data/MC agreement these variables were corrected using a histogram remapping method with a sample of probes from Z events. Figure 3.22 shows the data/MC comparison before the correction is applied, both for barrel and endcap. The final training of the photon identification, presented in Section 3.3.2, was done using shower shape variables after the correction.

Some discrepancies between data and simulation were observed also for the photon identification output, and they were included in the systematic uncertainty. As a first step, the treatment of systematic uncertainty for the photon identification output variable was done in a way similar to Run 1 [13], shifting its value for every photon in the simulation by ± 0.03. In spite of that, a small discrepancy was observed in the low score tail. Since one of the main sources of systematic uncertainty on the diphoton MVA output derives from the photon ID MVA (see Section 4.8.2), the imperfect coverage on the photon ID tail is reflected in the diphoton BDT tail. In order to take this into account, it was decided to be conservative and to estimate systematic uncertainty combining the shift of ± 0.03 with a linear correction that expands the uncertainty at low BDT scores. The photon ID MVA score distribution, for data and MC, with the final systematic uncertainty is shown in Figure 3.23.

This way of estimating the systematic uncertainty is quite coarse, but before proceeding with more complex studies we need to understand the origin of the discrepancies between data and MC. The corrections applied to R_9, S_4 and SC η-width brought some improvement, but it is possible that little discrepancies present in other variables entering the photon identification training can affect the final output.

3.4 Summary

In this chapter the photon reconstruction and identification algorithms of CMS are presented, with a particular focus on the differences between the first and the second run of the LHC. Performances of Run 1 and Run 2 reconstructions from the photon identification point of view are compared and found to be very similar. The optimisation of the photon identification algorithm for the Run 2 H $\rightarrow \gamma\gamma$ analysis is then described. Performances of the photon identification at 13 TeV and a data-simulation validation are finally presented. The treatment of the systematic uncertainties is for the moment quite coarse and can be improved in the future. First of all it would be important to understand the origin of the discrepancies between data and MC. After that, it would be interesting to find a method for the systematics estimation that takes into account the correlations between the variables entering the photon ID MVA. Some work was started for that, but it is far from a conclusive result.

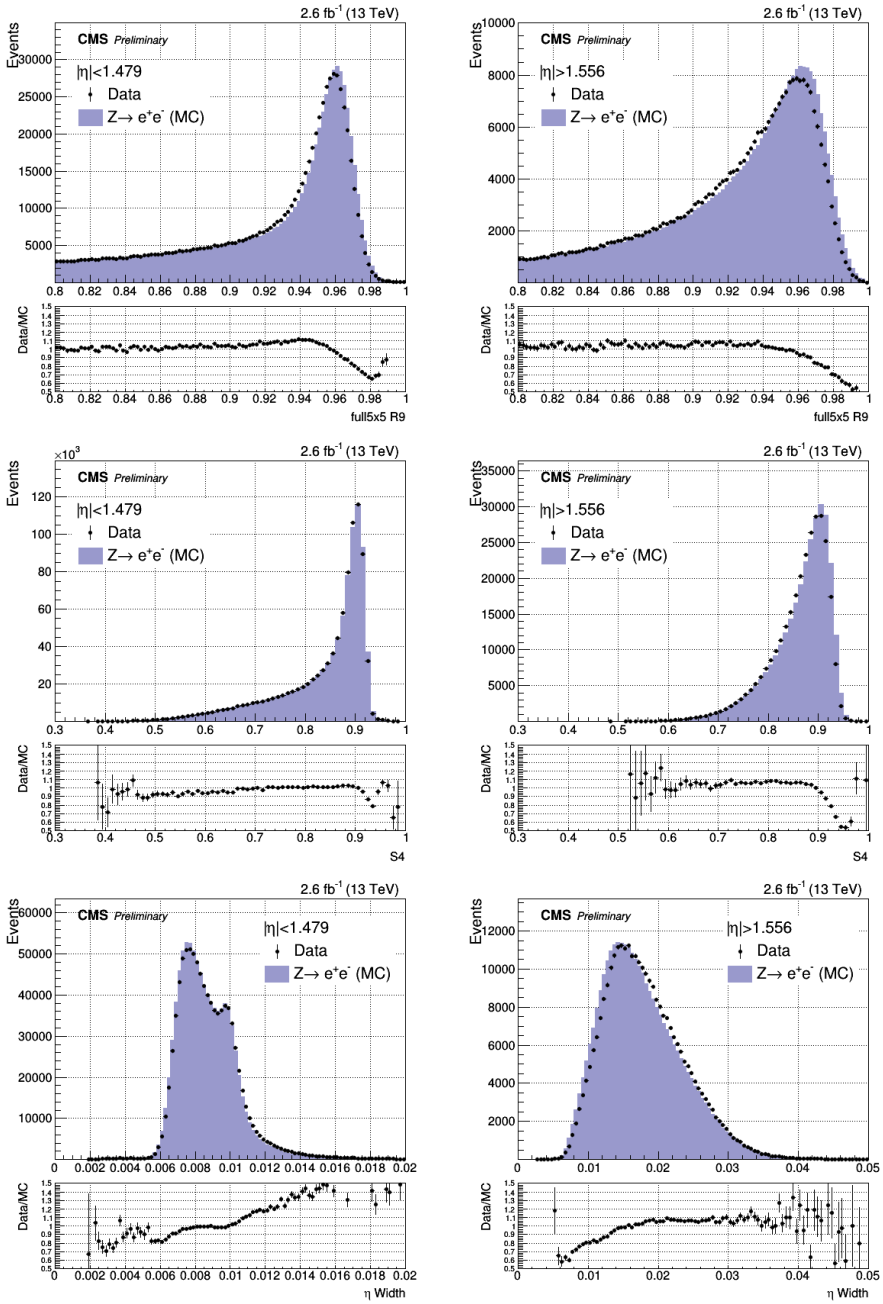


Figure 3.22: Data/MC comparison distribution for three shower-shape variables, R_9, S_4 and SC η-width, for both barrel (left) and endcap (right). Some discrepancies between data and MC are evident, in particular in the barrel.

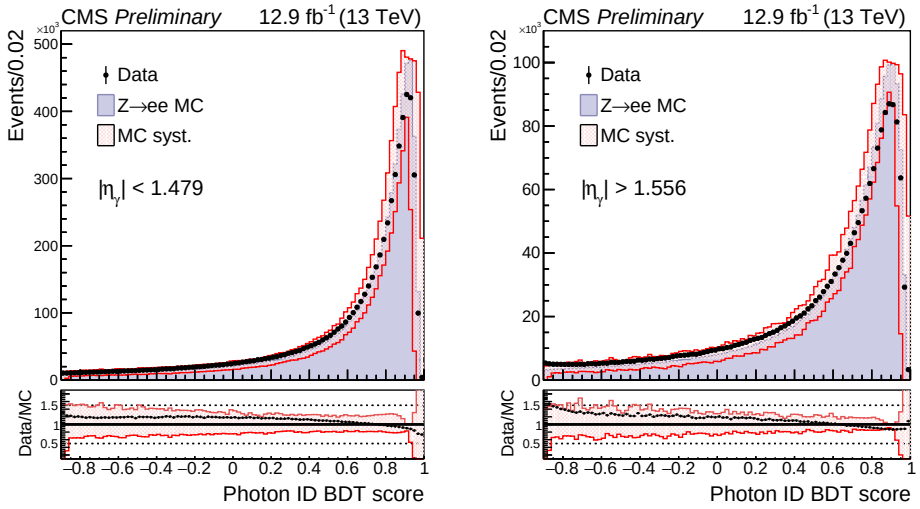


Figure 3.23: Photon ID MVA output distribution for $Z \to ee$ events in data and simulation, for photon candidates in the ECAL barrel (left) and endcaps (right). The event selection consists of the preselection (including the requirement ID MVA $>$ -0.9 with the electron veto inverted). The systematic uncertainty applied to the shape from simulation, corresponding to a shift of ± 0.03 in the value of the MVA output combined with a linearly increasing term, is represented by the hashed region.

Chapter 4

$H \to \gamma\gamma$ analysis at 13 TeV

In this chapter an overview of the analysis looking for an Higgs boson decaying to two photons at 13 TeV is presented. My contribution to this analysis concentrated on the photon identification development and on the study of its systematic uncertainties, and is presented in Chapter 3.
Another important contribution I gave to this analysis, as MC contact of the $H \to \gamma\gamma$ group, concern the production of all the simulation samples needed to perform the analysis.
This is the first $H \to \gamma\gamma$ analysis performed at 13 TeV after the restart of the LHC, and its results have been presented at ICHEP 2016. The main goal of the analysis is the rediscovery of the Higgs boson, waiting for more data to perform property studies.
Despite its small branching ratio of 0.23% for an Higgs boson mass of 125 GeV, the $H \to \gamma\gamma$ decay channel provides a clean final-state topology with an invariant mass peak that is reconstructed with great precision. For this reason the $H \to \gamma\gamma$ was one of the most important channels used in the discovery and first measurements of the Higgs boson, and in the LHC Run 2 this channel remains one of the best to perform a precise characterization of the Higgs properties.
The analysis strategy is similar to that used in Run 1. The sensitivity of the analysis is improved by categorizing events by mass resolution, signal-to-background ratio and production mechanism. Higgs production mechanisms other than gluon-gluon fusion (ggH) can be identified by selecting final state objects in addition to the diphoton pair. The events with additional objects are tagged as exclusive categories, while those that remain untagged are the inclusive categories identified as the gluon-gluon production mechanism. The inclusive events are categorized using a multivariate classifier, that creates categories based on photon kinematics, mass-resolution, as well as other inputs to indicate the signal-to-background ratio.
After the event classes are determined, the diphoton mass spectrum for each event class, and the corresponding Higgs signal model and background model, are constructed. The signal model is obtained from Monte Carlo simulated Higgs events, while the background

one directly from the data. The Higgs signal is finally extracted by statistical procedures based on simultaneous likelihood fit to the diphoton mass spectra over all event classes. The analysis is performed on 12.9 fb^{-1} of data collected during 2016 with 25 ns bunch spacing and magnetic field of 3.8 T. The analysis is performed in the invariant mass region 100 GeV $< m_{\gamma\gamma} <$ 180 GeV, blinding the region 115 GeV $< m_{\gamma\gamma} <$ 135 GeV.

4.1 Principles of Monte Carlo simulation and H $\to \gamma\gamma$ Monte Carlo samples

During the last two years I was in charge of the simulated samples production for the H $\to \gamma\gamma$ group.
In this section a detailed description of the techniques to simulate a hadronic collision event, based on the factorisation of hard and soft components, is given, along with the presentation of the simulated samples of the H $\to \gamma\gamma$ analysis.

4.1.1 High energy processes, hadron collisions

A hadronic collision is a complex phenomenon, which evolution can be described as follows:

1. The two beams collide at the intersection points. Each hadron is composed of quarks of different flavours and gluons, which carry fractions of its momentum. The hadron composition in terms of flavour and energy sharing is modelled by parton distribution functions (PDF);

2. The partons inside the colliding hadrons emit radiations, initiating a sequence of branching processes $q \to qg$, $g \to q\bar{q}$, $g \to gg$. Because of the large value of the strong coupling constant α_S, these splittings have a high probability to occur and this gives rise to the formation of initial-state parton cascades;

3. Two partons in the cascade enter the hard interaction, at a momentum transfer scale Q^2. The products of the hard scattering are the final-state elementary particles, partons, leptons and bosons, that characterise the event topology. Short-lived resonances, such as Z, $W^{\pm}$ and Higgs bosons, instantly decay into partons, leptons or photons. Even if the hard scattering subprocess is not observable, it determines the main properties of the collision event;

4. The outgoing partons (quarks and gluons) start branching and initiate final-state cascades;

5. After every branching in the initial and final-state showers, the momentum scale decreases down to the cutoff scale $\Lambda_{QCD} \sim 1$ GeV, where the perturbative theory is no more valid;

6. Below Λ_{QCD}, the strong interaction confines the partons into colourless hadrons. The confinement process is followed by the decay of the unstable particles. Therefore, through fragmentation and decay, the parton cascades evolve into jets of stable and meta-stable particles which are observable in particle physics detectors.

The modelling of a hadronic collision is factorized into subprocesses, each of them relatively easy to handle with the appropriate technique. This approach is adopted in the Monte Carlo generators, the main tool to describe and reproduce the phenomenology at a hadronic collider.
To do that, the *factorisation theorem* allows the independent treatment of the hard scattering and of the soft non-perturbative processes. For a proton-proton collision $p_A p_B \to X$, where X is a generic final state, it can be expressed by the formula:

$$\sigma_{AB} = \int dx_a dx_b f_a(x_a, Q^2) f_b(x_b, Q^2) \cdot \hat{\sigma}_{ab\to X} \tag{4.1}$$

where σ_{AB} is the total cross-section, x_a and x_b are the fractions of the proton momentum carried by the two partons a and b involved in the interaction, and $\hat{\sigma}_{ab\to X}$ is the hard partonic scattering cross section. Calculating the latter with the perturbative expansion, Equation 4.1 can be written as:

$$\sigma_{AB} = \int dx_a dx_b f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \cdot [\hat{\sigma}_0 + \alpha_S(\mu_R^2)\hat{\sigma}_1 + \alpha_S^2(\mu_R^2)\hat{\sigma}_2 + ...]_{ab\to X} \tag{4.2}$$

The $f_a(x_a, \mu_F^2)$ and $f_b(x_b, \mu_F^2$ terms are the parton distribution functions described in Section 4.1.2. The momentum transfer Q^2 is replaced by the *factorisation scale* parameter μ_F, which indicates the separation between the hard scattering and the soft process. In the perturbative expansion, the strong coupling constant α_S is evaluated at the *renormalisation scale* μ_R. This two scale parameters are unphysical. The μ_F and μ_R dependence of the parton density functions and of α_S is exactly compensated by the $\hat{\sigma}_i$ coefficients at all perturbative orders, resulting in the invariance of the σ_{AB} cross section under changes of their values. At a fixed order, instead, the dependences do not cancel out, and a specific choice of the scale parameter values is necessary for a cross section estimation. The sensitivity of the σ_{AB} prediction to variations of the scale parameters has to be accounted for as theoretical uncertainty.
Qualitatively, the factorisation theorem states that the hard scattering and the soft components of a hadronic collision can be disentangled, and independently modelled. The hard-scattering component for a specific process of interest can be solved in the context of the well known perturbation theory. Different methods are available, as described in Section 4.1.3. Part of the soft process, namely hadronisation and quark confinement, occurs instead at the momentum scales of non-perturbative QCD. Phenomenological models have to be used in this context, based on experimental data. Because of the process-independence of the soft-process phenomenology, these models have a general validity.

4.1.2 Parton distribution functions

The PDF functions in Equations 4.1 and 4.2 relate the dynamics of the partons entering the hard scattering to that of the colliding hadrons, by modelling the probability that a parton carries a fraction x of the momentum of the proton. The PDF depend on the momentum transfer Q^2, because of higher-order corrections from real and virtual gluon emission within the colliding protons. In good approximation, the PDF evolution as a function of the energy scale Q^2 of the process can be calculated by the Altarelli-Parisi, or DGLAP, equations, developed in the perturbative theory [46, 47, 48].
The x dependence is extracted from a global fit to data, including few thousands measurement points from deep inelastic scattering (DIS), Drell-Yan and jet production. Results are available at the tree-level, at the next-to-leading (NLO) and, only partially, at the next-to-next-to-leading (NNLO) order. The PDF extrapolations are affected by uncertainties, accuracy of the experimental data and of the analysis, uncertainty on the coupling constant α_S, and have to be taken into account as source of uncertainty in all theoretical predictions.
Various PDF sets are available: their main difference is the number of parameters of the model and the series of data used to fit the processes. The most common ones are CTEQ [49], MSTW [50] and NNPDF [51].

4.1.3 Steps in the event generation process

Monte Carlo event generation is used to simulate the final states of high-energy collisions in full detail down to the level of individual stable particles. The aim is to generate a large number of simulated collision events, each consisting of a list of final-state particles and their 4-momenta, such that the probability to produce an event is proportional to the probability that the corresponding actual event is produced in the real data. The event generation for a hadron-hadron collision is generally split into different steps by taking advantage of factorization theorem. An overview of the steps needed to obtain such a complete event is given in Figure 4.1. More details about the single steps in the event generation will follow in the subsequent paragraphs.

1. The computation of the hard subprocess (Matrix Element ME)

The first step in the event generation is the simulation of the hard subprocess. The hard process is defined by the collision of two particle beam constituents which interact with each other at a high momentum scale. Thus the strong coupling constant α_S is rather small for the hard subprocess, which can be described with perturbation theory and by a matrix element.
To describe the computation of the hard subprocess in more detail the collision of two protons, each consisting of partons (quarks and gluons, each with a colour charge), is

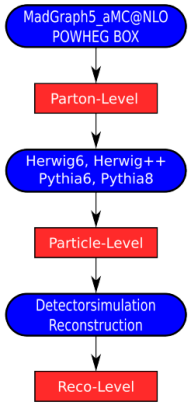


Figure 4.1: Overview of the steps in event generation.

illustrated in Figure 4.2. From a single proton only one high energetic parton participates in the hard interaction and produces two further outgoing fundamental objects (shown by two red outgoing lines), while the other partons (displayed by two black arrows) of the protons keep flying without participating in the main interaction. However these partons will become important later, because they form the so-called underlying event. Each of the particles occuring in the hard subprocess, which carries a colour charge, will be involved in the subsequent parton shower step.

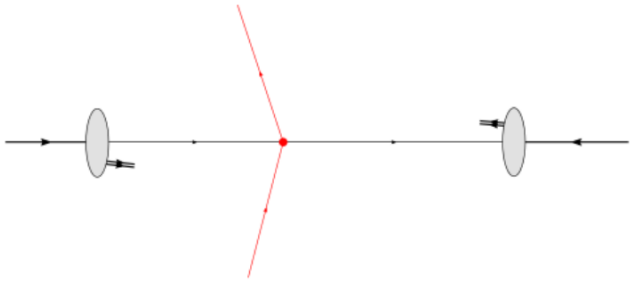


Figure 4.2: The hard subprocess in the event generation: two protons (displayed by gray ellipses) collide with each other. Thus one high energetic parton from each proton interacts with the other one in a hard interaction and produces two further fundamental objects (shown by two red outgoing lines). The other partons (displayed by two black arrows) from each proton are unaffected by the collision and keep on flying. Taken from Reference [52].

2. The parton shower step (PS)

After the hard collision a parton shower is used to evolve the event further. Due to the large momentum transfers during the hard subprocess step, the final-state particles

obtained from the matrix element have high energies. The strongly interacting partons (quarks and gluons) carrying a colour charge can emit QCD radiation in the form of gluons, similar to the electromagnetic shower generated by electrically charged particles. However, in contrast to QED radiation in which the uncharged photons, being the gauge bosons of QED, can produce only pairs of electrons, the gluons, being the gauge bosons of QCD, carry a colour charge themselves and hence can interact among each other by emitting further gluons. This leads to a parton shower which does not stop until the involved partons have decreased their energy by collinear parton splitting and/or soft gluon emission so much that they enter the hadronization phase. Different approximation schemes exist to describe the parton shower and, in principle, the showers represent higher-order corrections to the hard subprocess. However, we should keep in mind that the parton shower generally produces only low-energy additional radiation due to the collinear parton splitting and soft gluon emission and that for non-collinear parton splitting the parton shower approximation diverges. For this reason, the parton shower can fill phase-space for higher orders of perturbation theory, which are not covered by the matrix element of the hard process, but the emission of additional hard radiation is suppressed. Furthermore, a matching between the matrix element computation for the hard process and the parton shower is needed, otherwise some parts of the phase-space would be filled twice (see Section 4.1.5). The splitting of partons in the initial-state (before the hard interaction took place) and in the final-state (after the hard interaction) by emitting gluons is schematically illustrated in Figure 4.3.

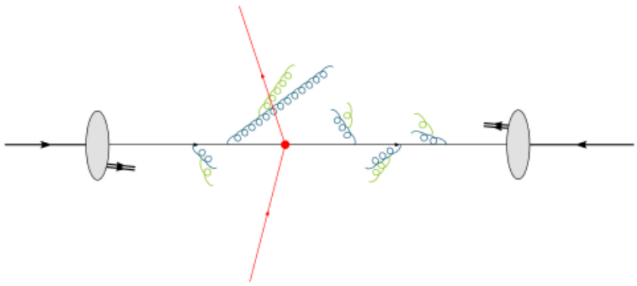


Figure 4.3: The parton shower step during the event generation: splitting of partons in the initial-state (before the hard interaction shown in red took place) and in the final-state (after the hard interaction) by emitting gluons (displayed in blue color) is shown. The newly emitted gluons (blue color) can radiate further gluons (green color). Taken from Reference [52].

3. The hadronization process

During the parton shower the involved partons lose energy by splitting and gluon emission, at the same time the strong coupling constant α_S rises due to its running. In this evolution

the dynamics reach a non-perturbative phase and at some point the coupling between the individual colour charged partons becomes strong enough to bound partons in colorless hadrons. The hadronization process begins roughly at an energy of 1 GeV (the exact value depends on the hadronization model) and cannot be described by the perturbative techniques currently available. For this reason, hadronization models based on empirical data are needed for the description of this formation of hadrons.
The cluster hadronization approach is represented in Figure 4.4: the quarks and gluons from the previous parton shower step form colour neutral clusters with corresponding colour partners (displayed as white ellipsoid objects). In a second step the clusters are rearranged and, if necessary, split into smaller clusters in a way that the clusters have a limited cluster mass and can decay into hadrons (shown in yellow color) using a simple decay model. The produced hadrons are the first particles in the event generation process which could theoretically be observed in nature. But most of these hadrons have only a short mean life time τ. Because of that only a small fraction of these hadrons is observed by particle detectors while most of them decay beforehand.

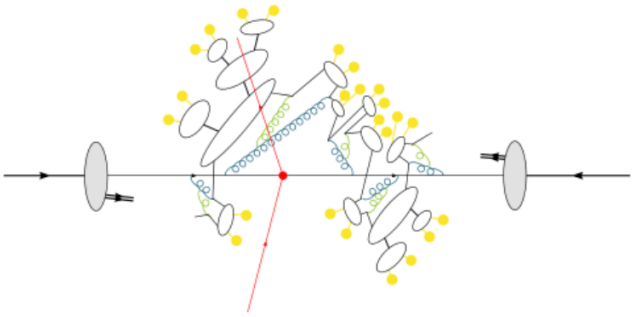


Figure 4.4: The hadronization process in the event generation: colourless clusters are formed by colour charged partons which have reached the hadronization scale by parton splitting and soft gluon emission. The clusters are rearranged and, if necessary, split into smaller clusters in a second step before the cluster decays to observable hadrons. Taken from Reference [52].

4. The hadron decay stage

After the formation of the hadrons a sequential decay stage (see Figure 4.5) follows because not all of the newly formed hadrons (shown in yellow) are stable. In this stage the excited and unstable hadrons decay into further hadrons until only long-living or stable hadrons are left over. For the simulation of high-energy hadron collisions like the LHC experiments it is important to include almost all possible excited hadronic states and their decay modes, since measurements done in previous experiments indicate that most of the final-state particles originate from these decaying hadrons. In some cases, this means that hadrons and decay modes not yet well established experimentally have to be modelled.

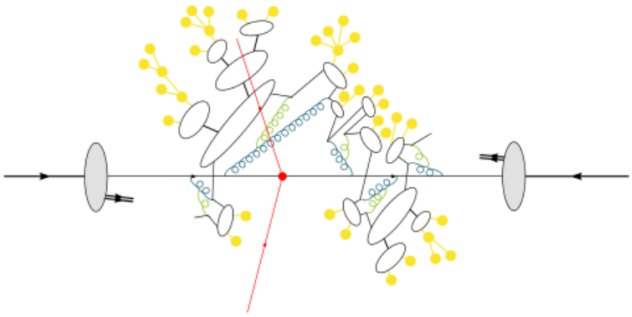


Figure 4.5: The hadron decay stage during the event generation: the excited and unstable hadrons, formed in the previous hadronization, decay into further hadrons until only long-living or stable hadrons are left over. Taken from Reference [52].

5. The underlying event

In the previous steps the interactions of the partons, which were part of the colliding protons but not participating in the hard subprocess, were neglected. As a first approximation it was assumed that these beam remnants keep on flying undisturbed. However they can also split and emit gluons themselves and take part in the hadronization process together with their emitted daughter partons. In Figure 4.6 the so-called underlying event is represented: the undisturbed flying beam remnants (shown in brown colour), consisting of partons, can split and emit gluons (blue colour) while they form hadronization clusters (white ellipsoidal objects), also with the partons originating from the hard subprocess and the subsequent parton shower step, before decaying into additional hadrons (displayed in yellow colour).
This ends the event generation per-se. However, in a full analysis the Monte Carlo event generation continues with the following steps:

6. Pile-up simulation

In the same bunch crossing there are multiple p-p interactions, which constitute the so-called pile-up. The pile-up is reproduced adding to the underlying event additional simulated events.

7. Detector simulation

In order to represent the response of the CMS detector with respect to the final state particles, a full detector simulation is performed, including the propagation of each particle in the magnetic field and the interactions of each particle with the passive and active elements of the detector. This simulation is performed using Geant 4 [53] and a detailed implementation of the CMS geometry. Interactions with the material are simulated for

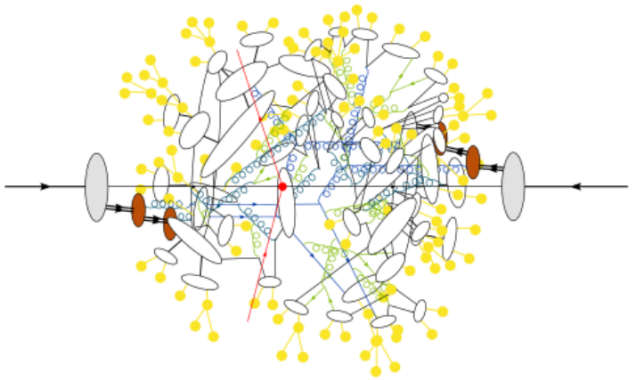


Figure 4.6: The underlying event in the event generation process: the beam remnants (shown in brown colour) consisting of partons which do not participate in the hard subprocess (red point) can split and emit gluons (in blue) while they form hadronization clusters (white ellipsoidal objects). Such hadronization clusters can also include partons from the hard subprocess and decay into hadrons (yellow colour). Taken from Reference [52].

each particle, including energy loss, Bremsstrahlung and photon conversions. For active detector elements, the simulated energy deposits are processed through a simulation of the readout electronics for each subdetector, including effects such as simulated noise.

8. Reconstruction and identification step

In a first step a particle detector only measures hits of the final-state particles in the single detector cells of their subdetectors. Afterwards tracks and particle objects can be reconstructed by using these informations applying reconstruction algorithms which make use of the specific properties of different particle types. In a last step, the obtained information is gathered and the particle objects are identified as particles.

4.1.4 Types of event generators

Monte Carlo event generators play an essential role in QCD modelling and in data analysis for high-energy physics. Different types of event generators exist, and in particular two types of event generators are used in the event generation for the LHC and will be explained in more detail.

GPMC event generators

The so-called general-purpose Monte Carlo (GPMC) event generators allow the complete simulation of high-energy physics processes. For this purpose, the GPMC event generators include low-order (LO and/or NLO) matrix elements for $2 \to 1$ and $2 \to 2$ process as well as parton shower, in which the shower evolution has to be independent from the details of

the hard scattering and relies only on a few features of the matrix element: the energies and flavours of incoming and outgoing partons in the matrix element as well as the overall energy scale Q^2 of the hard subprocess must be known by the successive parton shower. GPMC event generators include beside parton shower algorithms also hadronization models and, therefore, allow to describe physics processes starting from the matrix element up to the hadron formation and decay. Different GPMC event generators can be mainly distinguished by their choice of the evolution scale t (which models the progression of the parton shower), the implemented hadronization and hadron decay models, and the available matrix elements.
The main generators of this type are Herwig, Pythia and Sherpa. The last two are used in the $H \to \gamma\gamma$ analysis to model the background (see Section 4.1.6). In particular we used Pythia8 [54], which contains a wide range of hard-coded subprocesses and a highly developed multiple-interaction model for the underlying event. Sherpa [55] has two built-in matrix-element generators, AMEGIC++ and Comix, which allow to generate LO and NLO matrix elements for a wide range of subprocesses and are not limited to $2 \to 2$ processes. Additionally, Sherpa contains a multiple-interaction model which is loosely based on that of Pythia.

ME+PS and NLO+PS event generators

GPMC event generators have a big disadvantage: they rely on parton shower algorithms which are based upon a combination of collinear (small-angle) and soft (small-energy) approximations. This approximation proves to be inaccurate for hard, large-angle emissions. Thus the GPMC event generators make use of the so-called Matrix Element Corrections (MEC), which correct the emission of the hardest jet at large angles in a $2 \to 1$ or $2 \to 2$ process. Nevertheless, in the past decade, so-called Matrix Element and Parton Shower matching (ME+PS) event generators together with NLO and Parton Shower matching (NLO+PS) event generators were developed to improve the parton shower description of the hard scattering process and to get rid of the limitation to $2 \to 1$ and $2 \to 2$ processes. The main focus of these programs lays on the use of exact matrix elements for the hard subprocess. Due to the accurate description of the hard subprocess these generators strive for more advanced techniques for the generation of the matrix elements and the phase space calculation have to be applied resulting in more difficult computations. The matrix element event generators still depend on GPMC generators for the parton shower and hadronization, so that elaborated matching procedures between the matrix element generator and the parton shower are needed (see Section 4.1.5). In contrast to the ME+PS event generators, the NLO+PS event generators extend the accuracy of the generation of the basic process at NLO in QCD. For this purpose, the NLO+PS event generators contain also real emission and virtual loop matrix elements beside the Born matrix elements.
Two of the most common event generators of this type, MadGraph5_aMC@NLO [56] and

POWHEG Box [57], are used in the $H \to \gamma\gamma$ analysis, in particular to generate signal samples. These two event generators have a ME+PS operation mode as well as a NLO+PS operation mode.

MadGraph5_aMC@NLO is a fusion of the LO ME+PS event generator MadGraph5 and the NLO+PS event generator aMC@NLO. At the end of 2013 it was the first public event generator which allowed to perform a NLO computation and optional matching with a GPMC event generator without the need to include further external dependencies. MadGraph5_aMC@NLO additionally includes the MadSpin tool [58] for the proper treatment of the decay of parton-level events. If MadSpin is applied to the particle decay spin correlations between the decaying particles are preserved with very good accuracy.

POWHEG Box is a general framework for implementing NLO matrix element computations in GPMC event generators and makes use of the POWHEG method [59]. The particularity of POWHEG Box is that it cannot be used for arbitrary processes because for each process the LO and NLO matrix element computation of the hard subprocess has to be provided in the form of a program package. A large number of processes of general interest for high-energy physics were implemented and are now available in POWHEG Box.

4.1.5 Combining matrix element and parton shower: the jet matching

As previously discussed, the matrix element and parton shower techniques are appropriate for modelling different phase-space: the matrix element calculation is the most reliable for the hard scattering subprocess, while the parton shower is more suitable for soft and collinear radiation emissions and jet formation. Under these premises, it is evident that a combination (matching) of these two approaches would be the best solution for the extraction of theoretical predictions to compare to data.
The main obstacle to the practical implementation of the matching is the definition of the separation between the hard component of the process, to be solved with the matrix element calculation, and the soft and collinear emissions that instead have to be done by the parton shower. There is in fact an intrinsic ambiguity: an event with a given number N of jets can be produced either from a matrix element with N outgoing partons or from a matrix element with (N-1) outgoing partons plus a jet coming from parton shower. The two approaches are equivalent and lead to the same result. Since factorisation theorems are not rigorously applicable to complex final state topologies, a specific factorisation prescription, the *matching scheme*, is introduced, identifying on an event-by-event basis the approach that provides the best description of a given configuration. The aim of the matching scheme is to avoid double counting in the phase-space regions of overlap between the matrix element calculation and the parton shower. The most natural solution is to apply a cutoff, called *matching scale*: branchings occurring at a scale harder than the

cutoff are handled by the matrix element calculation, while softer radiation emissions are left to the shower program. The matching scale being unphysical, the resulting theoretical predictions should not depend, or at least show a small dependence on the choice of the cutoff.
Several matching schemes have been developed, such as the CKKW [60], the MLM [61] and the FxFx [62]. In MadGraph5_aMC@NLO the matching between the NLO matrix element computation and the parton shower is done by introducing so-called MC counterterms and a subtraction scheme. By adding or subtracting these MC counterterms during the computation of the parton-level short-distance cross sections the overlaps are taken into account and the parton shower can later be applied to the parton-level events. The necessary MC counterterms can be computed in a process-independent manner, but they depend on the particular parton shower and lead to a specific set of MC subtraction terms for each specific parton shower. Because of the MC counterterms, event samples generated by MadGraph5_aMC@NLO consist of a large number of events with positive weight, but also include a small number of events with negative weights. As a rule of thumb, the fraction of events with negative weights increases for processes with higher multiplicities. On the other hand, in POWHEG the hardest emission is generated with NLO accuracy at first and independently from the subsequent parton shower which is matched to the matrix element generator part. In this way, also for NLO event generation no subtraction terms are required.

4.1.6 H $\to \gamma\gamma$ Monte Carlo samples

The data sample used for this analysis corresponds to an integrated luminosity of 12.9 fb^{-1}, recorded at the LHC in pp collisions at a center-of-mass energy of 13 TeV.
Monte Carlo simulation is used to model Higgs to diphoton signal, after appropriate validation and corrections, in order to measure signal properties. Monte Carlo simulations of the background processes are used only to optimize the selection requirements and to train various multivariate discriminators, but are not used for the final analysis, and the results do not depend on a proper description of the background processes by the corresponding Monte Carlo simulation.
The signal samples produced for the standard H $\to \gamma\gamma$ analysis consist of all the four production modes, for seven different mass points (120, 123, 124, 125, 126, 127 and 130 GeV). The different mass points are used for the construction of a parametric signal model, as explained in Section 4.12.
Concerning the background, as already mentioned in Section 3.2.1, a large irreducible background is present from QCD diphoton production in both quark and gluon initial states. Leading order diagrams for the quark-initiated Born diphoton production and the gluon initiated box di-photon production are shown in Figure 4.7. Despite the fact that the gluon-induced process does not exist at tree-level, the contribution is comparable to the

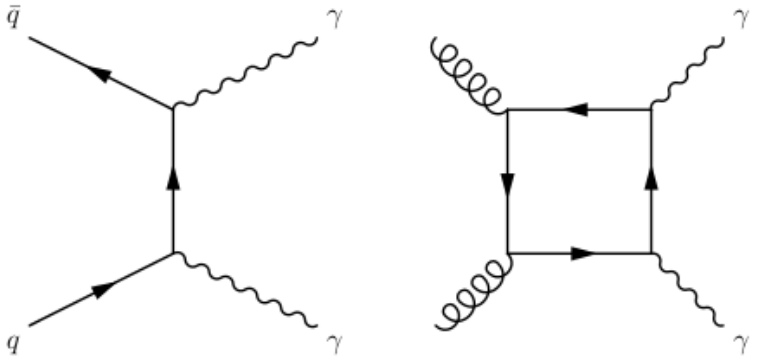


Figure 4.7: Leading order Feynman diagrams for QCD diphoton production from quark (left) and gluon (right) initial states.

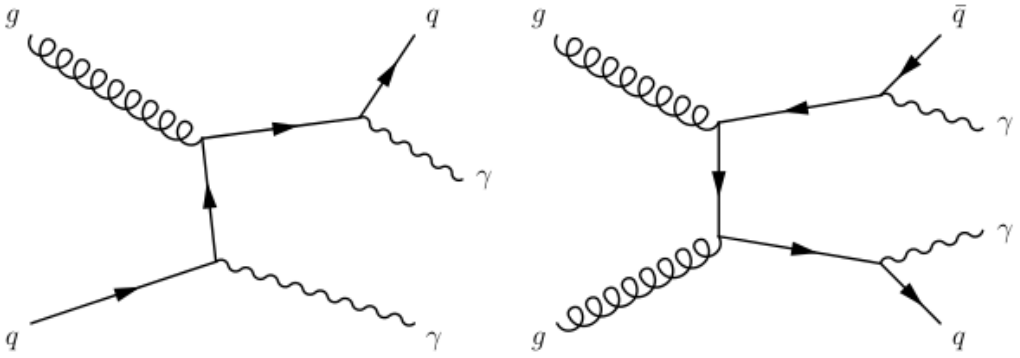


Figure 4.8: Example of tree-level Feynman diagrams for QCD diphoton production in association with one or two additional jets.

quark-induced production, given the higher gluon-gluon luminosity at the LHC as compared to quark-antiquark. An additional source of irreducible background are QCD di-jet or photon + jets production with additional photons produced by Initial State/Final State Radiation (ISR/FSR), example diagrams for such processes are shown in Figure 4.8. In addition to the irreducible background containing two prompt photons, there is a reducible background from QCD di-jet and photon-jet production where one or more of the reconstructed photons arise from a quark or gluon jet.

The software used for the matrix element step varies according to the availability of implementations for each desired process, using a mixture of POWHEG [57], MadGraph5_aMC@NLO [56] and Pythia 8 [54], where POWHEG and MadGraph5_aMC@NLO use NLO matrix elements, while Pythia uses LO matrix elements. In particular MadGraph5_aMC@NLO is the new version of both MadGraph5 and aMC@NLO that unifies the LO and NLO lines of development of automated tools within the MadGraph family.
The signal samples are generated both with POWHEG and MadGraph5_aMC@NLO, although it was decided to use MadGraph5_aMC@NLO in the analysis. The parton level samples are interfaced to Pythia 8 for parton showering and hadronization.
Different generators are used to generate the simulated samples for the background: Pythia 8, MadGraph5_aMC@NLO interfaced with Pythia 8, and the matrix element Sherpa generator which also generates the particle shower. The QCD diphoton prompt-prompt background, for example, is modeled with the Sherpa generator. It includes the born processes with up to 3 additional jets at LO accuracy as well as the box processes at LO. Prompt-fake and fake-fake backgrounds are instead modeled with Pythia 8.
In order to select the events which are likely to pass the later diphoton selection of the analysis, I developed a "double EM-enriched" filter for the production of the QCD di-jet and γ+jet samples. In this way computing power for the further simulation of interactions between the particles and the detector was saved. This filter requires a potential photon signal (electromagnetic activity), coming from photons, electrons, or neutral hadrons, with $p_T > 15$ GeV. In addition it is required that this potential photon signal has no more than two charged particles in a cone $\Delta R < 0.2$, mimicking a tracker isolation. These charged particles consist of electrons, muons, taus, pions and kaons. They are required to have $p_T > 1.6$ GeV and $|\eta| < 2.2$. For each event, the potential photon signals are coupled together to form double EM enriched objects. Just the couples passing particular cuts are kept. The DY sample is simulated with MadGraph5_aMC@NLO.
A complete list of simulated processes, generated at 13 TeV, as well as the corresponding matrix element generators are given in Table 4.1 Since the parton shower step may add additional photons to the event with respect to the matrix element, some care must be taken to avoid double counting of background processes. In particular, QCD photon + jet events with one additional photon added by the parton shower are already included in the Sherpa diphoton + jets sample. Because the matrix element is expected to describe the

Table 4.1: List of simulated processes and the corresponding matrix element generators.

Process	**Matrix Element**
Gluon Fusion Higgs	MadGraph5_aMC@NLO - POWHEG
Vector Boson Fusion Higgs	MadGraph5_aMC@NLO - POWHEG
W/Z Associated Production Higgs	MadGraph5_aMC@NLO - POWHEG
$t\bar{t}$ Associated Production Higgs	MadGraph5_aMC@NLO - POWHEG
Drell-Yan di-lepton + 0-2 jets	MadGraph5_aMC@NLO
QCD di-photon (gluon-gluon box and Born diagrams)	Sherpa
QCD Photon + jet Pt-20to40	Pythia
QCD Photon + jet Pt-40toInf	Pythia
QCD Di-jet Pt-30to40	Pythia
QCD Di-jet Pt-40toInf	Pythia

kinematics of such events better than the parton shower, these events need to be removed from the Pythia sample at analysis stage. Similarly, QCD di-jet events with two photons added by the parton shower are already included in the Sherpa diphoton + jets sample, and need to be removed from the Pythia sample.
The cross-sections for signal processes have been computed up to NNLO+NNLL and are documented in Reference [63]. The cross-sections for background processes, where used for optimization, are computed from the LO matrix element generators.

Pile-up conditions are simulated such that the running conditions of the 2016 run are covered. The pile-up scenario accounts for multiple pp collisions happening in the same bunch crossing as well as for 25 ns out-of-time pile-up in a window of [-12,+4] bunch crossings. The average number of pile-up events in data is 18.5. Events in the simulated samples are weighted such that the resulting pile-up distribution matches that of data. To validate the weighting technique that is applied to the simulation in order to match the actual pile-up events distribution observed in the data, the comparison of the number of reconstructed vertices is done between the data and the simulation after reweighting for a sample of $DY \rightarrow ee$ events. The distribution is shown in Figure 4.9.

4.2 Trigger

The events entering in this analysis must firstly pass a hardware level trigger (L1) followed by a software level trigger (HLT) decision. Since a certain number of events would be able to pass the analysis selection but not the diphoton trigger one, the trigger has an efficiency smaller than one, reducing the number of events which can enter the analysis. To measure the trigger efficiencies, one needs to evaluate separately the efficiency of the L1-seeding (L1 efficiency) and the efficiency of the HLT filters, provided that the L1 requirement has been satisfied. For efficiency measurements the tag and probe method on $Z \rightarrow ee$ data is used

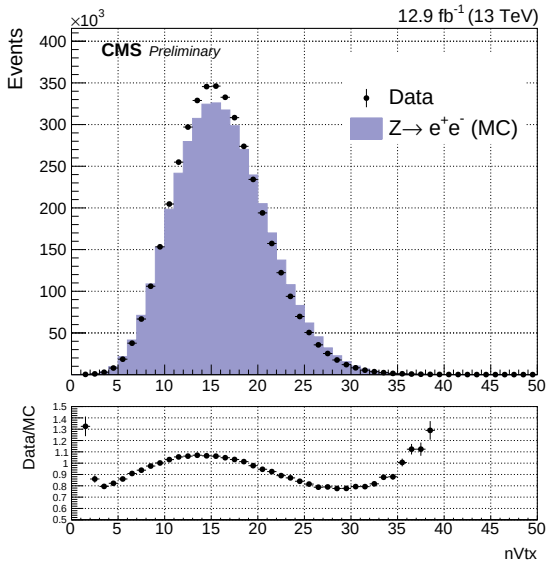


Figure 4.9: Distribution of the number of reconstructed vertices for $Z \to ee$ events in data (black dots) and simulation after having applied the re-weighting on the number of simulated pile-up events (blue histogram).

(see Appendix A for a description of the tag and probe method). The diphoton trigger and the measurement of its efficiency are described below.

4.2.1 Level 1 trigger

Each high level trigger diphoton path is seeded by at least one hardware level 1 electromagnetic candidate. Because of bandwidth limitations at the L1, 40 GeV is the lowest transverse momentum of single electromagnetic L1 candidates, giving a few percent inefficiency at the lowest transverse energy of the analysis selection. This is mitigated seeding the HLT paths also by a L1 pair of 22 and 10 GeV respectively. Figure 4.10 shows the L1 efficiency measured with the tag and probe technique on $Z \to ee$ data, as a function of the probe p_T.
Given an L1 seed, the ECAL clustering algorithm is performed by the HLT from the readout units overlapping a rectangle centered on the L1 candidate, extended from the rectangle of the L1 segment by $\Delta\eta \times \Delta\Phi = 0.14 \times 0.4$. The requirements of the HLT diphoton path are then applied.

4.2.2 High level trigger

The variables entering the HLT selection are grouped into general, isolation plus calorimeter identification (Iso+CaloId) and R_9 variables. The global criterion is applied firstly to all

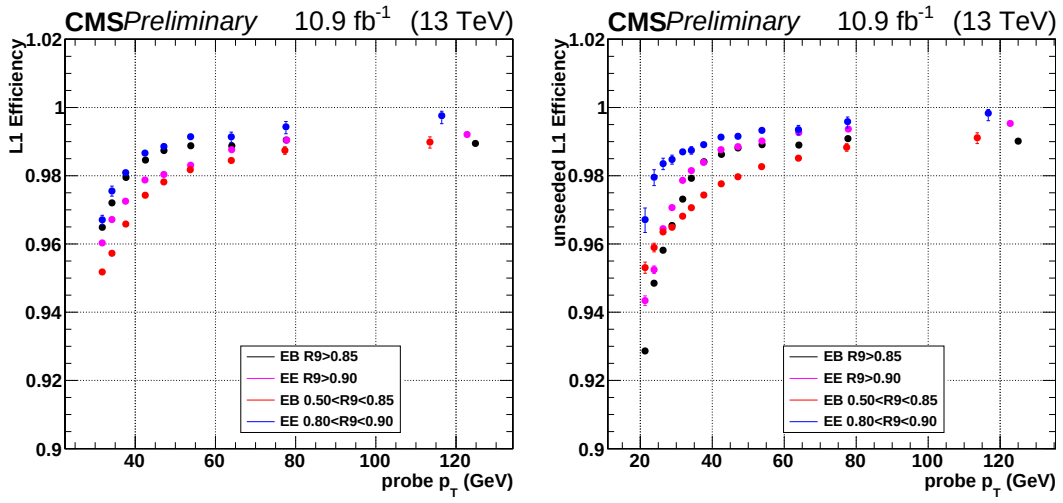


Figure 4.10: Trigger efficiency, as a function of offline probe p_T, measured on data for $Z \to ee$ events using the tag and probe technique. Efficiency of photons in the 4 cut-based analysis categories are shown. The plots correspond to the different p_T thresholds used to seed the diphoton HLT trigger, 22 GeV (on the left) and 10 GeV (on the right).

objects, then either the Iso+CaloId or the R_9 selection is required. The variables used are the following (with more details in Section 3.2.2):

- general variables: E_T of both photons, $m_{\gamma\gamma}$, H/E (with modified cone size R = 0.14),
- variables used in Iso+CaloId paths: $\sigma_{i\eta i\eta}$ ("full5x5"), ECAL PF Cluster ISO, Tracker isolation in a hollowcone (with modified R = 0.29 outer and R = 0.06 inner radii),
- variables used in R_9 paths: R_9 ("full5x5").

The L1-seeded leg of the HLT is required to have $E_T > 30$ GeV. The cluster must have $|\eta| < 2.5$ and $R_9 > 0.5$ (0.8) in EB (EE). In addition the cut H/E $< 0.12(0.1)$ in EB (EE) is applied. After that, the clusters are filtered by an $R_9 > 0.85$ (0.9) EB (EE) selection; if passing, the clusters continue to the unseeded step without going through the Iso+CaloId filters, if failing, the clusters must pass the Iso+CaloId selection, which consists of $\sigma_{i\eta i\eta} < 0.015$ (0.035) EB (EE) and ECAL isolation $< 6.0 + 0.012\ E_T$.

If at least one cluster passes the previous criterion, the entire ECAL is clustered. Two clusters, with one corresponding to the seeded HLT cluster, are now required on the unseeded leg: an $E_T > 18$ GeV cut is applied, the remainder of the selection is the same as for the seeded leg with the addition that both legs are required to pass tracker isolation $< 6.0 + 0.002\ E_T$.
The mass of the diphoton object is finally required to be above 90 GeV.

4.2.3 Trigger performance

The efficiency of the trigger selection is studied using the tag and probe technique. The efficiency of both the seeded and unseeded leg of the diphoton HLT path is measured on $Z \to ee$ data events. It is possible to select high purity $Z \to ee$ events requiring events to pass a tag and probe trigger and requiring an additional offline selection. Since the tag and probe trigger only requires a single electron, the second electron is an unbiased probe to measure the efficiency of the diphoton path. To account for the shower shape difference between electrons and photons (as well as the different η distributions for $Z \to ee$ and $H \to \gamma\gamma$), due to the material upstream of the ECAL, the entries to the efficiencies are weighted in R_9 and η from the respective simulated samples in order to match the $H \to \gamma\gamma$ distributions. This brings to an event migration to higher R_9 values, which gives an overall increase of the measured efficiency of the $Z \to ee$ events. The HLT efficiency of both the seeded and unseeded leg of the diphoton trigger with respect to the offline photon p_T is shown in Figure 4.11.

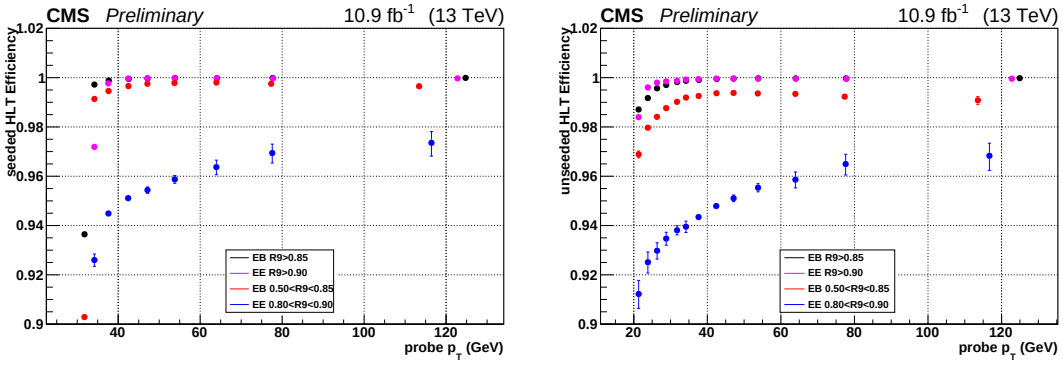


Figure 4.11: Trigger efficiency measured on data for $Z \to ee$ events using the tag and probe technique. The four untagged cut-based categories are shown for the seeded (left) and unseeded (right) HLT leg of the diphoton trigger.

For the majority of simulated samples used in this analysis the analysis trigger was not simulated. This means that there is no trigger applied on the simulated events in our analysis. After the analysis preselection is applied the trigger is reasonably efficient, but corrections must still be applied to the simulation to replicate the inefficiency of the trigger in data.

The efficiency measured on data, binned in E_T, R_9 and η, is used to scale the simulation. The scale factors are a combination of the HLT and L1 efficiency presented earlier in this section to properly simulate the seeding of the HLT by the L1. These scale factors are applied to the simulation to replicate the effect of the trigger.

Since the application of trigger scale factors affects both the signal and background yields

from simulation, a log-normal asymmetric uncertainty calculated from the tag and probe fits to the Z-peak is applied.

4.3 Photon energy correction

Different sets of corrections are necessary in order to achieve the best photon energy resolution. The first consists of crystal-level corrections needed to equalize the channel-to-channel response variations. The second, an high-level correction method called photon energy regression, takes into account finer effects, like the containment of the shower and the energy losses. This method is based on a multivariate approach and it provides a per-photon energy resolution estimator, which is used for the diphoton BDT as described in Section 4.8. Further differences between data and simulation are resolved using $Z \to ee$ with electrons reconstructed as photons. The energy scale is corrected in data and a smearing is applied to MC in order to have the best agreement between data and simulation for photon energy scale and resolution.
Figure 4.12 shows the comparison of simulation to data for the Z invariant mass plots in η, R_9 categories after all the energy corrections described above have been applied. A more detailed description on the photon energy corrections is given in Appendix B.

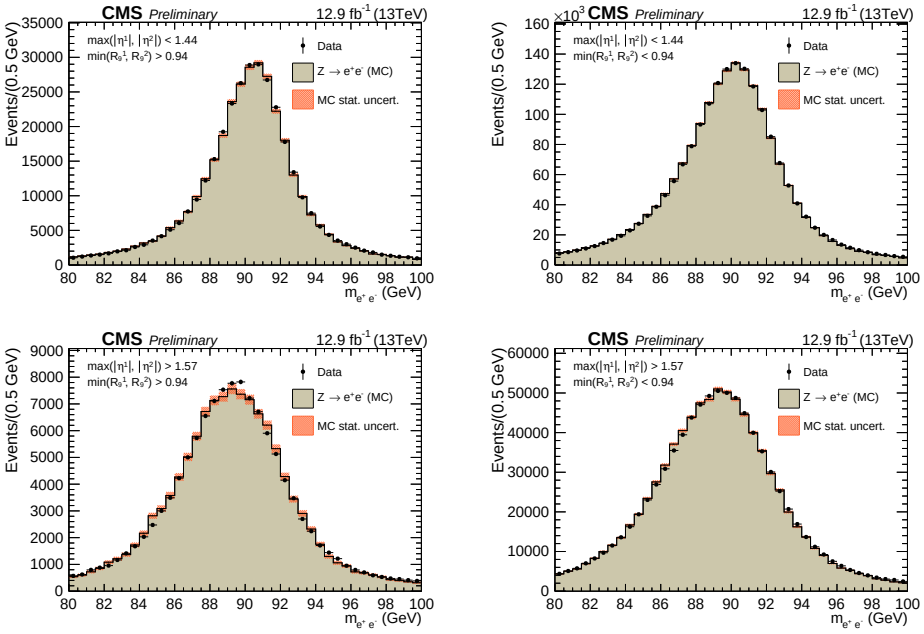


Figure 4.12: Comparisons of data to simulation for the invariant mass of di-electrons from Z boson decay reconstructed as photons. The events are splitted in η, R_9 categories. The comparison between the simulation (filled histograms) and data (black points) is shown.

4.4 Event preselection

For each event with at least two reconstructed photons, the diphoton pairs are first reconstructed by grouping the photons into all possible two photon combinations. A primary vertex is assigned to each diphoton pair, as described in Section 4.5. The momentum of each photon is built with its magnitude obtained from the corrected photon energy as described in Section 4.3 and its direction pointing from the selected vertex to the supercluster. A preselection designed to match the HLT requirements is then applied on each photon, in order to keep the same phase space in data and in simulation (where the HLT is not applied). The preselection, tuned to be tighter than the trigger, includes a cut on the acceptance of supercluster pseudorapidity measured with respect to the origin of the detector coordinate η_{SC}, a set of loose photon identification cuts against jets faking photons and an electron veto:

- the cuts on the acceptance of the supercluster pseudorapidity, which aim to exclude the transition region between the ECAL barrel and endcap and the region outside the tracker acceptance, are $|\eta_{SC}| < 1.442$ in the barrel or $1.566 < |\eta_{SC}| < 2.5$ in the endcap,

- the ECAL shower shape and isolation variables used in the loose photon identification cuts are the following: H/E, $\sigma_{i\eta i\eta}$ ("full5x5"), R_9 ("full5x5"), Photon ISO and Tracker ISO (both the isolations with a ρ correction applied to match the HLT).
  To apply these cuts, the photons are classified into four categories according to the photon supercluster location in the ECAL (barrel or endcap) and the value of R_9 (> 0.85 or ≤ 0.85 in the barrel, > 0.9 or ≤ 0.9 in the endcap). The photons in the barrel and endcap are treated separately because the geometry of the crystals and the amount of tracker materials in front are different in the two cases. The values of the loose photon identification cuts are summarised in Table 4.2,

- the electron veto is used to distinguish electrons from photons. A photon candidate is removed if its supercluster is matched to an electron track. To avoid rejecting the converted photons, the electron track is required to have no missing hits in the tracker before its first hit, and not to match an identified conversion.

In addition the leading photon (photon with the largest p_T) is required to have $p_T > 30$ GeV, while the subleading one (photon with the second largest p_T) must have $p_T > 20$ GeV.
Finally we apply a set of diphoton kinematic acceptance cuts, determined to select the phase space right above the trigger threshold and to define a region for the diphoton mass fit. The cuts include $m_{\gamma\gamma} > 95$ GeV, 100 GeV $< m_{\gamma\gamma} < 180$ GeV, $p_T^{\gamma 1}/m_{\gamma\gamma} > 1/3$ and $p_T^{\gamma 2}/m_{\gamma\gamma} > 1/4$, for the leading photon $\gamma 1$ and the subleading photon $\gamma 2$ respectively.

Table 4.2: The loose photon identification cuts for single photon preselection. The photons are divided into four categories according to the photon supercluster position in the ECAL (barrel or endcap) and to the R_9 value. The cut values vary with the photon categories.

	H/E	$\sigma_{i\eta i\eta}$ (5x5)	R_9 (5x5)	pfPhoIso	TrackerIso
EB; R9 $>$ 0.85	$<$ 0.08	–	$>$ 0.5	–	–
EB; R9 $\leq$ 0.85	$<$ 0.08	$<$ 0.015	$>$ 0.5	$<$ 4.0	$<$ 6.0
EE; R9 $>$ 0.90	$<$ 0.08	–	$>$ 0.8	–	–
EE; R9 $\leq$ 0.90	$<$ 0.08	$<$ 0.035	$>$ 0.8	$<$ 4.0	$<$ 6.0

Preselection efficiencies in the four photon categories are evaluated with the tag and probe technique using electrons from $Z \to ee$ events, for which the electron R_9 is rescaled to match photon R_9 distribution in H $\to \gamma\gamma$ simulation. Data and simulation efficiencies are then compared in order to calculate the appropriate scale factors. By definition the tag and probe technique using $Z \to ee$ events does not allow to measure the electron veto efficiency, which is instead measured independently using $Z \to \mu\mu\gamma$ events. Table 4.3 shows the efficiencies calculated on $Z \to ee$ events for data, simulation, and their ratio in four η, R_9 categories.

	Data			Simulation		Ratio	
	Eff.	Stat	Syst.	Eff.	Stat.	Eff.	Unc.
Barrel; R_9 $>$0.85	0.9451	0.0006	0.0192	0.9374	0.0007	1.0080	0.0192
Barrel; R_9 $<$0.85	0.8255	0.0012	0.0119	0.8258	0.0009	0.9960	0.0120
Endcap; R_9 $>$0.90	0.9099	0.0008	0.0212	0.9127	0.0010	0.9969	0.0212
Endcap; R_9 $<$0.90	0.4993	0.0018	0.0249	0.5024	0.0016	0.9938	0.0250

Table 4.3: Preselection efficiencies measured in the 4 photon categories using tag and probe with $Z \to ee$ events (for all preselection criteria except electron veto).

4.5 Diphoton vertex identification

In the decay of the Higgs boson into two photons the determination of the primary vertex associated with the signal is very important, because the vertex choice has a direct impact on the diphoton mass resolution (a wrong choice would worsen the resolution by about 1 GeV on average), and it is not trivial for two main reasons. Firstly, the final state unconverted photons are not detected in the tracker and the electromagnetic calorimeter alone

cannot be used for pointing, as it does not have a longitudinal segmentation. Secondly, the diphoton production vertex needs to be selected from an average of 18.5 pp collision vertices distributed in z with an RMS of about 4 cm. In order to discriminate between the diphoton production vertex and the pileup vertices, we use the knowledge that the total transverse momentum of the recoiling tracks associated with the diphoton production vertex roughly balances the diphoton transverse momentum plus the fact that a hard vertex produces high p_T tracks compared to a minimum bias one. Even if the balance is not exact as we do not have the association between neutral particles and vertices, it is possible to build some variables having different distributions for the recoiling tracks of the diphoton production vertex and those of the pileup vertices. The discrimination power provided by these variables can finally be exploited using a multivariate tool. In addition to the correlation between the kinematics of the recoiling tracks and that of the diphoton, when at least one of the photons is converted in the tracker the position of the conversion vertex, together with either the direction of the conversion momentum or the position of the ECAL supercluster, provides an extrapolation of the position of the diphoton vertex, which is used for the vertex selection. Therefore we train a BDT using the above information, to distinguish between the prompt vertex and the pileup vertices. The BDT assigns a score to each vertex according to how likely it is to be the $\gamma\gamma$ vertex, and the vertex with the highest score is selected. The effect of the vertex selection on the diphoton mass resolution is negligible with respect to the single photon energy resolution if the selected diphoton vertex is required to be within 1 cm in z from the actual diphoton vertex. A per-event probability to choose the right vertex is also determined, giving the full benefit of the excellent ECAL resolution.
A more detailed description on the diphoton vertex identification is given in Appendix C.

4.6 Photon identification

Photon identification for the H $\to \gamma\gamma$ analysis is explained in detail in Chapter 3.

4.7 Event classification

In order to improve the sensitivity of the analysis, diphoton events passing the preselection and having an output value of the photon identification BDT greater than -0.9 (see Section 3.3.2) for both photons are categorized by mass resolution, signal-to-background ratio and production mechanism. The production mechanisms which differ from gluon fusion can be identified by selecting final state objects in addition to the diphoton pair. The Vector Boson Fusion production is characterized by the presence of a pair of jets separated by a large rapidity gap. The tagging of VBF events increases the overall sensitivity of the analysis and the precision on the measured signal strength, and allows to measure the

coupling to vector bosons. The tagging of the ttH production mode, which is accompanied by b-quarks and possible charged leptons or additional jets, increases the sensitivity of the measurement of the coupling to vector bosons and top quark, and further probes the compatibility of the observed signal with a SM Higgs boson.
These events with additional objects are tagged as exclusive categories, while the untagged ones belong to the inclusive categories and identify the gluon fusion production mechanism. If more than one diphoton candidate is present, the candidate with the highest priority tag is selected. In the case of multiple diphotons with equal tag priority, the diphoton with the maximum scalar sum of photon transverse momentum is used for the analysis. The event tagging priority sequence is based on the signal purity and is as following:

- first, events with leptons from the leptonic or semi-leptonic top decays are selected (ttH leptonic tag),
- second, remaining events with jets from hadronic top decays are selected (ttH hadronic tag),
- then events with two forward jets are selected and divided into two VBF categories,
- finally remaining events, without additional objects, are classified in four different untagged categories (see later).

The inclusive events are categorized using a multivariate classifier. The multivariate diphoton classifier, described in the next section, divides the untagged events in four categories based on photon kinematics, mass-resolution and other inputs indicating the signal-to-background ratio.

4.8 Diphoton BDT

4.8.1 Classifier setup and performance

As already said in the previous section, an increase of the sensitivity of the analysis is obtained categorizing the events. In this way the "high-perfomance" categories are characterised by a higher signal-to-background ratio and a better mass resolution.
The untagged events are categorized using the output of a multivariate event classifier, implemented using a BDT, which goal is to assign high scores to events with two photons fulfilling the following criteria:

- signal-like kinematic characteristics,
- good diphoton mass resolution,
- photon identification BDT score in the "photon-like" region.

The events with two photons entering the diphoton classifier, must satisfy the preselection criteria described in Section 4.4, with an additional loose cut on the photon ID BDT output ($>$ -0.9) which is 99% efficient on signal events. In addition photons have to pass a diphoton mass-dependent requirement on the transverse momentum, that is $p_T > 1/3$ (1/4) $m_{\gamma\gamma}$ for the leading (subleading) photon.
The set of input variables is chosen such that the diphoton system mass cannot be inferred from these variables, in this way the classifier result cannot be biased by the specific mass of the signal used as training sample. For this purpose, dimensional variables are rescaled by the mass of the diphoton system.
The variables entering the diphoton event classifier are therefore the following:

- the transverse momenta of both photons, rescaled for the diphoton mass, $p_T^{1(2)}/m_{\gamma\gamma}$,
- the pseudorapidities of both photons, $\eta^{1(2)}$,
- the cosine of the angle between the two photons in the transverse plane, $\cos(\Delta\Phi_{\gamma\gamma})$,
- the identification BDT score for both photons,
- the per-event relative mass resolution estimate, under the hypothesis that the mass has been reconstructed using the correct primary vertex (σ_{rv}),
- the per-event relative mass resolution estimate, under the hypothesis that the mass has been reconstructed using an incorrect primary vertex (σ_{wv}),
- the per-event probability estimate that the correct primary vertex has been used to reconstruct the mass, based on the event-level vertex selection BDT as described in Section 4.5.

The per-event relative mass resolution estimate assuming the correct vertex is computed as the sum in quadrature of the per-photon energy resolution estimators of the leading and subleading photon as:

$$\sigma_{rv} = \sigma_m^{right}/m_{\gamma\gamma} = \frac{1}{2}\sqrt{\left(\frac{\sigma_{E1}}{E_1}\right)^2 + \left(\frac{\sigma_{E2}}{E_2}\right)^2}.$$

In order to correctly estimate the mass resolution for those events where an incorrect primary vertex is selected, the estimator of the mass resolution includes an additional term σ_{vtx} given by the displacement between the correct and the selected primary vertex. The distance between the selected vertex and the true one is distributed as a Gaussian with width $\sqrt{2}\sigma_Z^{beamspot}$ and the term σ_{vtx} can be computed analytically given the impact positions of the two photons in the calorimeter. The relative mass resolution under the incorrect vertex hypothesis is therefore computed as:

$$\sigma_{wv} = \sigma_m^{wrong}/m_{\gamma\gamma} = \sqrt{\left(\sigma_m^{right}/m_{\gamma\gamma}\right)^2 + \left(\sigma_m^{vtx}/m_{\gamma\gamma}\right)^2}.$$

In order to provide to the BDT training the information that signal-to-background is inversely proportional to mass resolution, the signal events entering the training are weighted as follows:

$$w_{sig} = \frac{p_{vtx}}{\sigma_{rv}} + \frac{1 - p_{vtx}}{\sigma_{wv}},$$

where p_{vtx} is the probability of the selected vertex being the right one estimated from vertex probability BDT, while σ_{rv} and σ_{wv} are the relative mass resolutions assuming the correct/wrong vertex is selected. In this way the events with better mass resolution get higher weights and appear more signal like.

The diphoton BDT is trained on samples simulating signal and background processes at 13 TeV. The signal sample consists of H $\rightarrow \gamma\gamma$ events at a Higgs mass of 125 GeV, with all four production processes weighted by cross section. The background sample consists of a proper mixture of prompt-prompt (from diphoton+jets sample), prompt-fake (from photon+jet and QCD di-jet samples) and fake-fake (from QCD di-jet samples) events.

The diphoton BDT discriminating power between signal and background events is summarised in Figure 4.13, where the ROC curve of the classifier and the normalised BDT output variable for signal and background are shown. Figure 4.14 shows the diphoton

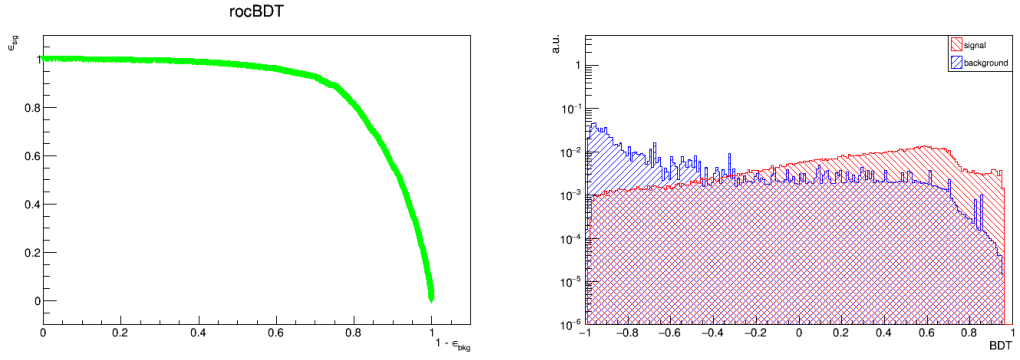


Figure 4.13: The diphoton classifier ROC curve (left) and the BDT output variable, normalised to unity, for simulated signal and background events (right).

invariant mass distribution, for both background and signal events, for different diphoton BDT ranges. It can be seen that for the background the distribution is smooth in all the diphoton BDT ranges, confirming the mass-blindness of the BDT classifier. The signal distribution, on the other hand, shows that events with narrower mass resolution tend to have higher BDT scores. The distributions of the diphoton BDT output for the different components of the signal and background samples are shown in Figure 4.15 for the mass region 100 GeV $< m_{\gamma\gamma} <$ 180 GeV. The BDT output variable is transformed in order to

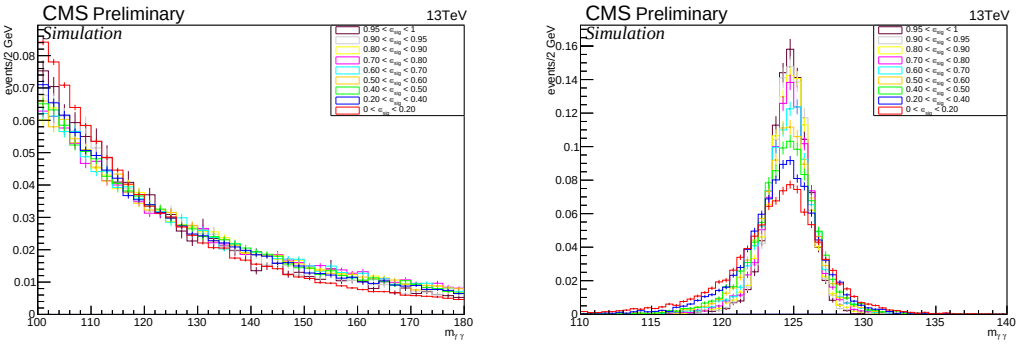


Figure 4.14: The diphoton mass distributions, normalised to unity in every diphoton BDT range, for background (top) and signal (bottom) events.

have a flat distribution for the cross-section weighted sum of the four signal samples, allowing to have a more immediate visualisation of the BDT performances. From the signal processes distribution it is evident that VBF, VH, and ttH processes tend to accumulate at higher BDT scores compared to gluon fusion, because of the harder Higgs p_T spectrum. On the other hand the distribution for the different background components shows that events contained in the diphoton+jets sample cluster in a region at high values of BDT score, since they are characterised by photons with good photon ID values and better resolution compared to prompt-fake and fake-fake components coming from γ+jet and QCD di-jet samples. Figure 4.15 also shows data-simulation comparison and the definition of the diphoton BDT-based untagged categories (described later in Section 4.8.3), with the black dashed lines showing the boundaries of the categories and the grey shaded area identifying the region of low BDT values where events are excluded from the analysis. The data and simulation are in good agreement.

4.8.2 Systematic uncertainties

In the H $\to \gamma\gamma$ analysis the background is modelled in a fully data-driven manner, so the result does not depend on the simulation prediction for the diphoton BDT output shape of the different background components. The signal modelling is driven instead by the simulation prediction, where corrections from simulation to data are applied. The two main sources of systematic uncertainty on the diphoton BDT output derive from the photon ID BDT and the per-photon energy resolution estimate from the regression.

The photon ID BDT output ranges from -1 to 1, with prompt photons tending to have values close to 1. The systematic uncertainty to this variable is assigned shifting its value for every photon in the simulation according to a transformation combining a shift of $\pm$

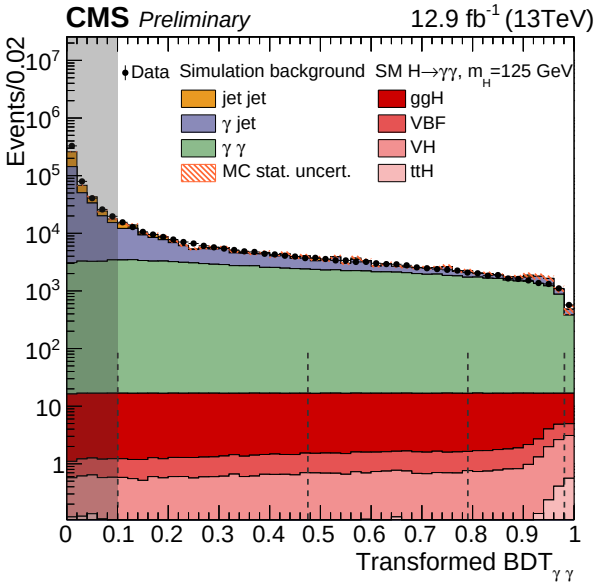


Figure 4.15: The transformed BDT output distribution is compared between simulation (stacked histograms) and data (black points). The transformation is done in order to have a flat distribution for signal events.

0.03 with a linear increase of the uncertainty for events having a low photon ID BDT score. Since larger values of the photon ID BDT tend to lead to a larger value of the diphoton BDT, the simultaneous translation of the photon ID BDT for both photons produces the maximal migration of events in the diphoton BDT output, which is then considered as a migration of the signal yield among the final event classes.

The per-photon resolution estimate is affected by imperfect modelling of the electromagnetic shower shape in the simulation. The diphoton BDT output is influenced by this quantity because of its impact on the mass resolution estimates, both under the right and wrong vertex hypotheses (σ_{rv} and σ_{wv}). The systematic uncertainty from imperfect modelling of σ_E/E is assigned shifting its value by $\pm 5\%$ for each photon. The diphoton BDT is expected to have lower values for larger energy resolution estimates and thus to produce the maximum event migration for simultaneous shifts.

The impact of these two sources of systematic uncertainty on the transformed diphoton BDT output variable is shown in Figure 4.16.

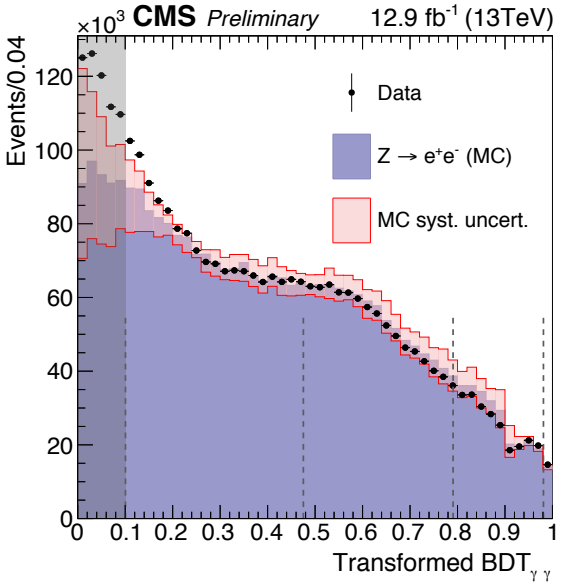


Figure 4.16: Data-simulation comparison for diphoton BDT output for electrons reconstructed as photons from the decay of the Z boson. The red band indicates the impact on the BDT output variable of the systematic uncertainties on the photon ID BDT variable and on σ_E/E.

4.8.3 Event categorisation using diphoton BDT output

The diphoton BDT is used to split the untagged events in categories. The boundaries of the categories in the BDT output spectrum are chosen as a result of an optimisation method, and it is found that splitting the events into four categories increases the expected significance of the analysis, while further splitting does not bring substantial additional gain.

The optimisation method, for an n-category analysis, initially places n boundaries at equal distances in signal quantiles from each other, given the two fixed boundaries at the edges of the spectrum -1 and 1. $n + 1$ categories are defined by the n boundaries, and only n are selected for the analysis and for the calculation of the expected significance, while the events falling in the lowest-BDT score category are not used. The positions of the boundaries are free to vary and are defined minimising the combined p-value [1]. In each category, signal and background simplified models are extracted from simulation. Both models are extracted through a fit to the diphoton invariant mass distribution in simulated samples.

[1] The p-value is the probability P, under assumption of a null hypothesis H_0, of obtaining a result as compatible or less with H_0 than the one actually observed. If H_0 is the background-only hypothesis, the p-value is the probability that the background fluctuates to the observed value. In case of an observed excess above the expected background, the p-value is used to estimate the significance of that excess.

The procedure is repeated from $n = 2$ up to $n = 6$; no larger values of n are explored since the gain in expected significance is only of the order of few 0.1% increasing n from 5 to 6, as shown in Figure 4.17. Given also the small gain achievable going from $n = 4$ to 5, the analysis is split into 4 generic categories. The positions of boundaries are shown in Figures 4.15 and 4.16 and depend on the luminosity.

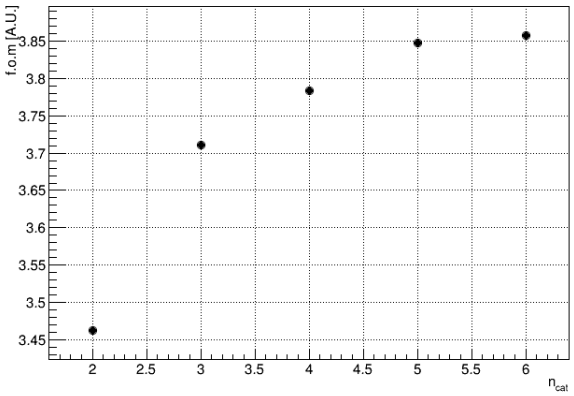


Figure 4.17: Expected significance as a function of the number of categories. The estimate of the expected significance is approximated since it is extracted from a simplified signal+background fit. It can be seen how the relative gain in expected significance is of the order of 1% after $n = 4$.

4.9 VBF tag

The events produced through the vector boson fusion (VBF) mode are characterised by the presence of two energetic jets, originating from the two scattering quarks, with large separation in η in addition to the photon pair in the final state. Even if the cross section of this production mode is more than ten times smaller than for gluon fusion, with an appropriate selection and the use of multivariate analysis tool, two categories of events with excellent signal-to-background ratio can be defined.
The VBF candidates are first preselected from the diphoton events by applying a set of loose cuts on dijet kinematics. To each VBF candidate is assigned a score from a kinematic dijet BDT, which provides a kinematic discriminator between the VBF events and both the background and the ggH events. Finally, events are further classified according to a BDT combining the output of the kinematic dijet BDT and of the diphoton BDT.
The jet definition, the dijet preselection cuts for the VBF candidates and the details about the kinematic dijet BDT and the combined BDT are provided below.

4.9.1 Jet definition

Jets are reconstructed from all particle flow candidates, clustering them through the anti-k_T algorithm [40] with a size parameter of $\Delta R = 0.4$. Furthermore, among the jet constituents the charged candidates associated with a vertex other than the selected vertex for the event are excluded (this is the so-called CHS technique).
For jets having $|\eta| > 2.5$, the tracker is no longer present and the CHS technique cannot be used. To cope with this problem a selection on the width of the jet is introduced: RMS < 0.03, with

$$RMS = \frac{\sum_{constituents} p_T^2 \Delta R^2}{\sum_{constituents} p_T^2},$$

where $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is between the constituent and the jet axis.
The value of 0.03 is chosen by maximising the significance of the kinematic dijet BDT.

4.9.2 Dijet preselection

As already said, in order to be tagged as VBF-like, the events are required to have at least 2 jets in addition to the photons. Only the two highest p_T jets are considered in the selection. The dijet preselection consists of a set of loose cuts: both jets are required to have $|\eta| < 4.7$, the p_T must be greater than 30 GeV for the leading jet and greater than 20 GeV for the subleading jet. Finally there is a requirement on the invariant mass of the dijet system, $m_{jj} > 250$ GeV.

4.9.3 Kinematic dijet BDT

The goal of the kinematic dijet BDT is to provide a kinematic discriminator between the VBF events and both the background and the ggH events. It is trained on simulated events: VBF H $\to \gamma\gamma$ sample with $m_H = 125$ GeV as signal, prompt-prompt, prompt-fake and fake-fake samples as background. The gluon fusion H $\to \gamma\gamma$ with $m_H = 125$ GeV is also used in the training as an additional background, in order to improve the purity of the dijet signal.
The variables entering the kinematic dijet BDT are the following:

- the transverse momenta of the leading and subleading photons divided by the invariant mass of the di-photon candidate: $p_T^{\gamma 1}/m_{\gamma\gamma}$ and $p_T^{\gamma 2}/m_{\gamma\gamma}$,
- the transverse momenta of the leading and subleading jets: p_T^{j1} and p_T^{j2},
- the dijet invariant mass m_{jj},
- the difference in pseudorapidity between the two jets $\Delta\eta_{jj}$,
- the *Zeppenfeld* variable, defined as the separation between the diphoton pseudorapidity and the average pseudorapidity of the dijet system: $|\eta_{\gamma\gamma} - (\eta_{j1} + \eta_{j2})/2|$,

- the difference in azimuthal angle between the dijet and the diphoton system, $\Delta\Phi_{jj\gamma\gamma}$.

4.9.4 Combined BDT and categorisation

The combined BDT is built using the kinematic dijet BDT, the diphoton BDT, and $p_T^{\gamma\gamma}/m_{\gamma\gamma}$ as input variables. Its goal is to maximally discriminate the VBF dijet signal from backgrounds using the information from all the relevant objects tagged in the event. $p_T^{\gamma\gamma}/m_{\gamma\gamma}$ is used as an input because it has a strong correlation to both the dijet BDT and the diphoton BDT.
The BDT is trained on simulated events: the signal sample is the same as for the kinematic dijet BDT, while the background sample consists of the same background events as for the dijet BDT but not the ggH events.
Figure 4.18, which shows the ROC curves for VBF signal efficiency vs efficiency for various backgrounds for both the dijet BDT and the combined BDT, demonstrates that the combined BDT is better than the dijet BDT at rejecting the background, especially QCD. In Figure 4.19 the dijet and the combined BDT outputs are shown for VBF signal, gluon fusion, the background simulation and data.

Two categories, a tight and a loose one, are defined by two selection requirements on the combined BDT. Their optimisation is done by first choosing the requirement that maximizes the $S/\sqrt{S+B}$ of the tight bin, then excluding that region and repeating the procedure for the loose bin. 69.97% of the events belonging to the tight category are VBF and 29.47% are ggH; for the loose category these two production modes make up 53.50% and 44.91% respectively of the total accepted signal. The remaining $\sim 1\%$ in each category comes from the other production modes.

4.10 ttH tag

The study of the ttH production channel is very important because it allows access to the coupling of the Higgs boson to the top quark, which, thanks to its high mass, might play a special role in the context of electroweak symmetry breaking.
Since the top quark decays almost always in a W boson and a b quark, the ttH final state is composed, in addition to the photon pair, by two b-jets along with additional jets or leptons coming from the decay of W bosons. Thus two different sets of selection criteria are used, optimised for the leptonic and hadronic W decays, corresponding to the leptonic and hadronic ttH tags.
The jets selected are as described for the VBF tag and are required to have $p_T > 25$ GeV and $|\eta| < 2.4$. In order to avoid the overlapping between the jets and one of the photons belonging to the photon pair, the jets must be separated from the photons with $\mathrm{R}(jet, \gamma) > 0.4$. Finally the Secondary Vertex algorithm is used for the b-jets identification.

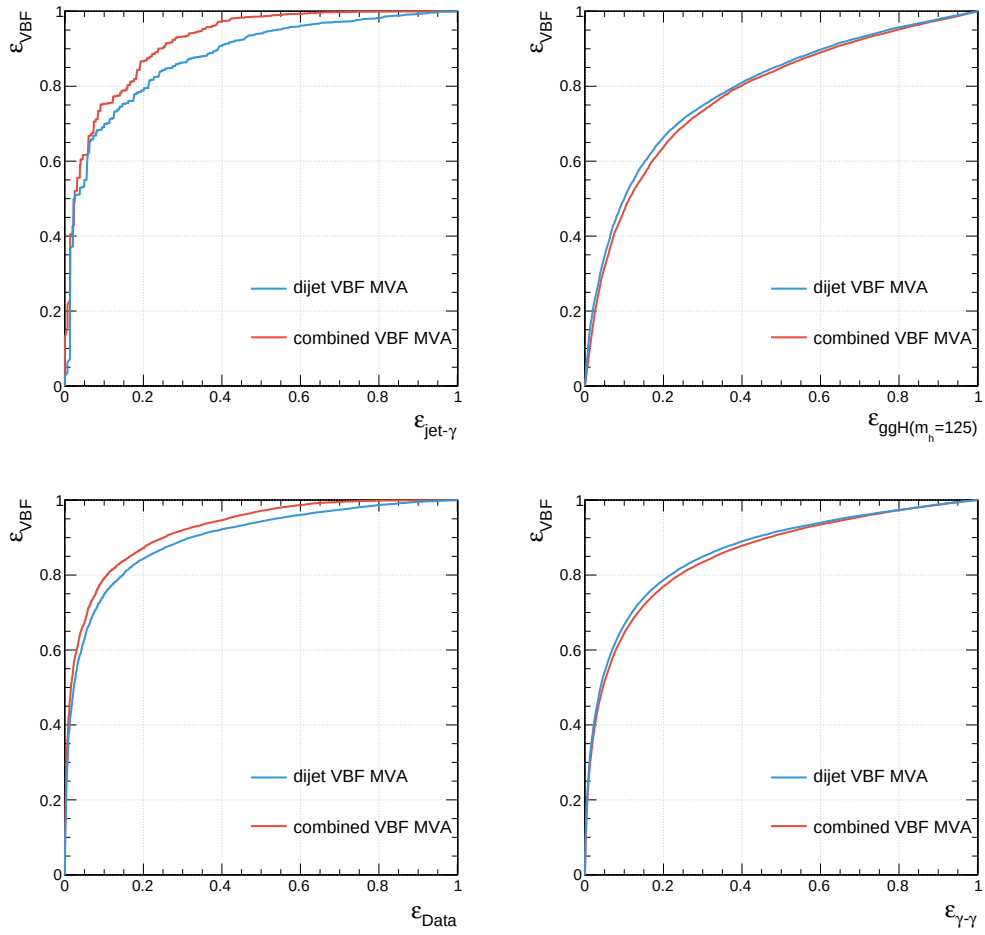


Figure 4.18: Efficiency in VBF signal simulation versus QCD efficiency (top left), versus gluon fusion efficiency (top right), versus preselected data efficiency (bottom left) and versus SM diphoton background efficiency (bottom right). All samples passed through the VBF preselection before entering the calculation. The combined BDT is better than the kinematic dijet BDT at rejecting backgrounds containing fake photons from QCD.

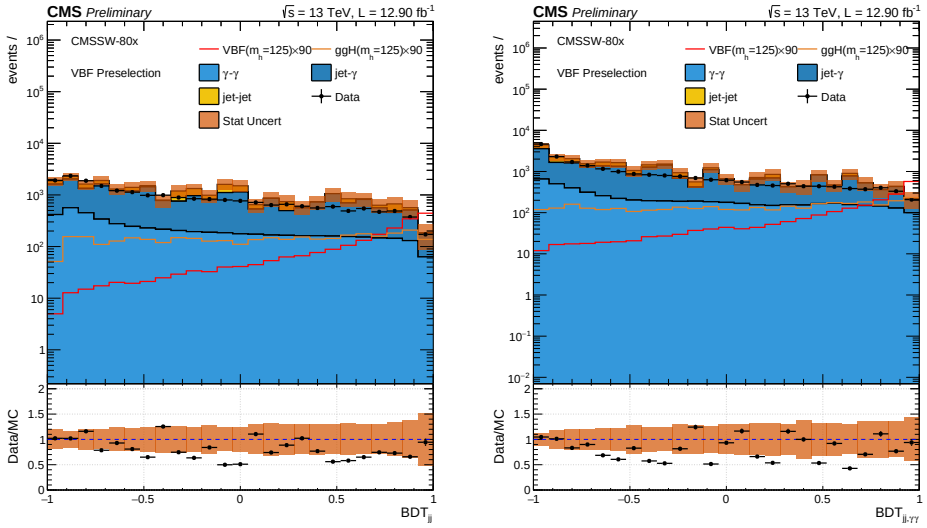


Figure 4.19: Dijet BDT (left) and combined BDT (right) output for VBF signal, gluon fusion, the different background components and data. The agreement between data and simulation is fair within the statistical uncertainty.

Since the simulation for events with two photons and multiple jets in the final state is not reliable, a control sample in data is defined to perform studies for the expected background. This control sample is built by selecting diphoton events where one of the photons is required to pass the preselection and photon ID BDT criterion (ID BDT > -0.9), while for the other photon, the photon ID BDT criterion is inverted and the preselection is not applied.

4.10.1 Leptonic tag

The leptonic tag is optimised for semi-leptonic and leptonic $t\bar{t}$ decays in ttH events, that is $t\bar{t} \to b\ell\nu_\ell \bar{b} q\bar{q}'$ and $t\bar{t} \to b\ell\nu_\ell \bar{b}\ell'\nu_{\ell'}$, where ℓ can be either a muon or an electron. The background is fairly low thanks to an high p_T isolated lepton in the final state.
The muons are requested to have $p_T > 20$ GeV, $|\eta| < 2.4$ and to pass criteria summarised in Table 4.4.
Electrons must have $p_T > 20$ GeV and $|\eta_{SC}| < 1.4442$ or $1.566 < |\eta_{SC}| < 2.5$. The electrons are also required to pass electron identification criteria summarised in Table 4.5. The electrons are vetoed if they are found to match converted photons.
The event selection for ttH leptonic category also requires additional criteria on the number of jets along with the presence of at least one isolated lepton. Events are first required to pass analysis preselection described in Section 4.4 and then:

Table 4.4: Muon selection criteria for ttH leptonic tag

Description	criterion
χ^2/n_{dof}	< 10
d_0 w.r.t. muon vertex	< 0.2 cm
d_z w.r.t. muon vertex	< 0.5 cm
Number of pixel hits	> 0
Number of tracker layers with hits	> 5
Number of muon station hits	> 0
Number of matched muon station segments	> 1
Combined relative PF isolation	< 0.25

Table 4.5: Electron selection criteria for ttH leptonic tag

Description	Value in barrel	Value in endcap
$\sigma_{i\eta i\eta}$ <	0.0103	0.0301
abs($\Delta\eta_{in}$) <	0.0105	0.00814
abs($\Delta\phi_{in}$) <	0.115	0.182
H/E <	0.104	0.0897
(1/E - 1/P) <	0.102	0.126
abs(d_0) <	0.0261	0.118
abs(d_z) <	0.41	0.822
Missing Inner Hit ≤	2	1
Relative Isolation with Effective Area <	0.0893	0.121

- leading photon $p_T > m_{\gamma\gamma}/2$,
- subleading photon $p_T > m_{\gamma\gamma}/4$,
- at least one selected lepton ℓ with $p_T > 20$ GeV,
- specific to the leptonic channel: the lepton should have $R(\ell, \gamma) > 0.4$,
- specific to the electron channel: $|m_{e,\gamma} - m_Z| > 10$ GeV, where m_Z refers to the mass of the Z boson,
- at least 2 selected jets in the event with $p_T > 25$ GeV, $|\eta| < 2.4$, $R(jet, \gamma) > 0.4$ and $R(jet, \ell) > 0.4$,
- at least one of the jets in the event has to be b-tagged with $p_T > 25$ GeV,

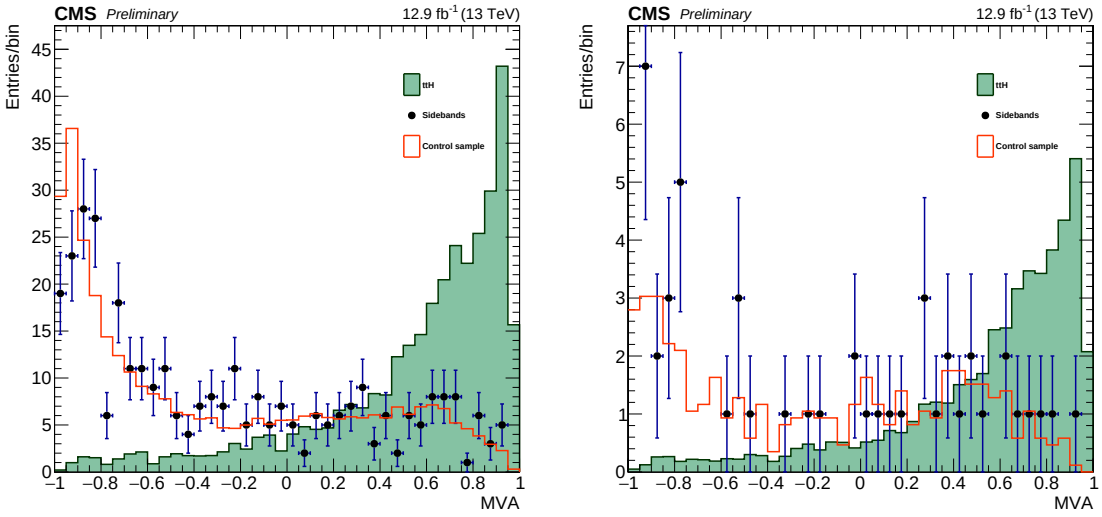


Figure 4.20: Distribution of diphoton BDT for signal, data sidebands and control samples in the hadronic (left) and leptonic (right) categories. The histograms are normalized to the area of the sidebands to compare the shape of the distributions.

- diphoton BDT $>$ -0.4, which gives a signal efficiency of 89%.

4.10.2 Hadronic tag

The hadronic tag is optimised for hadronic $t\bar{t}$ decays in ttH events, that is $t\bar{t} \to bq\bar{q}'\bar{b}q\bar{q}'$. Events passing the full analysis preselection described in Section 4.4 are required to pass the following selection requirements:

- photon pair selection with the highest sum of p_T,
- leading photon $p_T > m_{\gamma\gamma}/2$,
- subleading photon $p_T > m_{\gamma\gamma}/4$,
- no leptons (defined according to the leptonic tag),
- at least 5 jets in the event with $p_T > 25$ GeV,
- at least one of the jets in the event has to be b-tagged with $p_T > 25$ GeV,
- diphoton BDT > 0, which gives a signal efficiency of 90% and a background efficiency of 35%. These values were checked on control samples (see Figure 4.20).

4.11 Statistical analysis

As already mentioned in previous sections, events are classified into eight exclusive categories. Two categories allow to classify events produced in association with two jets, likely to originate from vector boson fusion (VBF), and are labelled "VBF Tag 0" and "VBF Tag 1", with the former being the most sensitive. Two further categories, "TTH Leptonic Tag" and "TTH Hadronic Tag", are used to classify events likely to have been produced in association with top quarks, decaying either leptonically or hadronically. Remaining events are categorised into four inclusive categories, labelled "Untagged 0, Untagged 1, Untagged 2, Untagged 3", ordered from the most to the least sensitive. Events which do not enter any of the above categories are discarded.

The overall strategy to interpret the data and to extract the Higgs signal is to perform maximum likelihood fits to the diphoton mass distribution simultaneously across all of the event categories. This requires models for the probability density functions (PDFs) of both the expected signal and the background in each category.

Simulated signal samples corresponding to each of the allowed Standard Model Higgs production processes (VBF, ttH, ggH, WH and ZH) are used to produce the signal model. Each sample is analyzed using the procedure described in the previous sections and is divided into the categories described above. This process is repeated for seven Higgs boson mass scenarios, 120, 123, 124, 125, 126, 127 and 130 GeV, referred to as the "mass points". These simulated signal samples are used to produce a parametric signal model, described in Section 4.12.

The background model is extracted directly from data, using the discrete profiling method described in Section 4.13.

4.12 Signal model

In each category MC simulation is used to extract signal efficiency $\times$ acceptance and signal mass PDF. A parametric signal model is built continuously for any value of the Higgs mass between 120 and 130 GeV. The signal model is derived from the signal simulation using an analytic function, the parameters of which are determined by fitting the simulated events in each category and at each of the seven Higgs mass points. Finally, the full signal model is defined by a linear interpolation of each fit parameter between the fitted mass values. For each category, the simulated events for each of the four production mechanisms are used as input to the fits. The analytic functions for each production mechanism are added together according to their relative cross-sections in the Standard Model to give the final function in each category.

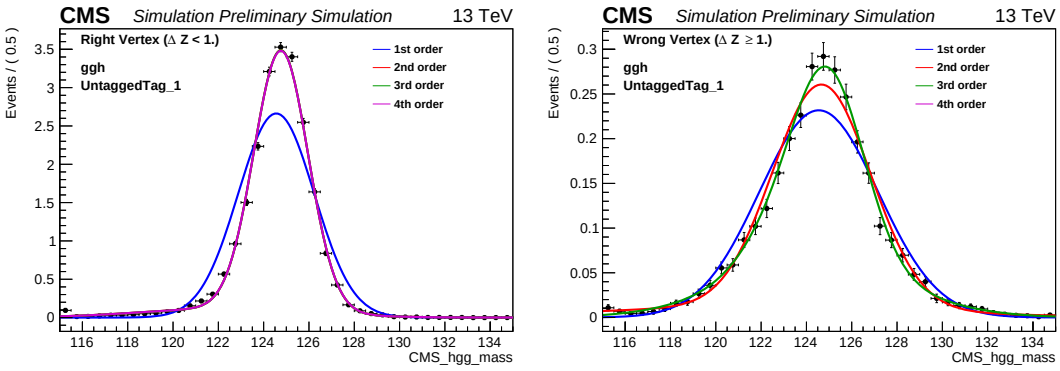


Figure 4.21: Fit results for the signal shape of simulated ggH events in the Untagged 1 category. The plots show the case when the right vertex (left) or the wrong vertex (right) is chosen. The black points are weighted events from simulated data and the lines represent the fits when 1 (blue), 2 (red), 3 (green) or 4 (pink) Gaussians are included in the fit. The mean, width and relative size of the Gaussians are allowed to vary independently. Similar plots are used to determine the number of Gaussians to use to fit each process and category.

Furthermore it is important to take into account that the $m_{\gamma\gamma}$ distribution changes considerably depending on whether the vertex associated with the diphoton candidate was correctly identified or not. In effect, when the correct vertex is identified the shape of the $m_{\gamma\gamma}$ distribution is dominated by the detector resolution and reconstruction, while when the incorrect vertex is chosen the signal shape is smeared significantly by the variation in the z position of the selected primary vertex with respect to the true Higgs production point. The cases where the right vertex (RV) and wrong vertex (WV) were chosen are fitted separately when constructing the signal model.

For each production mode, category and RV/WV scenario, the $m_{\gamma\gamma}$ distributions are fitted using a sum of at most four Gaussians. The width, mean and relative size of each Gaussian are left free in the fits to the simulation. A representative set of fits for events with correct and incorrect primary vertex selection, in one particular category for gluon-fusion production are shown in Figure 4.21.

The combined shape in each category for correct and incorrect vertex selection is constructed by adding the shapes for the two sub-components together, according to the correct vertex selection efficiency determined from each simulated sample. This efficiency is treated as another model parameter for the purposes of interpolation between mass points, although in practice the vertex selection efficiency does not vary much as a function of Higgs mass.

In order to facilitate the interpretation of the signal model in terms of a Standard Model Higgs production cross-section, and in order to facilitate the use of the signal model simultaneously across all of the event categories, the signal yield is parametrized in terms of a per class acceptance times efficiency, computed from each Monte Carlo sample. Figure

4.22 shows the efficiency × acceptance of the signal model as a function of m_H for all categories combined.

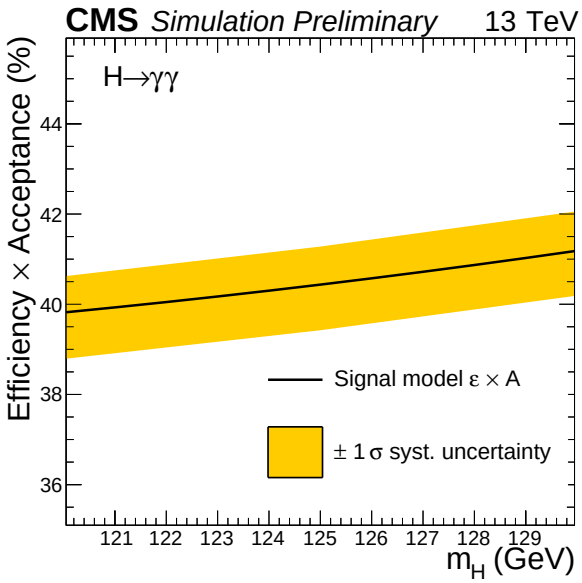


Figure 4.22: The efficiency × acceptance of the signal model as a function of m_H for all categories combined. The yellow bands indicate the effect of the systematic uncertainties.

The determination of the full set of signal model parameters at each Monte Carlo mass point is used to construct a signal model continuous in Higgs mass by performing a linear interpolation of each fit parameter, which gives rise to a smooth evolution of the signal shape. The final model for the 125 GeV Higgs mass scenario with all production processes summed as described previously and for the weighted sum of all categories is represented in Figure 4.23.
Systematic uncertainties corresponding to the smearing and scale of the individual photon energies, the fraction of events where the RV was correctly identified, the material corrections and ECAL crystal light yields are incorporated into the signal model as additional parameters to be treated as nuisances.

4.13 Background model

Since the level of background after selection is large, and a comprehensive Monte Carlo description is not available at NLO or beyond, the background is instead modelled in an entirely data driven manner. The model used to describe the background is produced by fitting various analytic functions to the $m_{\gamma\gamma}$ distribution, in the 100 to 180 GeV range. The data are then classified into the eight categories described previously.

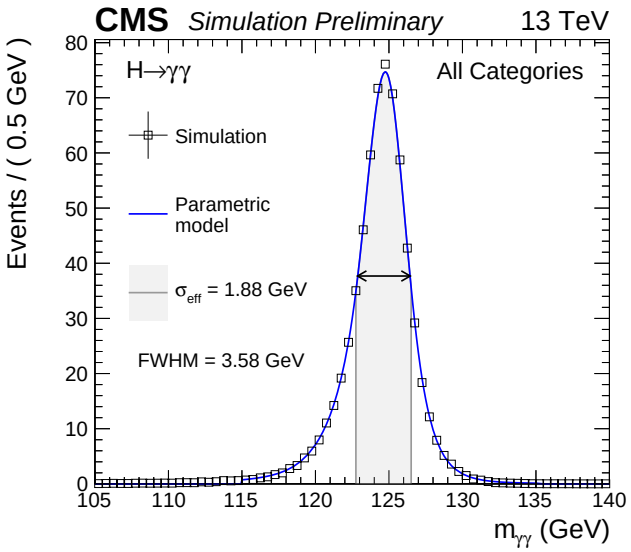


Figure 4.23: Full parametrized signal shape integrated over all event classes for the $m_H = 125$ GeV scenario at $\sqrt{s} = 13$ TeV. The black points represent weighted simulation events and the blue line is the corresponding model. Also shown are the σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution), FWHM and the corresponding interval.

The method used to do that is the discrete profiling or "envelope" method [64]. The discrete profiling method is designed as a way to determine the systematic uncertainty associated with choosing a particular analytic function to fit to the background $m_{\gamma\gamma}$ distribution. The method treats the choice of the background function as a discrete parameter in the likelihood fit to the data.
For each event category, four families of analytic functions are considered:

- Sums of exponentials:

$$f_N(x) = \sum_{i=0}^{N} p_{2i} e^{p_{2i+1} x},$$

- Sums of polynomials (in the Bernstein basis):

$$f_N(x) = \sum_{i=0}^{N} p_i b_{(i,N)}, \text{ where } b_{(i,N)} := \binom{N}{i} x^i (1-x)^{N-i},$$

- Laurent series:

$$f_N(x) = \sum_{i=0}^{N} p_i x^{-4+\sum_{j=1}^{i}(-1)^j (j-1)},$$

- Sums of power-law functions:

$$f_N(x) = \sum_{i=0}^{N} p_{2i} x^{-p_{2i+1}},$$

where for all k, the p_k are a set of floating parameters in the fits.
These functions cover a range of possible shapes for the background. For each event category, each functional form is fitted to the data, with the number of degrees of freedom being determined by an F-test [65], where terms are only included if they significantly improve the χ^2 probability of the fit.
When fitting these functions to the background $m_{\gamma\gamma}$ distribution, the value of twice the negative logarithm of the likelihood (2NLL) is minimized. A penalty is added to 2NLL to take into account the number of floating parameters in each candidate function. When making a measurement of a given parameter of interest, the discrete profiling method determines the envelope of the lowest values of 2NLL profiled as a function of the parameter of interest. The envelope obtained through this method will yield a broader curve than the 2NLL curve obtained from a single function choice. Figure 4.24 shows an example of the background fit with the different functions and the best fit background parametrization (assuming no signal), for two event categories. The uncertainties on the background shapes associated with the statistical uncertainties of the fits are shown by the 1 σ and 2 σ bands. These bands are obtained using an extended likelihood fit parametrised in terms of the background yield in a 1 GeV window, which is the size of the bins in the showed histogram. The signal model assuming $m_H = 125$ GeV is also overlaid on the background distribution.

4.14 Systematic uncertainties

Several types of systematic uncertainty are considered in this analysis. A summary of all the systematic uncertainties along with their treatment is presented in this section.
The systematic uncertainties related to the signal shape are treated differently depending on how they affect the $m_{\gamma\gamma}$ distribution. Those which modify the shape of the $m_{\gamma\gamma}$ distribution are generally built directly into the signal model as parametric nuisance parameters. If the shape of the $m_{\gamma\gamma}$ distribution is instead unaffected, the systematic variations are treated as log-normal uncertainties on the efficiency. Finally, for cases where the systematic has an effect on the input to one of the classification BDTs, the variation takes the form of a correlated log-normal uncertainty on the category yield, that is it is considered as a category migration systematic.

The systematic uncertainties considered in this analysis are listed below:

- Theory systematics:

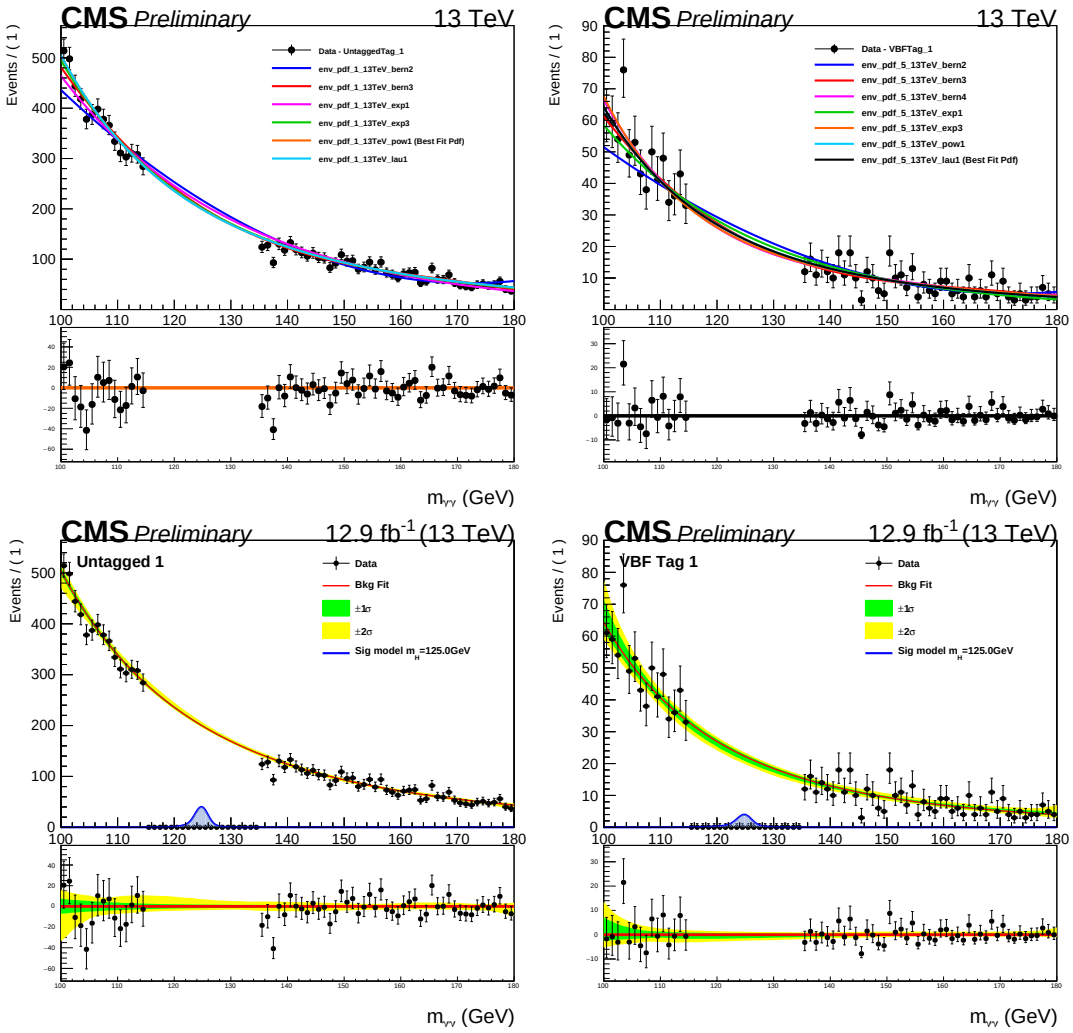


Figure 4.24: On the top, the set of functions chosen to fit the background using the discrete profiling method is shown, for the Untagged 1 (left) and VBFTag 1 (right) categories considered in this analysis. Four families of functions are considered. An F-test [65] is used to select representative functions from each of these four families in order to proceed with the discrete profiling. On the bottom, the best fit background parametrization plotted alongside the data is shown in the same two categories. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

- *parton density functions (PDF) uncertainties*: the uncertainty coming from the choice of PDF is assessed by estimating the relative yield variation in each process and category, after re-weighting the events of the simulated signal sample. The re-weighting is done according to PDF4LHC15 combined PDF set and NNPDF30 [66] using the MC2hessian procedure [67]. The category migrations are found to be less than 2%. The overall normalization variation is taken from [12].
- α_s *uncertainty*: the uncertainty on the value of the strong force coupling constant α_s is evaluated following the PDF4LHC prescription [68]. The overall variation in the relative event yield due to the α_s uncertainty is found to be at most 3.7%.
- *Underlying event and parton shower uncertainty*, corresponding to the choice and tuning of the generator: this systematic uncertainty is treated as an event migration systematics as it will mainly affect the jets in the analysis. The possibility that an event could move from one VBF Tag to another or from either VBF Tag to an inclusive category is assigned a systematic uncertainty of 7% and 9% respectively.
- *QCD scale uncertainty*, related to varying the renormalization and factorization scales: the uncertainties are taken as variations on the QCD parameters μ_R and μ_F. The overall effect on the normalisation is taken from [12] and effect on the relative category yield is found to be about 5-10%.
- *Uncertainty on the H→ $\gamma\gamma$ branching ratio*: it is estimated to be about 2% [12].
- *Gluon fusion contamination in VBF and ttH tagged categories*: the theoretical predictions for gluon fusion are not reliable in a regime where the Higgs boson is produced in association with a large number of jets. The uncertainty on the yield of gluon fusion events in the VBF tagged classes is estimated using the Stewart-Tackmann procedure, according to the LHC Higgs Cross Section Working Group [12] recommendations. The overall normalization is found to vary by 39% while migrations between the two VBF categories are of about 10%. The systematic uncertainty on the gluon fusion contamination in the $t\bar{t}H$ tagged classes is estimated taking into account the following contributions:
  - the uncertainty due to the limited size of the simulated sample is computed to be 10%.
  - the uncertainty coming from the parton shower modelling is estimated as the observed difference in the jet multiplicity between MADGRAPH5_aMC@NLO predictions and data in $t\bar{t}$ +jets events, with fully leptonic $t\bar{t}$ decays. This uncertainty is found to be of about 45% in the bins with the largest discrepancy ($N_{\text{jets}} \geq 5$) [69].

  - the uncertainty on the gluon splitting modelling is estimated by scaling the fraction of events from gluon fusion with real b-jets by the observed difference between data and simulation in the ratio $\sigma(t\bar{t}b\bar{b})/\sigma(t\bar{t}jj)$ at 13 TeV. This uncertainty implies a variation of about 18% in the yield of gluon fusion events.

- *Integrated luminosity*: the uncertainty on the integrated luminosity is estimated from data, and it is found to be of 6.2% on the signal yield.

- *Trigger efficiency*: the trigger efficiency is measured from $Z \to ee$ events using the tag-and-probe technique; the size of the effect on the event yields is of less than 0.1%.

- *Photon preselection*: the systematic uncertainty on the photon preselection is taken as the uncertainty on the ratio between the efficiency measured in data and in simulation; it ranges from 0.1% to 2.3% according to the photon category and results in an event yield variation up to 4% depending on the event category.

- *Vertex selection efficiency*: the systematic uncertainty in the vertex finding efficiency is taken from the uncertainty in the measurement of the corresponding data/simulation scale factor obtained using $Z \to \mu\mu$ events. It is handled as an additional nuisance parameter built into the signal model which allows the fraction of events in the right vertex/wrong vertex scenario to change. The size of the uncertainty of the vertex selection efficiency is 1.5%.

- *Energy scale and resolution*: scale and resolutions are studied with electrons from $Z \to ee$ events and then applied to photons. The main source of systematic uncertainty is the difference between electrons and photons in the interaction with material upstream the ECAL . Uncertainties are assessed by changing the R_9 distribution, the regression training (using electrons instead of photons) and the electron selection used to derive the corrections. The uncertainty on the additional energy smearing is computed propagating the uncertainties on the various $|\eta|$ and R_9 bins to the Higgs boson signal phase space. In both cases dedicated nuisance parameters are included as additional systematic terms in the signal model and result in less than 0.5% depending on the photon category.

- *Non-uniformity of the light collection*: the uncertainty on the response of the ECAL crystals has been slightly amplified with respect to Run 1 to account for the effect of larger transparency loss of the ECAL crystals. The size of the effect on the photon energy scale for 2016 data is estimated to be 0.07%.

- *Non-linearity*: the uncertainty associated with the non-linearity of the photon energy between MC simulation and data is estimated using Z boson decays to electron-

positron pairs. The effect is found to be 0.1% on the photon energy in all categories, except Untagged 0 in which it is 0.2%.

- *Geant4*: a small uncertainty is added to account for imperfect electromagnetic shower simulation in GEANT 4 [53]. A simulation made with an improved shower description, changes the energy scale for both electrons and photons. Although mostly consistent with zero, the variation is interpreted as a limitation on our knowledge of the correct simulation of the showers, leading to a further uncertainty of 0.05% on the photon energy.

- *Modeling of the material budget*: the uncertainty on material budget between the interaction point and the vertex, which affects the behaviour of electron and photon showers, is estimated with special simulation samples where the material budget is uniformly varied by ±5%. Its effect on the energy scale is at most 0.17%.

- *Shower shape corrections*: the uncertainty deriving from the imperfect shower shape modelling in simulation is estimated using simulated H $\to \gamma\gamma$ and $Z \to ee$ samples with and without shower shape corrections. The effect on the photon energy scale is found to be at most 0.064%.

- *Photon identification BDT score*: in order to cover the observed discrepancies between data and simulation, the uncertainty on the signal yields in the different categories of the analysis is estimated conservatively by propagating the uncertainty described in Section 3.3.3.

- *Per photon energy resolution estimate*: it is parametrized as a rescaling of the resolution estimate by ±5% about its nominal value.

- *Jet energy scale and smearing corrections*: this uncertainty is implemented as migration within VBF categories, within $t\bar{t}H$ categories and from tagged to untagged categories. Jet energy scale corrections (JEC) account for a 4-15% migration within VBF categories and 4-15% from VBF to untagged categories. The migration due to energy scale in $t\bar{t}H$ categories is about 5%. The jet energy resolution has an impact on the event migration smaller than 2%.

- *b-tagging efficiency*: the systematic uncertainty on the b-tagging efficiency is evaluated by varying the ratio between the measured b-tagging efficiency in data and simulation within their uncertainty [70]. The resulting uncertainty on the signal yield is found to be of about 2% in the lepton-tagged category and 5% in the hadronic-tagged category.

- *Lepton identification efficiency*: the uncertainty is computed, for both electrons and muons, by varying the ratio of the efficiency measured in data and simulation by its

uncertainty. The resulting difference in the selection efficiency for the $t\bar{t}H$ lepton-tagged category is less than 1%.

- *Background modelling*: the choice of background parametrization is handled using the discrete profiling method. This automatically leads to an uncertainty on the choice of background function as described in Section 4.13.

4.15 Results

Results are extracted performing a simultaneous binned maximum-likelihood fit to the diphoton invariant mass distributions in all the event classes over the range $100 < m_{\gamma\gamma} < 180$ GeV. The signal PDF is that described in Section 4.12, where the shape and normalization for each production mode are separately tracked and defined. The background is evaluated by fitting the $m_{\gamma\gamma}$ distribution in data, without reference to the MC simulation. Thus the likelihood to be evaluated in a signal-plus-background fit is

$$\mathcal{L} = \mathcal{L}(data|s(p, m_{\gamma\gamma}) + f(m_{\gamma\gamma})),$$

where p comprises those parameters of the signal, such as m_H or the signal strength, that are allowed to vary in the fit, $s(p, m_{\gamma\gamma})$ is the parametric signal model, and $f(m_{\gamma\gamma})$ the background fit function.
The test statistic, used to determine how signal-like or background-like the data are, is based on the profile likelihood ratio. The systematic uncertainties are integrated in the analysis via nuisance parameters and treated according to the frequentist paradigm. A description of the general methodology used in this analysis can be found in References [71, 72].

Figures 4.25 and 4.26 show the data, the background-only and the signal plus background model fit for each category used in this analysis. The 1 σ (green) and 2 σ (yellow) uncertainty bands shown for the background component of the fit include the uncertainty in the fitted parameters.

Figure 4.27 shows the diphoton mass spectrum with each event weighted proportionally to S/(S+B), where S and B are the numbers of expected signal and background events respectively.
Table 4.6 shows the expected number of signal events for each category. The total number is broken down by percentage contribution of each production mode. Also listed are the σ_{eff} and σ_{HM}. The former represents the smallest interval in the distribution which contains 68.3% of the entries. The latter represents the full width at half maximum divided by 2.355. Also listed in the table is the expected number of background events per GeV in the corresponding σ_{eff} window around 125 GeV, which is taken from the best-fit candidate

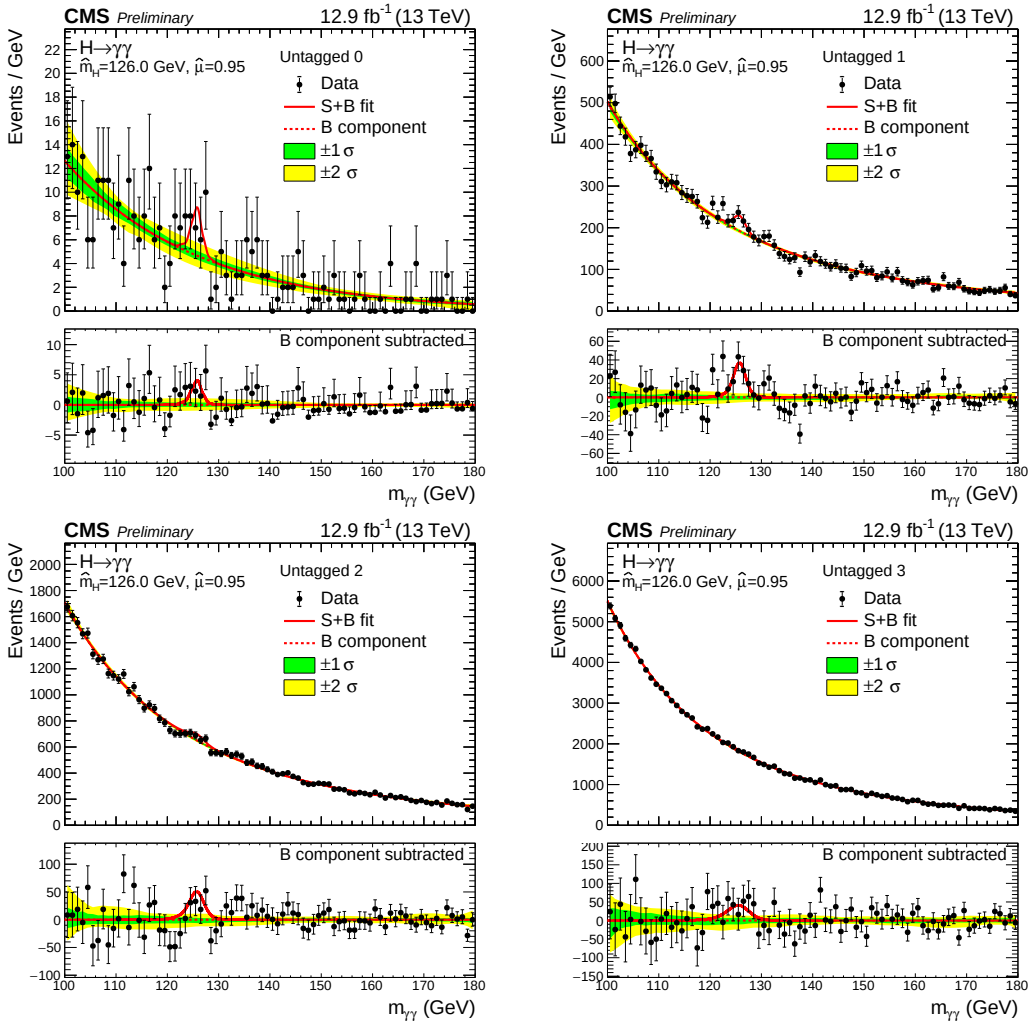


Figure 4.25: Data points (black), background-only fit (dashed red) and S+B model fit (red) in the four untagged categories are shown. The 1 σ (green) and 2 σ bands (yellow) include the uncertainties of the fit. The bottom plot shows the residuals after background subtraction.

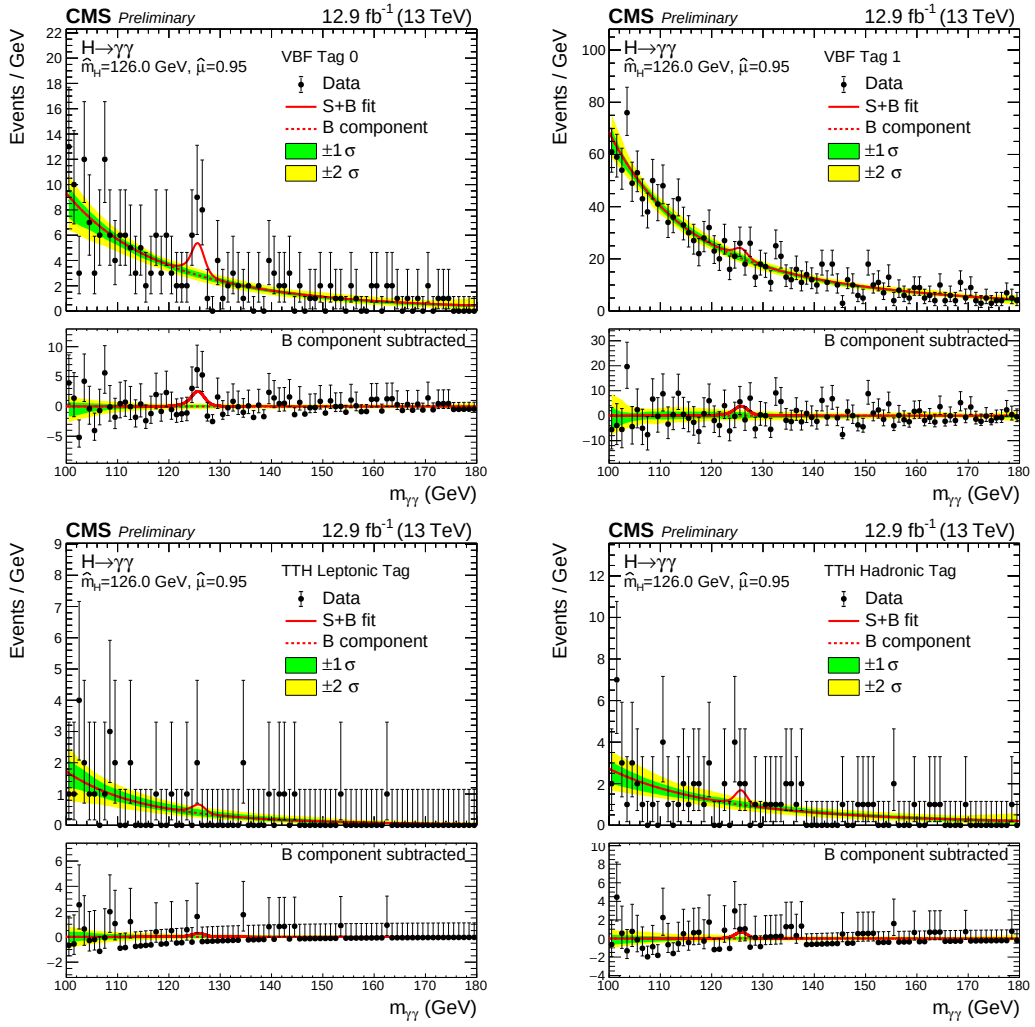


Figure 4.26: Data points (black), background-only fit (dashed red) and S+B model fit (red) in VBF and ttH categories are shown. The 1 σ (green) and 2 σ bands (yellow) include the uncertainties of the fit. The bottom plot shows the residuals after background subtraction.

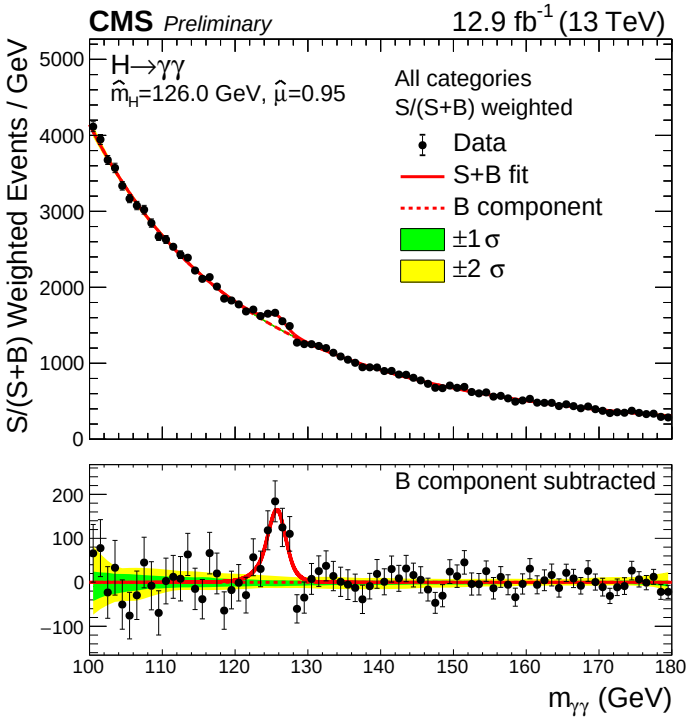


Figure 4.27: The diphoton mass spectrum where the categories are summed weighted by their sensitivity. Data points (black), background-only fit (dashed red) and S+B model fit (red) are shown. The 1 σ (green) and 2 σ bands (yellow) include the uncertainties of the fit. The bottom plot shows the residuals after background subtraction.

Event Categories	SM 125GeV Higgs boson expected signal								Bkg
	Total	ggh	vbf	wh	zh	tth	σ_{eff}	σ_{HM}	(GeV^{-1})
Untagged 0	11.92	79.10 %	7.60 %	7.11 %	3.59 %	2.60 %	1.18	1.03	4.98
Untagged 1	128.78	85.98 %	7.38 %	3.70 %	2.12 %	0.82 %	1.35	1.20	199.14
Untagged 2	220.12	91.11 %	5.01 %	2.18 %	1.23 %	0.47 %	1.70	1.47	670.44
Untagged 3	258.50	92.35 %	4.23 %	1.89 %	1.06 %	0.47 %	2.44	2.17	1861.23
VBF Tag 0	9.35	29.47 %	69.97 %	0.29 %	0.07 %	0.20 %	1.60	1.33	3.09
VBF Tag 1	15.55	44.91 %	53.50 %	0.86 %	0.38 %	0.35 %	1.71	1.40	22.22
TTH Hadronic Tag	2.42	16.78 %	1.28 %	2.52 %	2.39 %	77.02 %	1.39	1.21	1.12
TTH Leptonic Tag	1.12	1.09 %	0.08 %	2.43 %	1.06 %	95.34 %	1.61	1.35	0.42
Total	647.77	87.93 %	7.29 %	2.40 %	1.35 %	1.03 %	1.88	1.52	2762.65

Table 4.6: The expected number of signal events per category and the percentage breakdown per production mode. σ_{eff} and σ_{HM} are also provided as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events per GeV around 125 GeV is also listed.

background parametrization.

In Figure 4.28 the expected and observed significances for the observation of a standard model Higgs boson are shown as a function of m_H.

The local significance for the observation of a standard model Higgs boson at $m_H = 125.09$ GeV is 5.6 σ, while 6.2 σ was expected, and the maximum significance of 6.1 σ is observed at $m_H = 126.0$ GeV. A likelihood scan of the signal strength ($\mu = \sigma/\sigma_{SM}$) profiling all the nuisances is performed and it is shown in Figure 4.29. In this scan, the Higgs boson mass was profiled in the same way as other nuisances in the fit. The best-fit signal strength measured for all categories combined using this method is $\hat{\mu} = 0.95 \pm 0.20 = 0.95 \pm 0.17$ (stat.)$^{+0.10}_{-0.07}$ (syst.)$^{+0.08}_{-0.05}$ (theo.). If the Higgs boson mass is fixed to the Run 1 best-fit value $m_H = 125.09$ GeV, then the resulting signal strength is measured to be $0.91 \pm 0.20 = 0.91 \pm 0.17$ (stat.)$^{+0.09}_{-0.07}$ (syst.)$^{+0.08}_{-0.05}$ (theo.). Figure 4.30 shows the signal strength separately for each of the categories used in the analysis (top) and the signal strength split by process (bottom). Since this analysis does not have any categories which specifically target the VH production mode, μ_{VH} is set to 1.
In addition a two-dimensional likelihood scan of the signal strength $\mu_{ggH,t\bar{t}H}$ for fermionic production modes (ggH and $t\bar{t}H$) and $\mu_{VBF,VH}$ for vector boson production modes (VBF, ZH, WH), with the value of the parameter m_H profiled in the fit, is performed. Figure 4.31 shows the 68% and 95% confidence level contours, the best-fit values are $\mu_{ggH,t\bar{t}H} = 0.80^{+0.14}_{-0.18}$ and $\mu_{VBF,VH} = 1.59^{+0.73}_{-0.45}$.

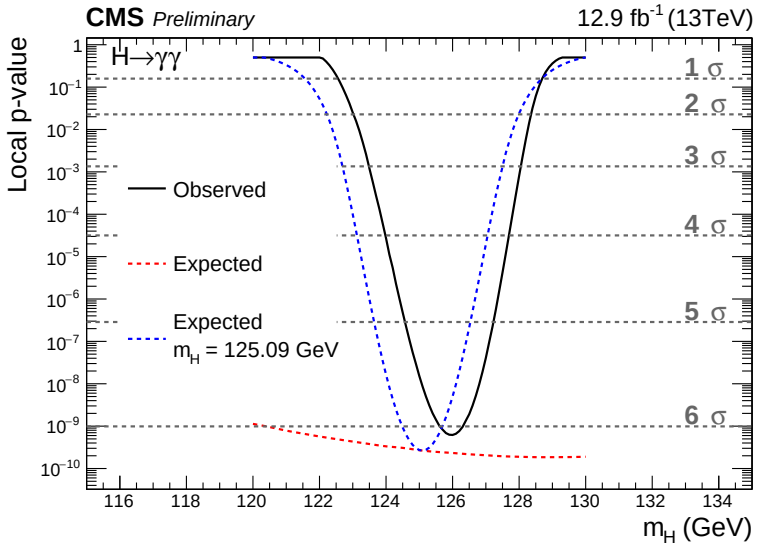


Figure 4.28: The observed p-value (black) is compared to the SM expectation across the fit range 120-130 GeV, where the SM Higgs boson is assumed to have a mass $m_H = 125.09$ GeV (blue). The red line shows the maximum significance for each mass hypothesis in the range 120 GeV $< m_H <$ 130 GeV.

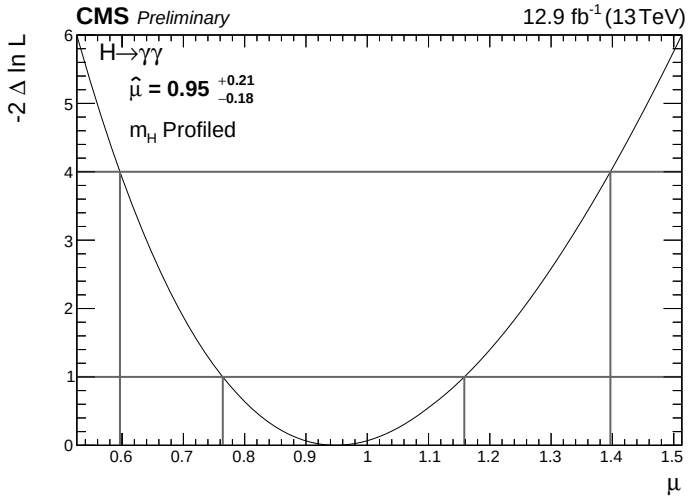


Figure 4.29: The likelihood scan for the signal strength modifier $\mu = \sigma/\sigma_{SM}$, where the value of the Standard Model Higgs boson mass is profiled in the fit.

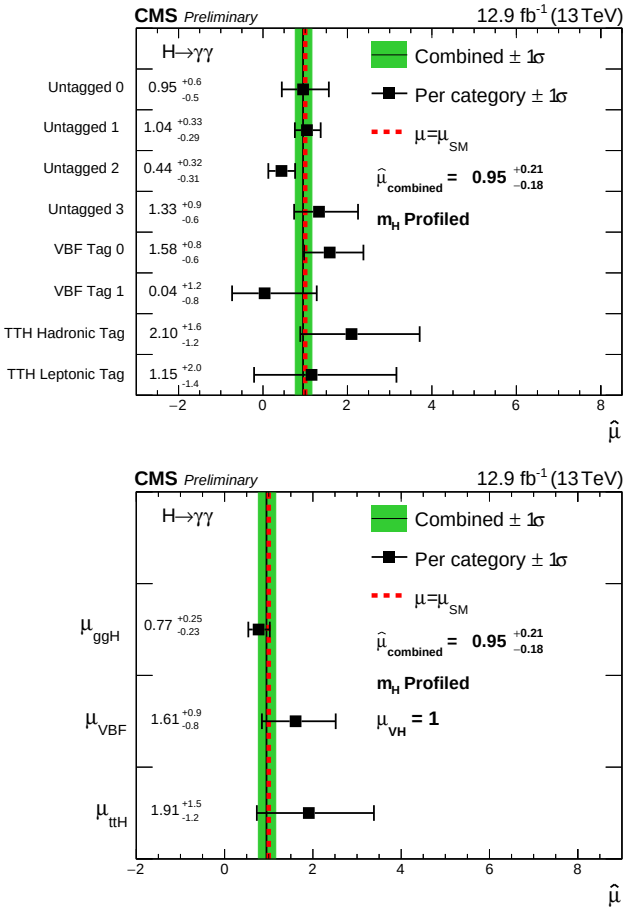


Figure 4.30: Signal strength modifiers (black points) measured in each category (top) and for each process (bottom) for profiled m_H, compared to the overall signal strength (green band) and to the SM expectation (dashed red line).

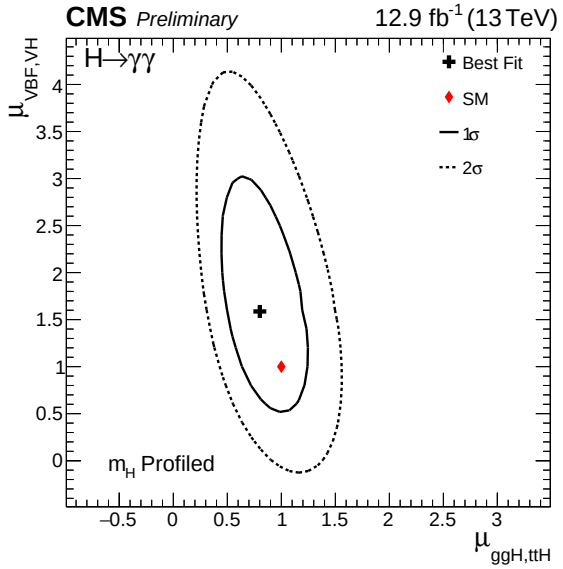


Figure 4.31: The 2-D best-fit (black cross) of the signal strength modifiers ($\mu = \sigma/\sigma_{SM}$) for fermionic (ggH, $t\bar{t}H$) and bosonic (VBF, ZH, WH) production modes compared to the SM expectations (red diamond). The Higgs boson mass is profiled in the fit. The solid (dashed) line represents the 1 σ (2 σ) confidence region.

4.16 Summary

In this chapter a general overview of the H $\to \gamma\gamma$ analysis using the first Run 2 data is presented. A particular focus is dedicated to the important role that Monte Carlo simulations have in the signal model building, selection optimisation and training of different multivariate analyses. This analysis brought to a rediscovery of the Higgs boson, with a significance greater than 5 σ. For the future, the VH tag and detailed studies to define the Higgs mass with its uncertainty should be included. Furthermore, studies of the Higgs spin-parity in the diphoton channel at 13 TeV could appear in the near future.

Chapter 5

Search for anomalous couplings of the Higgs boson to electroweak vector bosons in VBF production with H $\rightarrow \gamma\gamma$

5.1 Introduction

In this section, we will describe a preliminary study of the possibility to constrain HVV couplings in the VBF production, using the Higgs boson decay to 2 photons. This is done with the 8 TeV Run 1 CMS dataset (19.7 fb^{-1}), following closely the analysis designed to discover the Higgs boson in the 2 photons decay channel [13]. In this dataset, the sensibility to the VBF production per-se is small and therefore we do not expect to have strong constraints. Nevertheless this is an alternate and complementary approach to the one usually employed which is to study this coupling in the decay of the Higgs boson $H \rightarrow ZZ^*$ or $H \rightarrow WW^*$. We will see that despite the low event yield in the VBF channel, this new approach has an interesting sensitivity and should be pursued with larger dataset and a larger number of VBF events.

5.1.1 Theory

The observation of a new boson consistent with the Standard Model (SM) Higgs boson has been reported by ATLAS and CMS Collaborations in 2012 [8, 9]. Even if this discovery is an important step to complete the SM, there are still many questions to be answered and new physics may exist beyond the SM. There are two ways of performing searches for beyond the SM physics: the first one through direct searches for new physics, the second one by detailed studies of the properties of known particles. The precise study of the Higgs

boson properties allows to better understand its connection to the rest of the SM particles and can give a hint of the presence of some deviations to the SM.
In spite of a low branching ratio, the $H \to \gamma\gamma$ channel is very interesting because of the clear experimental signature thanks to an excellent diphoton mass resolution. This channel is important not only for the Higgs boson observation, but also to study the properties of this particle. In particular, as shown in Figure 5.1, if we consider the vector boson fusion production (VBF), this channel contains a coupling to the vector bosons HVV (where V is either a W or a Z boson) which can be studied.

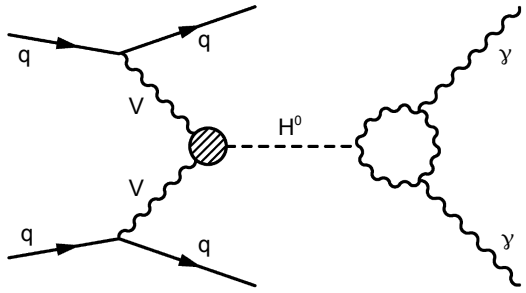


Figure 5.1: Higgs boson produced through the vector boson fusion and decaying to two photons.

Previous studies of anomalous couplings of the Higgs boson to vector bosons [73, 74, 75, 13, 76, 77], which aimed to measure the spin and parity of this new particle, have concluded that the Higgs is likely to be spin-0 with SM-like couplings. Thanks to these studies we know that the Higgs couplings are SM-like (within $\simeq 30$ % of the SM predictions), but the precision needs to be improved.

In this analysis we use a model-independent approach to determine anomalous couplings; we define a generic amplitude for a spin-zero boson coupling to vector bosons and we scan the coupling parameter space in order to determine the most likely value. Following the notation of previous CMS results [73], it is possible to write the general scattering amplitude that describes interactions of a spin-zero boson with gauge bosons, such as ZZ, WW, Zγ, $\gamma\gamma$ or gg, as

$$A(\mathrm{HVV}) \sim \left[a_1^{\mathrm{VV}} + \frac{\kappa_1^{\mathrm{VV}} q_{\mathrm{V1}}^2 + \kappa_2^{\mathrm{VV}} q_{\mathrm{V2}}^2}{\left(\Lambda_1^{\mathrm{VV}}\right)^2} \right] m_{\mathrm{V1}}^2 \epsilon_{\mathrm{V1}}^* \epsilon_{\mathrm{V2}}^* + a_2^{\mathrm{VV}} f_{\mu\nu}^{*(1)} f^{*(2),\mu\nu} + a_3^{\mathrm{VV}} f_{\mu\nu}^{*(1)} \tilde{f}^{*(2),\mu\nu} \quad (5.1)$$

where $f^{(i),\mu\nu} = \epsilon_i^\mu q_i^\nu - \epsilon_i^\nu q_i^\mu$ is the field strength tensor of a gauge boson with momentum q_i and polarization vector ϵ_i, while $\tilde{f}^{(i),\mu\nu} = \frac{1}{2}\, \epsilon^{\mu\nu\alpha\beta} f_{\alpha\beta}$ is the conjugate field strength tensor.

m_{V} is the mass of the vector boson, and Λ_1 is the scale of new physics which is a free parameter of the model. The tree-level SM-like contribution corresponds to $a_1^{\mathrm{ZZ}} \neq 0$ and $a_1^{\mathrm{WW}} \neq 0$, while there is no tree-level coupling to massless gauge bosons, that is $a_1^{\mathrm{VV}} = 0$ for Zγ, $\gamma\gamma$ and gg. In the SM $a_1^{\mathrm{VV}} = 1$ for ZZ and WW, and $a_2 = a_3 = 0$.
The pseudoscalar interaction (CP-odd state) corresponds to the a_3^{VV} terms, while the other terms describe the parity-conserving interaction of a scalar (CP-even state). The a_3^{VV} terms appear in the SM only at a three-loop level and are extremely small. The a_2^{VV} and Λ_1^{VV} terms appear in loop-induced processes and give small contribution $O(10^{-3} - 10^{-2})$ due to radiative corrections.

Since a_2 and a_3 are expected to be small, they can be studied separately. In this analysis we consider only the effects of the pseudoscalar coupling a_3, assuming $a_2 = 0$.
To scan the allowed hypotheses, the physical effects of such anomalous couplings are parameterized as effective cross-sections and phases. Following again the convention of previous CMS studies it is possible to define:

$$f_{a3} = \frac{|a_3|^2\sigma_3}{|a_1|^2\sigma_1 + |a_3|^2\sigma_3}, \qquad \phi_{a3} = \arg\left(\frac{a_3}{a_1}\right), \tag{5.2}$$

where σ_i is the cross section of the process corresponding to $a_i = 1$, $a_{j\neq i} = 0$.
$f_{a3} = 0$ corresponds to a pure scalar hypothesis, while $f_{a3} = 1$ corresponds to a pure pseudoscalar. In the absence of additional anomalous Higgs couplings, the signal strength parameter ($\mu = \sigma\mathcal{B}/\sigma\mathcal{B}|_{SM}$) is given by

$$\mu = \frac{|a_1|^2\sigma_1 + |a_3|^2\sigma_3}{\sigma_1}. \tag{5.3}$$

5.1.2 Analysis strategy

For this analysis we follow as closely as possible the CMS 8 TeV SM H $\to \gamma\gamma$ search in the VBF production mode [13]. We use the same techniques to ensure accuracy of the results; the photon energy regression, background modeling and the objects selected are unchanged, but the event selection, though similar, has been adapted: VBF events are selected with looser cuts with respect to Reference [13]. The event selection will be presented hereafter. An important aspect to take into account is that the events satisfying VBF selection criteria are partly real VBF events, but also coming from ggH production with two additional emitted partons. As shown in Figure 5.2, the latter event topology has the same final state as the real VBF one, with two forward jets and two isolated photons.
We will see in the following that the ggH processes are treated as a background, since in this event topology it is difficult to discriminate between 0^+ and 0^- events. Therefore the analysis has to cope with the following processes: VBF scalar (0^+), VBF pseudoscalar (0^-) and their interference, the ggH and the continuum background.

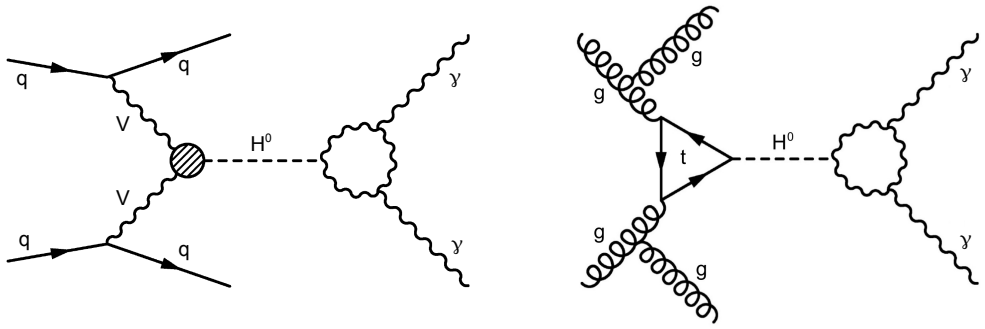


Figure 5.2: Vector boson fusion production (left), and gluon fusion production with two additional jets (right), with the H decaying to two photons.

For each event selected two different kinematic discriminants are built, using the Matrix Element Likelihood Approach (MELA), one to discriminate between VBF and ggH scalar components and one to distinguish VBF 0^+ from VBF 0^- hypotheses. Thanks to these discriminants it is possible to determine different regions of the phase-space enriched with a certain process and extract their Higgs signal yield from a fit to the diphoton mass. In this way we can infer the yields due to the different productions: VBF 0^+, VBF 0^-, ggH and the continuum. Finally we perform a scan of f_{a3}, comparing these yields to data, in order to constrain f_{a3}.

5.2 Data sample and simulated events

The data sample used in this analysis was recorded by the CMS experiment during the LHC Run 1 at a centre-of-mass energy of 8 TeV. A total of 19.7 fb^{-1} of data is analyzed, collected with diphoton triggers with asymmetric transverse energy thresholds and complementary photon selections, as in the SM search [13].

The signal MC samples are generated with the LO matrix element generator JHUGen [78, 79, 80], both for gluon fusion and vector boson fusion production mode, under the assumption of a SM-like, purely scalar (0^+) HVV coupling structure, as well as an anomalous, purely pseudoscalar (0^-) HVV coupling structure, and a 50/50 mix. In the mixed sample, 50% of the cross section comes from the a_1 term, and the other 50% of the cross section comes from the a_3 term in Equation 5.1. The phase Φ_{a3} is chosen to be zero.
The parton level samples are interfaced to PYTHIA6 [81] for parton showering and hadronization. Detector simulation is performed with GEANT4.
Two additional signal samples are produced with POWHEG interfaced to PYTHIA6, for

the scalar hypothesis 0^+ and both for ggH and VBF production modes. These samples are used in the analysis and in the study for the systematic uncertainties. All the signal MC samples are listed in Table 5.1.

Table 5.1: Signal MC samples used in the analysis.

Sample	HVV coupling
ggH+jj 0^+ at 8 TeV - JHU + pythia6	0^+
ggH+jj 0^- at 8 TeV - JHU + pythia6	0^-
vbfH 0^+ at 8 TeV - JHU + pythia6	0^+
vbfH 0^- at 8 TeV - JHU + pythia6	0^-
vbfH 50% 0^+- 50% 0^- at 8 TeV - JHU + pythia6	50/50
ggH 0^+ at 8 TeV - Powheg + pythia6	0^+
vbfH 0^+ at 8 TeV - Powheg + pythia6	0^+

5.3 Object and event selection

We follow closely the object selection used in the cut-based SM analysis for VBF categories [13]. The "cut-based" analysis does not use a multivariate techniques for selection or classification of events. The identification selection requirements are specific to the category, and use a subset of the discriminating variables that are used in the multivariate photon identification described in Reference [13].
In our analysis we use a looser VBF category than the one described in Reference [13]. First a selection on the two jets p_T and invariant mass is applied. In a second step several cuts are added, in order to improve the purity of the VBF selection: we cut on the difference between the pseudorapidities of the two jets, $|\Delta\eta_{jj}|$, the absolute difference in the azimuthal angle between the diphoton system and the dijet system, $|\Delta\Phi_{\gamma\gamma jj}|$ and the difference between the average pseudorapidity of the two jets and the pseudorapidity of the diphoton system, $|\eta_{\gamma\gamma} - (\eta_{j1} + \eta_{j2})/2|$. The values of the second set of cuts have been tuned in order to obtain the best expected sensitivity. The selection requirements are listed in Table 5.2.

Table 5.2: Summary of event selection criteria.

Variable	Requirement
p_T^{j1}	> 30
p_T^{j2}	> 20
m_{jj}	> 250
$\|\Delta\eta_{jj}\|$	> 2
$\|\Delta\Phi_{\gamma\gamma jj}\|$	> 2.6
$\|\eta_{\gamma\gamma} - (\eta_{j1} + \eta_{j2})/2\|$	< 2.5

As outlined in Section 5.1, after this loose VBF selection a contamination from ggH process is still present, more precisely $\sim 60\%$ of the selected events belong to VBF topology, while $\sim 40\%$ derive from ggH production.

5.4 Classifying Higgs boson production processes

5.4.1 Discriminating variables and 1D kinematic discriminants

As briefly described in Section 5.1, the analysis has to discriminate on one hand VBF (0^+) from VBF (0^-) processes, on the other hand ggH from VBF scalar processes. The kinematics of the production of the Higgs boson and the 2 jets in the $H \to \gamma\gamma$ channel are sensitive to its spin and parity. This is illustrated in Figures 5.3 and 5.4, where some kinematic variables of the system $H \to \gamma\gamma$ + 2 jets are shown after the VBF selection described in Table 5.2, for four different production/spin-parity hypotheses, VBF 0^+, VBF 0^-, ggH 0^+ and ggH 0^-. It is evident that a good discrimination between VBF 0^+ and VBF 0^- processes is present, followed by a fair discrimination between VBF 0^+ and ggH 0^+. It is therefore possible to exploit these differences in kinematic distributions building kinematic discriminants to distinguish any two spin-parity hypotheses. These discriminants, calculated with the Matrix Element Likelihood Approach (MELA) [80, 79, 78], take as input the diphoton and dijet kinematics and have the form $\mathcal{D}_{12} = \mathcal{P}_1/(\mathcal{P}_1 + \mathcal{P}_2)$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the probability densities corresponding to the two spin-parity hypotheses we wish to discriminate. Therefore we define two kinematic discriminants: M_{0^-} for the discrimination between a SM Higgs boson and a pure pseudoscalar state $\mathcal{J}^P = 0^-$ in the VBF production mode; M_{VBF} for the discrimination between VBF and ggH production under the hypothesis of a scalar Higgs boson. The discriminant calculation takes the diphoton and dijet systems kinematics as input, and for each event it calculates the theoretical differential cross section for the different processes (i in Equation 5.4) given the event kinematics $\bar{x}$, obtaining the probability density $\mathcal{P}_i(\vec{x})$:

$$\mathcal{P}(\vec{x}\,|i) = \frac{1}{\sigma_i}\frac{d\sigma_i}{d\vec{x}}. \tag{5.4}$$

The two discriminants that we obtain can thus be expressed as:

$$M_{0^-}(\vec{x}) = \frac{\mathcal{P}(\vec{x}\,|VBF\ 0^+)}{\mathcal{P}(\vec{x}\,|VBF\ 0^+) + \mathcal{P}(\vec{x}\,|VBF\ 0^-)}, \qquad M_{VBF}(\vec{x}) = \frac{\mathcal{P}(\vec{x}\,|VBF\ 0^+)}{\mathcal{P}(\vec{x}\,|VBF\ 0^+) + \mathcal{P}(\vec{x}\mid ggH\ 0^+)} \tag{5.5}$$

The distributions for these two discriminants, for simulated samples, are shown in Figures 5.5 and 5.6. It can be seen from Figure 5.5 that the separation between VBF 0^+ (blue histogram) and VBF 0^- (green histogram) hypotheses is very good. By construction the 0^+ distribution peaks at 1, and the 0^- distribution peaks at 0. In Figure 5.6 the discrimination between VBF scalar (blue histogram) and ggH scalar (red histogram) hypotheses is shallow. The fact that the ggH distribution is similar to the VBF one is due to our VBF

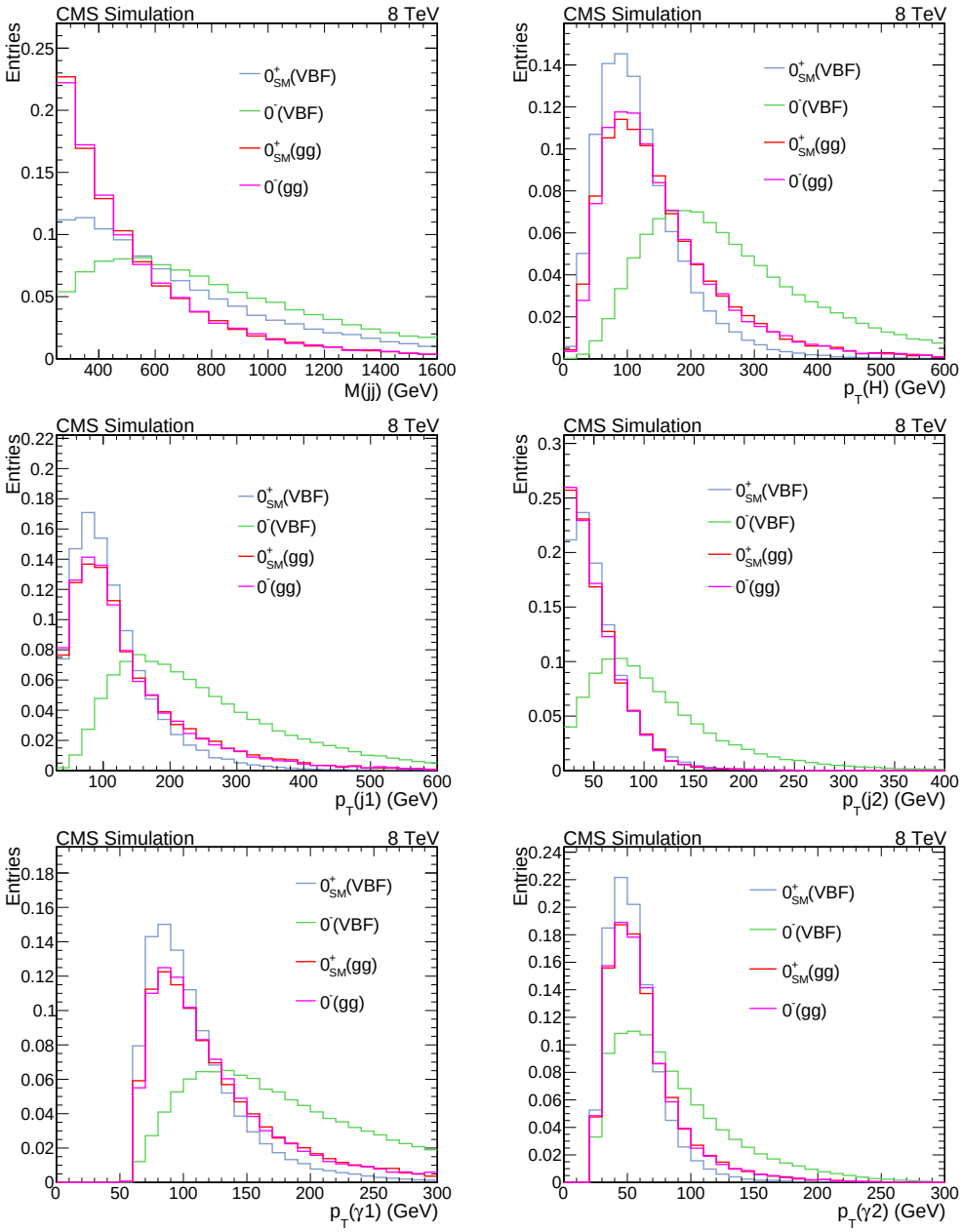


Figure 5.3: Distributions of kinematic variables of the diphoton-dijet system. The distributions are shown for four different production/spin-parity hypotheses, VBF 0^+ (blue), VBF 0^- (green), ggH 0^+ (red) and ggH 0^- (magenta). The VBF selection of the analysis was applied.

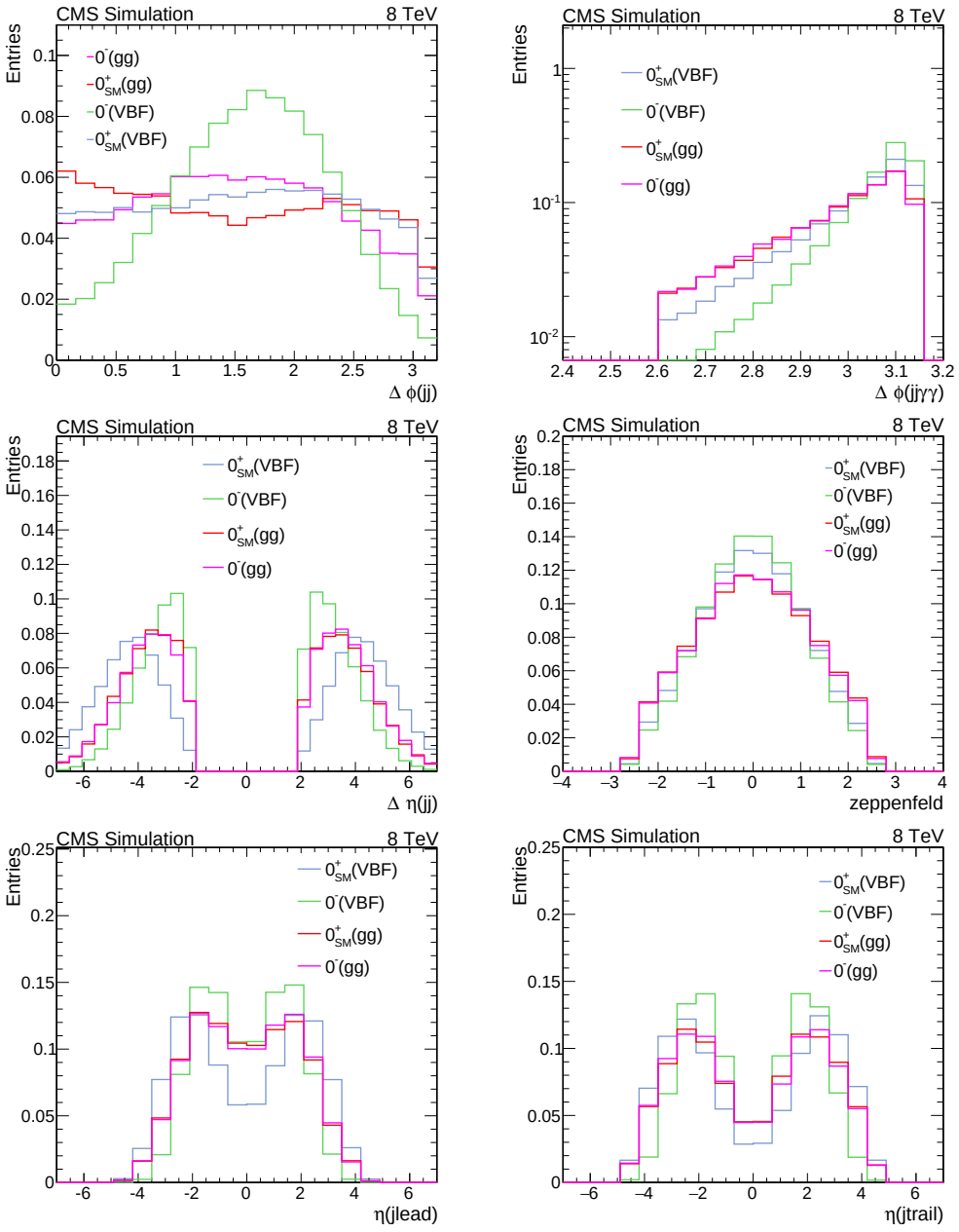


Figure 5.4: Distributions of kinematic variables of the diphoton-dijet system. The distributions are shown for four different production/spin-parity hypotheses, VBF 0^+ (blue), VBF 0^- (green), ggH 0^+ (red) and ggH 0^- (magenta). The VBF selection of the analysis was applied.

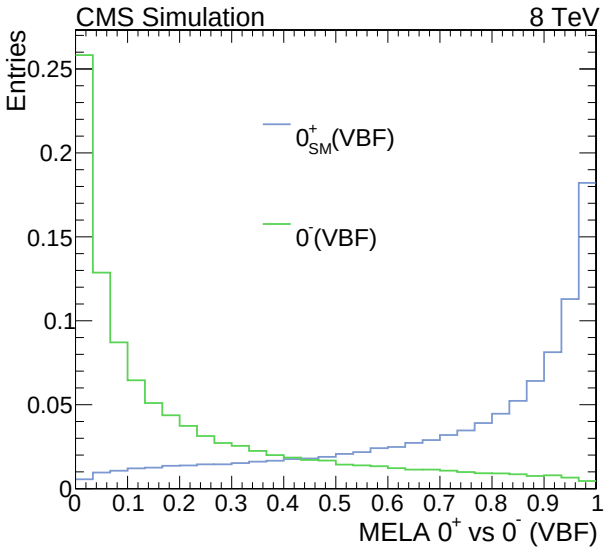


Figure 5.5: Simulated distributions for the kinematic discriminant sensitive to ratios of scalar to pseudoscalar components in the HVV vertex assuming a VBF production.

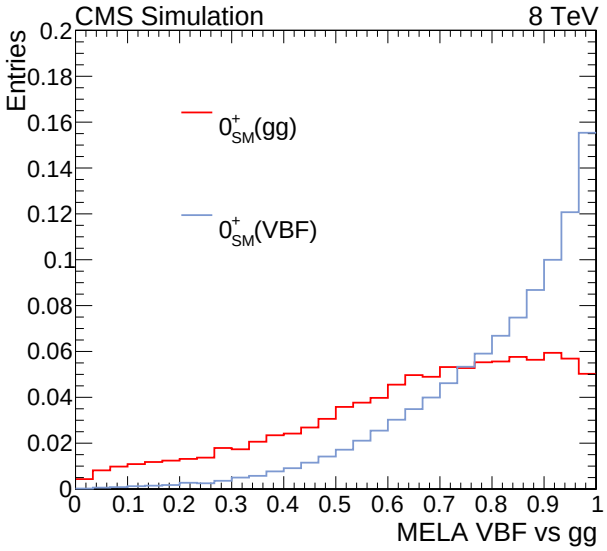


Figure 5.6: Simulated distributions for the kinematic discriminant for the discrimination between VBF and ggH production under the hypothesis of a scalar Higgs boson.

selection: ggH events passing the VBF selection are VBF-like.

In a first step we calculated these discriminants also for ggH scalar and pseudoscalar processes, in order to try to distinguish the two parity hypotheses also in the case of gluon fusion production mode. But as shown in Figure 5.7 the two parity distributions, in magenta (ggH 0^-) and in red (ggH 0^+), have very similar shapes for all discriminants. Therefore no attempt was made to separate ggH 0^+ and ggH 0^-. These two contributions were estimated together and simulated as ggH 0^+.

5.4.2 2D MELA

The two discriminants described in Section 5.4.1 can be used simultaneously to build 2D maps that allow us to determine different regions of the phase-space enriched with a certain process. The discriminant between VBF and ggH hypotheses constitutes the x axis, while the one between VBF scalar and pseudoscalar processes constitutes the y axis. In Figure 5.8 the 2D maps are presented for the different simulated processes and for data (in data the mass region between 115 and 135 GeV is excluded). From Figure 5.8 one can see VBF 0^+ simulated events (top left plot) populate the region on the top right of the 2D map, while the VBF 0^- events (top right plot) dominate the bottom right region. The ggH 0^+ simulated events are in the same region as VBF 0^+, but they have a broader distribution, as already seen for the 1D discriminants (Figure 5.6). The data are also distributed in the top right part of the 2D map, with a broad distribution.
As explained in the next section, thanks to the separation of the processes in different regions of the phase-space, the latter can be divided in 2D categories enriched with a certain process. The number and borders of these categories are optimised in order to maximize the analysis sensitivity, as explained in Section 5.7.

5.5 Treatment of the scalar - pseudoscalar interference

Since the 2 VBF production 0^+ and 0^- give raise to identical final states, the interference between the 2 has to be simulated. To infer the interference term we used three MC samples: pure VBF 0^+, pure VBF 0^- and $0^+/0^-$ mixed.
We define:

$$\begin{aligned} \sigma_1 &= |\mathcal{A}mp_1|^2 \\ \sigma_3 &= |\mathcal{A}mp_3|^2, \end{aligned} \tag{5.6}$$

where σ_1 and σ_3 are the cross sections, $\mathcal{A}mp_1$ and $\mathcal{A}mp_3$ are the amplitudes of the VBF 0^+ ($a_1 = 1$, $a_3 = 0$) and VBF 0^- ($a_1 = 0$, $a_3 = 1$) process respectively.
The amplitude of a generic $0^+/0^-$ mixed process can be written as:

$$\mathcal{A}mp_{mix} = a_1\mathcal{A}mp_1 + a_3\mathcal{A}mp_3, \tag{5.7}$$

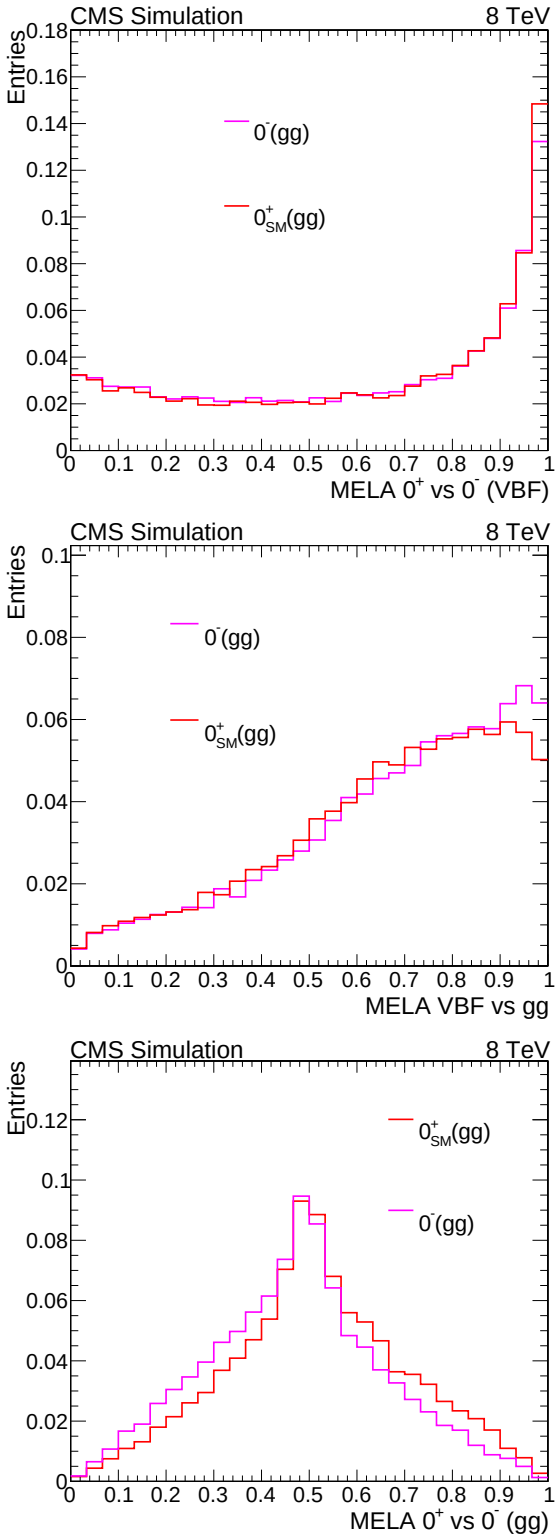


Figure 5.7: Simulated distributions for the ggH 0^+ and ggH 0^- processes for three kinematic discriminants: the one sensitive to ratios of scalar to pseudoscalar components in the HVV vertex assuming a VBF production (top), the one for the discrimination between VBF and ggH production under the hypothesis of a scalar Higgs boson (middle), and the one sensitive to ratios of scalar to pseudoscalar components assuming a ggH production (bottom).

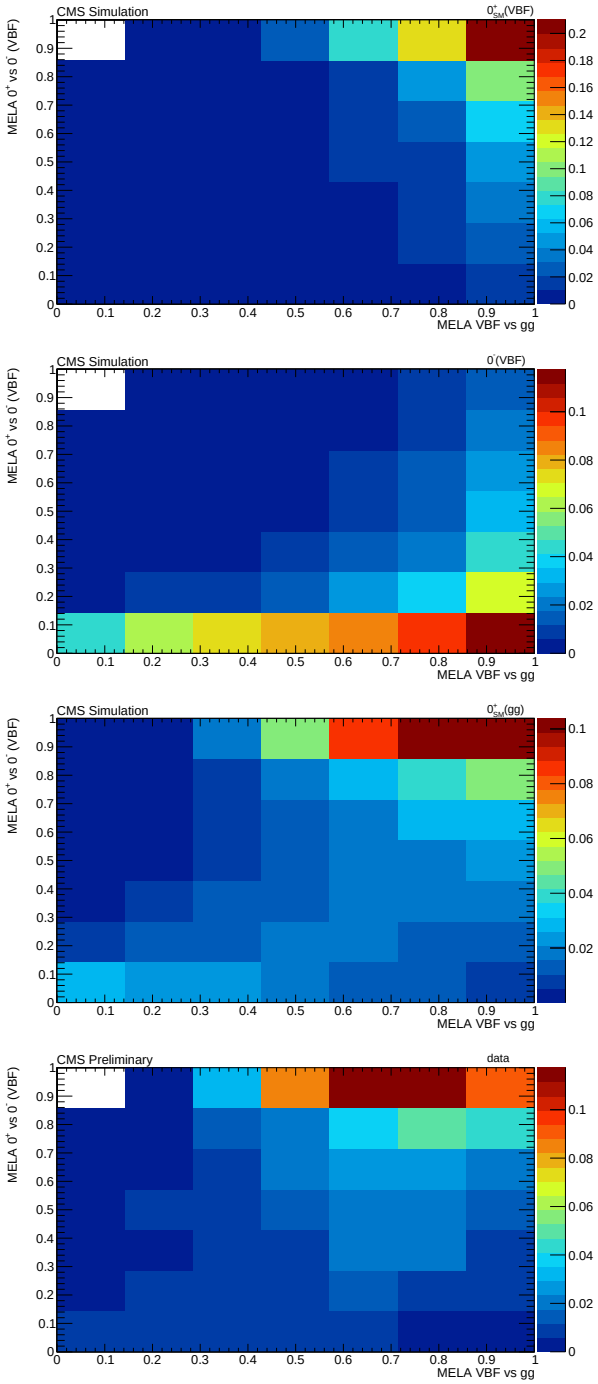


Figure 5.8: 2D maps composed by the two kinematic discriminants described in Section 5.4.1 and calculated for simulated samples, VBF 0^+, VBF 0^-, ggH 0^+, and data (the mass region between 115 and 135 GeV is excluded in data). The distributions are normalized to 1.

where a_1 and a_3 are the coupling constants of the 0^+ and 0^- processes, from Equation 5.1. The square of Equation 5.7 gives the cross section of a generic $0^+/0^-$ mixed process:

$$|\mathcal{A}mp_{mix}|^2 = \sigma_{mix} = |a_1|^2|\mathcal{A}mp_1|^2 + |a_3|^2|\mathcal{A}mp_3|^2 + 2|a_1||a_3||\mathcal{A}mp_1||\mathcal{A}mp_3|cos(\delta + \Phi), \tag{5.8}$$

where

$$\delta = arg\left(\frac{\mathcal{A}mp_3}{\mathcal{A}mp_1}\right), \quad \Phi = arg\left(\frac{a_3}{a_1}\right). \tag{5.9}$$

In the following we will use $\Phi = 0$.
Using Equation 5.6 one can write σ_{mix} as function of σ_1 and σ_3:

$$\sigma_{mix} = \sigma_1|a_1|^2 + \sigma_3|a_3|^2 + 2|a_1||a_3|\sqrt{\sigma_1\sigma_3}\ cos\delta. \tag{5.10}$$

After acceptance cuts from the analysis, the fiducial cross section is therefore given by replacing

$$\sigma_i \rightarrow \sigma_i A_i, \tag{5.11}$$

where A_i are the acceptances for the different processes.

Taking the acceptances into account, Equation 5.10 can be written, for any value of a_1 and a_3:

$$\tilde{\sigma}_{mix}(\vec{x}) = \sigma_{mix}(\vec{x})A_{mix}(\vec{x}) = \sigma_1 A_1(\vec{x})|a_1|^2 + \sigma_3 A_3(\vec{x})|a_3|^2 + 2|a_1||a_3|\sqrt{\sigma_1\sigma_3}\ A_{int}(\vec{x}), \tag{5.12}$$

where $A_i(\vec{x})$ are the acceptances for the different processes in each point of the phase-space, and $cos\delta$ has been absorbed by $A_{int}(\vec{x})$.

If we perform the integration on the phase-space, we obtain:

$$\int \tilde{\sigma}_{mix}(\vec{x})dx = \sigma_1 I_{A_1}|a_1|^2 + \sigma_3 I_{A_3}|a_3|^2 + 2|a_1||a_3|\sqrt{\sigma_1\sigma_3}\ \sqrt{I_{A_1}I_{A_3}}\ \epsilon, \tag{5.13}$$

where I_{A_i} are the integrals of the acceptances on the phase-space, and ϵ is defined as:

$$\epsilon = \int \frac{A_{int}(\vec{x})}{\sqrt{I_{A_1}I_{A_3}}}dx. \tag{5.14}$$

Using the integrals of the acceptances, Equation 5.12 can be expressed as:

$$\tilde{\sigma}_{mix}(\vec{x}) = \sigma_1 I_{A_1}\frac{A_1(\vec{x})}{I_{A_1}}|a_1|^2 + \sigma_3 I_{A_3}\frac{A_3(\vec{x})}{I_{A_3}}|a_3|^2 + 2\sqrt{\sigma_1 I_{A_1}|a_1|^2}\sqrt{\sigma_3 I_{A_3}|a_3|^2}\ \frac{A_{int}(\vec{x})}{\sqrt{I_{A_1}I_{A_3}}}. \tag{5.15}$$

Before going on, we define the fraction of the pseudoscalar component that takes into account the acceptances as:

$$\tilde{f}_{a3} = \frac{\sigma_3 I_{A_3}|a_3|^2}{\sigma_1 I_{A_1}|a_1|^2 + \sigma_3 I_{A_3}|a_3|^2}. \tag{5.16}$$

Defining

$$\tilde{s} = \sigma_1 I_{A_1} |a_1|^2 + \sigma_3 I_{A_3} |a_3|^2, \tag{5.17}$$

one can obtain:

$$\begin{aligned} \sigma_3 I_{A_3} |a_3|^2 &= \tilde{f}_{a3} \tilde{s} \\ \sigma_1 I_{A_1} |a_1|^2 &= (1 - \tilde{f}_{a3}) \tilde{s}. \end{aligned} \tag{5.18}$$

Now, inserting Equation 5.18 in Equation 5.15 and keeping in mind that $\frac{A_i(\vec{x})}{I_{A_i}}$ are probability densities, we obtain:

$$\tilde{\sigma}_{mix}(\vec{x}) = \tilde{s} \left[(1 - \tilde{f}_{a3}) P_1(\vec{x}) + \tilde{f}_{a3} P_3(\vec{x}) + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, \frac{A_{int}(\vec{x})}{\sqrt{I_{A_1} I_{A_3}}} \right], \tag{5.19}$$

where $P_1(\vec{x})$ and $P_3(\vec{x})$ are the probability densities of pure 0^+ and pure 0^- process respectively (after cuts).

Similarly we can write Equation 5.13 as:

$$\int \tilde{\sigma}_{mix}(\vec{x}) dx = \tilde{s} \left[1 + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, \epsilon \right]. \tag{5.20}$$

Finally we can obtain the probability density for a generic $0^+/0^-$ mixed process:

$$P_{mix}(\vec{x}) = \frac{\tilde{\sigma}_{mix}(\vec{x})}{\int \tilde{\sigma}_{mix}(\vec{x}) dx} = \frac{(1 - \tilde{f}_{a3}) P_1(\vec{x}) + \tilde{f}_{a3} P_3(\vec{x}) + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, P_{int}(\vec{x})}{1 + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, \epsilon}, \tag{5.21}$$

$$P_{mix}(\vec{x}) = \frac{(1 - \tilde{f}_{a3}) P_1(\vec{x}) + \tilde{f}_{a3} P_3(\vec{x}) + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, P_{int}(\vec{x})}{1 + 2\sqrt{(1 - \tilde{f}_{a3}) \tilde{f}_{a3}} \, \epsilon}, \tag{5.22}$$

where $P_{int}(\vec{x}) = \frac{A_{int}(\vec{x})}{\sqrt{I_{A_1} I_{A_3}}}$ is the probability density of the interference in each point of the phase-space, while ϵ, as defined in Equation 5.14, is its integral.
Equation 5.22 is valid for any $\tilde{f}_{a3}$. Therefore we can use any mixed sample to infer $P_{int}(\vec{x})$. This will allow to simulate $P_{mix}(\vec{x})$ for any f_{a3}. We used a mixed sample with $f_{a3} = 0.5$. We found that $I_{A_1} = 0.4$ and $I_{A_3} = 0.27$, and knowing that in the mixed MC sample $a_1 = 1$ and $a_3 = 3.2$, we found $\tilde{f}_{a3} = 0.4$. This value of $\tilde{f}_{a3}$ was used in Equation 5.22 to compute the contribution of the interference $P_{int}(\vec{x})$.
Starting from $P_{mix}(\vec{x})$, $P_1(\vec{x})$ and $P_3(\vec{x})$, which are the 2D maps of the mixed $f_{a3} = 0.5$, pure 0^+ and pure 0^- sample respectively, we computed ϵ in an iterative way, finding it very close to 0. $P_{int}(\vec{x})$ is shown in Figure 5.9. From Figure 5.9 one can see that $P_{int}(\vec{x}) \ll P_1(\vec{x}), P_3(\vec{x})$ in all the phase-space and it will be neglected in the following.

As mentioned above, because of our acceptance cuts, the fraction of pseudoscalar component that we observe ($\tilde{f}_{a3}$) is slightly different from that defined in Equation 5.2. Comparing Equations 5.2 and 5.16, one can obtain the expression of f_{a3} as a function of $\tilde{f}_{a3}$:

$$f_{a3} = \frac{\tilde{f}_{a3}}{\frac{I_{A_3}}{I_{A_1}}(1-\tilde{f}_{a3}) + \tilde{f}_{a3}}. \tag{5.23}$$

So constraining $\tilde{f}_{a3}$ corresponds to constraining f_{a3}.

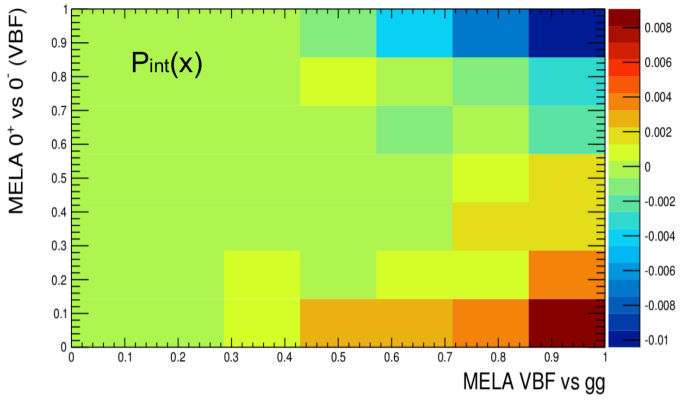


Figure 5.9: 2D map of $P_{int}(\vec{x})$, that is the value of the interference in each bin of MELA.

5.6 Analysis strategy and signal extraction

In order to extract the different signal, we use $M_{VBF} \equiv$ Mela VBF vs gg to discriminate VBF from ggH production, $M_{0^-} \equiv$ Mela VBF 0^+ vs 0^- to discriminate VBF 0^+ from VBF 0^- productions and the invariant mass of the two photons $m_{\gamma\gamma}$ to discriminate Higgs production from diphoton background.

Starting from Equations 5.17 and 5.19, the number of expected VBF events can be written as

$$\frac{N_{VBF}(\vec{x})}{\mathcal{L}} \equiv \left(|a_1|^2 \frac{1}{1-\tilde{f}_{a_3}}\right) \times \sigma_1 I_{A_1} \times P_{mix}(\vec{x}) \equiv \mu^{VBF} \times \sigma_1 I_{A_1} \times P_{mix}(\vec{x}) \tag{5.24}$$

where $\mathcal{L}$ is the luminosity and μ^{VBF} is defined as

$$\mu^{VBF} \equiv |a_1|^2 \frac{1}{1-\tilde{f}_{a_3}}$$

in such a way that in the SM, $\mu^{VBF} = 1$ and $\tilde{f}_{a_3} = 0$.

Therefore the total number of expected events in the sample can be parametrized in the 3 dimensional space (M_{VBF}, M_{0^-} and $m_{\gamma\gamma}$) and can be written as:

$$\begin{aligned} N(m_{\gamma\gamma}, M_{VBF}, M_{0^-})/\mathcal{L} = \mu^{VBF} \times \sigma^{SM}_{VBF} \times I_{A_1} \times \\ \Big[(1-\tilde{f}_{a_3})P_1(m_{\gamma\gamma}, M_{VBF}, M_{0^-};\theta) + \tilde{f}_{a_3}P_3(m_{\gamma\gamma}, M_{VBF}, M_{0^-};\theta)\Big] + \\ \mu^{ggH} \times \sigma^{SM}_{ggH} \times I_{A_{ggH}} \times P_{ggH}(m_{\gamma\gamma}, M_{VBF}, M_{0^-};\theta) + \\ N_{bkg} \times P_{bkg}(m_{\gamma\gamma}, M_{VBF}, M_{0^-};\theta) \end{aligned} \tag{5.25}$$

where we have redefined $\sigma^{SM}_{VBF} = \sigma_1$, σ^{SM}_{ggH} is the number of expected events from ggH production in the Standard Model and μ^{ggH} is a potential deviation to this expectation (due to theoretical uncertainties for instance or to the contribution of new processes), the interference term has also been neglected as aforementioned.

In order to simplify the extraction we do not fit the number of events in the 3D space, but rather we will define categories in the 2D space of (M_{VBF},M_{0^-}) and fit the $m_{\gamma\gamma}$ mass distribution in each of these categories. This does not change substantially Equation 5.25, but the 2D space (M_{VBF},M_{0^-}) is replaced by a 1 dimensional discrete function, the category number.

The model described by Equation 5.25 will be used as input to the standard CMS statistical treatment to extract the constraints on the different parameters. In addition several systematics need to be added, either affecting the total yield, or the migration in the 2D space (M_{VBF},M_{0^-}), or practically the migration of events between categories.

The following section describes the optimisation of these categories.

5.7 Categories optimisation

For each MELA discriminant separately we first performed 1D optimisation to obtain the best sensitivity, dividing the phase-space in two or more categories. The values obtained for the 2x1D optimisations are then combined to form the 2D categories. In the following the optimisation procedure is explained in more detail for the two discriminants.

5.7.1 Optimisation of VBF vs ggH discriminant M_{VBF}

We performed the optimisation of the 1D discriminant between VBF SM and ggH SM hypotheses. To do that, referring to the physics model presented in Equation 5.25, we set these conditions:

- $\tilde{f_{a3}}$ fixed at 0, since with this discriminant we want to distinguish between the VBF and ggH,
- μ^{ggH} is constrained to 1 within a theoretical systematic uncertainty of 32% (this value was taken from Reference [82, 13]),
- μ^{VBF} is left floated,
- the sensitivity to μ^{VBF} is considered as figure of merit.

The full range is first cut in 2 categories, trying several cuts between 0 and 1. For each cut value we compute the expected sensitivity performing a profile likelihood scan as a function of μ^{VBF} with Asimov toys [83], and we kept the value for which we obtained the best sensitivity. The sensitivity here was defined as the value of the statistical test (likelihood ratio) when testing $\mu^{VBF} = 0$, the higher is this number, the more likely it is to observe a VBF production signal. The best value for this first cut was found to be 0.91. After that, we took the best category of the two, that is the one with the highest signal over background ratio, and we performed a further split cut, dividing in this way the discriminant range in three 1D categories. Doing this we saw a little improvement in the expected sensitivity, in particular for the cut at value 0.97. We verified that the addition of more than three categories does not bring any improvement to the expected sensitivity. The final cut values kept for the VBF SM vs ggH SM discriminant are therefore 0.91 and 0.97. In Figure 5.10 the profile likelihood scan as a function of μ^{VBF} is shown for two 1D categories, with a cut at 0.91, and for three 1D categories, with cuts at 0.91 and 0.97.

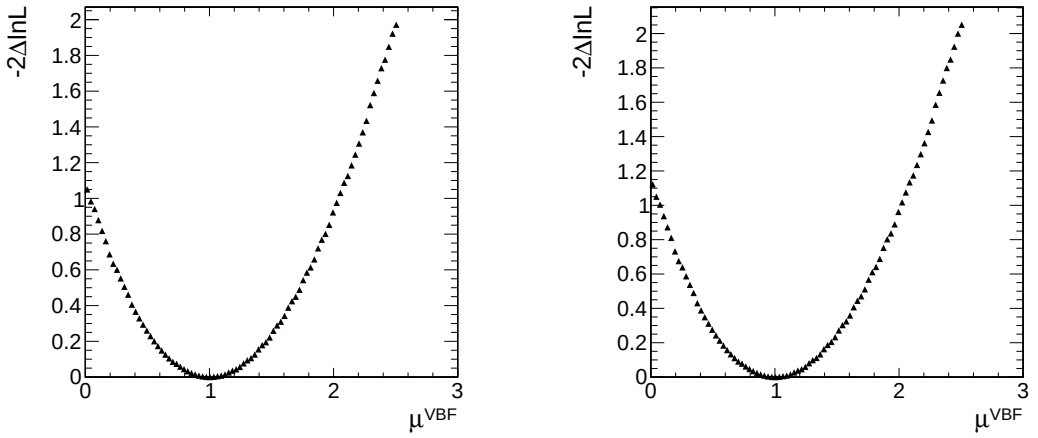


Figure 5.10: Profile likelihood scan for the optimisation of the VBF vs ggH discriminant. The triangles show the expected -2Δln$\mathcal{L}$ value as a function of μ^{VBF}, for two 1D categories on the left and for three 1D categories on the right.

5.7.2 Optimisation of VBF 0^+ vs VBF 0^- discriminant M_{0^-}

The optimisation of the 1D discriminant between VBF 0^+ and VBF 0^- hypotheses was done similarly. In Equation 5.25, we set the following conditions:

- μ^{ggH} is constrained to 1 within a theoretical systematic uncertainty of 32% (this value was taken from Reference [82]),
- μ^{VBF} is left floated,
- $\tilde{f_{a3}}$ is left floated between 0 and 1 in the fit, while $\tilde{f_{a3}} = 0$ in the true underlying model.

As for the VBF vs ggH discriminant, we started dividing the discriminant range in two 1D categories, trying several cuts between 0 and 1. For each cut value we computed the expected sensitivity performing a profile likelihood scan as a function of $\tilde{f_{a3}}$ with Asimov toys, and we kept the value for which we obtained the best sensitivity, that is 0.1. This sensitivity is defined as the value of the likelihood obtained when probing $\tilde{f_{a3}} = 1$, the higher is this number the more likely it is to exclude an anomalous production (when the VBF production is pure SM). A further categorisation does not bring any improvement to the expected sensitivity, so 0.1 is the only value kept as boundary of the final 2D categories.

5.7.3 Optimised 2D categories

After the independent optimisation of the two 1D discriminants, the discriminant values that give the best expected sensitivity are combined in 2D to form the final optimised categories. Given that the boundaries are 0.91 and 0.97 for the discriminant VBF vs ggH, and 0.1 for the discriminant VBF 0^+ vs VBF 0^-, we should have 6 categories. Nevertheless, the categories in the phase-space region delimited by (MELA VBF 0^+ vs VBF 0^-) < 0.1 were gathered in a single category, in order to increase the statistic in this region, where the background is already extremely small. Therefore we ended up with 4 categories, shown in Figure 5.11. Their numbering, from 0 to 3, will be maintained in the following.

5.7.4 Optimisation of the diphoton MVA cut

Once defined the 4 categories, we added a cut on the diphoton MVA. This quantity is very similar to the one presented in chapter 4 and allows to make the best use of the quality of the photons in the event, hence improving our sensitivity. To do this we performed a scan for diphoton MVA values going from -1 to 0.4. For this optimisation we set these conditions:

- $\tilde{f_{a3}}$ fixed at 0,

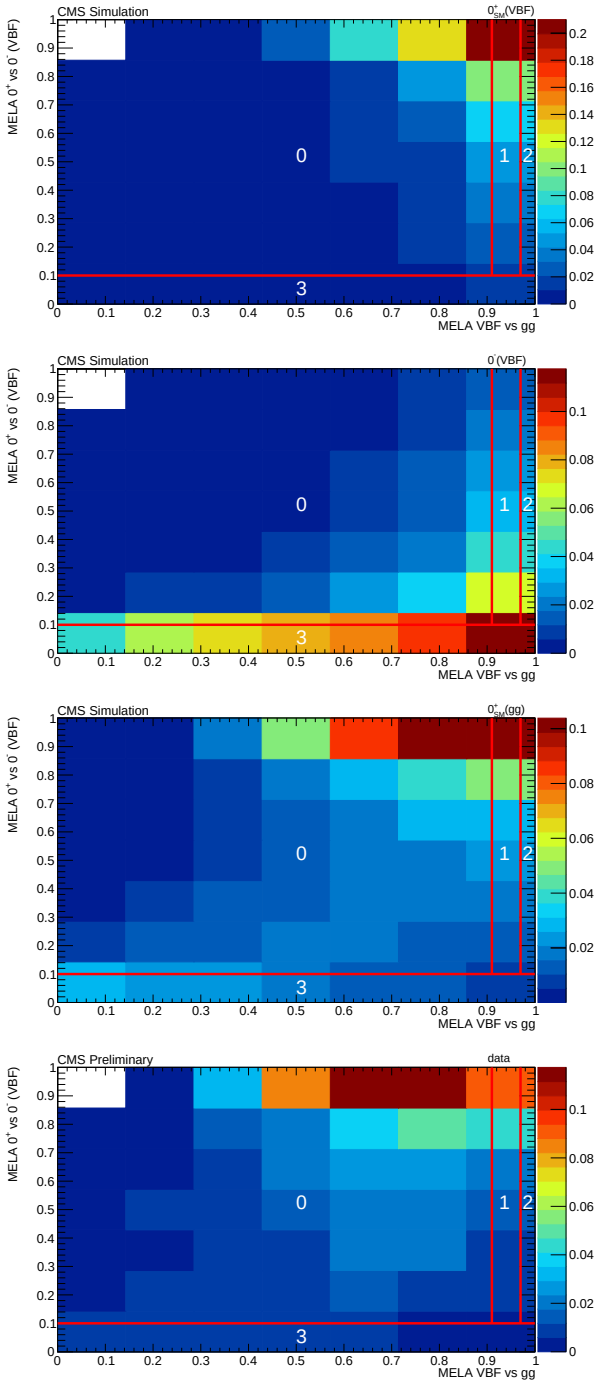


Figure 5.11: 2D maps composed by the two kinematic discriminants described in Section 5.4.1 and calculated for both simulated samples, VBF 0^+, VBF 0^- and ggH 0^+, and data (the mass region between 115 and 135 GeV is excluded in data). The 4 categories chosen after the optimisation are delimited by red lines and are numbered from 0 to 3.

- μ^{ggH} is constrained to 1 within its theoretical systematic uncertainty of 32% (this value was taken from Reference [82]),
- μ^{VBF} is left floated.

We then use the same procedure as for the 1D optimisation of M_{VBF}. As one can see in Figure 5.12, the sensitivity reaches a plateau for a cut on the diphoton MVA of 0. It was decided to cut at 0.2 on the diphoton MVA output, which retains a lot of statistic while having a high sensitivity.

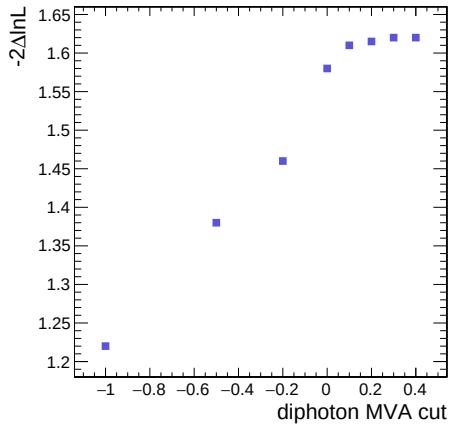


Figure 5.12: Likelihood value at 0 as a function of the cut on the diphoton MVA.

5.8 Bias study for the background function

Since we do not rely on MC to predict the background distribution, we follow the same procedure as the one described in the analysis [13], and we use a polynomial to fit the background. The order of the polynomial is chosen following the same procedure as in Reference [13] and is described below.
Suppose we obtain a set of measurements of a parameter τ, whose "true" or "generated" value is τ_g. The measurements are statistical fluctuations around τ_g and could, for example, follow an exponential time distribution

$$\frac{1}{\tau_g} e^{-t/\tau_g}. \tag{5.26}$$

If a histogram is produced, there would be Poisson fluctuations on the numbers in each bin. A fit to the data would give a value $\tau_m \pm \sigma_m$. Then, for a large number of events in the distribution, we would expect τ_m to be approximately Gaussian distributed about τ_g,

even though the distribution 5.26 is non-Gaussian. For many repetitions of this procedure, the quantity

$$g = \frac{\tau_m - \tau_g}{\sigma_m} \tag{5.27}$$

should follow a Gaussian distribution. The above definition is called a pull and can be used for checking the properties of a fitting algorithm with large numbers of pseudo-experiments. In our study we generated the expected number of background events, in a diphoton mass range between 100 and 180 GeV, according to a particular functional form, power law, exponential and polynomial of order between 1 and 4. Then we fitted this generated mass distribution with other functions. This procedure was repeated several thousands of time to get the distribution

$$N_{fitted} - N_{expected}, \tag{5.28}$$

where N_{fitted} is the number of events that the fit was able to extract, while $N_{expected}$ is the number of generated events. From this distribution we computed the bias as

$$\frac{mean(N_{fitted} - N_{expected})}{RMS(N_{fitted} - N_{expected})}. \tag{5.29}$$

The more the fitting function is appropriate to fit the generated distribution, the more the bias value is close to zero. So by construction when we generate and we fit the distribution with the same function, the bias value is zero.
As already mentioned, we were interested in choosing the best order of the polynomial function for each category. Therefore, in each category, we computed the bias for a model generated both with a power law and with an exponential, and fitted with a polynomial of order 1, 2, 3 and 4. The polynomial function with the lowest order which gives a bias $<$ 0.15 is chosen. The polynomials finally chosen for the different categories are the following:

- category 0: polynomial of order 3,
- category 1: polynomial of order 1,
- category 2: polynomial of order 1,
- category 3: polynomial of order 1.

The polynomial order decreases with the statistic present in the category. In Figure 5.13, as an example, the distribution $N_{fitted} - N_{expected}$ is shown, for category 0, for events generated according to an exponential function and fitted with an exponential, a power law, a polynomial of order 1 and 3. As expected, if the fit was done with an exponential or with a polynomial of order 3, the mean of the distribution is closer to zero.

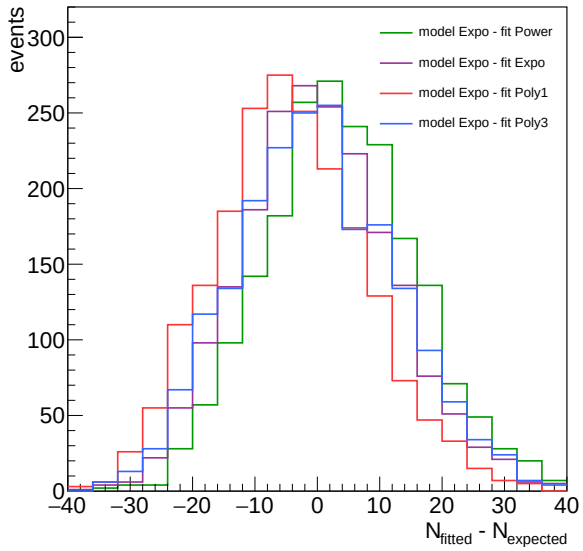


Figure 5.13: Distribution of $N_{fitted} - N_{expected}$ for category 0. Events are generated according to an exponential function and fitted with an exponential (violet), a power law (green), a polynomial of order 1 (red) and a polynomial of order 3 (blue).

5.9 Systematic uncertainties

Systematics can have 2 different effects: they can affect the overall yields of expected signal or the signal shapes, either via the mass distribution or via migration between the different signal categories. It has to be noted that for this analysis we are not sensitive to systematics affecting the overall yield of μ^{VBF}, since it is left floated in the extraction of the results. Nevertheless, main systematics affecting the overall yield have been also considered in order to extract μ^{VBF} for the case of $\tilde{f_{a3}} = 0$.
A summary of the systematic uncertainties considered is presented in this section.

Integrated luminosity

A 2.6% uncertainty was assigned to the overall luminosity [84, 85].

ggH + 2jets cross section uncertainty, QCD scale

The theoretical uncertainty on the ggH + 2 jets process cross section, coming from both uncertainties due to the missing higher orders and uncertainties related to the parton distribution functions, is of about 32% in the VBF categories (see Reference [82]). This is the dominant source of systematic uncertainty.

H+2jets kinematics description

For the final results we use the JHU MC sample for the VBF 0^- process, and the Powheg MC samples for the ggH 0^+ and VBF 0^+.
The distributions of the different discriminants for several generators (powheg, JHU, MiNLO,aMC@NLO) have been studied and are shown in Figures 5.14 and 5.15. This was done both at the generator and reconstruction level. The differences present between different generators create event migrations between categories, and leads to a large source of systematic uncertainties. To estimate this systematics we computed the ratio of the four categories yield from an alternate generator to the reference one. The deviation of the ratio from 1 in each category is taken as systematics (blue line in Figure 5.16). This was done at reconstruction level for VBF 0^+, VBF 0^- and ggH 0^+, comparing JHU and Powheg generators. Figure 5.16 shows the event yield in the 4 categories for the two generators and their ratio, for the different processes.
Since the Powheg 0^- sample was not available, we performed a 2D reweighting of the 2 kinematic discriminants in the VBF 0^+ JHU sample in order to match the distributions of the VBF 0^- JHU sample. Then we applied the 2D weights to the VBF 0^+ Powheg sample, obtaining in this way a VBF 0^- Powheg sample.
One may notice on distributions from Figures 5.14 and 5.15 that JHU generator is quite different from others especially for ggH production. This has been traced back to the factorisation scale, setup too high for JHU, which yields more radiations in JHU than in other MCs used in CMS. Therefore, the systematic uncertainty estimated as the difference between JHU and powheg is conservative.

Jet energy correction and jet energy resolution

The signal efficiency is affected both by the uncertainty on the jet energy scale and by the uncertainty on the jet energy resolution. We varied the jet energy scale (JEC) by shifting the scale by $\pm 1\sigma$, where σ is the full jet energy scale uncertainty. The uncertainty is evaluated as prescribed for the 2012 dataset.
The jet energy resolution is smeared by the level of disagreement between the resolution measured in data and in MC. The uncertainty on this over-smearing is taken as systematics. Both the JEC and JER uncertainties affect the analysis causing events to migrate from one category to another. We computed the ratio of the four categories yield for the shifted sample (both for shift up and down) over the four categories yield for the nominal sample. The deviation of the ratio from 1 in each category is taken as systematics. This was done for VBF 0^+, VBF 0^- and ggH 0^+. Figure 5.17 shows the event yield in the 4 categories for the shifted and nominal samples and their ratio, for the different processes, both for JEC and JER.

Systematics on diphoton MVA

Since our selection contains a cut on the diphoton MVA, we evaluated its systematic uncertainty. The two main sources of systematic uncertainty on the diphoton MVA output derive from the photon ID MVA and the per-photon energy resolution. Following prescription in Reference [13] the systematic uncertainty to photon ID MVA was assigned varying the photon ID MVA score of $\pm$ 0.01. The uncertainty in the per-photon energy resolution estimate was parameterized as a rescaling of the resolution estimate by $\pm$10% about its nominal value. These systematics are varied correlated for both photons. Figure 5.18 shows the event yield in the 4 categories for the shifted and nominal samples and their ratio, for the different processes, both for photon ID MVA and for photon energy resolution.

Systematics on photon preselection and jet identification

Following Reference [13], we assigned a systematics on the photon efficiency of 1%(2.6%) in the barrel (endcap). In a similar way, the efficiency on the jet identification is estimated to be known at 3-5% level from the CMS collaboration. Note that these sources of systematics affect mostly the total yield of μ^{VBF} and therefore not the final constraint on $\tilde{f}_{a3}$.

Systematics on photon energy scale and photon energy resolution

An uncertainty of 0.1% on the electromagnetic calorimeter energy scale is assigned, following prescription in Reference [13] and is correlated over all signal categories.
An additional smearing is applied as an additional constant term of the energy resolution, as reported in Table 5.3. The systematic error on this smearing is used to infer the systematics due to potentital data/MC disagreement on the photon energy resolution. The effect on the experimental width was found to be of the order of 0.1% correlated over the categories.

Table 5.3: The additional constant term determined from a comparison of data to Monte Carlo.

Photon category	$\Delta\sigma$(%)
EB, $\lvert\eta\rvert < 1$, $R_9 > 0.94$	0.03
EB, $\lvert\eta\rvert < 1$, $R_9 < 0.94$	0.02
EB, $\lvert\eta\rvert > 1$, $R_9 > 0.94$	0.09
EB, $\lvert\eta\rvert > 1$, $R_9 < 0.94$	0.02
EE, $\lvert\eta\rvert < 2$, $R_9 > 0.94$	0.08
EE, $\lvert\eta\rvert < 2$, $R_9 < 0.94$	0.04
EE, $\lvert\eta\rvert > 2$, $R_9 > 0.94$	0.04
EE, $\lvert\eta\rvert > 2$, $R_9 < 0.94$	0.05

Systematics on the Higgs boson mass

In the statistical treatment we use to measure μ^{VBF} and $\tilde{f}_{a3}$, we fix the Higgs boson mass to its world average 125.09 ± 0.24 GeV and let this mass vary within its uncertainty (0.2%).

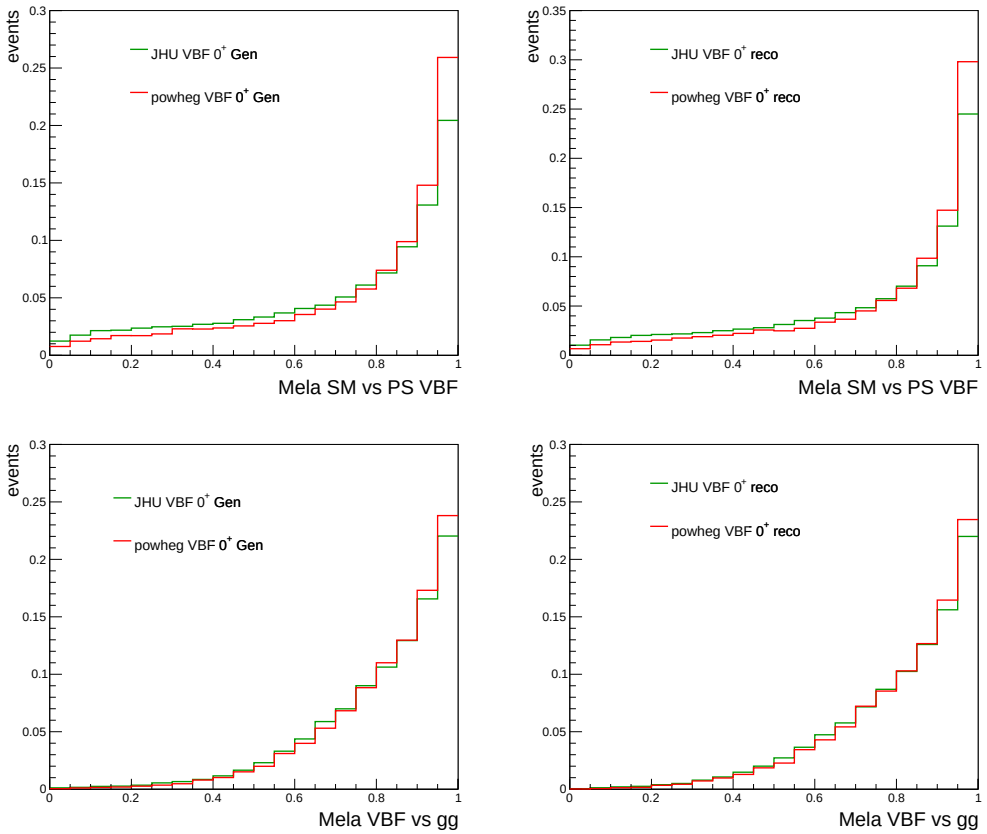


Figure 5.14: Distributions of the two kinematic discriminants, Mela SM vs PS VBF (top) and Mela VBF vs gg (bottom), for the VBF 0^+ process and both for generator (left) and reconstruction (right) level are presented. JHU and Powheg generators are compared.

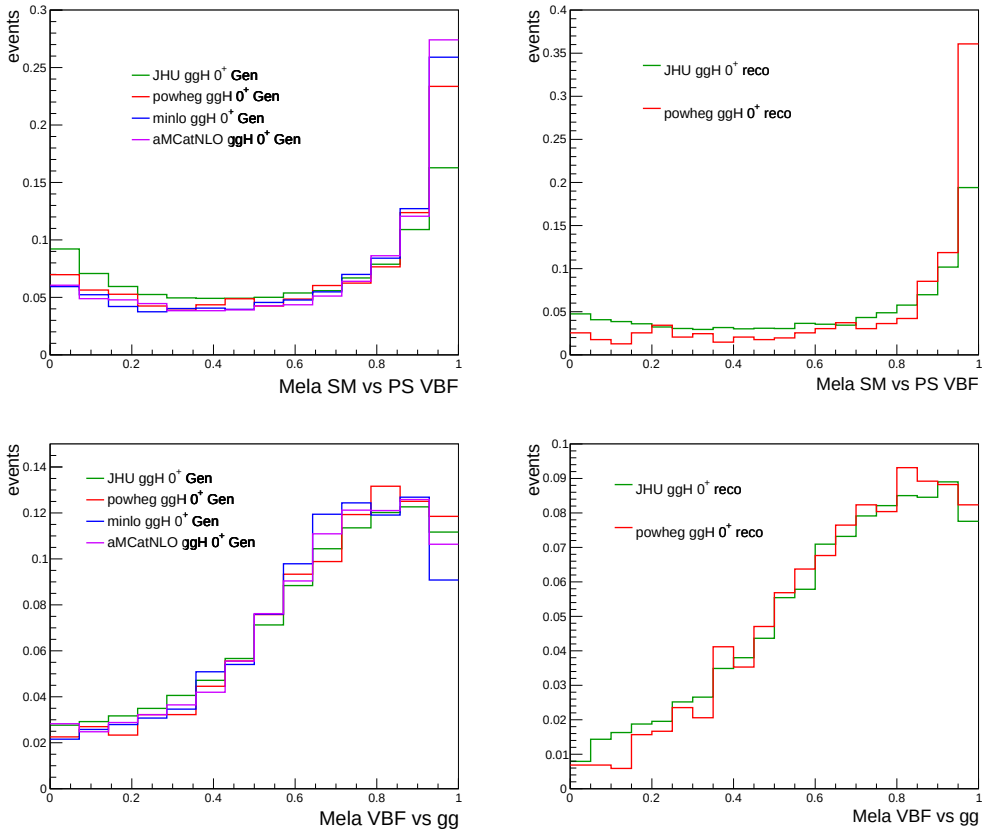


Figure 5.15: Distributions of the two kinematic discriminants, Mela SM vs PS VBF (top) and Mela VBF vs gg (bottom), for the ggH 0^+ process and both for generator (left) and reconstruction (right) level are presented. JHU, Powheg, MiNLO and aMC@NLO generators are compared at generator level, JHU and Powheg generators are compared at reconstruction level.

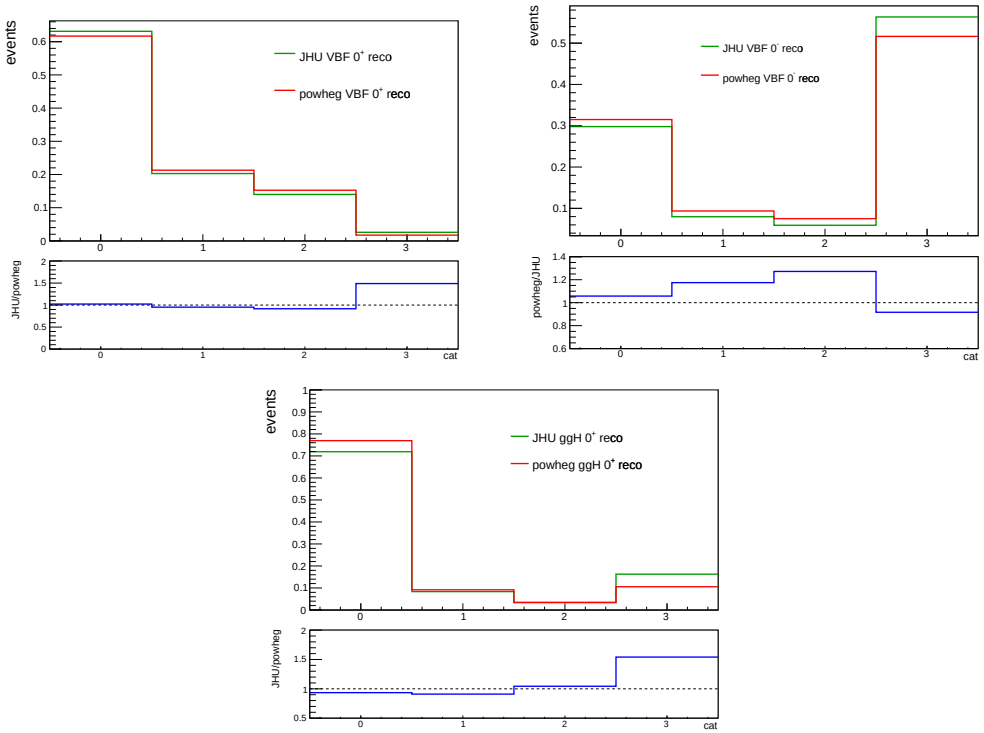


Figure 5.16: Distribution of the event yield in the 4 categories for the JHU (green) and Powheg (red) generators and their ratio, for VBF 0^+, VBF 0^- and ggH 0^+. The blue line is taken as systematics.

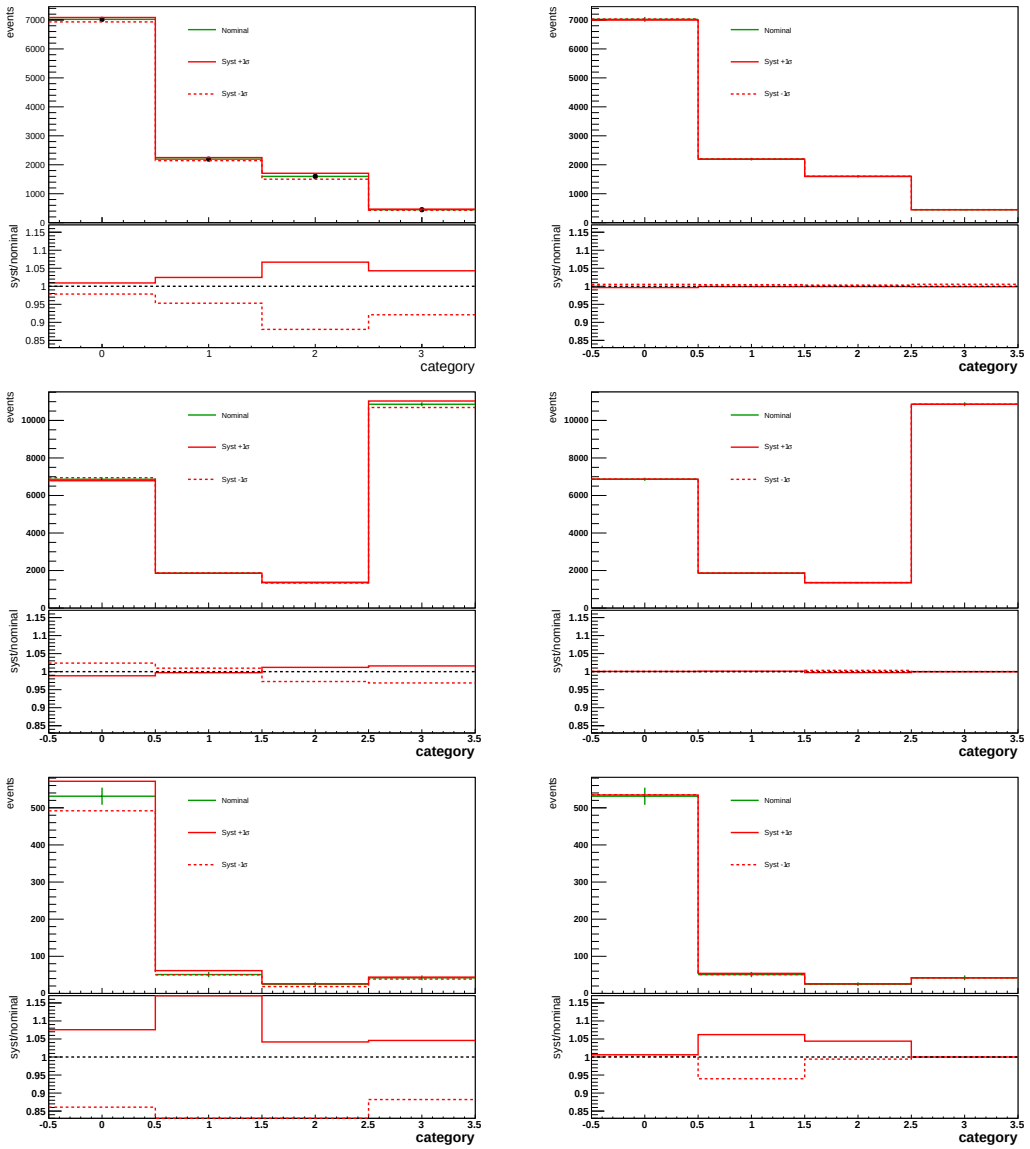


Figure 5.17: Distribution of the event yield in the 4 categories for the nominal sample and its variations of $\pm 1\sigma$. At the bottom the ratios between the nominal sample and the $\pm 1\sigma$ samples are also shown. Plots on the left represent the JEC, plots on the right the JER. Plots are done for the different processes VBF 0^+ (top), VBF 0^- (middle) and ggH 0^+ (bottom).

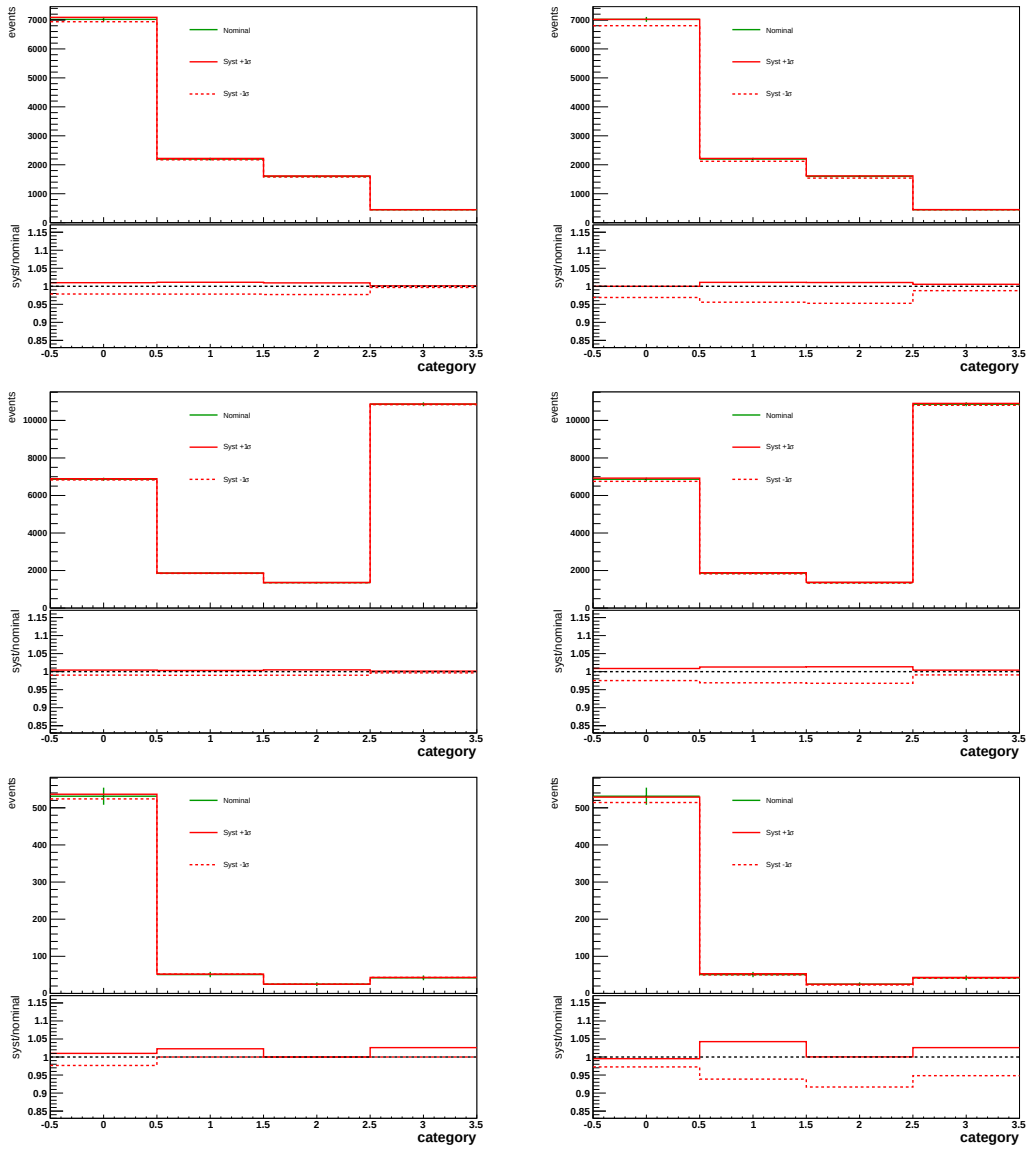


Figure 5.18: Distribution of the event yield in the 4 categories for the nominal sample and its variations of $\pm 1\sigma$. At the bottom the ratios between the nominal sample and the $\pm 1\sigma$ samples are also shown. Plots on the left represent the error derived from the photon ID MVA, plots on the right the error derived from the per-photon energy resolution. Plots are done for the different processes VBF 0^+ (top), VBF 0^- (middle) and ggH 0^+ (bottom).

5.10 Results

5.10.1 Expected and observed number of events

Table 5.4 shows the expected number of signal events and the corresponding number of events in data for each category. In Table 5.5 we provide the expected number of signal events and the corresponding number of events in data also for the cut-based analysis of Reference [13, 82] compared to the expected and observed number of events that we find making the same categorisation (2 VBF categories, one tight and one loose). One can see that we are able to reproduce the results of the standard analysis at 8 TeV [13, 82].

Table 5.4: The expected number of signal events and the corresponding number of events in data per category.

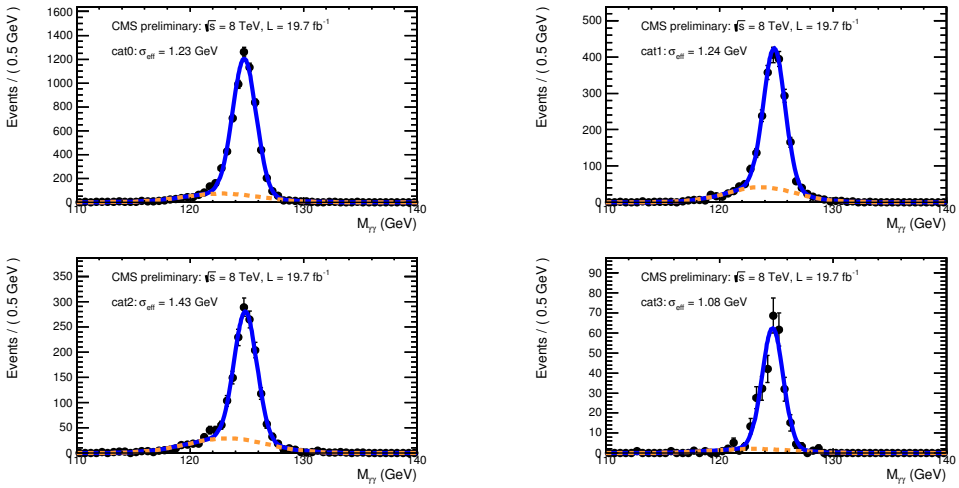
Category	VBF 0^+	VBF 0^-	ggH	bkg
cat 0	5.079	3.983	4.658	509
cat 1	1.758	1.078	0.4458	39
cat 2	1.277	0.7858	0.2181	16
cat 3	0.2276	6.301	0.3682	42

Table 5.5: The expected number of signal events and the corresponding number of events in data per category for cut-based analysis of Reference [13, 82] and for our analysis where we applied the same categorisation.

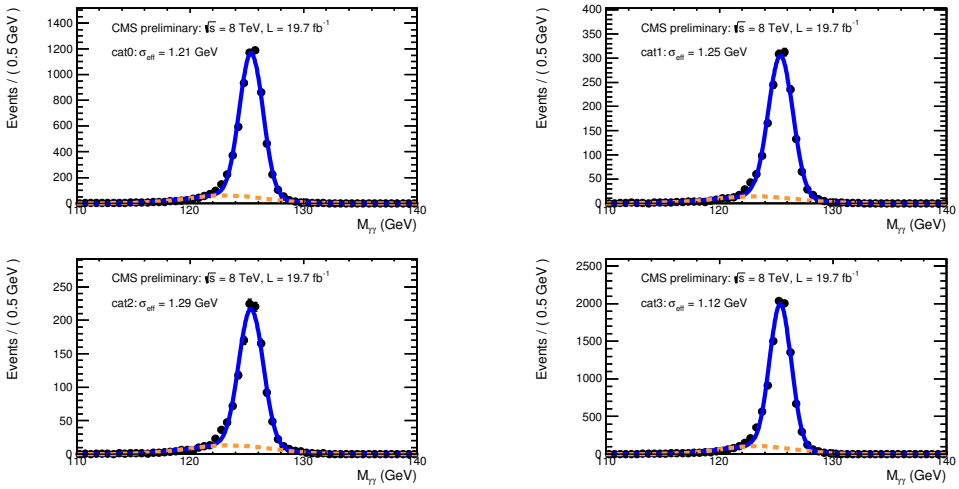
Category	VBF	VBF reference	ggH	ggH reference	bkg	bkg reference
cat tight	7.199	7.19	2.127	1.95	352	353
cat loose	5.273	5.2	5.567	5.15	1033	1029

5.10.2 Fit to the diphoton mass

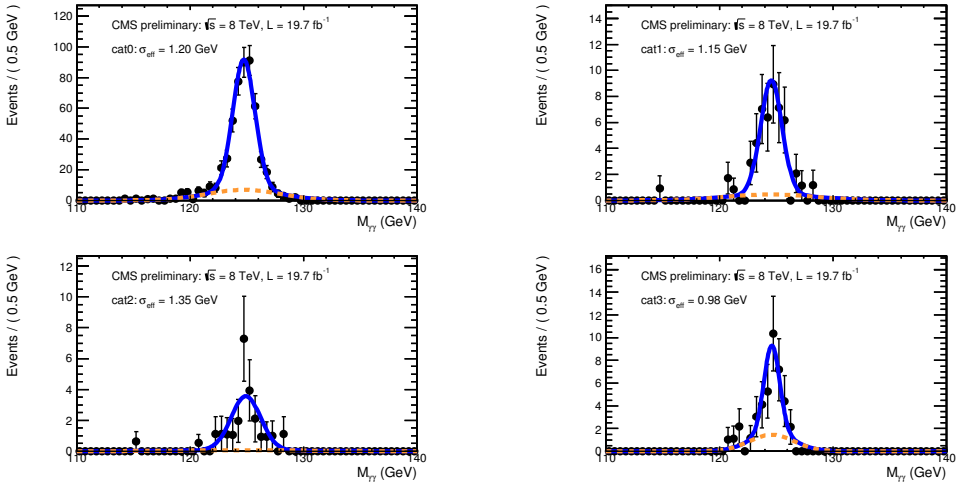
After the definition of the 4 final 2D categories and the choice of the background function for each of them, a fit to the diphoton mass was performed in each category. The simulated signal, for the processes VBF 0^+, VBF 0^- and ggH 0^+, is fitted with a double Gaussian. The data, that constitute our background, are instead fitted with a polynomial function of different order according to the category. The analysis was performed in the invariant mass range 100 GeV $< m_{\gamma\gamma} <$ 180 GeV, blinding the region 115 GeV $< m_{\gamma\gamma} <$ 135 GeV. With the fit to the diphoton mass we extracted simultaneously the yield for each process in each category. In Figure 5.19 the fits to the signal simulated samples in the 4 categories are shown, for the different processes, VBF 0^+, VBF 0^- and ggH 0^+. Figure 5.20 shows instead the unblinded fit to the data, along with the diphoton mass shapes of the simulated signals.



(a)



(b)



(c)

Figure 5.19: Fit results for the signal shape of simulated Higgs events in the 4 categories, for VBF 0^+ (a), VBF 0^- (b) and ggH 0^+ (c).

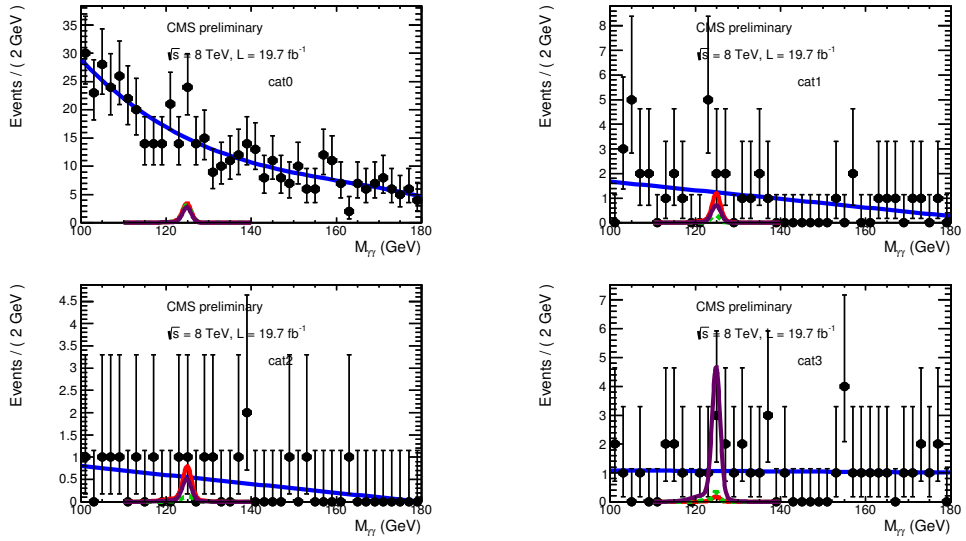


Figure 5.20: Fit to the data, in the invariant mass region 100 GeV $< m_{\gamma\gamma} <$ 180 GeV, in the 4 categories. The fitting function (blue solid line) is a polynomial of variable order according to the category (see Section 5.8). The diphoton mass shapes of the simulated signals are also shown, VBF 0^+ in red, VBF 0^- in violet and ggH 0^+ in green.

5.10.3 Parameter scan

We performed a scan over $\tilde{f_{a3}}$ values from 0 (pure scalar) to 1 (pure pseudoscalar). At each point in the scan, we maximised the binned likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \frac{n_i^{\text{sig}}(\vec{\zeta}) + n_i^{\text{bkg}}}{n_i!} e^{-n_i^{\text{sig}}(\vec{\zeta}) - n_i^{\text{bkg}}}, \tag{5.30}$$

where n_i^{sig} (n_i^{bkg}) is the expected number of signal (background) events and n_i is the observed number of events in bin i of N bins. The expected number of signal events depends on the anomalous coupling parameters, $\vec{\zeta}$. Systematic uncertainties described in Section 5.9 are introduced with nuisance parameters. Deviations from the global minimum in the negative log likelihood distribution, $-2\Delta\ln\mathcal{L}$, are used to quantify consistency of the pseudoscalar hypothesis with the data. Expected results are determined from a fit to an Asimov dataset with the SM Higgs included ($\mu = 1$).

5.10.4 1D scans

Figure 5.21 shows the likelihood scan as a function of μ for our analysis, where the cut-based categorisation of Reference [13, 82] has been applied, for the tight category on the left and for the loose category on the right. The black line represents the expected μ and the red line the observed one. Comparing our observed values in the two cut-based categories with the one of Reference [13, 82] shown in the plot at the bottom, one can see that we obtain similar results.

Figure 5.22 shows the likelihood scan as a function of μ^{VBF} for our analysis, performed with the cut-based categorisation of Reference [13, 82] (blue line) and with the categorisation done with the MELA discriminants (red line). Plot on the left shows the expected results, plot on the right the observed results. From the plot on the left one can see that the categorisation with the MELA discriminants has a better expected sensitivity, while plot on the right shows that the analysis with the cut-based categorisation has an observed result closer to the expected.

Figure 5.23 shows the likelihood scan as a function of $\tilde{f}_{a3}$ on the left, and as a function of f_{a3} on the right, where f_{a3} was obtained applying the transformation described in Equation 5.23. The dashed line represents the expected result, the solid line the observed result. One can see that we do not reach any exclusion of a pure pseudo-scalar state (which would be $-2\Delta NLL = 3.84$), nevertheless the analysis has some sensitivity to f_{a_3} despite the fact that we do not have sensitivity to observe SM VBF signal.

We also performed a projection for the Run 2 LHC data. Figure 5.24 shows the ratios of LHC parton luminosities, as a function of the mass, for different initial states. We are interested in the blue dashed curve (13 TeV/ 8TeV ratio for $q\bar{q}$ initial state) at a mass of around 700 GeV, that is the mass of the Higgs + dijet system. For this value of the mass, the parton luminosity ratio 13 TeV / 8 TeV is $\sim$ 2.5. From that we can deduce that 250 fb^{-1} at 8 TeV correspond roughly to 100 fb^{-1} at 13 TeV, if we assume that the background scales as the signal. We scale our process yields accordingly and we obtain the projection of the likelihood scan as a function of $\tilde{f}_{a3}$ for 250 fb^{-1} at 8 TeV, shown in Figure 5.25. From this figure one can see that it would be possible to exclude $\tilde{f}_{a3} > 0.22$ at 95% CL.
In Figure 5.26 we show, as a reference, the likelihood scan as a function of f_{a3} obtained by the $H \to ZZ$ channel combining data at 8 and 7 TeV. From this figure one can see that with our analysis at 13 TeV it would be possible to reach results similar to the $H \to ZZ$ ones at 7+8 TeV.

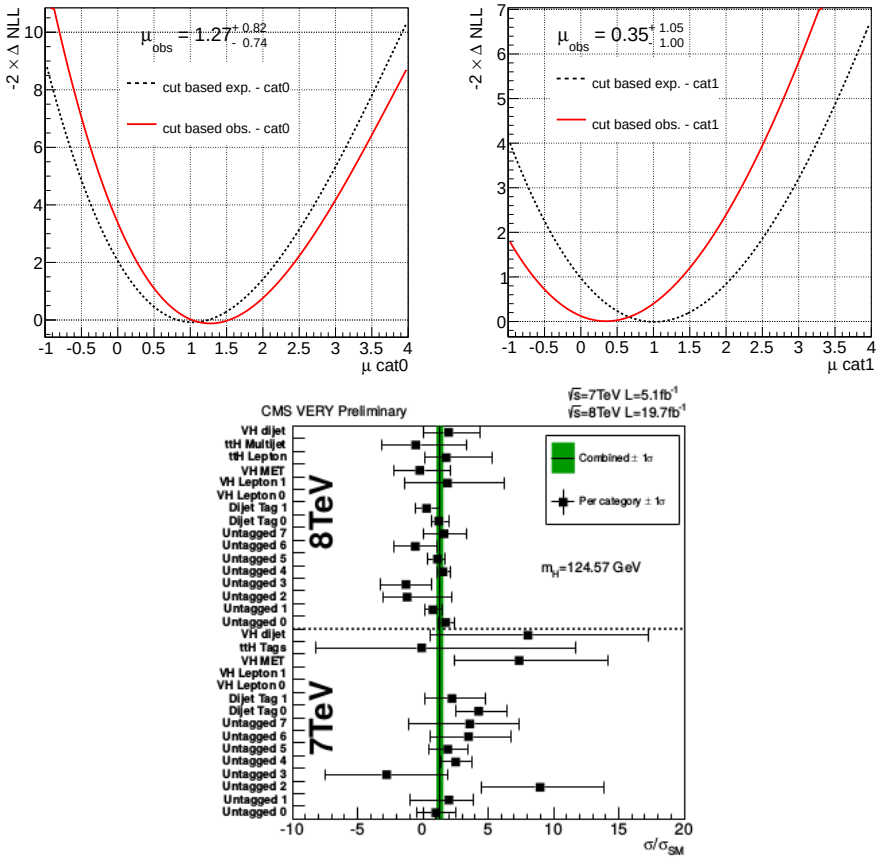


Figure 5.21: Likelihood scan as a function of μ for our analysis (top plots), where the cut-based selection of Reference [13, 82] has been applied, for the tight category on the left and for the loose category on the right. The black line represents the expected μ and the red line the observed one. The plot at the bottom shows the results of Reference [13, 82] at 8 TeV for the different categories. The results for the 2 Dijet Tag categories are similar to what we obtain.

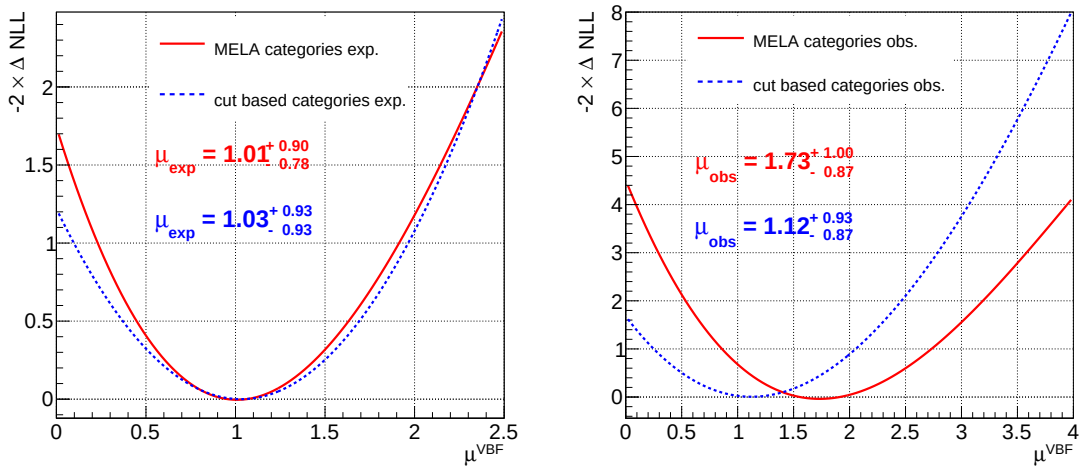


Figure 5.22: Likelihood scan as a function of μ^{VBF} for our analysis, performed with the cut-based categorisation of Reference [13, 82] (blue line) and with the categorisation done with the MELA discriminants (red line). Plot on the left shows the expected results, plot on the right the observed results.

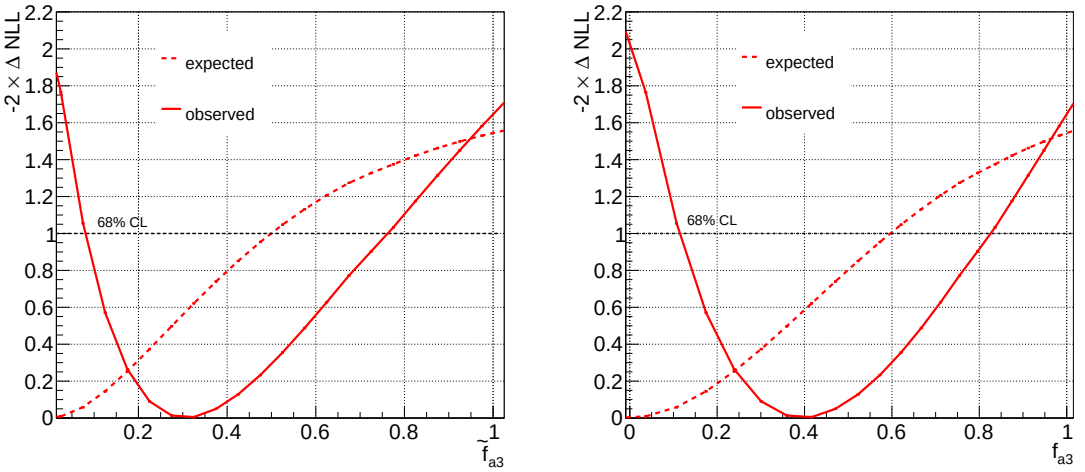


Figure 5.23: Likelihood scan as a function of $\tilde{f}_{a3}$ on the left, and as a function of f_{a3} on the right, where f_{a3} was obtained applying the transformation described in Equation 5.23. The dashed line represents the expected result, the solid line the observed result.

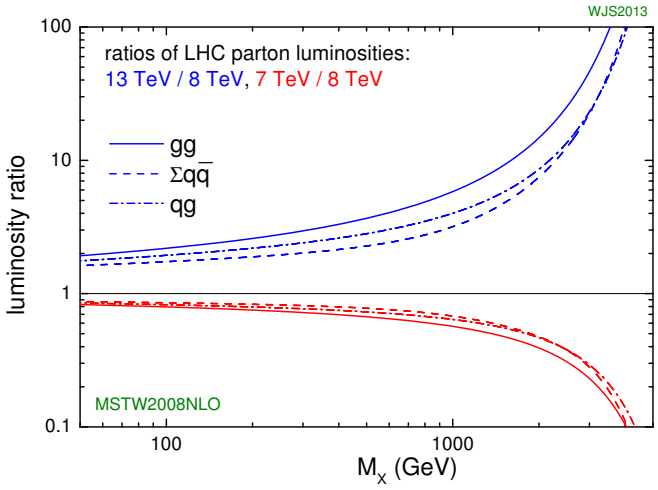


Figure 5.24: Ratios of LHC parton luminosities, 13 TeV/ 8 TeV in blue and 7 TeV/ 8 TeV in red, as a function of the mass, for different initial states.

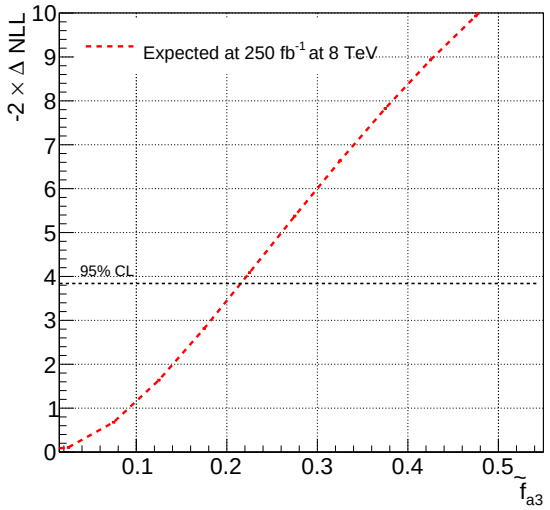


Figure 5.25: Projection of the limit on $\tilde{f}_{a3}$ for 250 fb^{-1} at 8 TeV, which would roughly correspond to 100 fb^{-1} at 13 TeV assuming that the background scales as the signal.

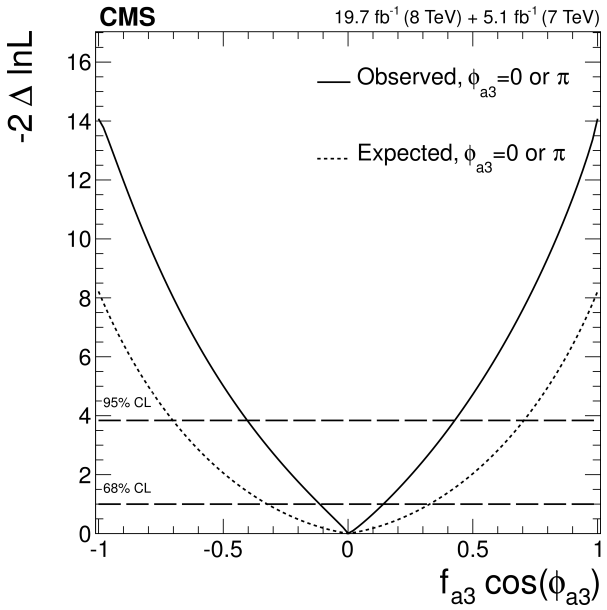


Figure 5.26: Likelihood scan as a function of f_{a3} obtained by the $H \to ZZ$ channel combining data at 8 and 7 TeV.

5.11 Summary

In this last chapter a feasibility study, having the aim of constraining the anomalous couplings of the Higgs boson to the vector bosons, is presented. The peculiarity of this study is the use of the Higgs boson production through vector boson fusion (VBF), instead of exploiting the decay of the Higgs in two vector bosons, as was done so far in other analyses. For this reason this is an alternate and complementary approach to the one usually employed. It was proven that despite the low event yield in the VBF channel, this new approach has an interesting sensitivity and should be pursued with larger dataset and a copier number of VBF events at 13 TeV.

Conclusions

The work presented in this thesis was carried out within the CMS Saclay group, working in the search of the Higgs boson in the decay channel H $\rightarrow \gamma\gamma$.
The main contributions of my studies throughout the period 2013-2016 are reported in Chapters 3, 4 and 5.

After the description of the Standard Model framework and the CMS detector, Chapter 3 describes in detail the photon reconstruction and identification at CMS. Despite many changes of the reconstruction algorithm between Run 1 and Run 2, the performance of the reconstruction from the photon identification point of view is found to be very similar between the two runs. The photon identification algorithm for the H $\rightarrow \gamma\gamma$ analysis is optimised for Run 2, using a multivariate analysis method, and its performance and validation studies are presented.

In Chapter 4 the H $\rightarrow \gamma\gamma$ analysis using the first Run 2 data is presented, with a particular accent on the simulation methods used to simulate the Monte Carlo samples used in this analysis. The analysis is performed with a dataset corresponding to an integrated luminosity of 12.9 fb^{-1} at 13 TeV. The observed significance for the standard model Higgs boson at $m_H = 125.09$ GeV is 5.6 σ, while 6.2 σ was expected, and the maximum significance of 6.1 σ is observed at $m_H = 126.0$ GeV. The best-fit signal strength relative to the Standard Model prediction is $\hat{\mu} = 0.95 \pm 0.20 = 0.95 \pm 0.17\ (\text{stat.})^{+0.10}_{-0.07}\ (\text{syst.})^{+0.08}_{-0.05}\ (\text{theo.})$ when the mass parameter is profiled in the fit, and $0.91 \pm 0.20 = 0.91 \pm 0.17\ (\text{stat.})^{+0.09}_{-0.07}\ (\text{syst.})^{+0.08}_{-0.05}$ (theo.) when it is fixed to $m_H = 125.09$ GeV.

Chapter 5 finally presents a feasibility study, having the aim of constraining the anomalous couplings of the Higgs boson to the vector bosons. This analysis is performed using the data collected at 8 TeV during Run 1 at the LHC, corresponding to an integrated luminosity of 19.7 fb^{-1}. It exploits the production of the Higgs boson through vector boson fusion (VBF), with the Higgs decaying to 2 photons, and it is an alternate and complementary approach to the one usually employed which is to study this coupling in the decay of the Higgs boson $H \rightarrow ZZ^*$ or $H \rightarrow WW^*$. It is found that even if at 8 TeV we do not reach any

exclusion of a pure pseudo-scalar state, the analysis has some sensitivity to f_{a_3}. Because of that this new approach should be pursued with larger dataset and a copier number of VBF events. A projection done for the Run 2 LHC data is also presented, showing that with 100 fb^{-1} at 13 TeV it would be possible to exclude a pure pseudo-scalar state at 95% CL.

Appendices

Appendix A
Tag and probe method

One well established approach to measure particle efficiencies is the so called tag and probe method. The tag and probe method uses the Z resonance to select an unbiased set of particles like electrons or muons. $Z \to ee$ events are selected with triggers requiring at least one electron. The tag electron is required to match a trigger level electron and to pass a tight selection requirement. This allows to select a very pure set of unbiased electrons/positrons. Indeed the other lepton from the Z decay, named probe lepton, is selected only with very loose constraints while still having a very pure sample (the invariant mass of the 2 leptons is required to be compatible with the Z mass to improve further the purity of the sample). One can then test identification criteria on this probe lepton in both data and simulated events. This allows to correct the simulation in order to reproduce the efficiency of the selection criteria observed in data.
The efficiency itself is measured by counting the number of "probe" particles that pass the desired selection criteria:

$$\epsilon = \frac{P_{pass}}{P_{all}},$$

where P_{pass} is the number of probes passing the selection criteria and P_{all} is the total number of probes counted using the resonance.

Appendix B
Details on photon energy correction

Photon energy regression

Different sets of corrections to ECAL reconstructed hits and photon energy are necessary in order to achieve the best photon energy resolution. The first consists of crystal-level corrections necessary to equalize the channel-to-channel response variations. The second, an high-level correction method called photon energy regression, is applied in order to take into account finer effects. In particular this method corrects for the containment of the

shower in the clustered crystals, and the energy losses of photons which convert in material upstream of the calorimeter. Starting from the raw supercluster energy, this technique aims to make the best prediction of the true photon energy, based on a multivariate approach. Furthermore it provides a per-photon energy resolution estimator, which is used for the diphoton BDT as described in Section 4.8. The regression is trained, separately for barrel and endcap, on photons in a sample of simulated events.

The input variables to the regression are the following (with more details in Section 3.2.2):

- shower shape variables: R_9, the number of reconstructed clusters, energy weighted η-width and Φ-width of the supercluster, $\text{cov}_{i\eta i\Phi}$, the ratio of the seed cluster energy to the supercluster energy and H/E. In the endcaps, preshower σ_{RR} is used as well. These variables are sensitive to global containment effects and provide information on the conversion probability and on the degree of showering,
- energy ratios within the seed cluster: these are sensitive to the local energy distribution and to the local containment effects,
- absolute position variables: seed crystal indices $i\eta$ and $i\Phi$, the difference between the seed crystal position and the supercluster position, variables comparing the seed crystal indices to the modules and supermodules boundaries in the barrel. These variables are sensitive in particular to the gap and crack energy loss,
- pile-up variables: number of reconstructed vertices and ρ, used to take into account potential energy additions due to pile-up.

A BDT is used to implement this regression, and it is trained to reproduce the probability density of E_{True}/E_{Raw} for any photon with input variables $\vec{x}$. In each leaf of the BDT the output E_{True}/E_{Raw} is fitted by a double-sided Crystal Ball function, whose parameters are functions of the input variables $\vec{x}$. The advantage of this technique is to be able to estimate simultaneously the energy of the photon and its median uncertainty. An example of the performance of the regression method is given in Figure 27, which shows the comparison of the ratios of the photon raw and corrected energies to the true energy, for photons from a simulated sample of Higgs to gamma-gamma events with $m_H = 125$ GeV.

Energy correction between data and simulation

The imperfect simulation of detector effects causes discrepancies in the scale and resolution of regression photon energy between data and Monte Carlo simulation. These discrepancies are dealt with by correcting the energy scale in data and by then determining a smearing to be applied to the simulated samples in order to have the best agreement between data

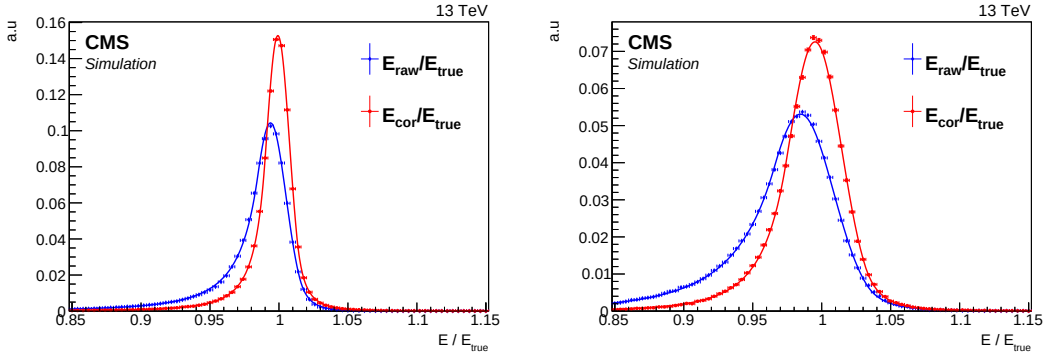


Figure 27: Ratio of photon raw energy to its true energy compared to the ratio of the photon energy corrected by the regression to its true energy, considering photons from a simulated Higgs to gamma-gamma sample (m_H = 125 GeV), in the barrel (left) and in the endcaps (right). The distributions are fitted with a double-sided Crystal Ball function.

and simulation for photon energy scale and resolution. The corrections are derived from $Z \to ee$ events from data and simulation with the electron energy esitmated in the same way as a photon energy, and performed in a multi-step procedure.

The supercluster energy scale is corrected for by varying the scale in the data to match the simulation in $Z \to ee$ events. The extraction of the energy scale corrections is done using two methods: the "fit method" and the "smearing method". The first one consists in performing an analytic fit to the Z invariant mass peak, built with the supercluster energies corrected by the regression. The fit is performed using a Breit-Wigner function convoluted with a Crystal Ball, where the Breit-Wigner function models the intrinsic distribution of $Z \to ee$, with the peak mass and width parameters fixed to the Particle Data Group values (m_Z = 91.188 GeV and Γ_Z = 2.495 GeV). The Crystal Ball function gives a reasonable description of the calorimeter resolution effects and bremsstrahlung losses in front of the calorimeter, and its parameters are left free during the fit. Data and simulated distributions are fitted separately and the fit results are compared to extract the scale offset. The relative shift between data and simulation is given by:

$$\Delta P = \frac{\Delta m_{data} - \Delta m_{MC}}{m_Z},$$

where $\Delta m_{MC(data)}$ is the fitted mean of the Gaussian core of the Crystal Ball function for simulation (data).
Since the data-simulation difference is time dependent and the time dependence is not the same in different pseudorapidity regions, the scale correction is extracted, with the fit method, per run range and per pseudorapidity region (4 bins, two in EB and two in EE). Then, once the scale has been corrected in the different categories, the residual data-MC

difference is addressed looking separately at the two R_9 bins (non-showering and showering electrons), in every pseudorapidity region, in order to factorize the effect of the material in front of the calorimeter. In this second step the scale corrections and the smearing of the MC energies are derived using the "smearing method".

The aim of the "smearing method" is to estimate more precisely the smearing to be applied to the MC energies. This method uses the invariant mass shape in the simulation as a pdf for a maximum likelihood fit instead of using an a-priori chosen function. The advantage is that all the known detector effects, reconstruction inefficiencies and Z kinematic behaviour are already taken into account in the simulation. The discrepancies between data and simulation are found to be successfully covered by a Gaussian smearing function. Another big advantage of this method with respect to the first one is that a larger number of electron categories can be handled, allowing to include also the events with the two electrons having different properties. The method is based on the maximization of the likelihood between the data and the smeared simulation in the $Z \to ee$ invariant mass distribution in each electron category. In the likelihood maximization, the electron supercluster energies are modified by applying a Gaussian multiplicative factor centered in $1 + \Delta P$ and with a $\Delta\sigma$ resolution, where ΔP is the energy scale correction and $\Delta\sigma$ is the additional constant term in the energy resolution.

The final energy scale correction is therefore derived as the product of the two corrections in the number of run ranges $\times$ 4 (pseudorapidity regions) $\times$ 2 (R_9) categories.

Systematic uncertainties

The systematic uncertainty in the determination of the energy scale correction and the additional energy smearing to be applied to the $H \to \gamma\gamma$ simulation is dominated by the difference between electrons and photons.
A variable particularly sensitive to this difference is R_9: its variation as a function of η reflects qualitatively the variation of the tracker material budget and is an indication of how much the material upstream from the calorimeter affects differently electrons and photons. The lower R_9 requirement is thus varied both for the $R_9 > 0.94$ and the $R_9 < 0.94$ categories. The maximum absolute difference among all these variations with respect to the nominal value, defined as Δ_{R_9}, is then used in the evaluation of the systematic uncertainty. Furthermore, one can apply to electrons the regression trained for electrons, and apply to electrons reconstructed as photons the regression trained for photons, in order to account for possible differences between electrons and photons (this systematic uncertainty is defined as Δ_{pho}).
Other possible sources of systematic uncertainty are investigated: in particular the event selection is varied using the medium and tight working points for the electron identification

and the requirement on electron E_T is varied from 20 GeV to 25 GeV. These systematic uncertainties are called respectively Δ_{sel} and Δ_{E_T}.
The final estimate of the systematic uncertainty is obtained adding all these effects in quadrature:

$$\Delta^{tot}_{syst} = \sqrt{\Delta^2_{R_9} + \Delta^2_{pho} + \Delta^2_{sel} + \Delta^2_{E_T}}$$

The systematic uncertainty is computed in the different R_9-η categories, both for the energy scale and the energy resolution. Its values vary between 0.15% and 0.5%.

Appendix C
Details on diphoton vertex identification

Vertex Selection BDT

The vertex selection BDT is trained to select the diphoton production vertex in an event, which has an average of 18.5 pp collision vertices as a consequence of pileup interactions. The training is performed using simulation for all Higgs production modes weighted according to their expected cross sections. The signal sample consists of right vertices, matched to the Higgs MC truth vertices, while the background sample consists of the pileup vertices. The most discriminating variables for the vertex identification, entering the BDT, are the following:

- $\sum_i |\vec{p}^{\,i}_T|^2$: the sum of the square of transverse momentum of each track associated with the vertex, $\vec{p}^{\,i}_T$. This quantity is expected to be larger for the diphoton vertex than for pileup vertices,
- $-\sum_i \left(\vec{p}^{\,i}_T \cdot \frac{\vec{p}^{\,\gamma\gamma}_T}{|\vec{p}^{\,\gamma\gamma}_T|}\right)$: the negative sum of projections of the $\vec{p}_T$ of the tracks on the diphoton $\vec{p}_T$. This quantity tends to be positive for the true vertex, as tracks recoil against the diphoton, and centered on 0 for wrong vertices,
- $(|\sum_i \vec{p}^{\,i}_T| - p^{\gamma\gamma}_T)/(|\sum_i \vec{p}^{\,i}_T| + p^{\gamma\gamma}_T)$: the asymmetry between the total momentum of the tracks attached to a given vertex and the modulus of the diphoton p_T. This quantity tends to have higher values for the true vertex and to peak at -1 for wrong vertices,
- $|z_{\rm PV} - z^{\rm conv}_{\rm PV}|/\sigma_{\rm conv}$, only for cases with at least one converted photon: the distance between the z position of the vertex and the estimated position of diphoton vertex from the conversion and normalized by the uncertainty of the estimation. This quantity is expected to be near 0 for the diphoton vertex while larger for pileup vertices.

Figure 28 shows the efficiency of choosing the vertex within 1 cm of the true vertex from Higgs to two photons simulation as a function of the p_T of the diphoton pair and as a

function of the number of vertices. This figure also shows separately the efficiencies for events with at least one converted photon (red) and for events where no photon is converted (blue). It is evident that the use of converted photons allows to improve the efficiency of the vertex identification algorithm.

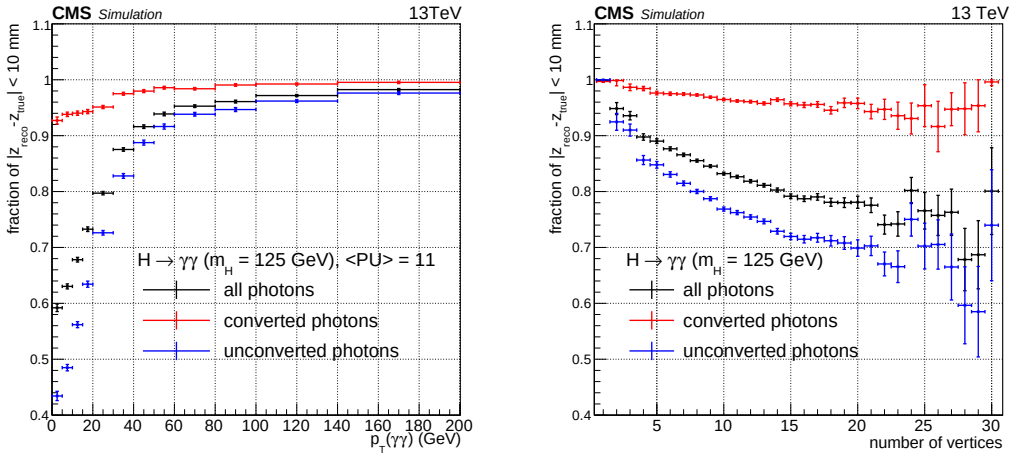


Figure 28: Efficiency of choosing the vertex within 1 cm of the true vertex from Higgs to two photons simulation as a function of the p_T of the diphoton pair (left) and as a function of the number of vertices (right).

Vertex probability BDT

Vertex probability BDT is trained to estimate the probability that the selected vertex is the correct diphoton vertex for each event. The criterion for being correct is that the distance between the selected vertex and the true diphoton vertex is within 1 cm in the z direction, in which case the resolution on the photon opening angle makes a negligible contribution to the diphoton mass resolution. Figure 29 shows the mass resolution worsening as a function of the z distance to the true vertex. For the events with the best energy resolution (two photons in the ECAL barrel with high R_9), the resolution worsens by about 15% when the vertex is between 1 cm and 1.5 cm of the true one, and by more than 20% between 1.5 cm and 2 cm.

The probability to choose the vertex within 1 cm of the true vertex is computed from the output of a second MVA obtained using a BDT method, as for vertex identification. It is used for the diphoton BDT as described in Section 4.8. The training is performed on a Monte Carlo simulation of H $\rightarrow \gamma\gamma$ events. The signal sample consists of events with correct vertex selected, and the background sample consists of events with wrong vertex selected. The input variables are the following:

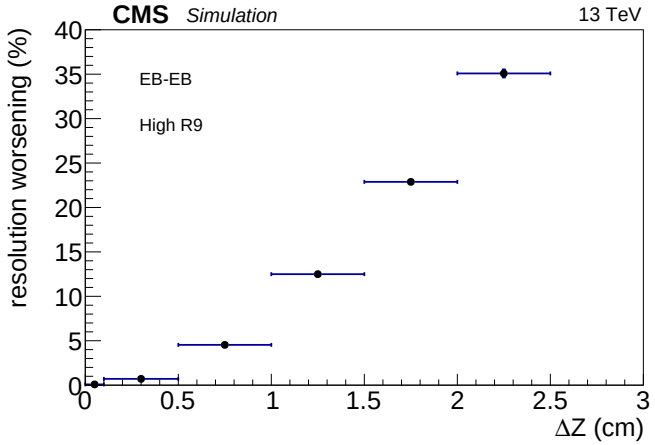


Figure 29: Mass resolution worsening in percent as a function of the z difference between the toy vertex and the true vertex. The plot is done for the events with two photons in the barrel and high R_9.

- the p_T of the diphoton system,
- the number of vertices,
- the values of the vertex identification BDT discriminant for the best three vertices,
- the Δz between the best vertex and the second and third choices,
- the number of converted photons (0, 1, or 2).

The BDT is then transformed to a probability computed for both converted and unconverted photons.

To measure the performance of both the vertex selection BDT and the vertex probability BDT, the diphoton vertex selection efficiency and the average vertex probability are evaluated on Monte Carlo simulated H $\to \gamma\gamma$ events at a Higgs mass of 125 GeV, in bins of $p_T^{\gamma\gamma}$. As shown in Figure 30, the average vertex probability along with uncertainty (blue band) predicts well the measured vertex selection efficiency (data points), and both increase with the increasing $p_T^{\gamma\gamma}$. The total vertex selection efficiency is ∼82% for the H $\to \gamma\gamma$ events at a Higgs mass of 125 GeV.

Systematic uncertainties

The systematic uncertainty in the vertex finding efficiency is taken from the uncertainty in the measurement of the corresponding data/simulation scale factor obtained using $Z \to \mu\mu$ events. In effect $Z \to \mu\mu$ events are used to validate the H $\to \gamma\gamma$ vertex identification

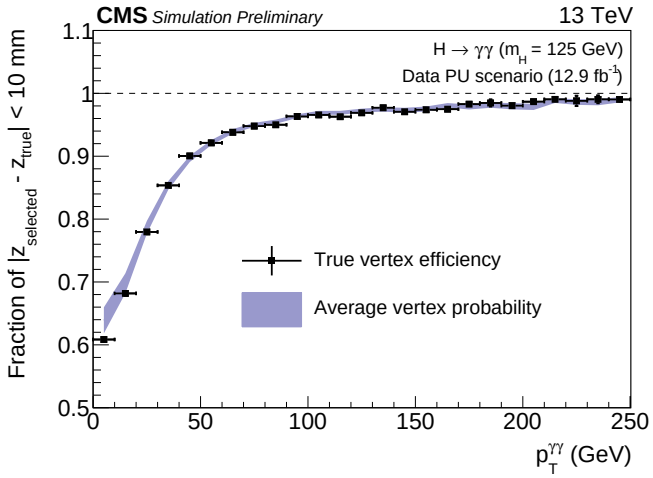


Figure 30: Comparison between the true vertex identification efficiency and the average estimated vertex probability as a function of the reconstructed diphoton p_T. All the production modes are included, the reweighting is done according to the cross-sections of the different production modes and with respect to pileup in data.

algorithm for the case of unconverted photons. To do that the vertices are first refitted without the muon tracks, mimicking a diphoton system. Then the muon tracks are removed from the tracks used by the identification algorithm and the efficiency of finding the good vertex is estimated both in data and simulation. Thus the vertex is selected within the vertex collection without muon tracks, while the true vertex is determined from the muon tracks.

Events with 2 muons are selected if they satisfy all tight identification criteria, except the ones involving vertex, and in addition the dimuon mass is required to be within 70 GeV and 110 GeV.

The data/simulation efficiencies ratio as a function of p_T, shown in Figure 31, is used to correct the efficiencies in $H \to \gamma\gamma$ simulation, and it is varied within uncertainties to estimate the associated systematics.

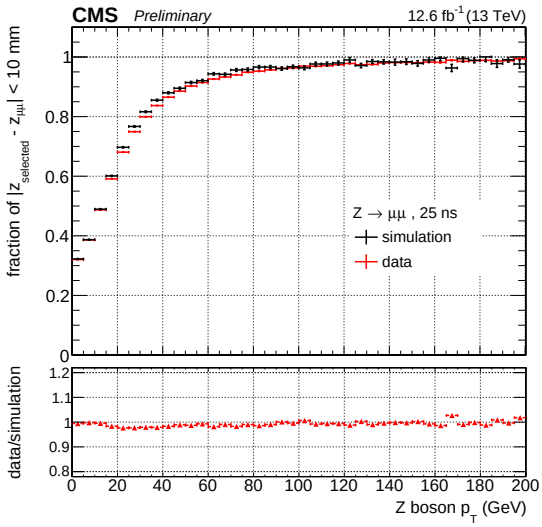


Figure 31: Efficiency as a function of p_T to find the vertex within 1 cm of the true one, using $Z \to \mu\mu$ events for data and simulation.

Résumé substantiel

Dans ce document, deux analyses des propriétés du boson de Higgs se désintégrant en 2 photons dans l'expérience CMS située auprès du LHC (Large Hadron Collider) sont présentées.

Le Modèle Standard, théorie quantique des champs qui a pour but la description des constituants élémentaires de la matière, a été testé expérimentalement de manière exhaustive. Développé dans les années 1960-1970, il fournit un élégant cadre mathématique qui décrit la façon dont les constituants fondamentaux de la matière intéragissent entre eux à travers les forces électromagnétique, faible et forte. De plus, il a expliqué et prédit avec succès nombre de résultats expérimentaux et une grande variété de phénomènes.
Le Modèle Standard de la physique des particules prévoit l'existence d'un seul boson scalaire de Higgs associé à la brisure spontanée de symétrie. La masse de ce boson est un paramètre libre de la théorie. De plus le champ de Higgs est considéré comme responsable de la masse de toutes les particules élémentaires connues. Cette particule, dont la recherche est l'un des objectifs principaux du collisionneur de hadrons LHC installé au CERN de Genève et de ses expériences, a enfin été découverte par les expériences ATLAS et CMS en 2012, avec une masse mesurée d'environ 125 GeV. Les objectifs principaux après la découverte du boson de Higgs sont la mesure de ses propriétés et la recherche de possibles déviations du Modèle Standard.
La production inclusive du boson de Higgs du Modèle Standard, suivie par la désintégration en 2 photons, est un des canaux les plus sensibles pour la recherche et l'étude d'un boson de Higgs de masse d'environ 125 GeV. En effet, malgré sa section efficace modeste, il a une signature expérimentale très claire grâce à une excellente résolution expérimentale de la masse diphoton.

L'expérience CMS est, avec ATLAS, un détecteur à large spectre. La caractéristique principale de CMS est un solénoïde supraconducteur, de 13 m de long et 6 m de diamètre, qui fournit un champ magnétique axial de 3.8 T. Le solénoïde contient la majorité des sous-détecteurs, alors que en déhors du solénoïde la culasse de retour de champ est équipée de détecteurs à gaz utilisés pour identifier les muons. Les trajectoires des particules chargées sont mesurées par le trajectographe, composé de pixels et strips en silicium, couvrant les pseudo-rapidités $|\eta| < 2.5$. Un calorimètre électromagnétique (ECAL), composé de crystaux de tungstate de plomb, et un calorimètre hadronique (HCAL), composé de cuivre et de scintillateurs, entourent le trajectographe et couvrent la région $|\eta| < 3$. Le calorimètre électromagnétique est composé de 75848 crystaux de tungstate de plomb, qui permettent une couverture en pseudorapidité $|\eta| < 1.479$ dans la région du tonneau et $1.479 < |\eta| < 3.0$ dans les régions des bouchons. Un détecteur de pied de gerbe, composé par 2 couches de capteurs de silicium intercalées avec un radiateur en plomb, est placé face aux bouchons. Un calorimètre à l'avant prolonge la couverture calorimétrique jusqu'à $|\eta| < 5$. La

calibration du ECAL utilise la symétrie axiale du flux d'énergie dans des événements de biais minimum, les désintégrations π^0, $\eta^0 \to \gamma\gamma$, $W \to e\nu$, et $Z \to ee$. Les variations de transparence des crystaux du ECAL dûes à l'irradiation pendant les périodes de prise de données sont monitorées constamment et corrigées, en utilisant de la lumière injectée par un laser. La résolution du ECAL est de 1% à $|\eta| = 0$ pour des photons non-convertis et se dégrade jusqu'à 4% pour des photons convertis à $|\eta| = 2.5$.

Le défi majeur du canal $H \to \gamma\gamma$ est d'observer un signal noyé dans un bruit de fond plusieurs ordres de grandeurs plus grand. Le bruit de fond peut être séparé en 2 composantes, une "irréductible" et une "réductible". La composante "irréductible" est constituée d'événements diphoton "prompt", c'est à dire un photon issu du vertex de l'intéraction forte. La composante "réductible" contient des événements dijet et γ + jet, où les jets sont identifiés comme des photons (fake). L'identification des photons est un enjeu capital de l'analyse, et elle a le but de réduire le bruit de fond "réductible", en général en appliquant des conditions d'isolation et en exploitant les différences entre les profils des gerbes dans le calorimètre pour distinguer entre un photon prompt et un couple de photons provenant de la désintégration d'un méson neutre.
Dans cette thèse les algorithmes de reconstruction et d'identification des photons sont présentés, avec une attention particulière aux différences entre le premier et le deuxième run du LHC, le premier run (Run 1) ayant été pris entre 2010 et 2012 avec une énergie dans le centre de masse de 7 puis 8 TeV, le deuxième (Run 2) ayant commencé en 2015 avec une énergie dans le centre de masse de 13 TeV. La principale nouveauté présente dans la reconstruction du Run 2 est l'utilisation de l'algorithme du flux de particules, qui permet une description globale de l'événement. Les performances des reconstructions du Run 1 et du Run 2 en ce qui concerne l'identification des photons pour le Run 2 sont comparées et on obtient des performances similaires. Ensuite l'algorithme d'identification des photons pour l'analyse $H \to \gamma\gamma$ est optimisé pour le Run 2. Pour ce faire, une méthode d'analyse multivariée, qui prend en considération des variables discriminantes liées au profil de la gerbe et à l'isolation, est utilisée. Les performances de l'identification des photons pour le Run 2 sont enfin étudiées et une validation données-simulation est effectuée. La Figure 32 représente l'efficacité pour le bruit de fond en fonction de l'efficacité pour le signal, en bleu pour la nouvelle identification des photons et en rouge pour l'identification utilisée dans l'analyse du Run 1 et appliquée aux échantillons du Run 2. Cette figure montre que la nouvelle identification a une meilleure performance par rapport à l'ancienne et démontre clairement l'avantage d'une identification des photons dédiée à la nouvelle énergie. La Figure 33 représente une validation données-simulation pour la variable de sortie de l'analyse multivariée. Cette variable est représentée, dans un interval de masse invariante $100 < m_{\gamma\gamma} < 180$ GeV, pour les données et les différentes composantes du bruit de fond. La somme des composantes du bruit de fond est compatible avec les données, mais quelques désaccords sont visibles pour des valeurs de la variable proches de 1. Ces désaccords sont

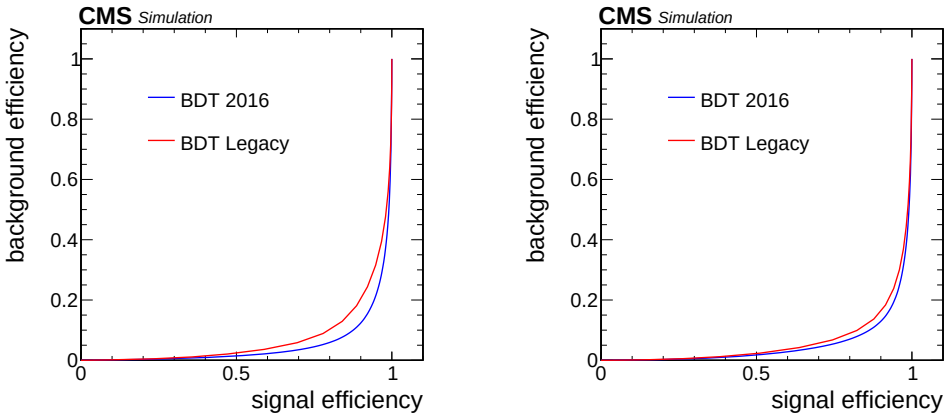


Figure 32: Courbe de l'efficacité de l'identification des photons pour le bruit de fond en fonction de l'efficacité pour le signal, en bleu pour la nouvelle identification et en rouge pour l'identification utilisée dans l'analyse à 8 TeV et appliquée aux échantillons à 13 TeV. La figure de gauche fait référence au tonneau, celle de droite aux bouchons.

traités par les incertitudes systématiques.

Les résultats principaux de l'analyse H $\rightarrow \gamma\gamma$ du Run 2 sont ensuite présentés, avec un accent particulier sur le rôle joué par les échantillons simulés dans la construction du modèle de signal, l'optimisation de la sélection et l'entrainement de différents discriminants. L'analyse H $\rightarrow \gamma\gamma$ présentée a été effectuée avec les données du Run 2 enregistrées en 2016, correspondantes à une luminosité intégrée de 12.9 fb^{-1}. Une catégorisation des événements est faite, afin de rendre maximale la signification statistique du signal et d'étudier les différents modes de production du boson de Higgs. La signification statistique observée pour le boson de Higgs du Modèle Standard à $m_H = 125.09$ GeV est 5.6 σ, pour une signification attendue de 6.2 σ, et la signification maximale de 6.1 σ est observée à $m_H = 126.0$ GeV. La Figure 34 représente la distribution de la masse diphoton où chaque événement est repondéré selon S/(S+B), où S et B indiquent respectivement le nombre d'événements de signal et de bruit de fond attendus. La Figure 35 représente la signification attendue et observée pour l'observation d'un boson de Higgs du Modèle Standard en fonction de m_H.

Enfin une étude de faisabilité ayant pour but de contraindre les couplages anomaux du boson de Higgs aux bosons de jauges est présentée. En effet, le boson de Higgs du MS doit avoir pour spin-parité 0^+. Si l'hypothèse 0^- est aujourd'hui rejetée (par l'étude de la désintégration $H \rightarrow ZZ \rightarrow 4\ell$), une faible contribution de spin-parité 0^- n'est pas exclue. Cette analyse exploite la production du boson de Higgs par fusion de bosons-vecteurs (VBF), avec le boson de Higgs se désintégrant en 2 photons. Les distributions cinématiques des jets

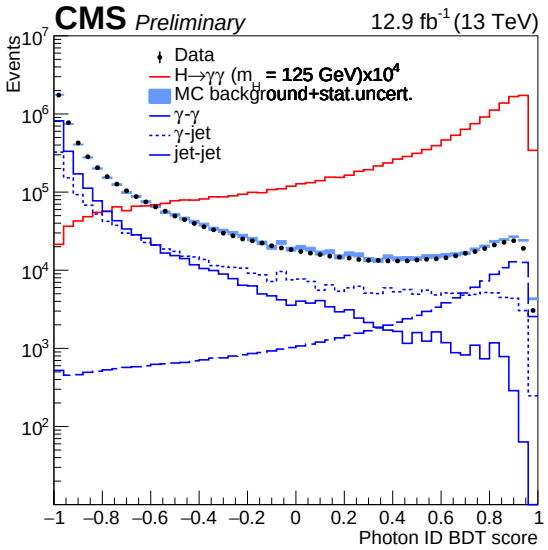


Figure 33: Variable de sortie de l'analyse multivariée pour photons appartenant à couples diphoton avec une masse invariante $100 < m_{\gamma\gamma} < 180$ GeV, pour les données (points) et pour la simulation du bruit de fond (histogramme cyan). Les distributions sont représentées aussi pour différentes composantes du bruit de fond, avec 2, 1 ou 0 photons prompt. La distribution rouge correspond aux événements de signal simulés.

et du Higgs, qui dépendent de l'hypothèse de spin-parité, sont utilisées pour construire des discriminants capables de séparer les différentes hypothèses de spin-parité. Ces discriminants permettent de définir différentes régions de l'espace des phases enrichies en signaux de spin-parité différente. Les différents nombres d'événements de signal sont extraits dans chaque région par un ajustement de la masse invariante diphoton, permettant de déterminer les contributions respectives des différents signaux et permettant ainsi de contraindre la production de boson de Higgs pseudo-scalaire (spin-parité 0^-). Cette analyse utilise les données collectées pendant le Run 1 du LHC, correspondant à une luminosité intégrée de 19.7 fb^{-1}. Avec cette quantité de données la sensibilité à la production VBF est faible, néanmoins cette approche est alternative et complémentaire à celle habituellement utilisée, qui étudie les mêmes couplages dans la désintégration du boson de Higgs $H \to ZZ^*$ ou $H \to WW^*$. Bien que le lot de données du Run 1 ne soit pas suffisant pour exclure un état pseudo-scalaire, l'analyse montre une sensibilité intéressante à exploiter avec plus de données, en particulier car il s'agit des couplages lors de la production du boson de Higgs. La Figure 36 à gauche représente le profil du likelihood, attendu et observé, en fonction de la fraction de composante pseudo-scalaire. La figure à droite montre la projection pour 250 fb^{-1} à 8 TeV.

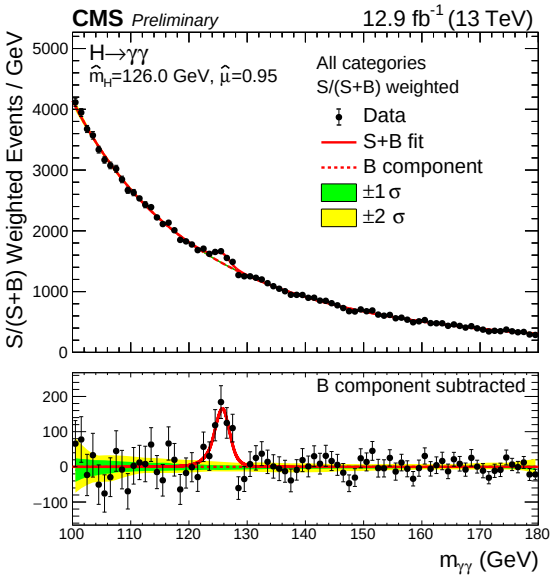


Figure 34: Distribution de la masse diphoton où chaque événement est pondéré par un poids selon S/(S+B). Les données (points noirs), l'ajustement sous l'hypothèse de seul bruit de fond (ligne rouge pointillée) et l'ajustement sous l'hypothèse S+B (ligne rouge continue) sont ici représentés. Les bandes 1σ (vert) et 2σ (jaune) contiennent les incertitudes de l'ajustement. La même distribution après la soustraction du bruit de fond est montrée en bas.

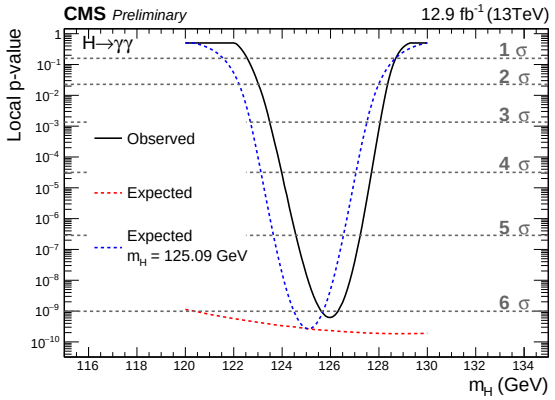


Figure 35: La signification observée (noir) est comparée à celle attendue dans l'interval 120-130 GeV, où on suppose que le boson de Higgs a une masse $m_H = 125.09$ GeV (bleu). La ligne rouge indique la signification statistique attendue pour chaque hypothèse de masse dans l'interval 120 GeV $< m_H <$ 130 GeV.

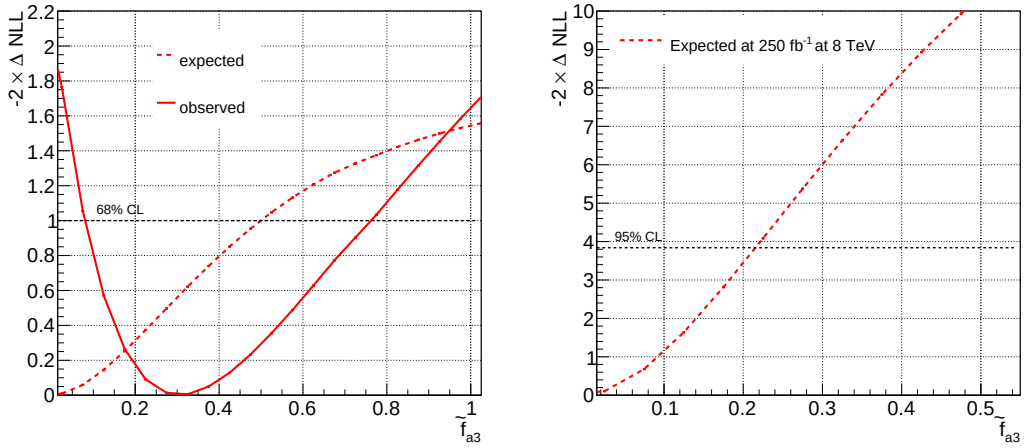


Figure 36: Profil du likelihood, attendu et observé, en fonction de la fraction de composante pseudo-scalaire qu'on veut contraindre (gauche) et likelihood scan attendu, en fonction de la même quantité, pour une projection correspondante à 250 fb^{-1} à 8 TeV (droite).

Bibliography

[1] F. Halzen and Alan D. Martin. Quarks and leptons: An introductory course in modern particle physics. 1984. New York, Usa: Wiley (1984) 396p.

[2] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961.

[3] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967.

[4] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321–323, 1964.

[5] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, 1964.

[6] E. Fermi. Trends to a Theory of beta Radiation. (In Italian). *Nuovo Cim.*, 11:1–19, 1934. [,535(1934)].

[7] S. Dawson. Introduction to electroweak symmetry breaking. In *Proceedings, Summer School in High-energy physics and cosmology*, pages 1–83, 1998.

[8] Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, and Ernest Aguilo. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716(arXiv:1207.7235. CMS-HIG-12-028. CERN-PH-EP-2012-220):30–61. 59 p, Jul 2012.

[9] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012.

[10] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. 2011.

[11] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 2. Differential Distributions. 2012.

[12] J R Andersen et al. Handbook of LHC Higgs Sections: 3. Higgs Properties. 2013.

[13] Vardan Khachatryan, Albert M Sirunyan, and Tumasyan. Observation of the diphoton decay of the Higgs boson and measurement of its properties. *Eur. Phys. J. C*, 74(arXiv:1407.0558. CMS-HIG-13-001. CERN-PH-EP-2014-117):3076. 79 p, Jul 2014. Comments: Replaced with published version. Added journal reference and DOI.

[14] Serguei Chatrchyan et al. Measurement of the properties of a Higgs boson in the four-lepton final state. *Phys. Rev.*, D89(9):092007, 2014.

[15] Vardan Khachatryan et al. Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV. *Eur. Phys. J.*, C75(5):212, 2015.

[16] Georges Aad et al. Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments. *Phys. Rev. Lett.*, 114:191803, 2015.

[17] Dario Buttazzo, Giuseppe Degrassi, Pier Paolo Giardino, Gian F. Giudice, Filippo Sala, Alberto Salvio, and Alessandro Strumia. Investigating the near-criticality of the Higgs boson. *JHEP*, 12:089, 2013.

[18] Vardan Khachatryan et al. Constraints on the Higgs boson width from off-shell production and decay to Z-boson pairs. *Phys. Lett.*, B736:64–85, 2014.

[19] Vardan Khachatryan et al. Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV. *Phys. Rev.*, D92(1):012004, 2015.

[20] M. Sullivan and J. Guy. Snls3: Constraints on dark energy combining the supernova legacy survey three-year data with other probes. *The Astrophysical Journal*, 737(2):102, 2011.

[21] A. D. Sakharov. Violation of CP Invariance, c Asymmetry, and Baryon Asymmetry of the Universe. *Pisma Zh. Eksp. Teor. Fiz.*, 5:32–35, 1967. [Usp. Fiz. Nauk161,61(1991)].

[22] C. Jarlskog. Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of Maximal CP Violation. *Phys. Rev. Lett.*, 55:1039, 1985.

[23] M. E. Shaposhnikov. Possible Appearance of the Baryon Asymmetry of the Universe in an Electroweak Theory. *JETP Lett.*, 44:465–468, 1986. [Pisma Zh. Eksp. Teor. Fiz.44,364(1986)].

[24] I. Aitchison. *Supersymmetry in Particle Physics: An Elementary Introduction.* Cambridge University Press, 2007.

[25] Jing Shu and Yue Zhang. Impact of a CP Violating Higgs Sector: From LHC to Baryogenesis. *Phys. Rev. Lett.*, 111(9):091801, 2013.

[26] Lisa Randall and Raman Sundrum. A Large mass hierarchy from a small extra dimension. *Phys. Rev. Lett.*, 83:3370–3373, 1999.

[27] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN, Geneva, 2004.

[28] *Technical proposal*. LHC Tech. Proposal. CERN, Geneva, 1994. Cover title : CMS, the Compact Muon Solenoid : technical proposal.

[29] *The Tracker Project Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1998.

[30] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[31] Amos Breskin and Rüdiger Voss. *The CERN Large Hadron Collider: Accelerator and Experiments*. CERN, Geneva, 2009.

[32] Serguei Chatrchyan et al. Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV. *JINST*, 8:P09009, 2013. [JINST8,9009(2013)].

[33] *The Hadronic Calorimeter Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[34] G. Acquistapace et al. CMS, the magnet project: Technical design report. 1997.

[35] *The CMS muon project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[36] Vardan Khachatryan, Sirunyan, et al. Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s}$ =8 TeV. *J. Instrum.*, 10(arXiv:1502.02702. CMS-EGM-14-001. CERN-PH-EP-2015-006):P08010. 59 p, Feb 2015. Comments: Submitted to JINST.

[37] Vardan Khachatryan, Sirunyan, et al. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. *J. Instrum.*, 10(arXiv:1502.02701. CERN-PH-EP-2015-004. CMS-EGM-13-001):P06005. 63 p, Feb 2015. Comments: Replaced with published version. Added journal reference and DOI.

[38] Emilio Meschi, T Monteiro, Christopher Seez, and Pratibha Vikas. Electron Reconstruction in the CMS Electromagnetic Calorimeter. Technical Report CMS-NOTE-2001-034, CERN, Geneva, Jun 2001.

[39] Joanna Weng. A Global Event Description using Particle Flow with the CMS Detector. In *Proceedings, 34th International Conference on High Energy Physics (ICHEP 2008)*, 2008.

[40] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[41] E. S. Pearson J. Neyman. *On the Problem of the Most Efficient Tests of Statistical Hypotheses.* Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706):289-337, 1933.

[42] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis. *ArXiv Physics e-prints*, March 2007.

[43] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics Vol. 29, No. 5*, October 2001.

[44] Mingming Yang. Observation and Measurement of a Standard Model Higgs Boson-like Diphoton Resonance with the CMS Detector. 2015.

[45] I. Antcheva et al. ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization. *Comput. Phys. Commun.*, 180:2499–2512, 2009.

[46] Guido Altarelli and G. Parisi. Asymptotic Freedom in Parton Language. *Nucl. Phys.*, B126:298–318, 1977.

[47] Yuri L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov. Phys. JETP*, 46:641–653, 1977. [Zh. Eksp. Teor. Fiz.73,1216(1977)].

[48] V. N. Gribov and L. N. Lipatov. e+ e- pair annihilation and deep inelastic e p scattering in perturbation theory. *Sov. J. Nucl. Phys.*, 15:675–684, 1972. [Yad. Fiz.15,1218(1972)].

[49] Pavel Nadolsky, Jun Gao, Marco Guzzi, Joey Huston, Hung-Liang Lai, Zhao Li, Jon Pumplin, Dan Stump, and C. P. Yuan. Progress in CTEQ-TEA PDF Analysis. In *Proceedings, 20th International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS 2012)*, pages 417–420, 2012. [,417(2012)].

[50] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the LHC. *Eur. Phys. J.*, C63:189–285, 2009.

[51] Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Juan Rojo, and Maria Ubiali. Impact of

Heavy Quark Masses on Parton Distributions and LHC Phenomenology. *Nucl. Phys.*, B849:296–363, 2011.

[52] S. Gieseke. Monte Carlos – Lecture 1, in: DESY MC school 2012, Hamburg, 2012.

[53] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506:250–303, 2003.

[54] Peter Skands Torbjorn Sjostrand, Stephen Mrenna. A Brief Introduction to PYTHIA 8.1. 2007.

[55] T. Gleisberg, Stefan. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *JHEP*, 02:007, 2009.

[56] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[57] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010.

[58] Pierre Artoisenet, Rikkert Frederix, Olivier Mattelaer, and Robbert Rietkerk. Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations. *JHEP*, 03:015, 2013.

[59] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 11:070, 2007.

[60] S. Catani, F. Krauss, R. Kuhn, and B. R. Webber. QCD matrix elements + parton showers. *JHEP*, 11:063, 2001.

[61] Johan Alwall et al. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *Eur. Phys. J.*, C53:473–500, 2008.

[62] Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *JHEP*, 12:061, 2012.

[63] LHC Higgs Cross Section Working Group. https://twiki.cern.ch/twiki/bin/view/ lhcphysics/cernyellowreportpageat13tev.

[64] P. D. Dauncey, M. Kenzie, N. Wardle, and G. J. Davies. Handling uncertainties in background shapes: the discrete profiling method. *JINST*, 10(04):P04015, 2015.

[65] R. A. Fisher. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 1922.

[66] Federico Demartin, Stefano Forte, Elisa Mariani, Juan Rojo, and Alessandro Vicini. The impact of PDF and alphas uncertainties on Higgs Production in gluon fusion at hadron colliders. *Phys. Rev.*, D82:014002, 2010.

[67] Stefano Carrazza, Stefano Forte, Zahari Kassabov, José Ignacio Latorre, and Juan Rojo. An unbiased Hessian representation for Monte Carlo PDFs. Aug 2015.

[68] Jon Butterworth et al. PDF4LHC recommendations for LHC Run II. *J. Phys.*, G43:023001, 2016.

[69] CMS Collaboration. First measurement of the differential cross section for $t\bar{t}$ production in the dilepton final state at $\sqrt{s} = 13$ tev. CMS Physics Analysis Summary CMS-PAS-TOP-15-010, CERN, 2015.

[70] Serguei Chatrchyan et al. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013.

[71] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011.

[72] Serguei Chatrchyan et al. Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Lett.*, B710:26–48, 2012.

[73] Vardan Khachatryan et al. Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV. 2014.

[74] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, and W. Adam. Measurement of the properties of a higgs boson in the four-lepton final state. *Phys. Rev. D*, 89:092007, May 2014.

[75] Serguei Chatrchyan et al. Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states. *JHEP*, 1401:096, 2014.

[76] Georges Aad et al. Evidence for the spin-0 nature of the Higgs boson using ATLAS data. *Phys.Lett.*, B726:120–144, 2013.

[77] Combined CDF and D0 Constraints on Models for the Higgs Boson with Exotic Spin and Parity. 2014.

[78] Ian Anderson, Sara Bolognesi, Fabrizio Caola, Yanyan Gao, Andrei V. Gritsan, et al. Constraining anomalous HVV interactions at proton and lepton colliders. *Phys.Rev.*, D89:035007, 2014.

[79] Sara Bolognesi, Yanyan Gao, Andrei V. Gritsan, Kirill Melnikov, Markus Schulze, et al. On the spin and parity of a single-produced resonance at the LHC. *Phys.Rev.*, D86:095031, 2012.

[80] Yanyan Gao, Andrei V. Gritsan, Zijin Guo, Kirill Melnikov, Markus Schulze, et al. Spin determination of single-produced resonances at hadron colliders. *Phys.Rev.*, D81:075022, 2010.

[81] S. Mrenna T. Sjostrand and P. Z. Skands. Pythia 6.4 physics and manual. *JHEP*, 0605, 2006.

[82] The CMS Collaboration. Measurements of the Higgs boson at 125 GeV in the two photon decay channel in the Full LHC Run 1 Data at CMS. *CMS Analysis note*, 2014.

[83] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J.*, C71:1554, 2011. [Erratum: Eur. Phys. J.C73,2501(2013)].

[84] CMS Collaboration. Absolute Calibration of the Luminosity Measurement at CMS: Winter 2012 Update. 2012.

[85] CMS Collaboration. CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update. 2013.

Remerciements

Arriver au bout d'une thèse de doctorat est une grande réussite qu'on ne peut pas réaliser tous seul. Le fait d'être entouré par des personnes qui t'aident, te conseillent et te soutiennent est fondamental, et je peux dire d'avoir eu cette chance pendant ces trois années de thèse.
Je tiens donc à remercier...

Les membres de mon jury de thèse, pour avoir accepté de lire mon manuscript et pour vos commentaires utiles à son amélioration. Un remerciement particulier aux rapporteurs, Josè et Lucia, pour avoir analysé de façon très détaillée mon travail;
Fabrice, pour ce parcours long qui a mené à de beaux résultats. Merci pour tout ce que tu m'as enseigné, avec patience et dévouement. Merci pour m'avoir poussée dans les moments plus délicats, et pour m'avoir laissé de la liberté dans d'autres situations. Je te remercie en particulier pour le dévouement avec lequel tu m'as suivie dans la dernière partie de la thèse. J'ai beaucoup apprécié ce travail d'équipe pour arriver à obtenir des résultats dont on peut être satisfait;
Patrick, pour ta disponibilité et tes encouragements, ceux pour la "dernière ligne droite" en particulier. Même si tu étais loin de Saclay tu as toujours été présent dans les moments importants avec tes conseils et tes mots gentils;
Gautier, pour avoir accepté de faire partie de mon jury de thèse malgré la quantité de boulot que tu as dans cette période. Merci pour ton support, tes conseils et tes enseignements le long de ces trois ans, j'ai pris beaucoup de plaisir à parler de physique avec toi, tu as un enthousiasme contagieux et tu fais partie du petit groupe de personnes qui parlent de physique avec les yeux qui brillent;
Carlotta, parce que tu n'as pas été "seulement" une excellente collègue, mais surtout une amie. Merci pour les 2 ans qu'on a passés ensemble, merci pour tes enseignements et ta patience. Tu as toujours eu à coeur le bon développement de ma thèse, et je peux dire que tes conseils, tes encouragements dans les moments plus difficiles ont été très importants. J'espère que tes efforts pour construire les bases de notre analyse sont récompensés par le résultat qu'on a enfin obtenu;
Le groupe CMS-Saclay..., pour m'avoir permis de vivre ces 3 ans de thèse dans une

ambiance "de famille". Je voudrais remercier en particulier Marc, Julie, Federico, Philippe et Amina, pour votre grande disponibilité et gentillesse;
... et ses doctorants et post-docs, Inna, Giulia, Clément et Saranya (et Pierre, qui est presque devenu un doctorant de CMS). J'ai passé vraiment de très beaux moments avec vous. C'est joli de voir les étudiants d'un groupe si unis, je crois que c'est très important pour pouvoir travailler avec sérénité. Donc continuez avec vos tea-break même sans la "Super 3rd year PhD", je serai avec vous avec la pensée. Ah j'oubliais...bonne chance pour vos thèses!
Le SPP tout..., pour m'avoir accueillie avec enthousiasme et m'avoir permis de connaître plein de gens cool. Un remerciement particulier à mon parrain Bertrand, ton sourire pouvait améliorer les journées les plus sombres, et à Laurent, ce que tu fais pour les jeunes du SPP est vraiment louable. Merci aussi à Georges et Martine, pour votre dévouement et disponibilité;
... et ses doctorants, pour votre amitié, les beaux moments passés ensembles, les repas du vendredi et les sorties à Paris. Sans oublier les discussions de physique...
Les amis de la Sirène et du Paris Downtown, parce que c'est une chance de pouvoir se relaxer en jouant avec vous dans les endroits les plus beaux de cette ville magnifique;
Ma famille, pour m'avoir toujours soutenue dans mes décisions et dans ma volonté de "me balader autour du monde". Je souhaite bonne chance à mon frère pour ses études (je suis sûre que tu iras bien plus loin que ta soeur!);
Enfin, merci pour ton sourire qui a été le soleil de ces dernières années.

Title: Higgs boson production in the diphoton decay channel with CMS at the LHC: first measurement of the inclusive cross section in 13 TeV pp collisions, and study of the Higgs coupling to electroweak vector bosons

Abstract: In this document two analyses of the properties of the Higgs boson in the diphoton decay channel with the CMS experiment at the LHC (Large Hadron Collider) are presented. The document starts with a theoretical introduction of the Standard Model and the Higgs boson physics, followed by a detailed description of the CMS detector. Then, photon reconstruction and identification algorithms are presented, with a particular focus on the differences between the first and the second run of the LHC, where the first run (Run 1) took place from 2010 to 2012 with a centre-of-mass energy of 7 and then 8 TeV, while the second run (Run 2) started in 2015 with a centre-of-mass energy of 13 TeV. Performances of Run 1 and Run 2 reconstructions from the photon identification point of view are compared. Then the photon identification algorithm for the $H \to \gamma\gamma$ analysis optimised for Run 2 is presented. To do that a multivariate analysis method is used. Performances of the photon identification at 13 TeV are finally studied and a data-simulation validation is performed. Afterwards, the $H \to \gamma\gamma$ analysis using the first Run 2 data is presented. The analysis is performed with a dataset corresponding to an integrated luminosity of 12.9 fb^{-1}. An event classification is performed to maximize signal significance and to study specific Higgs boson production modes. The observed significance for the Standard Model Higgs boson at $m_H = 125.09$ GeV is 5.6 σ, while 6.2 σ was expected, and the maximum significance of 6.1 σ is observed at $m_H = 126.0$ GeV.
Finally a feasibility study, having the aim of constraining the anomalous couplings of the Higgs boson to the vector bosons, is presented. This analysis is performed using the data collected at 8 TeV during Run 1 at the LHC, corresponding to an integrated luminosity of 19.7 fb^{-1}. This analysis exploits the production of the Higgs boson through vector boson fusion (VBF), with the Higgs decaying to 2 photons. The kinematic distributions of the di-jet system and the Higgs, which depend from the spin-parity hypothesis, are used to build some discriminants able to discriminate between different spin-parity hypotheses. These discriminants allow to define different regions of the phase-space enriched with a certain spin-parity process. The Higgs boson signal yield is extracted in each region from a fit to the diphoton mass, allowing to determine the contributions of the different processes and then constrain the production of a pseudo-scalar (spin-parity 0^-) Higgs boson.

Keywords: LHC, CMS, Higgs boson, photon identification, Higgs coupling, vector boson fusion