SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of Fisica “Giuseppe Occhialini”

Dottorato di Ricerca in / PhD program Fisica e Astronomia
Ciclo / Cycle XXXI

TITOLO TESI / THESIS TITLE

Study of the Higgs boson associated production with a vector boson in the Higgs boson diphoton decay channel with the CMS detector

Cognome / Surname Ciriolo Nome / Name Vincenzo

Matricola / Registration number 736088

Tutore / Tutor: Prof. Tommaso Tabarelli de Fatis

Coordinatore / Coordinator: Prof. Marta Calvi

ANNO ACCADEMICO / ACADEMIC YEAR 2017-2018

Summary of the Ph.D. work

This Ph.D. thesis presents a study of the associated production of the Higgs boson with a vector boson (VH) in proton-proton collisions, using data collected at a center-of-mass energy of 13 TeV with the CMS experiment in 2016 for an integrated luminosity of 35.9 fb^{-1}. The study is restricted to final states where the Higgs boson decays in two photons (H$\rightarrow \gamma\gamma$). A key feature of the CMS experiment is a hermetic and homogeneous electromagnetic calorimeter (ECAL) based on lead-tungstate scintillating crystals, which provides excellent energy reconstruction and identification capabilities for the photons from the Higgs boson decays.

After an introduction on the experimental and theoretical landscape that motivates this study (Chapter 1) and an overview of the CMS experiment (Chapter 2), the experimental work performed to ensure the best quality of the CMS ECAL reconstruction is described in detail (Chapter 3). The analysis method and the results are then described (Chapter 4), including the discussion of optimization steps that were identified and will be relevant for future developments. This analysis, included in a paper recently submitted for publication, represents the first result on the measurement of the VH production process in the H$\rightarrow \gamma\gamma$ final state with the LHC Run 2 data. An abstract of the thesis with the indication of the main aspects of my work are summarized below.

The discovery of the Higgs boson by the ATLAS and CMS collaborations in 2012 represented the culmination of the long path toward the experimental confirmation of the standard model of particles and interactions (SM). A full understanding of the SM structure requires a thorough characterization of the Higgs boson properties, which is one of the main goals of the CMS and ATLAS experiments. At the current level of experimental precision, no deviations from the SM predictions have been observed. The amount of data provided by the LHC Run 2 (above 100 fb^{-1}) and the planned High-Luminosity phase of the LHC (HL-LHC) will enable to investigate the Higgs boson properties at increasing precision, giving access to rarer processes. The VH process is the third most probable Higgs boson production mechanism at the LHC, with a cross-section at a centre-of-mass energy of the collisions of 13 TeV, of about 2.1 pb. It mainly occurs at tree-level through the emission of a Higgs boson from a virtual vector bosons (V*) produced by the interaction between two quarks ($q\bar{q} \rightarrow V^* \rightarrow VH$). This process is characterized by the production of Higgs bosons of high transverse momentum. This feature allows to exploit the VH process to probe the Higgs boson interactions in a range of momentum transfer, where the effects due to physics beyond the SM (BSM) can arise. On the experimental side, the decay products of the vector boson provide a strong handle for the rejection of background processes. In particular, in the leptonic decay channels of the V, the dominant QCD background at the LHC is dramatically suppressed.

The Higgs boson diphoton decay channel represents a good channel to study the Higgs boson properties, thanks to its clean experimental signature in the detector made up of two high-energy, isolated photons and the excellent performance of the CMS electromagnetic

calorimeter (ECAL). The analysis results are extracted from a fit of the diphoton invariant mass spectrum, where a narrow peak due to the H$\rightarrow \gamma\gamma$ signal is reconstructed over a large continuous smoothly-falling background due to events with a pair of photons in the final state, or γ+jets and multijet events, with jets misidentified as photons.

The analysis strategy employs a categorization aiming at selecting events produced via a particular production mechanisms, namely the vector boson fusion (VBF), associated production with a top quark pair (ttH), and VH, on the base of additional particles reconstructed in the final state. If the event can not be tagged,it is categorized according to the expected diphoton invariant mass resolution, in order to build categories with high signal-over-background ratio, which are enriched in events where the Higgs boson is produced via gluon-gluon fusion.

The photon energy resolution and the identification are crucial for the analysis performance. The ECAL was designed to achieve an energy resolution about 1% for photons from the Higgs boson diphoton decay. In this range of photon energy, one of the leading contribution to the energy resolution of the ECAL is due to the accuracy in the intercalibration of the channel-by-channel response. This is performed from data through different methods. During my Ph.D., my main contribution to the ECAL activity regarded the channel intercalibration and the monitoring of the ECAL energy scale with high-energy electrons from W and Z bosons.

The intercalibration is performed by constraining the electron energy measured in the ECAL to the reference value provided by the momentum measured by the CMS tracker. High-energy electrons are used also to track the evolution in time of the ECAL energy scale through the monitoring of the position of the E/p peak as a function of time. During the LHC Run 2, the ECAL exhibited also a sizeable drift of the response. Time-dependent effective corrections were derived with high-energy electrons. Their impact on the ECAL scale stability improved either the energy resolution or the photon identification performance, that plays a key role in analysis with photons in the final state.

In the context of the H$\rightarrow \gamma\gamma$ analysis, I focused on the study of the categories targeting the VH production mechanism. In particular, my contribution dealt with the optimization of the category targeting VH events with the V boson decaying into a pair of quark and the category aiming at the selection of WH events in the W$\rightarrow l\nu$ channel. The former targeted the analysis of the data collected in 2016 with proton-proton collisions at a center-of-mass energy of 13 TeV, corresponding to an integrated luminosity of 35.9 fb^{-1}. These results are included in a paper that is at the last steps toward the publication on journal at the time of writing. It was the first result including the VH production process in the H$\rightarrow \gamma\gamma$ channel with the LHC Run 2 data. The statistical significance of the observed and expected excess with respect to the absence of the VH production hypothesis are 2.4 σ and 1.2 σ respectively. The signal strength (μ), defined as the ratio between the measured cross-section and the SM prediction, was measured to be μ_{VH}=$1.21^{+0.58}_{-0.51}$. The coupling of the Higgs boson to vector bosons, which is involved in the VH mechanism, was measured within the k-framework, which interprets the deviation of the measurement from the SM expectations as multiplicative modifiers of the couplings predicted by the SM. The results of the analysis were also interpreted in the Simplified Template Cross-Section (STXS) framework, where ratios between the observed and the SM cross-section in mutually exclusive portions of the phase-space are measured. The ratio related to the VH production in the V$\rightarrow$qq' channel was measured to be $5.1^{+2.5}_{-2.3}$. The WH leptonic tag was subject of a longer-term study targeting the analysis of the entire LHC Run 2 dataset. In the 2016 analysis, the categorization followed a cut-based strategy. The potential improvement in background rejection provided by a multivariate

approach, thanks to a more effective exploitation of the kinematics features of the signal. The naive significance of the category was considered as a figure of merit. Thanks to the optimized categorization an improvement in the naive significance about 35% was assessed. The projected value with an integrated luminosity of 100 fb^{-1} (the reference luminosity for the Run 2 dataset) is about 2.48σ.

Contents

Chapter 1

Introduction

In this chapter, the standard model of particles and interactions is briefly introduced. Its formulation as a gauge theory and the derivation of fundamental forces will be described. The phenomenology of the physics of the Higgs boson at the LHC is discussed, with focus on the associated production channel with a vector boson and the decay in the diphoton channel.

1.1 The Standard Model of particles and interactions

Throughout the XX century, technological progress has lead to the construction of more and more sophisticated and precise experimental apparatus, which have enabled physicists to probe the nature of matter and the interactions occurring among particles at a fundamental level. In the field of experimental, high-energy physics, this has meant the construction of more and more sophisticated and powerful particle accelerators and precise detectors that has culminated in the construction of the Large Hadron Collider (LHC) [1] at CERN, together with the detectors installed at the LHC interaction points.

From the theoretical point of view, the research activity has lead to the definition of a comprehensive model, called *Standard Model of particles and interactions* (SM), through which it is possible to describe the particles and interactions that have been observed experimentally.

The SM is a gauge theory, developed in the context of quantum field theory, that coherently describes 3 of the 4 known fundamental interactions: the strong, the weak and the electromagnetic force. For what concerns the gravitational interaction, its formulation within the quantum field theory and in agreement with the general relativity theory [2] and the experimental observation has not been successfully developed yet. The particles predicted by the SM are divided into two groups, according to their spin: they are called *bosons*, if they have integer spin and *fermions* if they have half-integer spin.

1.1.1 Fermions

Fermions follow the Fermi-Dirac statistics[3, 4] and they are the building blocks of matter. They are categorized in *leptons* and *quarks*, according to the interactions they undergo, which are determined by their quantum numbers. All fermions are subject to the weak force, and only electrically charged fermions to the electromagnetic force. Quarks, that also carry a *color* charge, interact through the strong force too. The fermions are arranged in three families, as reported in Table 1.1

Name	Mass	Name	Mass
Leptons			
electron (e)	0.511 MeV	electron neutrino (ν_e)	< 2 eV
muon (μ)	105.7 MeV	muon neutrino (ν_μ)	< 190 eV
tauon (τ)	1777 MeV	tauon neutrino (ν_τ)	< 18.2 MeV
Quarks			
up (u)	2.2 MeV	down (d)	4.7 MeV
charm (c)	1.27 GeV	strange (s)	96 GeV
top (t)	173.21 GeV	bottom (b)	4.18 GeV

Table 1.1: Summary of the fermions predicted by the SM and their measured mass.

1.1.2 Bosons

Bosons are integer spin particles and, in the SM, they are the mediator of the three fundamental forces described by the model:

- Photons (γ): carriers of the electromagnetic force. The photon is massless and neutral.
- W^+, W^-, Z^0 bosons: mediator of the weak force. The W^+ and the W^- bosons are electrically charged, while the Z^0 boson is neutral. They are massive, with masses: m_W=80.363±0.020 GeV and m_Z=91.1876±0.0021 GeV [5].
- 8 gluons: mediator of the strong force. They are massless, electrically neutral. They carry a color charge, implying that they interact with themselves through the strong force.

1.1.3 The Lagrangian of the SM

The Lagrangian formalism is used to describe particles and their interactions. At the basis of the standard model, there is the concept of gauge invariance [6]. In general, the invariance of a Lagrangian under a global continuous gauge symmetry implies the existence of a conserved current, and an associate conserved charge, through the Noether's theorem [7]. In a gauge theory, each generator of the group of symmetry is associated with a vector boson, which, if the symmetry is unbroken, has a vanishing mass; otherwise, it is massive.

In quantum field theory, fermions are described by four-component spinors ψ:

$$\psi = \begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix} \tag{1.1}$$

where ψ_L and ψ_R are the left and right chirality components, respectively. The fermion Lagrangian for a free particle is:

$$\mathcal{L} = \bar{\psi} i \gamma_\mu \partial^\mu \psi \tag{1.2}$$

The Lagrangian of the SM is built on a group of three gauge symmetries $SU_c(3) \times SU_L(2) \times U_Y(1)$, where the subscript c stands for color, the L means that the interaction occurs only between left-handed fermions, and Y stands for hypercharge. The fermion free-particle Lagrangian (eq. 1.2) is not invariant for local transformations of this group. The terms of the Lagrangian describing the interaction between forces and particles are

derived by requiring the invariance of the free-particle Lagrangian. At this purpose, the standard derivative is substituted with a *covariant derivative*, which includes an extra-term containing a new field that transforms under the gauge transformation in such a way that the local gauge invariance is restored.

The $SU_C(3)$ group generates the strong interaction, which is mediated by eight massless particles called *gluons*. The associated conserved charge associated with this group is called *color*. Fermions with a non-null color charge are called *quarks* and are subject to the strong interaction. Since the gluons have a color charge, they are subject to self-interactions, giving rise to 3-gluons and 4-gluons vertices.

On the experimental side, isolated colored particles have never been observed, while quarks have been observed only in colorless bound states of two quarks (*mesons*) or three quark (*barions*). This behavior is explained by the concept of asymptotic freedom [8], which implies that the coupling gets weaker as the energy increases and the distance decreases, while it becomes stronger at longer distances.

The $SU(2)_L \times U(1)_Y$ group generates the electroweak interaction. The conserved quantity associated with the U(1) group is called *hypercharge*. The SU(2) group of symmetry is related to the weak interaction. Contrary to U(1) and SU(3) groups, the symmetry of this group is broken, implying that its associated vector bosons have mass. The conserved quantity associated with the generators of this group is called *isospin* (T) and the gauge bosons are the W^i bosons, with i=1, 2, 3.

A fundamental point to address in building the weak interaction Lagrangian is the parity violation, that is observed experimentally in phenomena involving the weak force. This question was already addressed by Fermi in his description of the nuclear β-decay, hypothesizing an interaction of the type V-A, where V is a vector current and A an axial current. In the formulation of the weak interaction in the formalism of gauge theories, the violation of parity is introduced via a different coupling of the force mediators to the left-handed and the right-handed components of fermions. In particular, the W bosons interact only with the left-handed component of fermions, while the Z boson interacts with both the right-handed and the left-handed component. Left-handed fermions are coupled in weak isospin doublets, while the right-handed components are singlet as represented in 1.3.

$$\begin{pmatrix} e \\ \nu_e \end{pmatrix}_L, e_R \tag{1.3}$$

The physical weak bosons W^+ and W^- are a linear combination of the W^1 and W^2 gauge bosons (eq. 1.4). Also, the photon and the Z boson are obtained as a combination of the W^3 and the gauge boson associated with the U(1) group (B_μ) that can be represented as a rotation (eq. 1.4) of an angle θ_W called Weinberg's angle, whose precision measurement is a critical test for the SM. Its most accurate measurement was performed at LEP at the Z pole: $sin^2(\theta_W) = 0.23152 \pm 0.00016$ [9], which is in agreement with the SM prediction.

$$\begin{aligned} W_\mu^\pm &= \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp i W_\mu^2\right) \\ \begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} &= \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix} \end{aligned} \tag{1.4}$$

The mixing is also evident in the combination of hypercharge and third component of the weak isospin to obtain the conserved quantity Q:

$$Q = T_3 + \frac{Y}{2} \tag{1.5}$$

where Q is the electric charge, T_3 is the third component of the weak isospin and Y is the weak hypercharge.

In the historical form of the SM, neutrinos are massless. Nevertheless, experimental evidences of neutrino oscillations require neutrinos to be massive, and the SM can be easily extended to accommodate for massive neutrinos. The right-hand neutrino in this theory is sterile and is not subject to any of these three forces.

A recap of all the particles predicted by the SM, together with their quantum numbers and interaction is reported in table 1.2

Leptons	Q	T_3	Y	Interaction	Quarks	Q	T_3	Y	Interactions
e_L, μ_L, τ_L	-1	-1/2	-1	EM, Weak	u_L, c_L, t_L	2/3	1/2	1/3	Strong, EM, Weak
ν_L^e, ν_L^μ, ν_L^τ	0	1/2	-1	Weak	d_L, s_L, b_L	-1/3	-1/2	1/3	Strong, EM, Weak
e_R, μ_R, τ_R	-1	0	-2	EM, Weak	u_R, c_R, t_R	2/3	0	4/3	Strong, EM, Weak
ν_R^e, ν_R^μ, ν_R^τ	0	0	0	None	d_R, s_R, b_R	-1/3	0	-2/3	Strong, EM, Weak

Table 1.2: Summary of the lepton and quarks of the SM and their charges, and the different kind of interactions they are undergo.

1.1.4 The Higgs mechanism

Within the framework exposed above, all the particles have to be massless, since the introduction in the SM Lagrangian of a mass term of the form $m\psi\bar{\psi}$, containing both the left-handed and the right-handed component of the fields, which transform in different ways, would break the gauge invariance. To give mass to the weak gauge bosons preserving the gauge invariance, Englert, Brout [10], Higgs [11], and others proposed mechanisms based on the spontaneous electroweak symmetry breaking (EWSB) of a local gauge symmetry. This mechanism can be applied to any renormalizable quantum field theory, as the SM.

In this model, the masses of the gauge bosons are generated by a scalar field ϕ, called the Higgs field. The term in the Lagrangian that includes the Higgs field and generates the mass of bosons is of the form:

$$\mathcal{L}_{Higgs} = \left| \left(i\frac{\partial}{\partial x_\mu} + g_2 \tau_i W_\mu^i + g_1 \frac{Y}{2} B_\mu \right) \phi \right|^2 - V(\phi) \tag{1.6}$$

where τ_i are the Pauli matrices, W^i and B are the gauge bosons associated with the $SU(2)_L$ and $U(1)_Y$ symmetry groups, respectively, and $V(\phi)$ is the Higgs field potential, computed as:

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \frac{\lambda}{2}(\phi^\dagger \phi)^2 \tag{1.7}$$

where μ is the real parameter and λ the quartic coupling representing the Higgs self-interaction.

The Higgs field is defined in such a way that it only affects $SU(2)_L$ group symmetry and not the U(1) group, in order to leave the photon massless. The chosen complex field is of the form:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2}\phi^+ \\ \phi^0 + i a^0 \end{pmatrix} \tag{1.8}$$

with ϕ^0 and a^0 being the CP-even and CP-odd neutral components, and ϕ^+ being the complex charged component of the Higgs doublet. The shape of the Higgs potential depends on μ^2: if $\mu^2 > 0$, the potential has a unique minimum at $\phi = 0$, while for $\mu^2 < 0$, it has an infinite number of global minima, as depicted in figure 1.1.

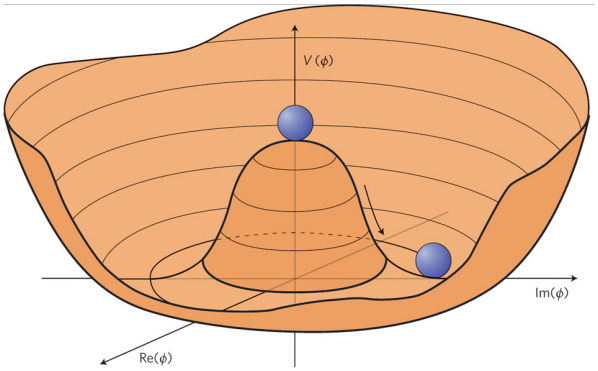

Figure 1.1: Representation of the Higgs potential shape. At high energy the particles sit on the center of the "hat" and do not interact with the Higgs field. At lower energies, the symmetry is broken since the particles lie on a particular configuration in the valley of the potential [12].

When one particular configuration among the infinite possible states is reached, the symmetry is said to be spontaneously broken, since the SM Lagrangian is invariant under SU(2) transformations, while the Higgs potential ground state, which can be written as, $\phi_0 = \frac{1}{\sqrt{2}}\begin{bmatrix}0\\ \nu\end{bmatrix}$ is not. The addition of this new field implies four additional degrees of freedom of the system. For the Goldstone's theorem [13], for each broken continuous symmetry, a massless Goldstone boson is produced. In the Higgs mechanism, these additional degrees of freedom due to the Goldstone's bosons are absorbed by the electroweak bosons, that acquire mass.

$$m_W^2 = \frac{g^2 v^2}{4}, \qquad m_Z^2 = \frac{(g'^2 + g^2)v^2}{4} \tag{1.9}$$

where g and g$'$ are the SU(2) and U(1) gauge couplings. The fourth degree of freedom results in a massive spin-zero particle, called, in the SM, Higgs boson (H).

The vacuum expectation value (vev) of the Higgs field can be expressed as:

$$|\phi_0^2| =< 0|\phi_0|0 >= -\frac{\mu^2}{2\lambda} = \frac{v^2}{2} \tag{1.10}$$

Its value is fixed by the Fermi constant G_F: vev=$\sqrt{2}(G_F)^{1/2}$ ≈246 GeV [5]. For what concerns fermions, their masses can not be generated in the same way. Fermions acquire mass through the Yukawa interaction with the Higgs field described by:

$$\mathcal{L}_{Yukawa} = h^u_{ij}\bar{q}^i_L u^j_R \tilde{\Phi} + h^d_{ij}\bar{q}^i_L d^j_R \Phi + h^e_{ij}\bar{l}^i_L e^j_R \Phi \tag{1.11}$$

where $\tilde{\Phi}$ is the charge-conjugated Higgs field, q_L (l_L) and u_R, d_R (e_R) are the quark (lepton) SU(2) doublets and U(1) singlets, respectively, and $h^{e,u,d}_{i,j}$ are the Yukawa couplings with the i,j=1, 2, 3 flavor indexes. After the EWSB, the Yukawa mass Lagrangian reads:

$$L_{mass} = m^u_{ij}\bar{u}^i_L u^j_R + m^d_{ij}\bar{d}^i_L d^j_R + m^l_{ij}\bar{e}^i_L e^l_R + c.c \tag{1.12}$$

with fermions interacting with the Higgs field with a coupling λ_f proportional to their mass and the vev:

$$\lambda_f = \frac{m_f\sqrt{2}}{vev} \tag{1.13}$$

with the Higgs field vev equal to 246 GeV. However, it should be noticed that the couplings are not fixed by any theoretical principles and must be fixed *ad hoc* based on experimental measurement.

The mass of the Higgs boson is a free parameter of the theory and has been measured by the ATLAS and CMS experiments to be $m_H = \lambda \cdot \nu = 125.01 \pm 0.21(\text{stat}) \pm 0.11(\text{syst})$ GeV [14].

1.2 Study of the Higgs boson at the LHC

1.2.1 Higgs production processes at the LHC

The Large Hadron Collider (LHC) at CERN is a proton-proton collider, which currently collides protons at a center of mass energy $\sqrt{s}$ of 13 TeV . The ongoing data-taking period is called Run 2 and it ended in October 2018 after having collected an integrated luminosity exceeding 100 fb^{-1}. After a 2-years long period of shutdown, during which an upgrade of the accelerator elements will take place, the LHC will begin its Run 3, with a higher center-of-mass energy of the collisions of 14 TeV. In a longer-term perspective, a High-Luminosity phase of the LHC (HL-LHC) [15], also called LHC Phase 2, is planned to begin in 2026. Thanks to the upgrade of the LHC components, the ultimate instantaneous peak luminosity will possibly reach $7.5 \cdot 10^{34} \text{cm}^{-2}\text{s}^{-1}$, resulting in a potential integrated luminosity of 4500 fb−1 at the end of the HL-LHC life.

At the LHC, the SM Higgs boson can be produced through a variety of processes:

- Gluon-Gluon fusion (ggH) (top left diagram of figure 1.2): in this production process, two gluons fuse into a Higgs boson through an intermediate loop, in which the main contribution is due to top quark and W boson. Since the gluon luminosity is very high at the LHC collision conditions, the gluon-gluon fusion process is the most probable one, with a cross-section at $\sqrt{s}$=13TeV TeV computed at the Next-to-leading Order accuracy in QCD equal to about 44 pb.

- Vector Boson Fusion (VBF)(top right diagram figure 1.2): it is the second most probable production mechanism of a Higgs boson at the LHC, with a cross-section an order of magnitude lower than ggH. At leading order, two quarks of the interacting protons emit two vector bosons which fuse to create a Higgs boson. From the experimental point of view, the presence in the final state of two jets in the forward region with a high dijet invariant mass, together with decay products of the Higgs boson, which are produced mainly in the central region of the detector, provide a strong handle to identify this process. The cross-section of this process is currently known at next-to-next-to-leading order (NNLO) in QCD and next to leading order (NLO) in electroweak accuracy.

- Associated production with a vector boson (VH)(bottom left diagram of figure 1.2): in the associated production with a vector boson, also called *Higgstrahlung*, two quarks fuse into a virtual W or Z boson, which emits a Higgs boson. This is the third most probable production mechanism at the LHC, a factor 2 less probable than VBF. This production process is described in more detail in section 1.3. The cross-section is known at the same level of accuracy of the VBF: NNLO in QCD and NLO in EW corrections.

- Associated production with a top quark pair (ttH)(bottom right figure 1.2): this mechanism is a hundred times less probable than ggH. However, the presence in the

final state of the product of the top quark pair decay, jets from the bottom quark hadronization in addition to either leptons or jets, help identifying the signal. This channel is of particular importance since it enables to directly measure the Yukawa coupling of the Higgs boson to the top quark.

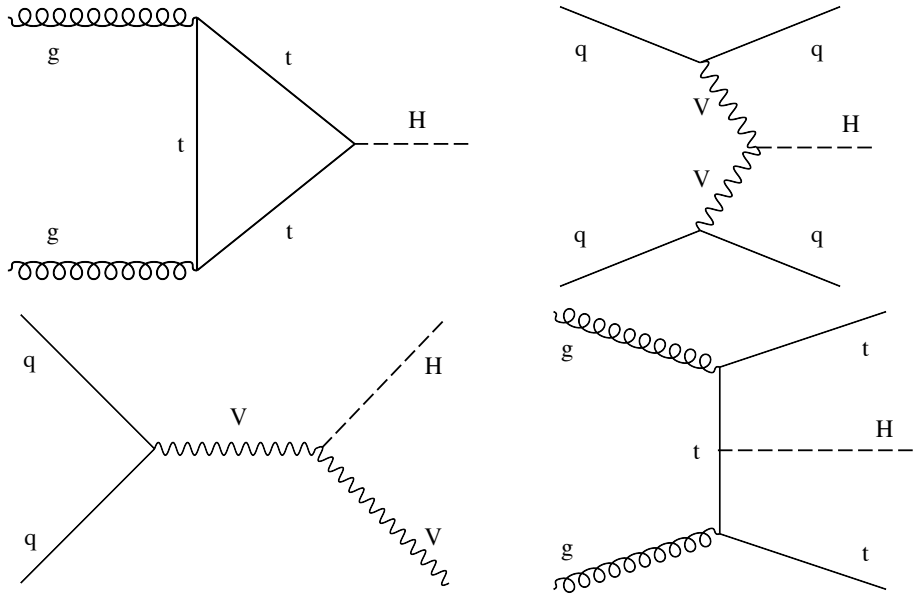

Figure 1.2: Tree level Feynman diagrams of the four dominant Higgs boson production mechanisms at the LHC.

The cross-sections of the main production process at the LHC of a SM Higgs with mass 125 GeV, in collisions with a center of mass energy of 13 TeV are summarized in table 1.3.

In addition to the four main processes described above, there are several subdominant production mechanisms of the Higgs boson at the LHC. Given their low cross-sections, their study requires the entire LHC Run2 dataset (above 100 fb^{-1}) or even the data collected during the high luminosity phase of the LHC (4500 fb^{-1}). Among these, the Higgs boson production in association with a single top quark can provide precious information about the sign of the top Yukawa coupling, since an order of magnitude difference in the cross-section is expected between the two sign hypothesis. The computed cross-section for this process is about 90 fb at the center of mass energy of the collisions of 14 TeV (the reference energy for the LHC operation planned to start in 2021). The study of the Higgs boson associated production with a pair of bottom quarks (bbH) require a large amount of data, which will be provided only during the HL-LHC phase, since the m_B/vev ratio suppresses its cross-section, which is approximately 600 fb.

1.2.2 Higgs boson decay channels

The Higgs boson couples to all the massive particles of the SM, resulting in a large variety of possible modes of decay. In addition, the Higgs boson can decay to massless particles, such as photons, through loops. Its total decay width is entirely determined by its mass, together with the relative branching fractions. In figure 1.4, the branching ratios in several decay channels are plotted as a function of the Higgs boson mass in a range between 120 and 130 GeV. The values for a SM Higgs boson with m_H=125 GeV are reported in table 1.4.

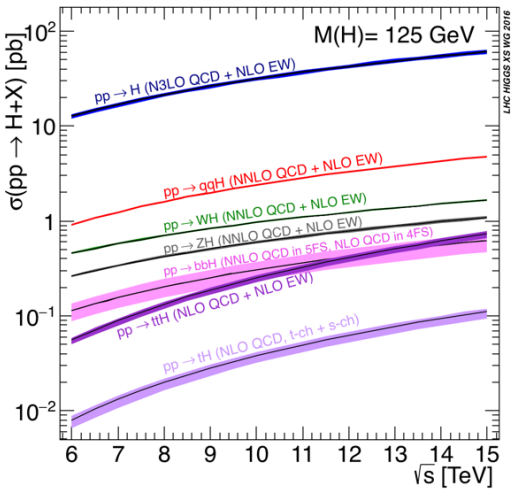

Figure 1.3: SM Higgs boson production cross-section as a function of the center-of-mass energy of the collisions [16].

Process	σ (pb)
ggH	$44.11^{+11\%}_{-11\%}$
VBF	$3.78^{+2\%}_{-2\%}$
WH	$1.37^{+2\%}_{-2\%}$
ZH	$0.88^{+5\%}_{-5\%}$
ttH	$0.51^{+9\%}_{-13\%}$

Table 1.3: Cross-sections of the four dominant production mechanisms of the SM Higgs boson with m_H=125 GeV.

The most probable decay channel is to a pair of bottom quarks. However, the reconstruction of this final state is massively affected by the copious production of jets due to QCD processes at the proton-proton collisions provided by the LHC. Nevertheless, this decay mode can be successfully identified in boosted topologies, such as in the VH process, that provide additional handles for the signal identification, thanks to the peculiar features of the reconstructed final state. The decay to a pair of W bosons has the second higher branching ratio. Efficient and precise measurement of leptons momentum enhances the sensitivity of the W boson leptonic channels. However, the presence in the final state of one or more neutrinos, which gives rise to missing transverse energy (MET) in the event reconstruction, prevent from the reconstruction of the invariant mass. Other channels, such as charm quark pairs or gluon pairs, are overwhelmed by the multijet background at the LHC and have not been targeted by any analysis. The branching ratio of the decay to a pair of muons is too low and can not be observed at the LHC. It is one of the main goals of the physics program of the HL-LHC and it would represent the first observation of the coupling between the Higgs boson and a second generation fermion [17].

The discovery of the Higgs boson in 2012 was driven by the ZZ* decay channel, with both the Z bosons decaying leptonically, and the diphoton decay channel [18], which are characterized by a full reconstruction of the final state and a high resolution in the mass measurement. In the analysis targeting the ZZ* channel, the background is largely rejected thanks to the requirement of four high transverse momentum leptons in the final state. The diphoton decay channel, which will be discussed in more detail in section 1.3, relies on the clean experimental signature in the detector and on the performance of the electromagnetic calorimeters in photon reconstruction and energy measurement, to reach an excellent diphoton invariant mass resolution, which enhances the analysis sensitivity.

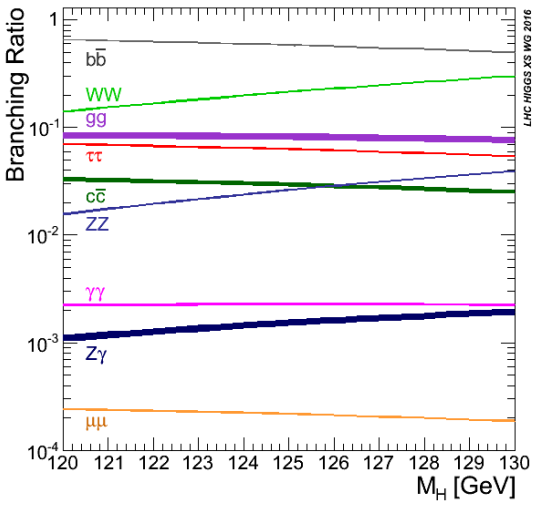

Figure 1.4: Branching ratios of the SM Higgs boson as a function of the mass of the Higgs boson.

Channel	Branching ratio	Uncertainty
$b\bar{b}$	$5.84{\cdot}10^{-1}$	$^{+3.2\%}_{-3.3\%}$
WW	$2.14{\cdot}10^{-1}$	$^{+4.3\%}_{-4.2\%}$
$\tau\bar{\tau}$	$6.27{\cdot}10^{-2}$	$^{+5.7\%}_{-5.7\%}$
ZZ	$2.62{\cdot}10^{-2}$	$^{+4.3\%}_{-4.1\%}$
$\gamma\gamma$	$2.27{\cdot}10^{-3}$	$^{+5.0\%}_{-4.9\%}$
Zγ	$1.53{\cdot}10^{-3}$	$^{+9.0\%}_{-8.9\%}$
$\mu\mu$	$2.18{\cdot}10^{-4}$	$^{+6.0\%}_{-5.9\%}$

Table 1.4: Branching ratios of the most probable decay channels of the SM Higgs boson with m_H=125 GeV.

1.3 The VH production and the diphoton channel

The associated production of a Higgs boson with a vector boson is the third most probable mechanism to produce a Higgs boson at the LHC. The cross-sections of the WH and ZH production are calculated at the NNLO in QCD and NLO in EW accuracy. They are 1.38 pb and 0.88 pb, for the WH and ZH mechanism, respectively, resulting in a total VH cross section of 2.2 pb. In this process, two quarks fuse into a virtual vector boson, that emits a Higgs boson. In the ZH production, a secondary contribution comes from gluon-initiated production, whose Feynman diagram is depicted in figure 1.5.

The study of this decay mode contributes to the measurement of the Higgs boson coupling to vector bosons, which is currently measured with a precision of about 10% [18], mainly driven by the study of the H→WW decay channel. However, the VH channel is an

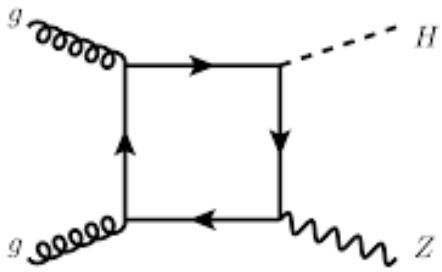

Figure 1.5: Feynman diagram of the gluon-initiated ZH production mechanism.

intriguing channel to study with the amount of data that the LHC and the HL-LHC will provide. In fact, since it occurs at leading order via an s-channel diagram, resulting in a boosted topology, this production mechanism enables to probe the structure of the VVH vertex in a high transfer-momentum range, where momentum-dependent contributions from physics beyond the standard model (BSM) may arise.

Categories targeting the VH production mode have been included in several analyses studying the Higgs boson, from its discovery in 2012 [18], to the current analyses that investigate its properties.

In the recent observation of the coupling of the Higgs boson with the bottom quark, announced by both the ATLAS [19] and CMS [20] collaborations, the VH channel played a crucial role. The analyses extensively profited from the peculiar features of the kinematics of the VH production and the additional particles in the final state due to the V decay, to dramatically improve the background rejection, due to multijet background processes that result in final states with b-jets

Currently, the VH production process has been observed by the CMS experiment in the combination of several analyses targeting different final states, with a statistical significance greater than 3 σ. The ratio of the measured cross-section over the expectation from the SM, has been measured to be $2.18^{+0.58}_{-0.55}$ and $0.87^{+0.44}_{-0.42}$, for the WH and ZH production respectively. The couplings involved in the VH mechanism have been measured at the 10% level precision [21].

In the study of the Higgs boson in the diphoton decay channel, dedicated categories are defined in the analysis to select events produced via Higgstrahlung. This decay occurs at tree level through a loop, whose main contribution comes from top quarks (right diagram of figure 1.6) and W bosons(left diagram of figure 1.6).

On the experimental side, it is characterized by a clean experimental signature in the detector: two isolated, high-energy photons resonating on the Higgs boson mass. The excellent performance of the CMS electromagnetic calorimeter (ECAL), which will be presented in chapter 3, in terms of photon identification and energy measurement is crucial to enhance the sensitivity of this channel and compensates the low branching ratio (about 0.2%).

In the analysis targeting this decay channel, a narrow peak in the diphoton invariant mass spectrum is searched over a large falling background, mainly due to diphoton, γ+Jets and multijet events, with jet fragments misidentified as photons. The invariant mass of the two photons is computed as:

$$m_{\gamma\gamma} = \sqrt{2E_{\gamma 1}E_{\gamma 2}(1-\cos\theta)} \tag{1.14}$$

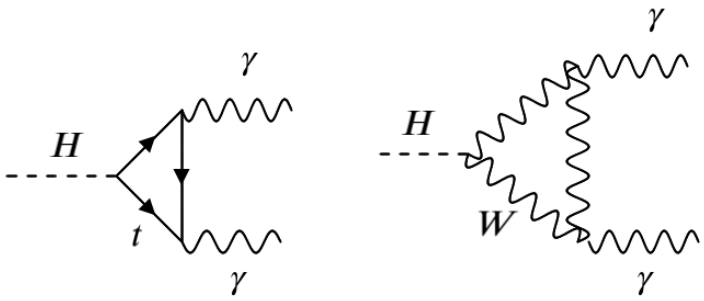

Figure 1.6: Feynman diagram at tree level of the Higgs boson decay in a pair of photons.

where $E_{\gamma 1}$ and $E_{\gamma 2}$ are the energy of the two photons and θ is the opening angle between their directions. The width of the Higgs boson resonance predicted by the SM is about 4 MeV, implying that the width of the reconstructed peak is fully dominated by the experimental resolution, which is about 1 GeV. However, the precise measurement of the energy of the photons is not sufficient to ensure an excellent invariant mass resolution, because of the contribution due to the estimation of the opening angle between the two photons, which is driven by the correct localization of the diphoton vertex position along the beam axis (z-coordinate). The accuracy in the reconstruction of the z-coordinate of the vertex for which the impact on the $m_{\gamma\gamma}$ resolution is negligible with respect to the contribution due to the resolution on the photons energy measurement is about 1 cm.

This final state is one of the most sensitive channels at the LHC for a SM Higgs boson with $m_H = 125$ GeV and was crucial for the discovery of the Higgs boson in 2012. Furthermore, given the excellent resolution on the diphoton invariant mass (approximately 1 GeV), it is one of the channels, together with the $H \to ZZ^* \to 4l$ decay that is employed to measure the mass of the Higgs boson [14].

In the study of the VH production in the Higgs boson diphoton decay channel, the decay products of the vector boson provide an important means through which rejecting the background. According to the decay of the vector boson, different final states arise:

- hadronic decay ($V \to qq'$): either a W or a Z boson decay to a pair of quarks, which are reconstructed in the detector as a shower of particles (jet). The final state is completely reconstructed, enabling to exploit kinematic relations between the diphoton system and the dijet system for background rejection.

- Leptonic decays ($W \to l\nu$ and $Z \to ll$): the leptons produced in the decay of the W or Z boson are very useful in suppressing the background contamination, which is mostly due to misidentified jets, as already mentioned.

- $Z \to \nu\nu$: this decay channels is characterized by a large missing transverse energy in the final state, which can be exploited for background discrimination

Chapter 2

The experimental apparatus and physics objects reconstruction

In this chapter, a brief description of the Large Hadron Collider and of the CMS subdetectors will be given. The electromagnetic calorimeter will be described in detail in the next section. A more detailed description of the CMS electromagnetic calorimeter and the technique employed to reconstruct and measure the energy of photons is given in chapter 3.

2.1 The LHC

The Large Hadron Collider (LHC) [1] is the largest and most powerful particle accelerator ever built. It is located at CERN, at the border between France and Switzerland and its circumference is about 27 km long. It is housed about 100 m underground in the tunnel which had been previously occupied by the Large Electron-Positron collider (LEP) [22], which has played a crucial role in the investigation of the electroweak sector of the Standard Model.

The LHC can produce collisions up to a center of mass energy of 13 TeV, at a frequency of 40 MHz. Its primary goal was the investigation of the electroweak symmetry breaking, firstly, with the search for the Higgs boson, and then with the measurement of its properties. The energies achieved during the collisions allow to probe the SM up to the TeV energy scale, and to search for phenomena predicted by beyond standard model (BSM) theories The same apparatus is used also to accelerate and collide ions, in order to reproduce the interactions and the status of matter that occurred very little time after the universe birth.

The LHC is made up of two rings with clockwise and anticlockwise circulating beams. They intersect in four interaction points, where four experiments are located: A Toroidal LHC Apparatus (ATLAS) [23] and the Compact Muon Solenoid (CMS) [24], which are two general purpose experiments, A Large Ion Colliding Experiment (ALICE) [25], which is dedicated to heavy ion physics and the Large Hadron Collider beauty experiment (LHCb) [26], whose focus is the study of flavor physics and CP violation. In order to bend the high energy proton beams, more than 1000 dipole magnets, long about 15 meters, made of copper-clad niobium-titanium are employed. The focalization of the beams is ensured by almost 400 quadrupole magnets, each about 5 meters long. In the proximity of the interaction points, stronger quadrupole magnets are used to squeeze the beams and maximize the probability of interactions. The magnets installed at the LHC must be

operated at very low temperature in order to be in the superconductive regime , necessary to be able to provide the needed magnetic field. Hence, all the magnets are cooled by means of liquid helium-4 at a temperature of 1.9 K. Before being injected in the LHC, the proton beams goes through a multiple step chain of acceleration [27], which is sketched in figure 2.1: at the beginning the protons are extracted from a bottle of gaseous Hydrogen and accelerated up to a 50 MeV energy by a linear accelerator called LINAC. Then the protons are packed in bunches and accelerated up to an energy of 26 GeV in the Proton Synchrotron (PS). Then, the beams organized in bunches are injected in the Super Proton Synchrotron (SPS), where they reach an energy of 450 GeV. After this last step, the beams are injected in the LHC.

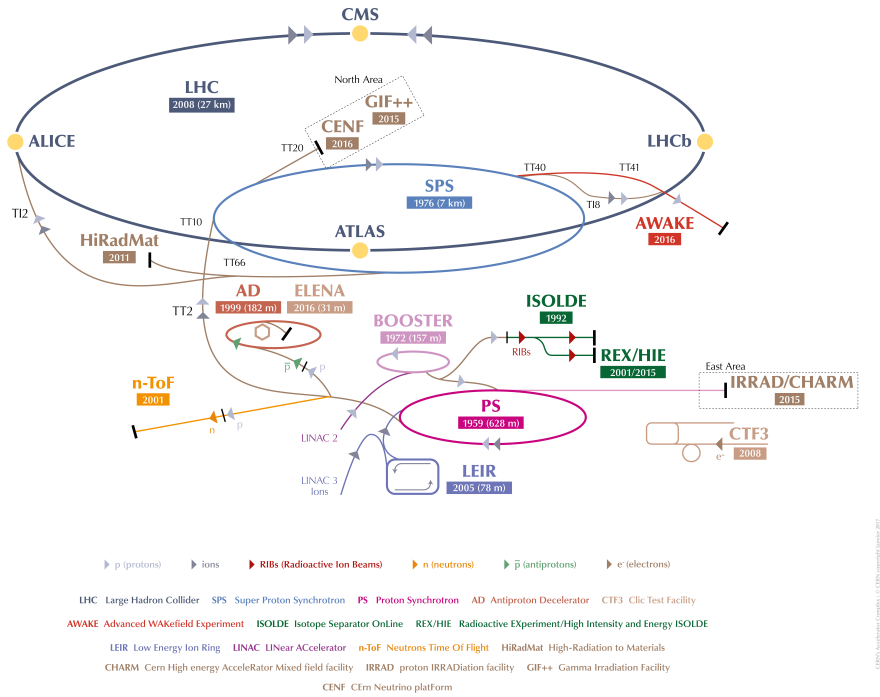

Figure 2.1: The CERN accelerators complex. Protons are first extracted from a hydrogen tank and and accelerated up to an energy of 50 MeV by a linear accelerator (Linac 2). The Proton synchrotron booster (BOOSTER) and Proton synchrotron (PS) push the energy up to 1.4 GeV and 25 GeV respectively. Protons are then injected into the Super proton synchrotron (SPS), where they are accelerated to 450 GeV and finally injected into the LHC. Others machines are present at CERN complex to provide dedicated beams to various experiments. Moreover, the same apparatus complex is used also to accelerate heavy ions [28].

The beams are made up of about $3{\cdot}10^3$ bunches, each containing order 10^{10} protons. The bunches collide at the interaction points every 25 ns.

2.2 The Compact Muon Solenoid

The Compact Muon Solenoid [24] is a general purpose detector installed along the LHC ring. The detector design was driven by the choice of the magnetic field configuration, necessary for the measurement of the transverse momentum of charged particles. CMS employs a superconducting solenoid, 13 meters long and with an inner radius of 5.9 m. It is able to produce a magnetic field of 3.8 T, which enables the measurement of the momentum of muons with a few percent resolution up to transverse momentum about 1 TeV.

The structure of the CMS follows the typical layout of an experiment at a collider , with a central cylindrical region called barrel, and two endcaps, that aims at the full coverage of the region around the interaction point. There are 5 subdetectors: a silicon pixel tracker in the proximity of the interaction point, whose main purpose is the reconstruction of vertices, a silicon strip tracker, which measures the momentum of charged particles, an electromagnetic calorimeter, which measures the energies of photons and electrons, and an hadronic calorimeter. These three subdetectors are enclosed inside the magnet solenoid. Beyond the return steel yoke, there is the muon system, which detect muons, which are able to penetrate all the previous detectors. The subdetectors are arranged concentrically around the interaction point. A sketch of the CMS detector is reported in figure 2.2, while a transverse plane section is pictured in figure 2.3

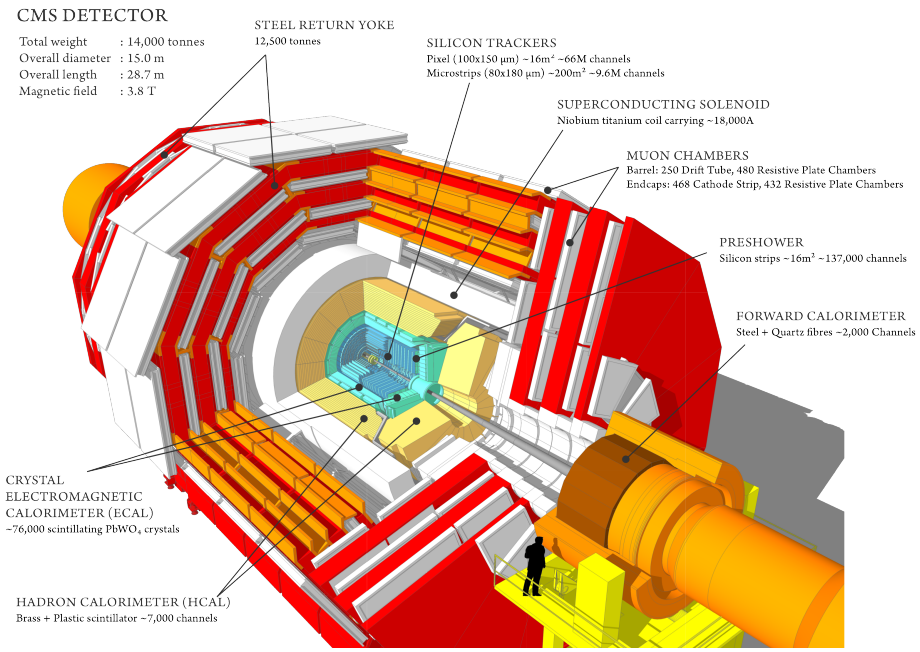

Figure 2.2: 3-dimensional sketch of the CMS detector, each subdetector is indicated [29].

The unprecedented experimental environment of the collisions at the LHC put several constraints on the design of the subdetectors. The high flux of particles arising from the LHC collisions calls for the employment of radiation resistant detectors, which otherwise would face a loss of performance with time because of radiation damage, especially in the

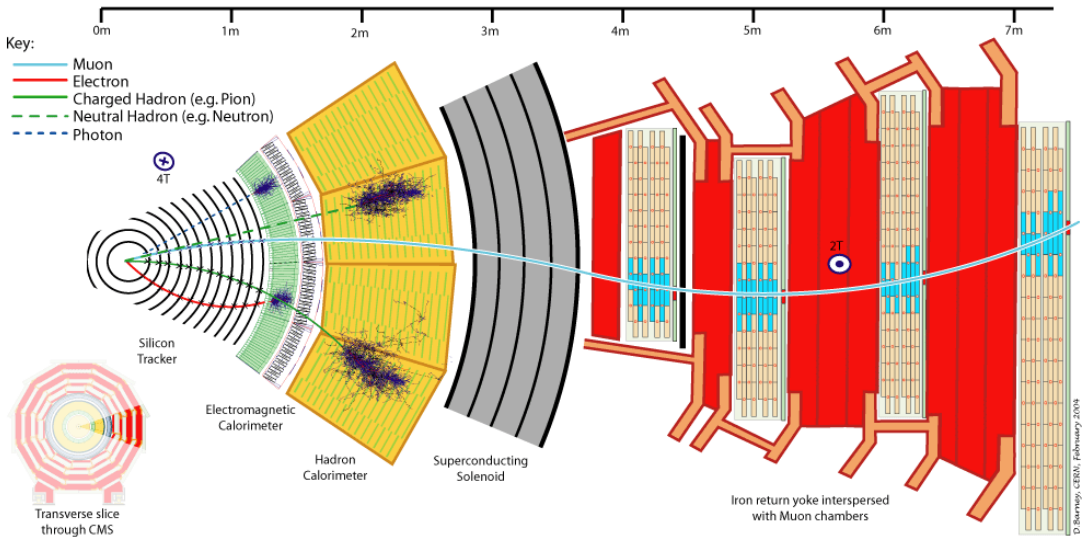

Figure 2.3: Slice of the CMS detector in the x-y view. The interaction that each kind of particle (photons, electrons, muons, charged and neutral hadrons) is detected in each of the CMS subdetectors is depicted.

subdetectors closer to the interaction point, such as the pixel and the silicon strip tracker, and in the forward region. Furthermore, the detectors must be also provided with a great spatial granularity, to avoid the overlap in the detector among particles produced in an interesting interaction and particles produced in concurrent soft interactions, known as pileup, occurring in the same bunch crossing. Given the bunched structure of the beam, also a good time resolution is required to be able to distinguish particles coming from subsequent bunch crossings, spaced in time by 25 ns.

The cartesian coordinate system adopted for the CMS has the origin of the axis placed in the nominal interaction point, which coincides with the geometrical center of the detector. The x axis points radially internally, while the y axis points upward. The x and the y axis define the transverse plane, where the azimuthal angle ϕ, which ranges from -π to π, is measured starting from the x axis. The z axis coincides with the beam axis. The polar angle θ is measured starting from the z axis and takes values between 0 and π. The *pseudorapidity* is a spatial coordinate commonly used in experiments at colliders, since it is invariant under boost along the z-axis. It is derived from the polar angle, which is the angle (θ) with respect to the beam axis, as η=-ln[tan(θ/2)]. The transverse component of the particle energy and momentum, E_T and p_T respectively, are invariant under longitudinal boosts and are computed as: E_T=Esin θ=E/ cosh η. The imbalance of the total transverse energy in each collision is referred to as missing transverse energy, denoted by either MET or $\not{E}_T$. In the ϕ-θ plane, a boost invariant distance between objects can be defined as:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{2.1}$$

2.2.1 The tracking system

The CMS tracker system [30] is the detector placed closest to the beam-line. Its purpose is to reconstruct the trajectories of charged particles and measure their transverse momentum from the curvature in the transverse plane, induced by the 3.8 T magnetic field produced by the CMS solenoid, which is parallel to the beam axis. The disclosure of the

trajectory from a straight line is measured by a quantity called *sagitta* (s):

$$s = \frac{0.3}{8}\frac{BL^2}{p_T} \tag{2.2}$$

where B is the magnetic field intensity, L is the track length. The resolution on the measurement of the transverse momentum depends on the accuracy on the reconstruction of sagitta, as can be seen in equation 2.3

$$\frac{\sigma_{p_T}}{p_T} \approx \frac{\sigma_s}{s} = \frac{8}{0.3BL^2} \cdot p_T \cdot \sigma_s \tag{2.3}$$

To achieve a good resolution, an intense magnetic field (B) and a long reconstructed track length (L) are necessary. The accuracy of this measurement worsen for high transverse momentum, since their curvature in the transverse plane is lower.

The tracker sub-detectors employ solid state silicon detectors and it is composed of two main parts: an inner part, which employs silicon pixels and an outer part, with a coarser granularity, made up of silicon strips. The layout of the CMS tracker is showed in figure 2.4.

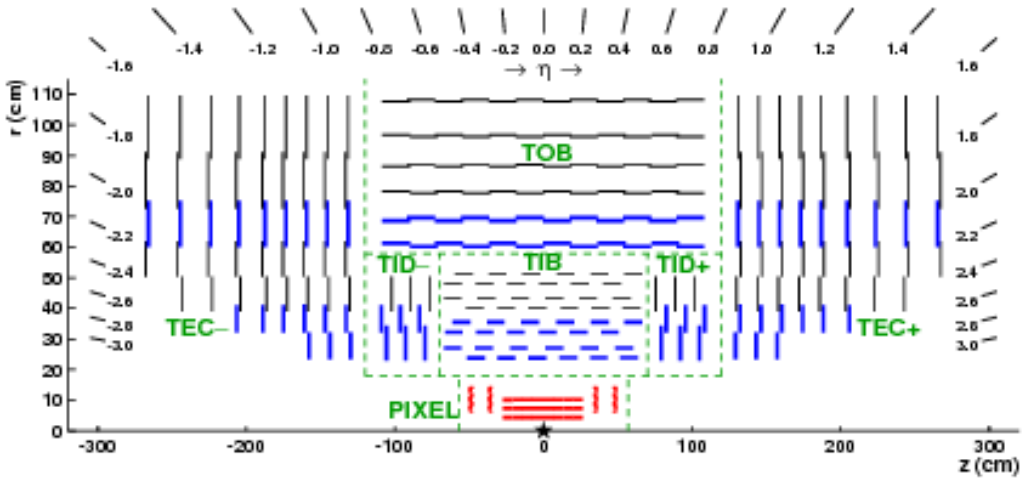

Figure 2.4: Schematic view of a half of the CMS tracker in the r-z plane. The center of the tracker system, corresponding approximately with the interaction point is highlighted by a star. The pixel modules are showed in red, thick lines. Strip modules that provide bidimensional determination of the hit position are showed in thin black lines, while the ones providing tridimensional reconstruction are showed in thick, blue lines. These modules are actually made up of a pair of modules that are rotated one with respect to the other by an "stereo" angle. Modules within each layer are arranged in the r-z plane such that no gap is present.

The innermost part of the tracking system is made up of silicon pixels, each with a surface of 100×150 μm^2, in order to provide a precise reconstruction of the trajectory of particles also at the expected, high fluency that it is expected to face, about 10^6 particles/(cm^2 s) at a distance of 8 cm from the interaction point. The pixel detector in the barrel region

is made up of three layers, placed at radii of 4.4, 7.3, and 10.2 cm, and a length along the z-axis of 53 cm. In the endcap regions, two pixel layers are located at $|z| = 34.5$ and 46.5 cm. In total, the pixel detector covers an area of about 1 m^2 with 66 million channels arranged in 1440 modules. The coverage in pseudorapidity is up to $|\eta|$=2.5, with almost 100% efficiency up to $|\eta|$=2.1, where it starts degrading up to almost 50% at $|\eta|$=2.5. The spatial resolution of the single hit is about 10μm for the coordinates in the transverse plane, and 15μm along the z-axis, in the barrel region, while in the endcap region, it is about 15 and 20 μm respectively.

The outer part of the tracking system is made up of silicon strips. It is divided in two sets of layers in the barrel region, called Tracker Inner Barrel (TIB) and Tracker Outer Barrel (TOB) and two sets of disks in the endcaps, called Tracker Inner Disk (TID) and Tracker EndCao (TEC). In total, it features more than 9 million of strips arranged in more than 15000 modules, for an area of 198 m^2. The single hit point resolution is about 30 μm in the r-ϕ plane and 300 μm along the z-axis.

The amount of silicon used in the tracker detectors, together with the electronics and its cables, and the cooling services, results in a huge amount of material placed in front of the calorimeter system. The estimated amount of material as a function of pseudorapidity can be seen in the plot of figure 2.5. It ranges from $\approx$ 0.4 radiation length (X_0) at $\eta = 0$, to a maximum value of 1.8 X_0 at $\eta \approx 1.5$, near the barrel-endcap transition region.

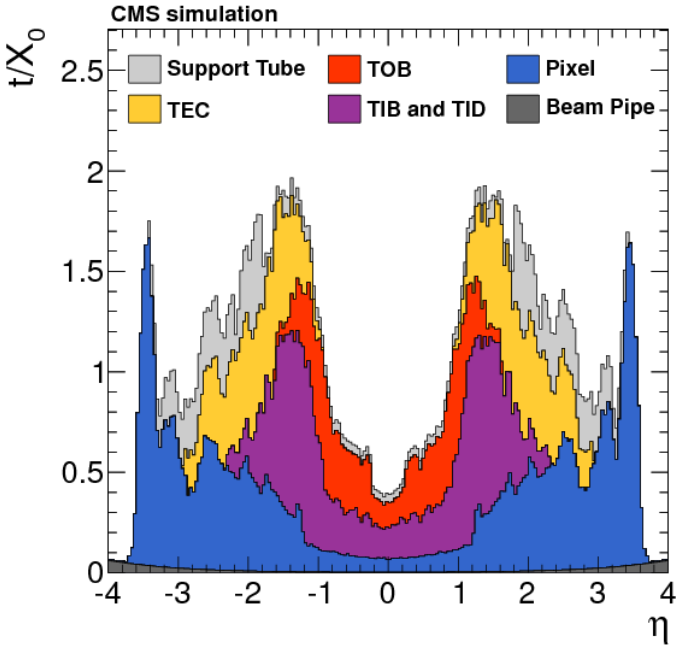

Figure 2.5: Material thickness (t) expressed in radiation length (X_0) traversed by a particle produced at the interaction point as a function of its pseudorapidity.

The effect due to the material placed in front of the calorimeters is one of the main sources of uncertainty in the measurement of the energy of electrons and photons by the electromagnetic calorimeter. In fact, electrons going through the tracker material, loose energy via bremsstrahlung radiation and photons may convert into an electron-positron

pairs before reaching the electromagnetic calorimeter, potentially spreading energy on a wide region along ϕ.

2.2.2 The electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [31] is a homogeneous, hermetic calorimeter made up of 75848 lead tungstate ($PbWO_4$) crystals read out by photodetectors. It is divided in two parts: the ECAL barrel (EB) and two endcaps (EE). A view of the ECAL in the r-z plane, showing the layout of the calorimeter and the arrangement of the crystals in both EB and EB, is reported in figure 2.6.

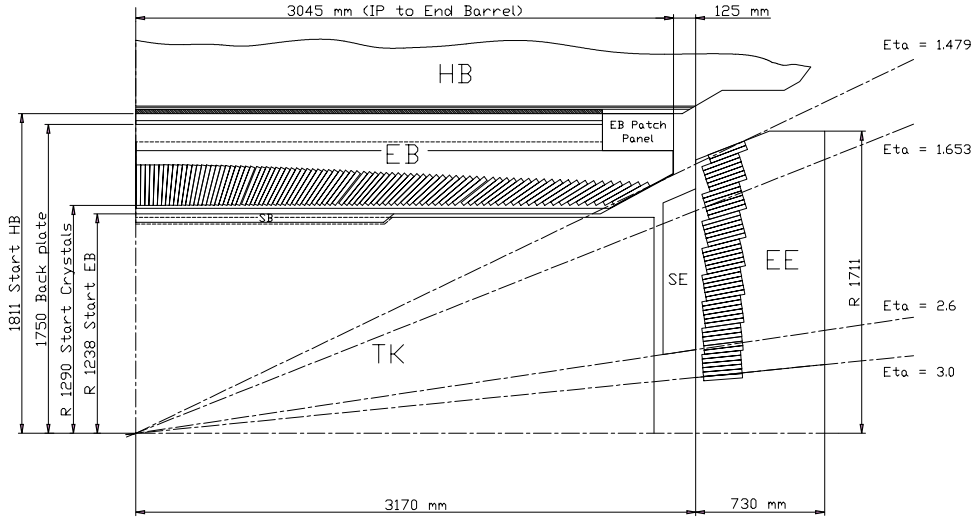

Figure 2.6: Scheme of the ECAL and the preshower in the z-r plane. Only one fourth of the whole detector is depicted.

The crystals are arranged in a quasi-projective geometry, with their axis tilted of 3° with respect to the direction to the nominal center of the detector, in order to prevent particles produced at the interaction point to pass through gaps between crystals without being detected.

The choice of the crystal to employ and the detector design were driven by the requirements necessary to enhance the sensitivity of the search for the SM Higgs decay in the diphoton channel. The ECAL plays a key role not only in the H$\rightarrow \gamma\gamma$ analysis, but also in every analysis which involves high energy photons or electrons in the final state. The $PbWO_4$ crystals are fast scintillators ($\sim$ 2 ns rise time), about the 80% of the scintillation light is emitted in 25 ns, making them suitable for the operation at the LHC bunch crossing frequency. Moreover, thanks to its short radiation length ($X_0 = 0.89$ cm), electromagnetic showers can be absorbed in short lengths, allowing to build a compact calorimeter. Thanks to the little Moliere radius of this crystals (2.2 cm), the longitudinal development of showers is limited and a good separation among different showers is achieved.

The EB is made up of 61200 crystals, 23 cm long (25.8 X_0), and a frontal face of 22×22 mm^2, arranged in 36 supermodules, each containing four modules consisting of 400 or 500 crystals. The detector surface is at a distance of 1.3 m from the interaction point. The

ECAL is provided with a high spatial granularity, with a single channel dimension in the η-ϕ plane of 0.00175×0.00175. The EB provides a pseudorapidity coverage up to $|\eta| = 1.479$. In this region, the scintillation light is read out by avalanche photodiodes (APD), which are characterized by a large quantum efficiency (70-80% at λ=420 nm), compensating the low light yield, about 50 γ/MeV, of the $PbWO_4$ crystals. Moreover, APDs are fast, radiation resistant and insensitive to effects due to the magnetic field. Two APDs with active area of 5×5 mm^2 are coupled to the crystals and read at the same time. At nominal operation bias voltage, their gain is about 50, with a dependence on the bias voltage: $\Delta G/\Delta V = 3.1\%/V$, which demands for a bias voltage system able to control bias voltage variations at the level of millivolts, which result in a contribution to the energy resolution below 0.1%. Furthermore, both the crystals and the APDs response are affected by temperature variations (-2%/C° at 18 C°). In order this effect to result in a negligible contribution to the ECAL energy resolution, the temperature variations oscillation is kept below 0.05 C°, thanks to a dedicated cooling system.

The EE consists in two endcaps, with 7324 crystals each, divided in two dees, placed at |z|=3.10m. They provides coverage in the pseudorapidity region $1.479 < |\eta| < 3.0$. The crystals have a length of 22 cm (24.7 X_0) and a frontal face of 28.62×28.62 mm^2. APDs are not radiation resilient enough to withstand the level of radiation of the endcap region. Hence, Vacuum Photo-Triodes (VPT) are used. Their gain, which is around 8, is lower than the APDs but the larger effective area covered by the VPTs compensates it. The effect of temperature variations on the VPTs in negligible, therefore a stability at the level of 0.1 C° is sufficient.

Even though the $PbWO_4$ are radiation resistant, they are known to suffer from loss of optical transmission due to radiation damage. This loss is measured by means of a laser monitoring system. Laser pulses are injected at the same time in the crystals and in a PN diode, which is used as a reference. The ratio between the signal measured by the APD, or VPT in the endcaps, and the reference PN is to compute a time-based correction (LC) used to account for the loss of transparency of the crystals. . More details about the ECAL laser monitoring system and the way the LC are computed are given in section 3.2.

The ECAL is complemented in the endcap region $1.663 < |\eta| < 2.8$, by a preshower detector (ES), placed in front of the EE. Its aim is to improve the discrimination between genuine high energy photons and pairs of collimated photons produced by the decay of a neutral meson, mainly π^0 and η^0, which may be reconstructed in the detector as a single high-energy photon. It employs two layers of lead absorbers, for a total radiation length of about 3 X_0, in which the electromagnetic shower begins, interspersed by silicon strips, as active element.

2.2.3 The hadronic calorimeter

The purpose of the hadronic calorimeter (HCAL) [32] is to measure the energy of charged and neutral hadrons. When hadrons interact with matter, a hadronic shower made up of several particles is produced. In order to precisely measure the energy deposited by such showers of particles, a hadronic calorimeter has to feature a good granularity, and an extended coverage with a good hermeticity. The hadronic calorimeter resolution is typically worse than electromagnetic calorimeters, because of the possible presence of neutrinos produced in the weak decay of hadrons, which escape detection, and the energy absorbed in nuclear interactions which is not measured by the calorimeter.

The HCAL dimension was constrained by the ECAL, whose external radius is R=1.77 m and the magnet coil, placed at R = 2.95 m. The HCAL consists in a sampling calorime-

ter, with brass absorbers, but the first and last element, which are in stainless steel for structural robustness reasons, and scintillating tiles as active element. The scintillation light is read out by means of wavelength shifter fibers and hybrid photodiodes (HPD). The HCAL is divided into two parts: the barrel (HB), which covers the range in pseudorapidity between 0 and 1.3 and two endcaps (HE), with coverage up to $|\eta|$=3. The granularity of the HCAL is coarser than the ECAL one, and it is in the $\eta \times \phi$ plane 0.087×0.087 in the HB and 0.17×0.17 in the HE. A high spatial granularity is not necessary because of the different characteristics of hadronic showers. The absorber effective thickness increases with the polar angle θ as $1/\sin\theta$, from about 5.8 interaction length (λ_I) to $\approx 10.6\ \lambda_I$ at the end of the barrel. In the endcaps, it is constant and measures about 10 λ_I. An additional module called HCAL outer (HO) is placed in the barrel region $|\eta| < 1.26$, outside the magnetic coil, in order to improve the containment of very high-energy jets. The HCAL acceptance is extended in pseudorapidity thanks to a forward calorimeter (HF), which provide coverage in the region 3< $|\eta|$ <5.2, placed at $|z|$=11.2. The detector used in the more central region of the HCAL is not radiation resistant enough to survive in such a forward region, hence a more radiation resilient quartz-based detector, made of quartz fibers embedded in steel elements, is used. The light is then read by photomultiplier tubes. A longitudinal view of the HCAL is shown in figure 2.7.

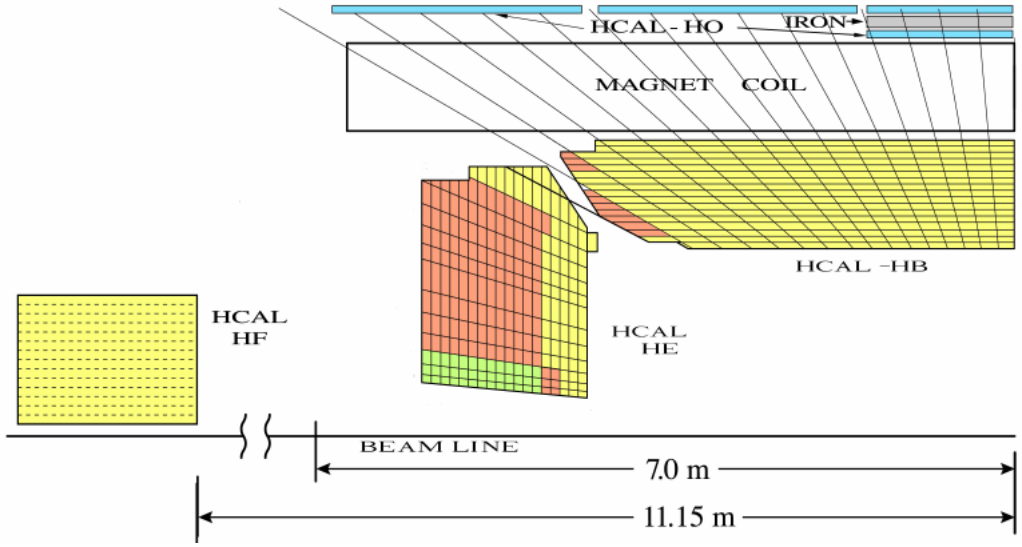

Figure 2.7: Schematic view of the HCAL in the z-r plane. Only one fourth of the detector is depicted.

2.2.4 The muon system

The CMS muon system [33] is designed to measure the momentum of muons. It is located beyond the magnetic coil and it is embedded in the CMS magnetic field return iron yoke. The usage of the return field is particularly important to improve the measurement of muons with momentum above 1 TeV, for which the tracker measurement is not accurate. Three detector different technologies are employed: drift tube (DT), cathode strip chambers (CSC), and resistive plate chambers (RPC). The choice of the detectors was driven

by the necessity for fast detectors (<1 ns) for trigger purpose, and spatial resolution for muon trajectory reconstruction.

The DT are used in the barrel only, where the muon fluency and the background is low. They are arranged in two layers, with an anode parallel do the beam axis, which provide spatial measurement in the r-ϕ plane. An additional anode perpendicular to the beam axis completes the position reconstruction providing the z-coordinate of the hit.

The CSC are multiwire chambers, with an anode segmented into strips. Since the charge collection by a wire, induces also a signal on the cathode strips, a 2-dimensional reconstruction of the detected particle can be performed, exploiting both the signals from the wire and from the strips.

The RPC are thin double-gaped chambers. Their spatial resolution is rather poor if compared to DT and CSC. Nevertheless, thanks to a fast charge collection, they have a very good time resolution, better than 1 ns, which is exploited for trigger purposes. Furthermore, it is also exploited for the identification of the correct bunch crossing time.

The layout of the muon system is shown in figure 2.8.

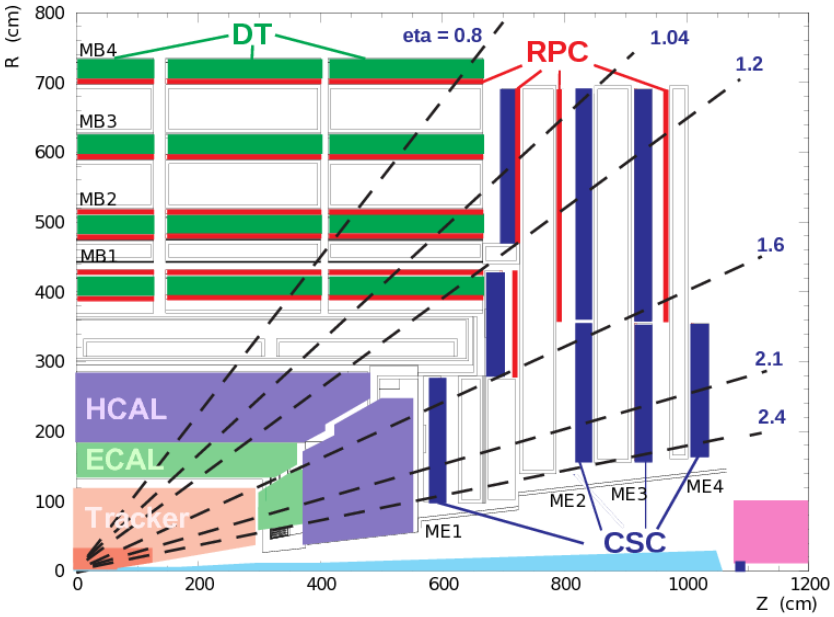

Figure 2.8: Scheme in the z-r plane of one fourth of the CMS muon system. The drift tube chambers (DT) are showed in green, the Resistive Plate Chambers (RPC) are showed in red, and the Cathode Strip Chambers (CSC) are showed in blue.

The barrel region consists in four layers at a distance between 4 and 7 m from the beam line. They host four DT chambers and six RPC modules. The modules are geometrically arranged in such a way that a muon produced at the interaction point, crosses at least three modules, in order to avoid inefficiency in the reconstruction. In the endcaps, four CSC and three RPC (up to $|\eta|$=1.6) modules are employed. The muon stations in the endcaps are placed perpendicularly to the beam line.

The muon momentum resolution varies with pseudorapidity. An overall 5-8& at 10 GeV and 20-40% at 1 TeV momentum resolution is achieved using only information from the muon system. If the tracker is also exploited in the reconstruction, the momentum resolution improves considerably to 1-1.5% at 10 GeV and 6-17% at 1 TeV.

2.2.5 The trigger system

The rate at which information can be recorded for offline analysis ($\sim$ 1 kHz) is by far lower than the LHC bunch crossing frequency (40 MHz). Therefore, a trigger system to select bunch crossings with a possible interesting interaction has been developed. The CMS trigger system [34] is a two-stage system: a first hardware level, called Level-1 (L1) and a consequent layer, which is implemented at software level, called High Level Trigger (HLT). After the HLT, events are transmitted to the storage elements at an average rate of 1 kHz.

The L1 trigger reduces the rate from 40 MHz to 100 kHz. It consists of two separate flows, which are depicted in figure 2.9: a calorimeter trigger, which exploits information from the ECAL and the HCAL, and a muon trigger, which employs reconstruct muons using hits in all the muon system.

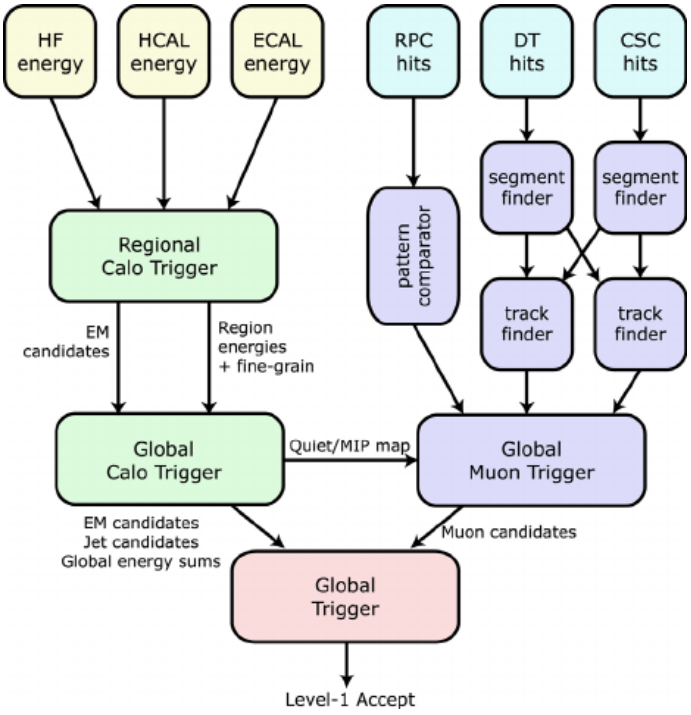

Figure 2.9: Scheme of the workflow of the CMS L1 trigger system.

The L1 decision must be made in a very short time (4μs). Therefore, the track reconstruction is not included in the L1 algorithms, since it takes a longer time to be performed. Furthermore, the ECAL is read out with a coarser granularity, corresponding to matrices of 5×5 crystals (*trigger tower)*, which match the HCAL granularity. Muon are reconstructed from the combined information of all the muon system detectors, DT, CSC, and RPC.

The L1 calorimeter trigger reconstructs physics object such as electrons, photons, tau leptons, jets, or global quantities such as the scalar sum of the transverse momentum of all the jets above a threshold (H_T) or $\not{E}_T$, over a certain threshold. The result of the muon and calorimeters decision is passed to a global trigger element (GT), which makes the final decision. If the rate of a L1 algorithm is too high and can not be lowered by

tuning its thresholds, it can be prescaled by a factor n, meaning that only a fraction 1/n of the events satisfying the algorithm are accepted and passed to the HLT step.

The HLT layer is fully implemented at software level. It exploits the complete readout of the CMS detector, performing the same event reconstruction as the one performed for offline analyses, but in an optimized configuration for being fast. At the HLT, all the physics objects are reconstructed, including track reconstruction and also more complex algorithms. It relies on the concept of path, which is an ordered sequence of reconstruction and filter modules. They are ordered according to their complexity: faster and lighter algorithms are run at the beginning, in order to filter out an event at the earliest possible step and save both time and CPUs. All the paths are run in parallel, but common modules or sequence can be shared among different paths. As for the L1 trigger, also at the HLT a trigger can be prescaled, in order to reach a total average rate of 1 kHz.

Furthermore, in addition to the paths targeting hard scattering interactions, there is a number of utility paths, used for instance for detector studies. For the aim of this paths, the global information from all the CMS sub-detectors is not necessary, therefore only the information from a particular sub-detector or, possibly, only in a particular region of it, is saved.

2.3 Physics objects reconstruction

The raw data provided by the CMS subdetectors in the interesting collisions selected by the trigger system have to be reconstructed and interpreted in order to build the physics objects used in offline analyses, such as photons, electrons, muons, jets, missing transverse energy. In this section, the Particle Flow (PF) [35] algorithm, which is the core of the event reconstruction at CMS, and the reconstruction and identification at CMS of the particles in the final state considered in the analysis described in this thesis, such as electrons, muons, jets, and missing transverse energy, are outlined. Also, the reconstruction and identification algorithms regarding photons are described at a level of detail sufficient for a fruitful reading of the H$\rightarrow \gamma\gamma$ analysis described in the section 4 of this thesis. The reader interested in the details of the photon energy reconstruction necessary to achieve excellent energy resolution essential for high-performance analysis, either on the ECAL side, which represents one of my main contribution during the Ph.D, or the offline reconstruction side, can find them in chapter 3.

2.3.1 The Particle Flow algorithm

The Particle Flow (PF) [35] algorithm is at the core of the event reconstruction at CMS. The algorithm consists in a smart combination of the information coming from all the CMS subdetectors to reconstruct particles. The fundamental blocks of the particle flow are tracks reconstructed by the tracking system and energy deposits in the calorimeter. The muon stations placed outside the magnet coils are also used to improve the muon reconstruction.

Clusters of energy in the calorimeters not linked to any track are the fundamental elements to reconstruct photons and neutral hadrons. First, photons are identified. Energy clusters in the ECAL are considered photon candidates, if they are isolated from either other energy clusters in the calorimeter or tracks reconstructed by the tracker and extrapolated to the calorimeter surface. Furthermore, the ratio between the energy deposited in the ECAL and in the HCAL must be below a threshold. Conversions of photons occurring in the material placed in front of the ECAL are accounted for by using also

information provided by the tracker. The remaining clusters with no linked tracks, are used to reconstruct neutral hadrons. Electrons are identified by a cluster of energy in the ECAL linked with a track pointing at it, with a compatible momentum. Charged hadrons are reconstructed from one or more cluster of energy in the calorimeters, connected with a reconstructed track in the tracker, and no signal in the muon system. The muons are reconstructed from either track segments in the tracker system or the muon stations.

2.3.2 Electron reconstruction

The electron reconstruction at the CMS [36] profits from the particle flow algorithm, combining the information from the tracker and the ECAL to identify and measure the energy of electron with the ultimate precision. The measurement of the energy in the ECAL is based on a clustering algorithm (more details are given in section 3.3), which collects the energy deposits in several channels of the ECAL, including also the energy radiated by electrons via bremsstrahlung while crossing the material upstream of the ECAL. From the point of view of the clustering algorithm there is no difference between a supercluster due to a photon or an electron.

The reconstruction and linking of the electron track to the supercluster are key steps in the electron reconstruction. Two approaches are used for track-seeding: one based on the ECAL and another one on the tracker. In the ECAL-based approach, the search for hits in the innermost layers of the tracker is driven by the energy and position of the supercluster, under either the electron or positron assumption. The performance of this approach is good for isolated, high-p_T electrons. Nevertheless, for electrons inside a jet or for low-p_T electrons this approach shows a poor reconstruction efficiency. In the former case, because of isolation criteria, and in the latter case, because the clustering algorithm is not able to fully recover the radiated energy, spoiling the supercluster energy and position estimation, preventing the track to be successfully linked to the supercluster.

In the tracker-based approach, the starting point is the standard iterative algorithm used in the track reconstruction at the CMS [37]. This algorithm exploits the Kalman Filter (KF) [38] for reconstructing the trajectory of charged particles. This algorithm is able to successfully reconstruct the track of low radiating electrons, but it fails in the case of the emission of hard photons. Tracks selected according to the number of hits and the χ^2 of the re-fitted using an algorithm more suitable for electrons, called Gaussian Sum Filter (GSF) [39], which is able to accommodate the change in the electron track trajectory due to the emission of high energy photons. A final selection is based on the output of a dedicated boosted decision tree that takes as input variables related to both the GSF and the KF fit.

The track seeds obtained with both the approaches are provided to the full electron tracking algorithm, which employs a more refined version of the GSF algorithm. The combined usage of both the methods, result in an improvement of the electron tracks reconstruction, down to a transverse momentum of 2 GeV.

Finally, the energy of the electrons is measured by combining the energy and momentum estimation provided by the ECAL and the tracker, respectively. In the plot of figure 2.10, where the energy resolution as a function of the energy of the electron is plotted, for the ECAL and the tracker separate measurements, and for the combined one, it is possible to clearly see the improvement in the energy resolution. In particular, it is also possible to notice the different behavior of the two detectors, resulting in a precise estimation of the electron energy over a wide energy range.

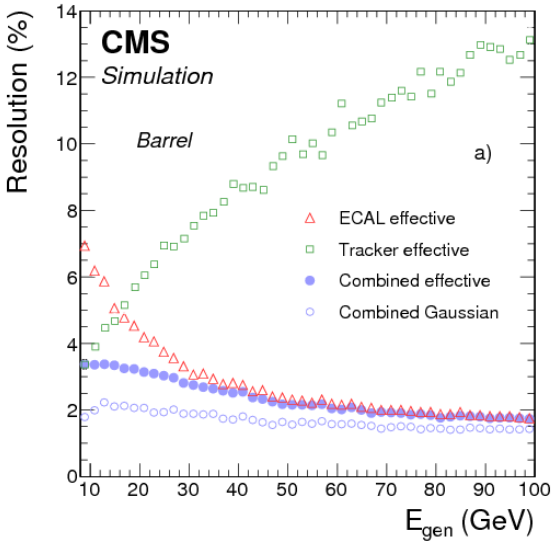

Figure 2.10: Effective resolution on the electron energy measurement in simulated events using only the supercluster corrected energy in the (red triangles), only the momentum measured in the tracker (green squares) and their combination (blue full circles). Also the resolution considering only the gaussian core of the distribution is reported (blue open circles).

Electron identification

The identification of electrons at CMS is pursued either with a cut-based or an MVA-based approach.

In the cut-based approach, a set of subsequent requirements on variables related to the shape of the energy cluster reconstructed in the ECAL , mainly meant to reject fake electrons due to jets, and on the quality of the associated track are applied with different thresholds used in the EB and EE. The variables used in the cut-based identification and the particular thresholds of the working point required for electrons considered in the H$\rightarrow \gamma\gamma$ analysis described in this thesis are:

- energy-weighted covariance of the supercluster along the η direction in terms of channels ($\sigma_{i\eta i\eta}$) <0.0103 and <0.0301 in the EB and EE, respectively
- H/E: ratio of the energy reconstructed in the HCAL in correspondence of the supercluster in the ECAL < 0.104 and <0.0897 in the EB and EE, respectively
- distance along η between the energy weighted position of the supercluster and the extrapolation of the track to the ECAL surface ($\Delta\eta_{SC-trk}$) < 0.105 and < 0.00814 in the EB and EE, respectively
- distance along ϕ between the energy weighted position of the supercluster and the extrapolation of the track to the ECAL surface ($\Delta\phi_{SC-trk}$) < 0.115 and < 0.182 in the EB and EE, respectively
- difference between the inverse of the energy reconstructed in the ECAL and the momentum measured by the tracker $|1/E - 1/p| < 0.102$ and < 0.126 in the EB

and EE, respectively

- number of missing hits in the inner layer of trackers <2 and <1 in the EB and EE, respectively
- transverse impact parameter of the track with respect to the associated vertex < 0.0261 and < 0.118 in the EB and EE, respectively
- track impact parameter along the z-coordinate of the track with respect to the associated vertex <0.41 and < 0.822 in the EB and EE, respectively

In the MVA approach, a boosted decision tree (BDT) is employed. The training is performed on simulated Z+Jets events, in bins of electron pseudorapidity and transverse momentum. For the training step, reconstructed electrons are considered as signal if they are geometrically matched with a generated electron, otherwise they are considered as background. The set of variables given as input is an extension of the set of variables listed above. For instance, the track-cluster matching variables are computed at the vertex and at the ECAL surface, more information related on the cluster-shape and the supercluster substructure are added.

Also, a variable called isolation is defined to quantify the energy flow around the direction of the electron. It is computed as the sum of the transverse energy of the particles (photon, charged and neutral hadrons) in a cone of radius $\Delta R=\sqrt{\Delta\eta^2 + \Delta\phi^2}=0.4$, where $\Delta\eta$ and $\Delta\phi$ are the distances between the particles and the electron along the eta and phi coordinates, respectively:

$$Iso = \frac{1}{p_T^e}\sum p_T(Chargedhadrons_{PV}) + \frac{1}{p_T^e} max(0, \sum E_T(Neutralhadrons) + \sum E_T(Photons) - \rho \times A_{eff}) \tag{2.4}$$

where ρ is the event-by-event average density of energy in the detector and A_{eff} is the effective area of the isolation cone, estimated through the FASTJET technique [40, 41, 42]. A selection based on this variable is effective in reducing the contamination due to either jet with high electromagnetic component, faking an electron, or electrons produced in semileptonic decays of mesons containing a bottom or a charm quark. In fact, in both the cases, a sizable amount of energy is expected in the proximity of the electron direction due to the other component of the hadronic shower.In the cut-based identification, the isolation variable computed in a cone of radius ΔR=0.3 is required to be less than 0.0893 and 0.121 in the EB and EE, respectively.

2.3.3 Muon reconstruction

The muon reconstruction does not directly exploit the PF algorithm. The muon system of the CMS is able to ensure high performance in terms of reconstruction efficiency and accuracy in the energy measurement, over the full detector acceptance (up to $|\eta|$=2.4). A high purity in the muon stations is achieved thanks to the shielding due to the calorimeters and the magnet return yoke. Moreover, the reconstruction of muons also profits from the superior resolution of the CMS tracker to reach high precision in the measurement of muon energy.

Muons are reconstructed following different approaches:

- standalone muons: the muon is reconstructed using only information provided by the detector of the muon system. Hits in the Drift-Tube Chamber (DT) or Cathode

Strips Chamber (CSC) detectors are clustered and used to seed the reconstruction of the muon trajectory. Subsequently, also hits in the Resistive Plate Chamber (RPC) detectors are included.

- tracker muons: tracks reconstructed in the tracker with transverse momentum higher than 0.5 GeV and total momentum exceeding 2 GeV are extrapolated to the muon system. If the extrapolated track is compatible with at least one track segment in the muon system, the inner track is qualified as tracker muon. The matching is performed in a x-y coordinate system lying on the plane transverse to the beam axis, where x is the better measured coordinate. The track is considered to be matched to a segment if the difference between the x-coordinate of the extrapolated track and the segment in the muon detectors is less than 3 cm, or if the ratio between this difference and its uncertainty is less than 4.

- global muons: a matching between each standalone muons and tracker muons is performed. The two tracks are considered matched if their parameters projected onto a common surface are compatible. In the case of positive matching, the hits of the inner track in the tracker and the ones in the muon detectors are combined and simultaneously fitted to constitute a global muon. The improvement in the momentum resolution with respect to the tracker-only muons is sizable, especially for muons with transverse momentum higher than 200 GeV.

The global muon approach has a high efficiency for muons with $p_T > 10$ GeV. For muons with lower transverse momentum, the efficiency is lower because of the multiple scattering they undergo in the magnet return yoke, which prevents the track to meet the reconstruction requirements. For these muons, the tracker-muon approach is more efficient, since the matching with just one track segment in the muon system in sufficient for the reconstruction.

Thanks to the high efficiency of both the tracker and the muon detectors, muon produced within the geometrical acceptance are reconstructed with 99% efficiency.

Muon identification

Muons reconstructed with the procedures listed above are identified on the base of variables related to the quality of the reconstructed track in the tracker and the muon system, and the compatibility of the reconstructed muon trajectory with the vertex they origin from. The variable used for muon identification and the requirements for muons entering the H$\rightarrow \gamma\gamma$ analysis are:

- χ^2/ndof of the global muon track fit, where ndof is the number of degrees of freedom the fit, must be less than 10

- transverse impact parameter of the track reconstructed in the tracker with respect to the muon vertex is required to be less than 0.2 cm

- the longitudinal impact parameter (d_z) of the track with respect to the muon vertex < 0.5 cm

- number of hits in the pixel detector > 0

- number of tracker layers with at least one hit > 5

- number of hits in the muon chamber that are included in the global muon track > 0
- at least one hit in the muon detectors must be included in the fit of the global muon track
- number of muon segments in different muon stations matching the muon track > 1.

The muon isolation variable is essential for discriminating muons inside jets due to semileptonic decays of hadrons containing bottom or charm quarks. It is computed summing all the energy contribution of particles lying in a cone with radius 0.4 around the muon direction, as:

$$I = \frac{1}{p_T^\mu}\sum p_T(Chargedhadrons_{PV}) + \frac{1}{p_T^\mu} max(0, \sum E_T(Neutralhadrons) + \sum E_T(Photons) - \frac{1}{2}\sum P_T(ChargedHadrons_{PU}) \quad (2.5)$$

where the term $\frac{1}{2}\sum P_T(ChargedHadrons_{PU})$ accounts for the contribution of energy deposits due to neutral particles produced in pileup interactions, which is assumed to be a half of the energy carried by charged hadrons associated to pileup vertices. For the muons considered in the H$\rightarrow \gamma\gamma$ analysis, it must be less than 0.25.

Another isolation variable, called mini-isolation (I_{mini}), is also defined. The main difference lies in the radius of the isolation cone used in the computation, which varies according to the p_T of the muon:

$$R_{mini} = \frac{10}{min(max(p_T^\mu, 50), 200)} \quad (2.6)$$

This variable is particularly suited to be used in the environment typical of ttH events, where several physics objects such as jets and photons are reconstructed in the final state, and a cone with a fixed radius gives sub-optimal performance in terms of signal efficiency.

2.3.4 Jet reconstruction

The quarks and gluons produced in the hard energy proton-proton collisions have a non-null color charge. However, due to color-confinement, quarks and gluons can not exist freely and undergo a process called hadronization, through which quarks and gluons form color-less hadrons [8]. From the experimental point of view, only the product of the hadron fragmentation can be observed. Their experimental signature in the detector is a collimated shower of particles, called jet, whose energy and direction depends on the energy and direction of the initial parton. Hence, jets are reconstructed from clusters of energy deposits in the calorimeters and tracks in the tracker, in a delimited region of the detector [35].

Several jet clustering algorithms have been developed, with the properties of the reconstructed jet depending on the particular procedure. At the CMS, the standard algorithm is the anti-k_t algorithm [43], since it meets the desired requirements of robustness against collinear gluon splitting and the emission of soft particles. This algorithm considers all the candidate reconstructed using the PF algorithm and two "distances" between particles, defined as:

$$d_{ij} = min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right)\frac{\Delta R_{ij}^2}{R^2} d_{iB} = \frac{1}{p_{Ti}^2} \quad (2.7)$$

where R_{ij} is the distance in the η-ϕ plane. Then, starting from the i-th particle, the j-th particle is considered, and if $d_{ij} < d_{iB}$, the j-th particle is merged to the i-th, to form a proto-jet. This procedure is iterated until no candidate meeting the $d_{ij} < d_{iB}$ requirement is left in the event. Then, the proto-jet is promoted to a jet. The procedure ends when all the particles in the event are associated to a jet.

The jet momentum is calculated as the vectorial sum of its constituent. A set of corrections [44] needs to be applied to improve the energy reconstruction of the jet, and make them usable in the analyses. They are:

- pileup correction: charged hadrons produced in pileup interactions are removed thanks to a technique called charged hadron subtraction (CHS) [45], which relies on the excellent spatial resolution of the CMS tracker. This technique enables to identify charged particles whose track is not compatible with the primary vertex of the event and exclude it from the jet clustering procedure. The impact of neutral particles due to pileup and overlapping to the jet in the detector is minimized by means of another correction, which is dependent on the average density of energy in the event and the pseudorapidity of the jet. Also, a technique, called Pile-Up per Particle Identification (PUPPI) [46], is used in some analyses. It weights the energy of each jet components according to its probability to arise from a pileup vertex.

- True response: it aims at correcting non uniformities of the CMS subdetectors. Correction factors are derived from MC simulation by comparing the reconstructed jet momentum and the generated one, in bins of p_T and η.

- Residual corrections: corrections aiming at equalizing the jet response along η are derived from dijet events, in which the p_T response of a probe jet, whose pseudorapidity is unconstrained, is equalized to the one in the central region ($|\eta|$ <1.3), used as reference, in bins of p_T of the dijet system. Then, the absolute scale is derived from γ+Jets and Z+Jets, with the Z boson decaying into electron-positron or muon pairs. The jet energy scale is derived by comparing the transverse momentum of the jet with the one of the reconstructed Z boson.

Energy deposits due to particles produced in pileup interactions can be clustered by the jet clustering algorithm, giving raise to jets, whose transverse momentum can exceed the typical thresholds used in CMS analyses to select low-energy jets (about 30 GeV), as for instance the search for the Higgs boson production in the VBF topology. At this purpose, in addition to the CHS techniques cited above, another orthogonal algorithm aiming at the identification of jets due to pileup particles, called PUJETID [47], is used. Furthermore, this is the only tool for pileup rejection in the region beyond $|\eta|$=2.5, which is not covered by the CMS tracker, implying that the CHS technique can not be used. It is based on the output of a boosted decision tree (BDT) to identify jets due to pileup according to the shape of the jet and the compatibility of the tracks of its charged components with the primary vertex of the event.

The hadrons containing a bottom quark usually have large mass and lifetime, of the order of picoseconds, resulting in a tight energy spectrum of they decay products and in a path in the detector of the order of few millimeters before decaying. The charged particles produce in the decay do not point to the primary vertex of the event, and enable to reconstruct a secondary vertex, corresponding to the point where the b-hadrons decay occurred. Several algorithms based on multivariate technique are used at CMS to infer the likelihood of a jets to arise from a bottom quark decay [48], using as input variables information related to the track impact parameter, possible secondary vertex compatible

with the track of the jet, and the jet constituents. The output of the b-jet identification algorithm is a continuous distribution in the range between 0 and 1.

2.3.5 Missing transverse energy

The missing transverse energy ($\vec{\not{E}_T}$) [49] is defined as the negative vectorial sum of the quadrimomenta of all the candidates in the final state reconstructed with the PF algorithm:

$$\vec{\not{E}_T} = -\sum_{i}^{N_{PF}} \vec{E}_T^i \tag{2.8}$$

Under the assumptions that all the particles produced in the interaction are reconstructed, the $\vec{\not{E}_T}$ corresponds do the vectorial sum of the transverse momenta of undetectable particles, hence, this variables is the experimental signature of processes producing neutrinos, or other not interacting particles predicted by beyond standard model theories, such as neutralinos. However, this assumption is not fully verified at CMS, because of the limited geometrical acceptance and inefficiencies in particle reconstruction, threshold used in the detector, and finite energy resolution. In order to improve the accuracy in the estimation of $\vec{\not{E}_T}$, the corrections applied to the jet momenta are propagated in the $\vec{\not{E}_T}$ computation.

2.3.6 Photon reconstruction

Photons are reconstructed from clusters of energy in the ECAL, obtained aggregating the energy deposited in several channels of the calorimeter. Starting from a local maximum of energy deposited in a channel, if it exceeds a threshold, it is used to seed the cluster reconstruction. Then, channels with at least one corner in common with the cluster are added to it, if their energy exceed a second threshold, whose purpose is to prevent contribution from noise in the detector readout electronics to be included in the supercluster.

Reconstruction of tracks from converted photons

Photons originated in the luminous regions traverse a sizable amount of material before reaching the ECAL, implying that they have a non-negligible probability to convert into an electron-positron pair, which can be as large as the 60% in the region where the material upstream of the ECAL is maximum (figure 2.11).

The clustering algorithm described earlier in this section is able to reconstruct the energy of converted photons. Moreover, the R9 variable can be employed to distinguish them. However, the reconstruction of the tracks of the electron-positron pair from a photon conversion [50] can provide further information and improve the event reconstruction. In particular, in the diphoton decay of the Higgs boson, the conversion tracks can be used to reconstruct the direction of the photon, enhancing the accuracy in the location of the primary vertex along the beam axis, which is one of the factors impacting the diphoton invariant mass resolution.

The search for tracks arising from photon conversions is based on the standard electron tracks reconstruction algorithm used at CMS, which is briefly described in section 2.3.2. Tracks identified as possible electron candidates are preselected according to the quality of the track fit and the number of missing hits. In addition to the standard algorithm requirements, specific selections are applied to identify tracks from photon conversions:

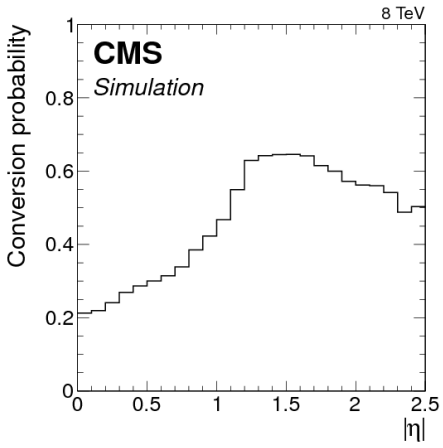

Figure 2.11: Fraction of photons converting before the last three layers of the tracker as a function of pseudorapidity. The results are obtained from a simulated sample of H$\rightarrow \gamma\gamma$ events. The vertex location information is extracted from the simulation information.

- the angular separation of the tracks in the longitudinal plane, measured as $\Delta \cot \theta$, must be less than 0.1. This selection is equivalent to requiring that the tracks are approximately parallel, as expected from the conversion of a photon, since it is massless. This requirement aims at discriminating photon conversions from decays of massive mesons, nuclear interactions or fake vertices due to misreconstructed tracks, for which this condition is not verified.

- the two-dimensional distance of minimum approach between the two tracks must be positive, to remove intersecting helices

- the point where the two tracks are tangent must lie within the tracker

Track pairs passing these selections are fitted to a common vertex, through a 3D-constrained kinematical fit, that imposes the two tracks to be parallel in both the transverse and the longitudinal planes. The track pair is retained, if the fit converges and the resulting χ^2 is below a threshold. The transverse momentum of the pair is measured by refitting the track with the vertex constraint. The p_T obtained in this way must be higher than an energy-dependent threshold, which is set at 10 GeV, in the energy range of photons produced in the Higgs boson diphoton decay. If more than one electron-positron pair compatible with a supercluster in the ECAL is reconstructed, the track pair with the smallest distance $\Delta R=\sqrt{\Delta\eta^2 + \Delta\phi^2}$ between the track pair momentum direction and the supercluster position is chosen.

Photon identification

At the proton-proton collisions provided by the LHC, the production of neutral mesons, such as π^0 and η^0, is copious. Their decay into a pair of photons represents one of the main source of reducible background for analyses involving photons in the final state. In fact, at the transverse energy of interest for the analyses currently performed at the CMS, including the H$\rightarrow \gamma\gamma$ analysis, the two photons produced in such decays are collimated and impinge the ECAL surface within a very short distance, giving raise to a single supercluster in the detector, which can be eventually identified as a single, high-energy

photon. The goal of the photon identification algorithms is to identify these fake photons and reject them, reducing this important source of background. At CMS, two approaches are adopted for this purpose: one is based on the application of a set of subsequent requirements, and the other one employs a multivariate technique. In both cases, the identification relies on variables related to the isolation and shape of the cluster in the ECAL. The cut-based selection is based on three variables:

- R9: the ratio between the energy reconstructed in a 3×3 matrix of crystals centered on the most energetic crystal (seed) of the supercluster and the energy of the supercluster. This variable is used to discriminate between unconverted and converted photons, that populate the high-region and the low-region of the R9 distribution, respectively. It provides further discrimination power against fake photons for two reasons: firstly, electromagnetic showers due to photon pairs from π^0 decays tend to have lower R9 values, since the energy is spread over a wider region, secondly because there is a higher probability that at least one of the pair of photons from the π^0 decay converts.

- $\sigma_{i\eta i\eta}$: energy-weighted extension along η of the 5×5 matrix of crystals centered on the seed of the supercluster, measured in terms of crystals;

- ratio between the energy reconstructed in the HCAL behind the ECAL supercluster;

- Charged isolation with respect to the right vertex: sum of the transverse momentum of charged particles lying in a cone of size R=0.3 around the photon direction reconstructed with respect to the primary vertex of the event;

- Charged isolation with respect to the wrong vertex: sum of the transverse momentum of charged particles lying in a cone of size R=0.3 around the photon and associated to the vertex resulting in the largest isolation sum;

In the MVA approach, the identification of photons is based on the output of a Boosted Decision Tree (BDT). In addition to the variables described above, an additional set of variables describing the cluster shape are given in input to the BDT:

- $E_{2\times2}/E_{SC}$: the ratio between the energy reconstructed in a 2×2 matrix containing both the seed and the second most energetic channel in the supercluster, and the energy of the supercluster

- $\text{cov}_{i\eta i\phi}$: covariance along η and ϕ of the crystals within a 5×5 matrix centered on the seed of the supercluster

- Supercluster width along η

- Supercluster width along ϕ

- Preshower σ_{RR}: standard deviation of the shower spread in the preshower in the x-y plane (used only in the region covered by the preshower detector $1.65< |\eta| <2.8$)

along with three additional variables which enhance the performance of the BDT:

- η of the supercluster

- ρ: average density of energy reconstructed in the detector. Furthermore, this variable is used to correct all the isolation variables in order to account for the contribution due to pileup.

- E_{ES}/E_{raw}: the ratio between the energy reconstructed in the preshower detector in correspondence of the photon-related supercluster, and the supercluster raw energy.

The BDT is trained using a MC simulation sample of γ+jet events, where the reconstructed photon matched to the true photon in the simulated event is considered as signal, while the reconstructed photon candidates not matched to a simulated true photon associated with the hard-scattering interaction is considered as background. Photon candidates are required to have $p_T > 18$ GeV and to be reconstructed within the ECAL fiducial region: $|\eta| < 1.479$ in the EB and $1.556|\eta| < 2.5$ in the EE. In order to avoid any dependence of the BDT output on the photon kinematics, the p_T and η of the signal photon candidates are reweighted to reproduce the distributions observed in the background photons.

Electrons are rejected according to the presence of hits in the tracker which are compatible with the trajectory of a charged particle impinging the ECAL surface in the proximity of the photon cluster reconstructed in the ECAL. Two vetoes are defined, with different level of stringency: the "conversion-safe electron veto" and the "pixel-seed electron veto". The first one is satisfied if no tracks with a hit in the inner layer of the pixel detector is compatible with the photon cluster in the ECAL, but tracks associated to a photon conversion vertex. The inefficiency on photons achieved with this method is almost totally limited to photons converting in the beam pipe. A more severe rejection of electrons can be achieved applying the "pixel seed veto". It rejects photons, if they are compatible with the extrapolation to the ECAL surface of a track seed, consisting in at least two hits in the pixel detector.

Chapter 3

Photon energy measurement in the ECAL

In this chapter, the steps necessary to reconstruct photons and electrons, and measure their energy with high precision are presented. The extraction of the signal from the ECAL and its treatment to derive an energy measurement are described in section 3.1 and 3.3. The techniques used for the ECAL channel intercalibration, which impact the constant term of the calorimeter energy resolution, and monitoring of the energy scale are described in section 3.4 and section 3.6. More details are provided about the intercalibration and monitoring with high-energy electrons, presented in sections 3.5 and 3.7, which are one of my main contributions to the ECAL activity. Finally, the corrections used to enhance the photon energy resolution are presented.

3.1 Signal reconstruction

The scintillation light produced when an electron, or a photon, traverses the ECAL $PbWO_4$ crystals, is read out by a couple of photodetectors glued on the rear face of the crystals.The analog signal produced by the photodetectors is shaped and pre-amplified by a multi-gain preamplifier. The output is digitized by a 12 bit ADC operating at 40 MHz frequency, which records ten consecutive sample, used for the shape reconstruction. A wide dynamic range, from 50 MeV to 3 TeV, is reached thanks to three independent amplifiers, which operate in parallel and provide different gains equal to 1, 6, and 12. The ADC range saturates for deposit of 1.7 and 2.8 GeV in a single channel, in the EB and EE, respectively [31]. The signal reconstruction algorithm used during the LHC Run 2 operations is called *multifit* [51]. A total of 10 pulse shapes are considered, corresponding to the in-time signal, and up to 9 pulse shapes accounting for deposits in the crystals due to particles produced in the five previous and four following bunch crossings. The recorded samples are simultaneously fitted with all the ten pulse shapes via a χ^2 minimization, according to the non-negative-least squares technique. The pulse shapes relative to each bunch crossing are assumed to be identical except for a shift in time of multiples of $\pm$ 25 ns, which is the time interval between two consecutive intersections of the beams. This technique aims to minimize the contribution to the measured in-time amplitude, due to energy deposits coming from particles produced in adjacent bunch crossing (*out-of-time pileup*), and it is particularly effective for low energy deposits, for which the effect of out-of-time pileup has a more significant impact. Two examples of fitted shapes in simulated events are showed in figure 3.1 for the EB (left) and the EE (right).

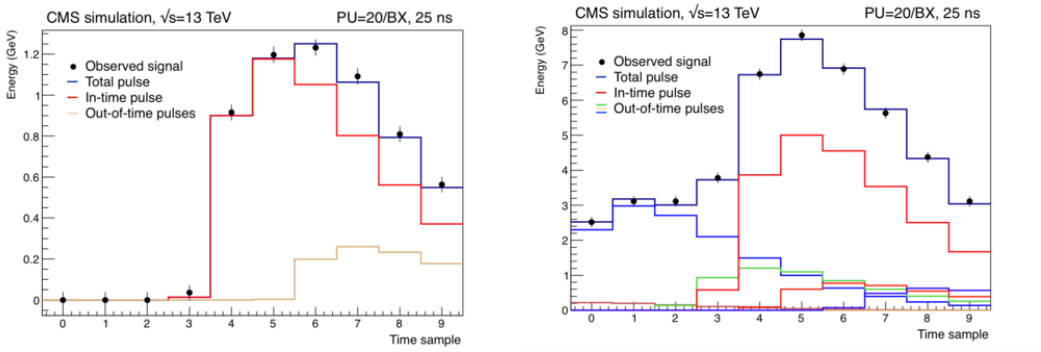

Figure 3.1: Example of fitted pulses in simulated events with 20 average pileup interactions and 25 ns bunch spacing, for a signal in the EB (left) and in the EE (right). Dots represent the 10 digitized samples, the red distribution represents the fitted in-time, while the other color distributions refer to out-of time pulses with positive amplitude. The dark blue histograms represent the sum of all the fitted contributions. [52]

3.2 The ECAL laser monitoring system

The irradiation of the ECAL $PbWO_4$ crystals under electromagnetic radiation causes the creation of color centers inside the crystal, that induce a loss of optical transmission in the same wavelength range where the peak of the scintillation light emission lies. However, the formation of color center is reversible and spontaneous annealing occurs also at room temperature, leading to recovery of optical transmission. On the other hand, exposure to hadrons can induce permanent damage of the crystal. Nevertheless, the impact of such kind of radiation damage on the energy resolution is negligible with respect to the design value, up to an integrated luminosity of 500 fb^{-1} [53].

The loss of optical transmission is monitored and measured *in-situ* during the LHC collisions through a laser monitoring system [54], which employs two sources of laser light: one with a wavelength corresponding to the $PbWO_4$ emission peak of scintillation light λ=447 nm (blue laser) and one with a wavelength λ=495 nm (green laser). A third light source with wavelength in the infrared region, which is insensitive to the crystals loss of transmission is used for performing stability studies regarding the photodetector response. The light is injected in the crystals during a 3 μs period every 90 μs in the LHC beam structure, during which no actual bunch crossing occurs, allowing a full scan of the ECAL in about 30 minutes. The laser pulse is split and distributed with optical fibers to the crystal and to a PN silicon diode, which provides a reference measurement of the laser pulse intensity. In the barrel region, the pulse is injected in the crystal from the front face, while the pair of APDs is placed on the crystal rear face. In the endcap, the laser pulse is injected from an angle of the rear face, implying that the light is reflected before being collected by the VPT.

Either in the EB or the EE, the transparency of the crystal is tracked by comparing the intensity of the pulse measured by the photodetectors and the one provided by the reference PN diode. Crystals are grouped into monitoring regions, made up of 100 or 200 channels, which are connected to the same PN diode, as depicted in figure 3.2.

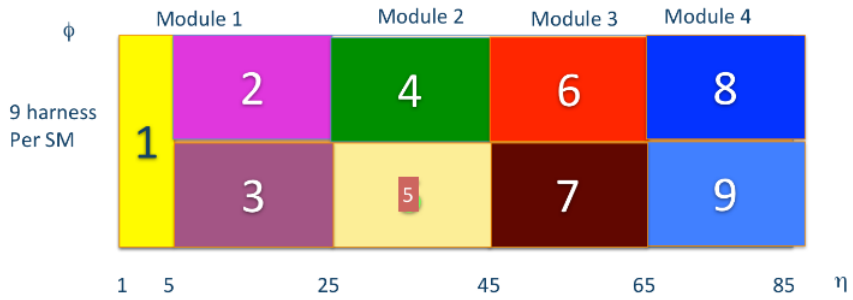

Figure 3.2: Scheme of the ECAL laser monitoring system. A pair of silicon PNs is used in each block to measure the intensity on the laser light injected in each crystal.

Since the scintillation light is produced isotropically, its optical path inside the crystal is, on average, longer than the light injected by the monitoring system. This difference is taken into account when computing the correction factors used to account for the transparency loss. It is possible to derive a relation between the intensity of the laser light and the scintillation light measured by the photodetectors. For an electromagnetic shower that would result in a measured amplitude S_0, the actual measured intensity S can be expressed as:

$$S = S_0 e^{-\frac{\Lambda_S}{\lambda_s}} \tag{3.1}$$

where Λ is the optical path of light and λ is an attenuation coefficient, depending on the crystal transparency. For the laser light, a relation of the same form but with different parameter holds:

$$R = R_0 e^{-\frac{\Lambda_R}{\lambda_R}} \tag{3.2}$$

combining these equations, it is possible to write:

$$\frac{S}{S_0} = \left(\frac{R}{R_0}\right)^{\alpha} \tag{3.3}$$

where $\alpha = \frac{\Lambda_S \lambda_R}{\Lambda_R \lambda_S}$ is a parameter depending on the particular crystal. The average value of α for the EB and EE was measured during beam tests before the CMS installation [55, 56]. The crystals of the ECAL were produced by two manufacturers: BCTP from Russia, which are the majority of the total ECAL crystals, and SIC from China. The α of the crystals by BTCP was measured to be 1.52 and 1.16, in the EB and in the EE, respectively, while the α of SIC crystals was measured to be 1, regardless from the detector region. Finally, the correction factor (LC) compensating for the transmission loss is computed as

$$LC(t) = \left(\frac{R(t)}{R(t_0)}\right)^{\alpha} \tag{3.4}$$

The reference response $R(t_0)$ is the one measured at the beginning of 2011. The evolution of R(t), since 2011, is shown in figure 3.3.

3.3 Energy measurement

The primary purpose of the ECAL is to measure the energy of photons and electrons with high precision. The front face of the ECAL crystals is comparable to the $PbWO_4$

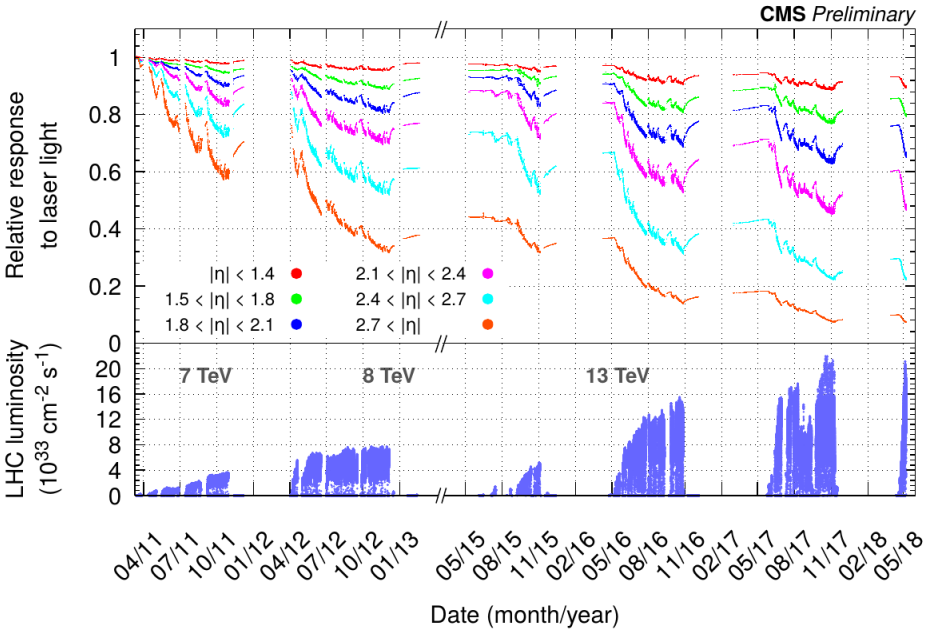

Figure 3.3: Relative response to laser light (440 nm in 2011 and 447 nm from 2012 onwards) injected in the ECAL crystals, measured by the ECAL laser monitoring system, averaged over all crystals in bins of pseudorapidity, for the 2011, 2012, 2015, 2016, 2017 and 2018 data taking periods, with magnetic field at 3.8 T. The response change observed in the ECAL channels is up to 10% in the barrel and it reaches up to 50% at $|\eta| \sim 2.5$, the limit of the tracker acceptance. The response change is up to 90% in the region closest to the beam pipe. The recovery of the crystal response during the periods without collisions is visible. These measurements, performed every 40 minutes, are used to correct the physics data. and include measurements taken up to May 2018. The bottom plot shows the instantaneous LHC luminosity delivered during this time period.

Moliere radius (1.959 cm), implying that an electromagnetic shower produced by a high-energy photon or electron, can span over several crystals. On average, a photon, or an electron, deposits 94% of its energy in a 3×3 matrix of crystals, and 97% in a 5×5 matrix of crystals centered on the one hit by the particle. Furthermore, while traversing the material upstream of the ECAL, photons can convert into electron-positron pairs, with the resulting charged particles bent and spread along the ϕ-coordinate by the strong CMS magnetic field, and electrons can radiate energy via bremsstrahlung radiation. Hence, to correctly measure the original energy of electrons and photons, it is essential to collect energy deposits, which can span along ϕ, in different regions of the detector.

The algorithm used during the LHC Run 2 is able to recover such energy. It starts from the localization of local maxima of energy deposited in the ECAL channels. If the energy reconstructed in the crystal is above a fixed threshold, different in the EB and EE, the channel is used to seed the cluster reconstruction. Then, starting from the seed, crystals are gathered to form *basic clusters* by aggregating channels with at least a corner in common with the cluster and with an energy above a threshold equivalent to twice the level of the electronic noise. Basic clusters are subsequently merged to form a *supercluster* if they lie inside the area delimited by two parabolas centered on the most energetic channel of the supercluster (seed channel), whose parameters are functions of η. This way to gather channels and merge clusters results in a mustache-like shape of the

supercluster, which is more accentuated in the high $|\eta|$ region, where the electromagnetic shower shape extension is not negligible not only along the ϕ direction but also in η.

The energy of the supercluster is computed as:

$$E_{e,\gamma} = F_{e,\gamma} \left[G \sum_i (C_i \cdot S_i(t) \cdot A_i) + E_{ES} \right] \tag{3.5}$$

where the index i runs on the crystals belonging to the supercluster, and the other terms represent:

$F_{e,\gamma}$: correction factor, computed by means of multivariate techniques. It compensates shower non containment, energy loss in the material upstream of the ECAL and non-perfect clustering. The difference in the interaction of electrons and photons when traversing the material in front of ECAL and the slightly different electromagnetic shower development inside the crystals are taken into account by computing different corrections for electrons and photons.

G: scale coefficient used to convert the signal output of the readout chain in ADC count, into energy expressed in GeV. Two different values are used for barrel and endcap. It is derived from the comparison of the energy scale in data and in MC simulation, using as a reference the electron-positron invariant mass reconstructed in Z→ee events.

C_i: channel intercalibration coefficient that accounts for channel-to-channel different response (more details are provided in section 3.4)

S_i: factor accounting for the time evolution of the crystal response due to loss of transparency induced by electromagnetic radiation damage. The procedure for its calculation is described in section 3.2.

A_i: signal amplitude of the i-th channel, expressed in ADC counts, estimated by means of the multifit algorithm described in 3.1.

E_{ES}: energy reconstructed in the preshower (only used in the region covered by the preshower detector)

3.4 Intercalibration

The ECAL energy resolution was measured at dedicated beam test [57] prior the CMS installation. It is parametrized by:

$$\frac{\sigma_E}{E} = \frac{A}{\sqrt{E(GeV)}} \oplus \frac{B}{E(GeV)} \oplus C \tag{3.6}$$

where A, B, and C represent, respectively:

the stochastic term: it accounts for shower containment variations, and fluctuations in the gain of the photodetectors. It was measured at beam tests to be 2.8%.

the noise term: it represents the contribution to the energy resolution due to the electronics noise of the readout chain. It was measured to be 12%, corresponding to an equivalent noise of 42 MeV.

the constant term: it includes several contribution, such as the non-uniformity of light collection, energy leakage from the rear face of the crystal, channel-to-channel response variation and instability of the response.

The ECAL was designed to achieve an unprecedented energy resolution, better than 1%, for photons produced by the Higgs boson decay (on average E_T ~60 GeV). At this energy scale, the energy resolution is entirely dominated by the constant term, which must be kept below 0.5% to preserve the required ECAL energy resolution. This demands for high accuracy in the computation of the channel intercalibration constants (IC) and stability of the ECAL response. Four methods can be used to derive the IC:

- ϕ-symmetry: this method is based on the assumption that in a large sample of soft interaction, the average energy deposited in each crystal at fixed pseudorapidity (η-ring at CMS) is the same. Therefore, observed channel-to-channel inequalities are ascribed to differences in channel response. The intercalibration is performed modifying the IC of the channels belonging to each η-ring, to equalize the energy measured in each crystal to the average of the channels at the same pseudorapidity. The main advantage of this method is that a moderate amount of data is necessary to compute the IC, typically in the 100-500 pb^{-1} range. On the other hand, the accuracy of the IC obtained by this method is systematically limited to a few percents, because of the non-perfect knowledge of the amount and arrangement of the material in front of the ECAL, which breaks the ϕ-symmetry. However, given that these systematic effects are constant in time, it can be used to track any possible time evolution of the IC.

- $\pi^0 \to \gamma\gamma$ and $\eta \to \gamma\gamma$ decay: it is based on an iterative procedure. The diphoton invariant mass distribution is fitted to a Gaussian distribution, accounting for the signal peak, plus a fourth-order polynomial function, describing the background. The value of the mass extracted from the fit is compared to the nominal π^0 and η mesons mass, and the IC of the channels included in the photon superclusters are updated at each iteration to correct the value of the mass obtained from the fit. Also for this method, a small amount of data is sufficient to derive the IC with good precision ($\sim$ 1% level), thanks to the high production rate of π^0 and η mesons at the LHC. The precision of the IC computed by this method is limited by systematics effect mainly due to the shower containment effect impacting the energy measurement. Furthermore, in the endcaps, the selection efficiency is low because of pileup and electronic noise in the readout chain of the detector.

- high energy electrons from W and Z bosons decay: also this method is based on an iterative procedure. The energy of the supercluster reconstructed in the ECAL is compared to the momentum measurement provided by the tracker system, which is used as a reference. The IC or the channels belonging to the supercluster are iteratively updated, constraining the E/p ratio to be 1. Contrary to the methods described above, a higher amount of data, corresponding to an integrated luminosity of about 10 fb^{-1} is necessary to achieve a statistical precision (defined as in section 3.4) at the level of few per-mill. However, the impact of systematic uncertainties on this method is by far lower.

- Z→ee events: It follows an iterative procedure. For each channel, the dielectron invariant mass distribution is built using for the measurement of the energy of the electron and the positron the information provided by the supercluster reconstructed

in the ECAL, without combining it with the momentum of the track measured in the tracker. The m_{ee} is fitted to a crystal-ball function [58], describing the signal shape, plus a polynomial function, for the background. The value of the Z boson mass extracted from the fit is compared to the nominal Z boson mass. The IC of the channels involved in the superclusters are updated at each iteration, constraining the mass obtained from the fit to the Z boson nominal mass.

In order to achieve the best possible accuracy, the sets of IC derived by each of the methods described above are combined per each η-ring, according to a weight equal to the precision of each algorithm.

All these methods provide the intercalibration of channels belonging to the same η-ring. The relative scale between different η-rings is derived through an independent procedure, which employs Z→ee events..

3.5 Intercalibration with electrons

The intercalibration of the ECAL employing high energy electrons produced in W and Z bosons leptonic decays compares the energy of the supercluster reconstructed in the ECAL to the momentum measured by the tracker system. For this reason, it also called E/p method.

3.5.1 The Algorithm

The iterative algorithm at the core of this method was originally developed to calibrate the L3 experiment [59] at the Large Electron-Positron collider (LEP) [22]. The idea at the basis of the intercalibration procedure is that for high energy electrons, the E_{SC}/p_{trk} ratio must be equal to one. Therefore, the algorithm modifies the IC of the ECAL channels belonging to the electron supercluster to constrain the E_{SC}/p_{trk} ratio to be as close as possible to its ideal value of 1. Since the momentum measured by the tracker is taken as a reference, it is assumed not to be affected by any biases. This caveat implies that any experimental bias in the momentum measurement must be corrected before the intercalibration procedure and residual effects considered as a source of systematic uncertainty.

At each iteration, the IC are computed through the following equation:

$$ic_i^N(i\eta, i\phi) = ic_i^{N-1}(i\eta, i\phi) \times \frac{\sum_{j=1}^{N_e} w_{ij} \cdot f\left(\frac{E_{SC}}{p_{trk}}\right)_j \cdot \left(\frac{p_{trk}}{E_{SC}}\right)}{\sum_{j=1}^{N_e} w_{ij} \cdot f\left(\frac{E_{SC}}{p_{trk}}\right)} \tag{3.7}$$

where i represent the i-th channel and the index j runs on the number of electrons used in the intercalibration procedure. The other terms are:

N: index of the iteration

$ic_i^{N-1}(i\eta, i\phi)$: intercalibration constant of the i-th channel, identified by its coordinates ($i\eta$, $i\phi$ in the barrel, and ix, iy, in the endcap) computed in the previous iteration.

w_{ij}: the fraction of the supercluster energy of the j-th electron deposited in the i-th channel. It depends on the IC and is updated at each iteration.

$f\left(\frac{E_{SC}}{p_{trk}}\right)_j$: weight which represents the probability of measuring a supercluster energy E_{SC} and a momentum p_{trk} in the tracker.

$\left(\frac{p_{trk}}{E_{SC}}\right)$: ratio between the momentum measured by the tracker and the energy of the electron supercluster reconstructed in the ECAL.

The calibration procedure must carefully treat dead channels. In fact, since an electromagnetic shower due to an electron interacting in the ECAL crystals develops across several channels, in the case of a dead channel in the supercluster, part of the electron energy would not be reconstructed, implying the supercluster energy to be systematically underestimated. The algorithm would compensate it, increasing the IC of the other crystals of the supercluster, leading to a bias in the IC computation. To avoid this effect, crystals in the proximity of not working channels, or matrices of 5×5 channels, called trigger towers (TT), are excluded from the procedure. In particular, a 3×3 matrix of crystals centered on the dead channel is excluded in the case of a single dead channel, while a 7×7 matrix of channel is excluded in the case of a whole dead TT.

The L3 algorithm is iterative, and the IC are updated at each iteration. A convergence criterion must be defined in order to assess the number of necessary iterations.The standard deviation of the distribution of the difference of the IC of each channel computed at the N^{th} step and the N-1^{th} step is taken as a figure of merit of the convergence of the algorithm. When convergence is achieved, the variation of the ICs at each step is expected to be low and not to variate with further iterations. The typical behavior is reported in the graph of figure 3.4, that refers to the intercalibration of the EB in 2017 data. The necessary number of iterations necessary to reach convergence is usually between 10 and 15.

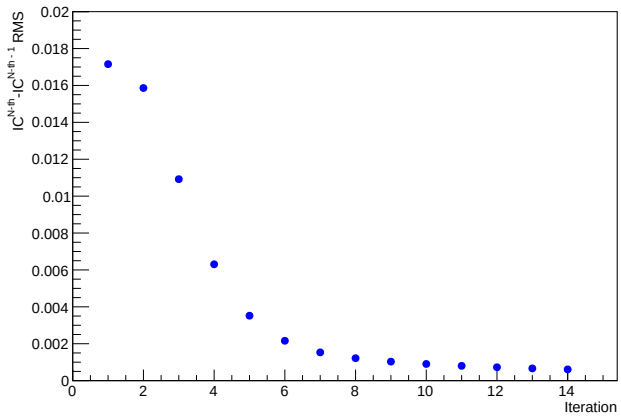

Figure 3.4: RMS of the distribution of difference of the IC computed at iteration N and N-1 for each channel as a function of the number of iteration. Convergence is reached after approximately 12 iterations.

In order to assess the accuracy of the obtained set of IC, the whole sample is divided into two independent sub-samples. The algorithm is run independently on the two sub-samples. For each η-ring, the standard deviation extracted from a Gaussian fit of the distribution of the difference of the IC of each channel, obtained with the two sub-samples, normalized to their sum, as in equation 3.8, is considered as a figure of merit of the intercalibration, called *statistical precision*, since in the ratio, the systematic contribution

Variable	Loose ID EB	Loose ID EE	Tight ID EB	Tight ID EE
H/E	0.104	0.0897	0.0597	0.0615
$\Delta\eta_{SC-trk}$	0.0105	0.00814	0.00926	0.00724
$\Delta\phi_{SC-trk}$	0.115	0.182	0.0336	0.0918
$\sigma_{i\eta i\eta}$	0.0103	0.0301	0.0101	0.0279
d_0	0.0261	0.118	0.0111	0.0351
d_z	0.41	0.822	0.0466	0.417
Isolation	0.0893	0.121	0.0354	0.0646
Missing inner hits	2	1	2	1
Conversion veto	True	True	True	True

Table 3.1: Identification requirements for electrons used in the calibration with the E/p method. Both the Tight and the Loose working points are reported, for electrons reconstructed in the EB and in the EE.

to the uncertainty cancels out, leaving only the statistical one.

$$\frac{ic_{even}(i\eta, i\phi) - ic_{odd}(i\eta, i\phi)}{ic_{even}(i\eta, i\phi) + ic_{odd}(i\eta, i\phi)} \tag{3.8}$$

The set of ICs obtained from the algorithm are rescaled by convention such that their average value in each η-ring is 1, resulting in an intercalibration only among channels at the same pseudorapidity.

3.5.2 Event selection

The selection of a pure sample of high energy electrons is the first step in the intercalibration procedure. The standard CMS primary dataset collecting events with one electron, called SingleElectron, and two potential electron candidates, called DoubleEG, are used. A set of selections slightly tighter than the one applied at trigger level are used. The identification of electrons relies on the shape of the reconstructed cluster in the ECAL and on its compatibility with an electron track reconstructed in the tracker. Moreover, electrons are required to be isolated. The definitions of the variables used for identification and the definition of the isolation variable are described in section 2.3.2.

Two working points for the identification of electrons, labeled in the following as Loose ID and Tight ID, are defined. The values of the selection employed for each of them, in the barrel and the endcap regions, are summarized in table 3.1

Events from the decay of W bosons are selected requiring the following selections:

- Exactly one electron reconstructed within the tracker acceptance ($|\eta|$ <2.5), with transverse momentum greater than 30 GeV, satisfying the tight ID criteria reported above
- A veto on the presence of other electrons satisfying the loose ID criteria
- A large amount of missing transverse energy ($\not{E}_T$), due to the escaping neutrino coming from the W boson decay, is required The $\not{E}_T$ of the event must be greater than 25 GeV
- W boson transverse mass, computed as $M_T = \sqrt{(2\not{E}_T E_T^e) \cdot (1 - \cos\Delta\phi)}$, with θ the angle between the electron and $\not{E}_T$ in the transverse plane, greater than 50 GeV

Z→ee events are selected trough the following criteria:

- An electron-positron pair, with both the particles satisfying loose ID criteria. If more than two pairs pass the selection, the pair with the highest p_T is used
- The dielectron invariant mass is required to be greater than 55 GeV

3.5.3 Further E/p selection

The E/p method relies on an a precise measurement of the electron momentum by the tracker. The typical E_{SC}/p_{trk} distribution for electrons reconstructed in the EB is shown in figure 3.5

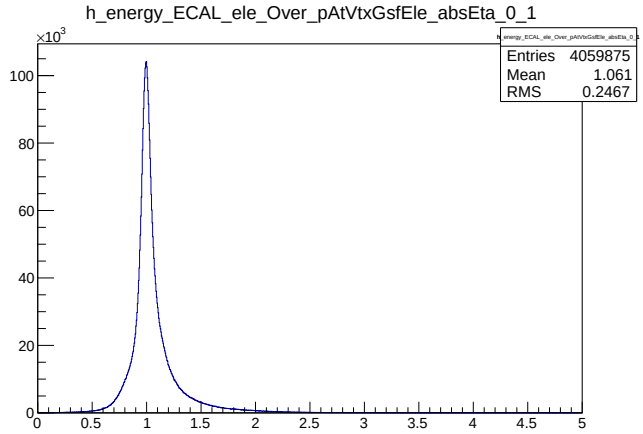

Figure 3.5: Typical distribution of E_{SC}/p_{trk} ratio in the region $|\eta| < 1$ in the EB.

The left and right tail of the distribution extends up to 0.5 and 2, respectively. Given the energy resolution of the ECAL for electrons satisfying the set of selections described in section 3.5.2 (at the percent level), such a significant deviation from the expected value of 1, is to be ascribed to a bad measurement of the momentum in the tracker. Including the events populating the tails in the E/p distribution would result in a worsening of the precision on the IC. Hence, a selection based on the E/p values itself is applied on each electron, at each iteration of the algorithm, in order to reject electron with a momentum poorly measured in the tracker and improve the accuracy of the intercalibration outcome. This is technically performed by modifying the weighting function $f(E_{SC}/p_{trk})$ in the calibration algorithm as in 3.9.

$$f\left(\frac{E_{SC}}{p_{trk}}\right) = \begin{cases} \frac{E_{SC}}{p_{trk}}, & \text{if } |E/p - 1| < \beta \\ 0, & \text{if } |E/p - 1| > \beta \end{cases} \tag{3.9}$$

where β is the width of the window in which electrons are retained.

The optimal value of β was assessed by running the intercalibration with several window widths, from 0.05 to 0.50, and using each set of IC to reconstruct a sample of Z→ee simulated events. The m_{ee} invariant mass peak was fitted to a Crystal-Ball function (CB) [58], accounting for detector effects, convoluted with a Breit-Wigner (BW) [60], representing the intrinsic width of the resonance. The standard deviation of the Gaussian

core of the distribution (σ_{CB}) was taken as a figure of merit of the optimization. This test indicated window widths of 0.15 and 0.20 in the EB and EE, respectively, as optimal values.

In principle, as possible drawbacks, this selection can induce either a loss of statistics or an artificial convergence on wrong IC values, since the E/p peak is artificially constrained around 1. This effect can be more significant for channels with a big initial miscalibration. The impact of the loss of statistics can be directly checked by comparing the statistical precision per η-ring achieved applying several values of the selection. The results obtained with data collected during 2015 are reported in figure 3.6, for several values of β, in the EB (top plot) and EE (bottom plot). In the inner part of the barrel (up to approximately η-ring 30), no significant difference can be observed among the different values of β., since the distribution of E/p is well centered around 1 and the tails of the distribution are not huge. Moreover, it was assessed that a statistics correspondent to an integrated luminosity of less than 10 fb^{-1} is sufficient to achieve less than 5 per-mill statistical precision, implying that the loss of statistics does not affect the precision of the intercalibration procedure. Divergences among the curves related to different width of the selection get larger and larger after η-ring 40. This can be explained by two effects: on one side, the selection aids in reducing the impact of electrons with a bad momentum measurement on the calibration precision, resulting in an improvement in the statistical precision. However, it can be noticed that for small values of β, such as 0.05 and 0.10 (green and red curves, respectively), the statistical precision gets worse. For what concerns the β=0.10 curve, its behavior is due to the loss of statistics that in the high η region of the barrel, where a huger number of electrons is necessary. The same applies to the green curve, up to η-ring 65, beyond which the curve gets stuck at a roughly constant values. This artifact is a side effect of the very tight selection, that does not allow the algorithm to accommodate for changes in the intercalibration constants, resulting in better statistical precision that does not correspond to a real improvement in the intercalibration precision.

The possible bias in the outcome of the calibration due to an artificial convergence of the algorithm on wrong values was evaluated using a MC sample of Z→ee events. An initial miscalibration of ±30% was applied in 8 modules over the total 16 modules in the EB and in 10 over 20 in the EE, before running the E/p algorithm. The distribution of the difference between the set of IC obtained from the algorithm and the true ones (IC_{true}) present in the MC simulation, was used to spot the presence of any biases. It was verified that, as expected in absence of bias, the distribution in different regions of the ECAL had a Gaussian shape with mean equal to 0. The RMS of the same distributions is also related to the accuracy in the IC computation. Its value as a function of the window width was plotted for different bins in pseudorapidity. The typical behavior in the barrel and the endcaps are reported in figure 3.7, in the left and right, respectively. The outcome of this test is in agreement with that observed in the previous study described earlier in this paragraph.

3.5.4 The Momentum Scale Calibration

The E/p method relies on the strong assumption that the tracker measures the momentum of electrons with no bias. Unfortunately, this is not true at CMS. The impact of this bias on the IC outcome was studied utilizing MC simulation. The intercalibration procedure

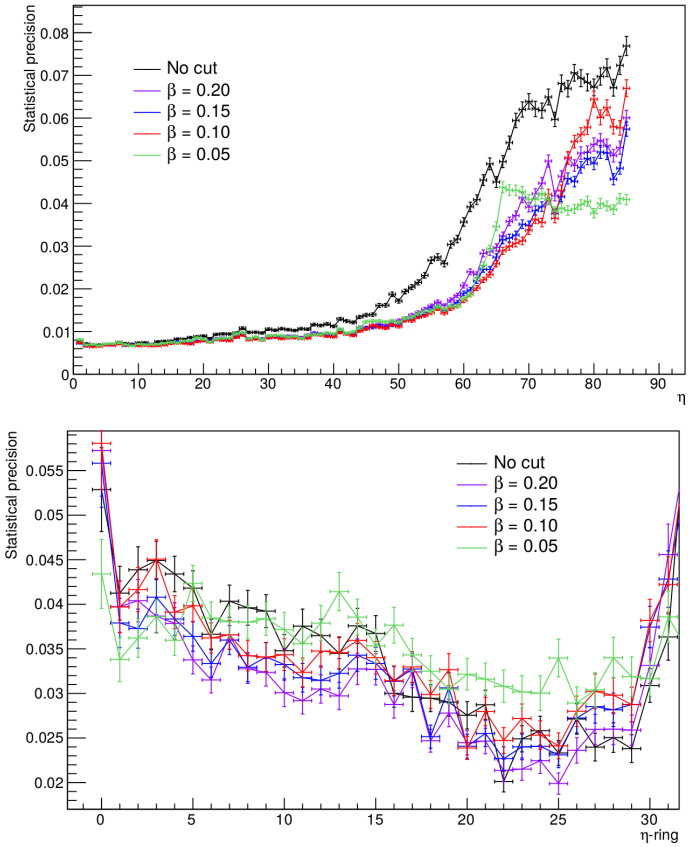

Figure 3.6: Statistical precision obtained with several widths of the selection on E/p, in the EB (top) and in the EE (bottom).

was run on a simulated sample of W plus jets events. The iη-iϕ map of the obtained IC is shown in figure 3.8 for the EB and EE, in the left and right plot, respectively.

In both the maps, a clear, systematic bias with a modularity along ϕ is visible. It is due to a biased measurement of the electron momentum in the tracker, probably caused by the tracker structure, which shows a similar ϕ-modularity.

In order to be able to achieve the required accuracy in the IC determination, these effects must be corrected. At this purpose, Z→ee events selected in data are exploited to derive a ϕ-dependent scale factor that equalizes the tracker momentum scale along the azimuthal angle. The ECAL is divided into 360 ϕ bins (1 bin for each degree in the ϕ angle). Each electron is assigned to a bin according to the position in the ECAL of the seed of its supercluster. For each bin, the dielectron invariant mass distribution is built using for the electron in the considered bin, the momentum measured by the tracker, and for the other electron the energy of the supercluster measured by the ECAL, as in equation 3.10

$$m_{ee}^2 = 4 \cdot E_{SC} \cdot p_{trk}(i\phi) \cdot \sin^2(\theta/2) \tag{3.10}$$

where the angle θ is the polar angle between the electron and positron tracks. A template distribution of the m_{ee} variable, divided by the nominal value of the Z boson mass, is built considering all the electrons, therefore integrating along the ϕ-coordinate, and using as energy measurement, the one provided by the ECAL only. Separate templates are built

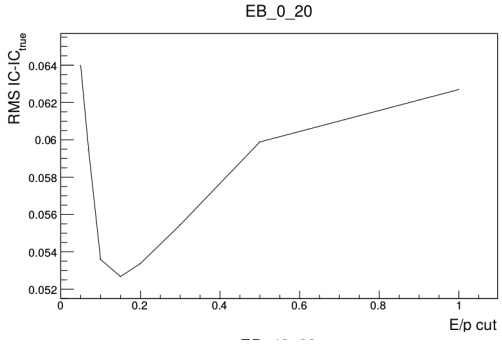

Figure 3.7: RMS of the distribution of the difference between the IC computed with the E/p method and the IC used in the MC simulation, as a function of the width of the selection on E/p applied in the algorithm.

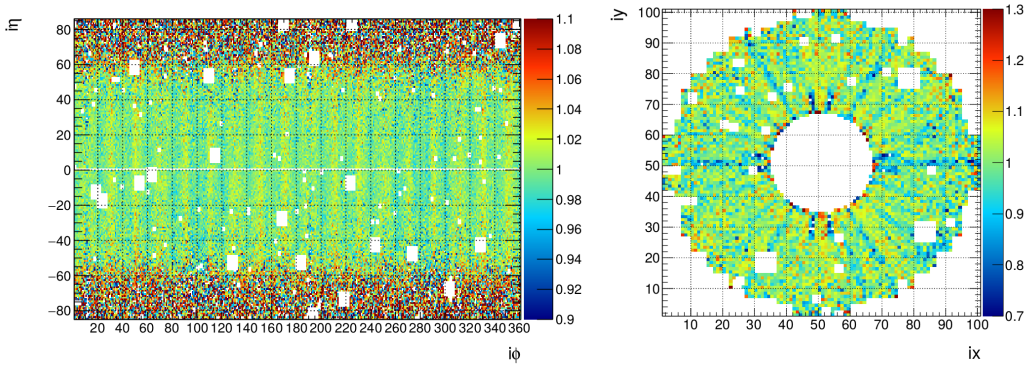

Figure 3.8: η-ϕ map of the IC constants obtained with the E/p calibration method in the EB (top) and EE (bottom) without performing a preliminary calibration of the momentum measured by the tracker. Biases periodic in ϕ due to tracker structures are clearly visible.

for the barrel and the two endcaps. Then, for each bin, the m_{ee}/m_Z distribution is fitted to a modified version of the global template, obtained as:

$$f(x,k) = N \cdot k \cdot F(kx) \tag{3.11}$$

where x is equal to m_{ee}/m_Z, N is the number of events in the considered ϕ bin, and k is a factor that quantifies the drift and rescale of the template distribution F(x) to fit the distribution in the ϕ-bin. This momentum calibration is performed for electrons and positrons separately, in order to account for any possible different bias due to the different curvature they undergo in the CMS magnetic field. An example of the result of the momentum scale calibration described above, obtained with data collected during 2017, is reported in figure 3.9, for the EB in the top plot and EE in the bottom plot. A clear modularity along ϕ is visible.

A scale factor up to the half percent in the barrel and two percent in the endcaps is extracted. It is used to correct the bias in the electron momentum measurement provided by the tracker system induced by its structures.

Even if at this step the energy measured by the ECAL is used to correct the bias in the electron momentum measured by the tracker, which is later used as a reference to intercalibrate the ECAL, this is not expected to induce any bias in the outcome of the

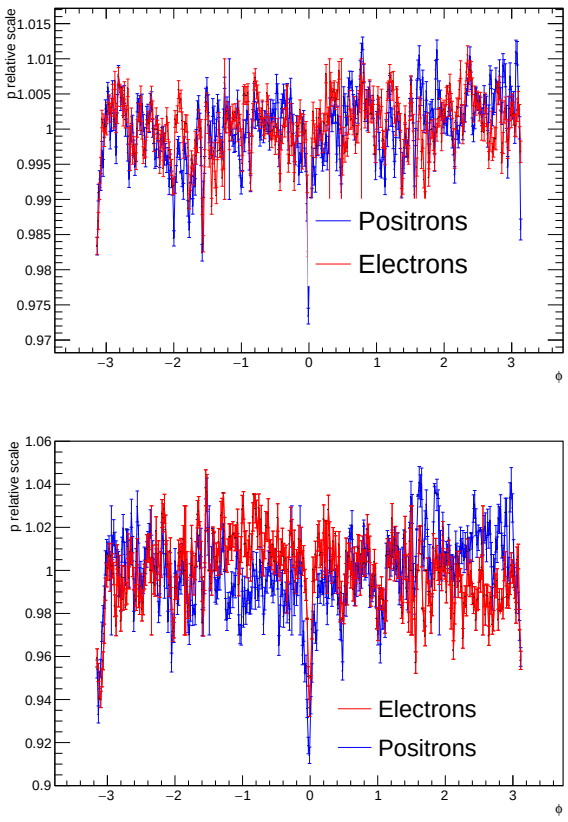

Figure 3.9: Correction factors used to account for biases in the momentum measured by the tracke, as a function of iϕ in the EB (top) and EE (bottom).

intercalibration, since the clusters considered in the di-electron invariant mass reconstruction are uniformly distributed in the ECAL, whose average response does not present any bias. Moreover, this approach enable to exploit the excellent energy resolution of the ECAL for electrons in the energy range typical of electrons produced in the Z boson decays (p_T >30 GeV), to improve the precision on the correction factors derivation.

3.5.5 ECAL intercalibration with 2017 Dataset

The intercalibration with high energy electrons is one of the methods employed to compute the IC in 2017. It played an important role, especially in the endcap region, since the π^0 method was not employable beyond $|\eta|$=2, because the reconstruction of the diphoton decay was very poor due to pileup and electronics noise of the detector readout chain.

The set of selections described in section 3.5.2, including the E/p selection of section 3.5.3 were applied to identify electrons from the W and Z boson leptonic decays, with a well-reconstructed momentum in the tracker system. The statistics available for the intercalibration corresponds to an integrated luminosity of 41.9 fb^{-1}. The map of the IC obtained with E/p method is reported in figure 3.10, for the EB in the top map and the EE plus and minus side, in the bottom left and bottom right maps, respectively.

The block structures arising in the map relative to the EB corresponds to the PN monitoring region of the laser system. During the run 2 of the LHC, time-dependent variations

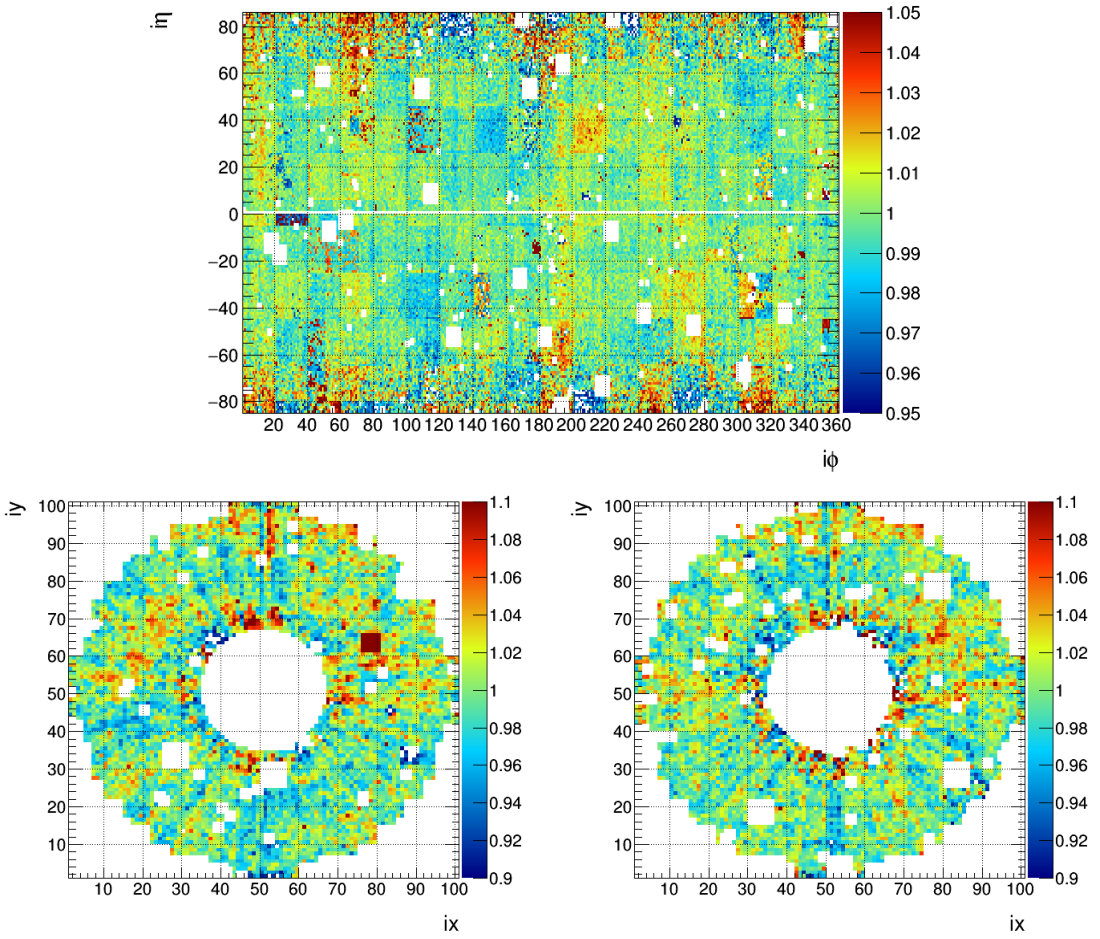

Figure 3.10: η-ϕ map and ix-iy maps of the IC obtained with the E/p method with the 2017 dataset, in the EB and in the EE, respectively.

of the channel response were observed and corrected through effective corrections derived with high energy electrons (the procedure is described in section 3.8). The modularity visible in the IC map reflects a different response at the beginning of the data-taking, for which the effective corrections do not account for. The statistical precision per each η-ring is shown in the plots of figure 3.11.

In the inner part of the EB, a statistical precision approaching 0.2% was achieved. A statistical precision better than 1% is reached for all the EB but the 4^{th} module, where the huge amount of material placed in front of the ECAL spoils the accuracy of the IC computation. The statistical precisions achieved in the two sides of the EE are compatible. They are at the level of 1% across all the η-rings but the first two rings, because of the amount of material upstream of the ECAL and the loss of part of the supercluster energy in the transition region between the EB and EE, and the last three rings, as a result of either a poorer energy resolution in the ECAL, mainly due to radiation damage of the crystals and noise in the detector readout, or a less accurate measurement of the electron momentum in the tracker.

The statistical precision of each of the method used in 2017 (electrons from W boson decays with the E/p method, Z→ee, and $\pi^0 \to \gamma\gamma$), together with their combination, is

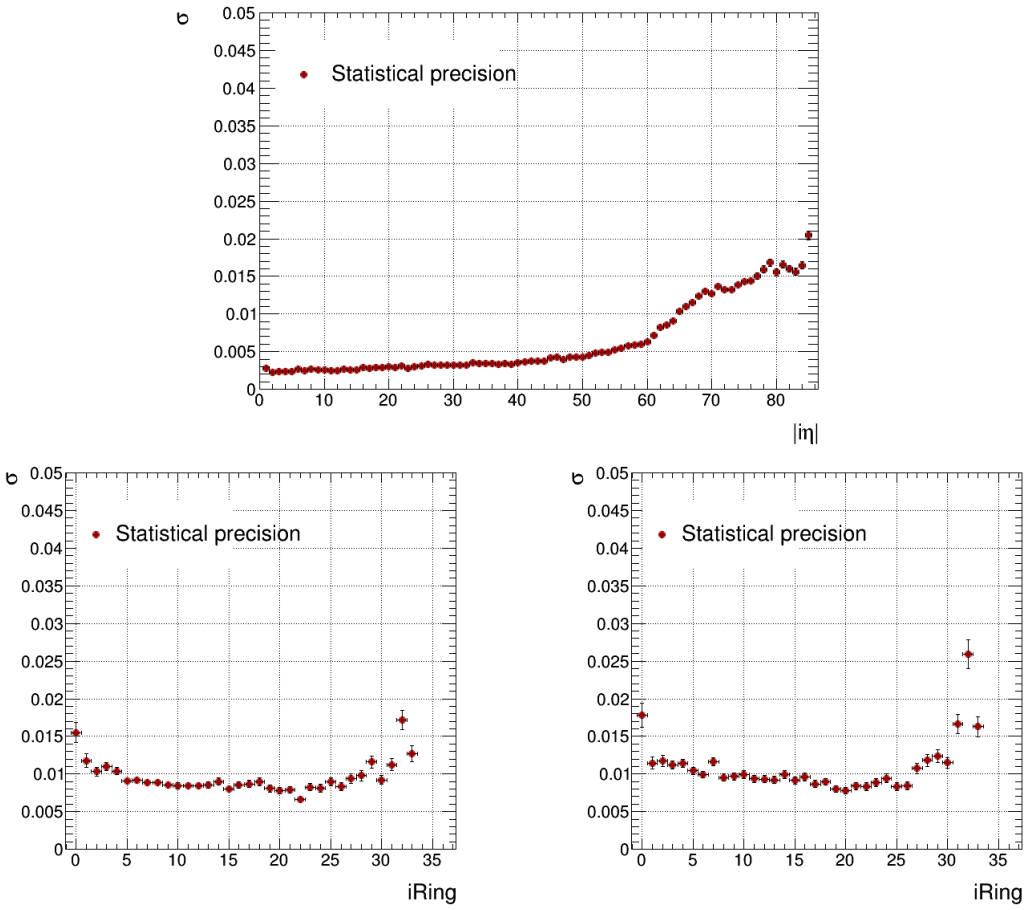

Figure 3.11: Statistical precision as a function of the η-ring achieved in the intercalibration with high-energy electrons with the 2017 dataset, in the EB (top) and in the EE (bottom).

showed as a function of pseudorapidity in figure 3.12.

The improvement in the energy resolution due to the more accurate intercalibration used in the reconstruction was assessed on Z→ee MC simulation events. In figure 3.13, it is possible to see the σ_{CB} obtained from an unbinned maximum likelihood fit of the e^+e^- invariant mass distribution, in bins of pseudorapidity, for low bremsstrahlung electrons in the left plot, and for high bremsstrahlung electrons in the right one. The improvement in energy resolution due to the dedicated calibration performed exploiting the full 2017 dataset is up to 10% in the EB, and 25% in the EE.

The residual discrepancy between the ECAL energy resolution in the MC simulation and data is cured by adding an additional Gaussian smearing to the energy measurement in the MC simulation. The necessary extra-smearing is derived by comparing the resolution in the m_{ee} distribution in Z→ee events in data and MC simulation.

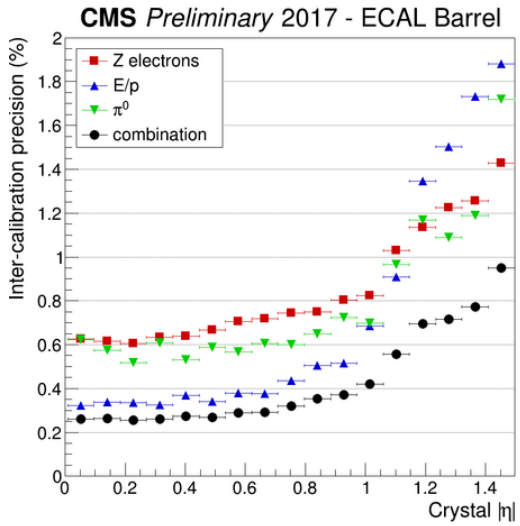

Figure 3.12: ECAL crystals intercalibration using the data collected in 2017. Precision of the channel inter-calibration, using energy deposits, as a function of the pseudorapidity η in the ECAL barrel detector. The precision for measuring the inter-calibration constants from Z→ee, $\pi^0 \to \gamma\gamma$ decays, and electrons arising from W and Z boson decays compared to the tracker response (E/p), is shown as a function of η in EB using 2017 data. The precision of the Z→ee and photon inter-calibrations is at the level of the systematic errors. The precision of the E/p inter-calibrations is still dominated by the statistical errors for $|\eta| > 1$. The black points represent the precision of the combination of the three methods (weighted average).

3.6 ECAL response monitoring

The time evolution of the ECAL channels response due to the transparency loss of the $PbWO_4$ crystals induced by radiation is tracked and corrected by a laser system of monitoring, which was presented in section 3.2. The correction factors derived from the laser system measurements are validated on data using the same concepts at the basis of the intercalibration methods:

ϕ-simmetry: this method is able to compute IC with a statistic of a few hundreds of inverse picobarn, implying that a set of IC can be derived every 1-2 days. The evolution of the IC of each channel is used to validate the the corrections measured by the laser monitoring system.

π^0: the relative variation of the π^0 mass measured in the diphoton decay channel is used to track the evolution of the response of the ECAL and validate the corrections computed by the laser system in the barrel region. This method can not be used in the EE, because in this region the reconstruction of the π^0 invariant mass is spoiled by the effects of electronics noise and pileup. Given the huge production of π^0 mesons at the collisions at the LHC, it is possible to check the stability of the ECAL response with a fine time granularity, greater than the typical duration of a fill of the LHC (about 12 hours).

high energy electrons from W and Z boson decays: the evolution of the E_{SC}/p_{trk} ratio is used to monitor the ECAL response stability and validate the laser corrections. This method is described in detail in section 3.7.

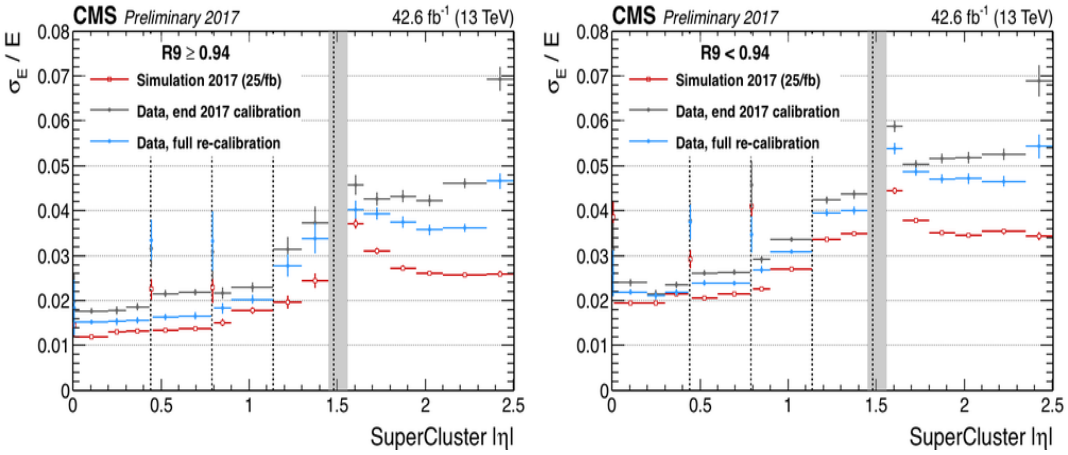

Figure 3.13: Relative electron (ECAL) energy resolution, unfolded in bins of pseudorapidity (η) for the barrel and the endcaps. Electrons from Z→ee decays are used. The resolution is showed separately for very low bremsstrahlung electrons (named golden, with R9>0.94 with R9 = E3x3 / ESC) and for bremsstrahlung electrons (R9 < 0.94). The relative resolution σ_E/E is extracted from an unbinned likelihood fit to Z→ee events, using a Voigtian (Landau convoluted with Gaussian) as the signal model. The resolution is plotted separately for data and MC events. The ECAL conditions used in the simulation reflect the status of the detector as predicted after 25/fb of data-taking in 2017. Conclusions: the resolution is affected by the amount of material in front of the ECAL and is degraded in the vicinity of the eta cracks between ECAL modules (indicated by the vertical lines in the plot). Also, the resolution improves significantly after a dedicated calibration using the full 2017 dataset (blue points) with respect to the end-of-year (EOY) 2017 calibration (gray points) for which only time dependent effects were corrected for.

Z→ee: the invariant mass reconstructed in di-electron events is used to track the evolution of the ECAL energy response.

3.7 ECAL response monitoring with electrons

The validation of the laser corrections and the monitoring of the ECAL energy response stability with high energy electrons rely on the same assumption driving the L3 intercalibration algorithm: for high energy electrons the E_{SC}/p_{trk} ratio is equal to 1. The same set of selections used in the intercalibration procedure, described in 3.5.2, are applied to identify a sample of suitable electrons from the W and Z boson decays. The first step of the procedure consists in the construction of a reference template distribution of E_{SC}/p_{trk}, using all the selected electrons. The same events are also collected in bins according to the time they were recorded. Each bin contains approximately 20000 and 15000 electrons, in the EB and EE, respectively. If the time distance between two consecutive events is longer than 24 hours, they are not included in the same bin, and a new one is created, in order not to merge events collected with different conditions of the detector. For each bin, the E/p distribution is built and fitted to a scaled version of the template distribution, defined as:

$$f(E_{SC}/p_{trk}, k) = N \cdot k \cdot F(k \cdot E_{SC}/p_{trk}) \tag{3.12}$$

where k is a factor quantifying how much the template distribution has to drift and scale to fit the distribution related to each time bin. The inverse of k is the quantity used to

monitor the stability of the ECAL response.

The time granularity achievable with this method is of order of some hours, enabling a monitoring of the response evolution also within an LHC fill (average duration $\sim$ 12 hours). The ECAL response stability obtained after a full reprocessing of 2016 data with improved reconstruction and calibration of the ECAL, including time-dependent corrections, in the EB and in the EE is showed in the top and bottom plot of figure 3.14, respectively

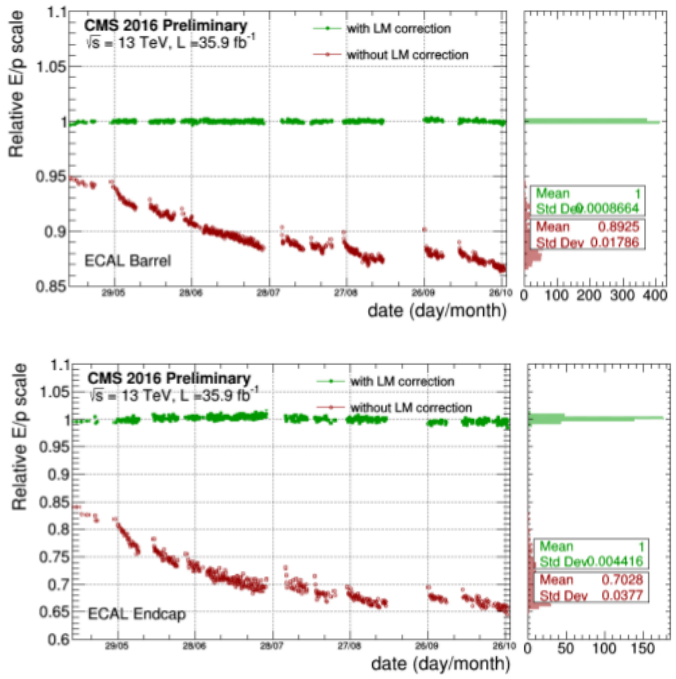

Figure 3.14: History plot for the reprocessed 2016 dataset of the ratio of electron energy E, measured in the ECAL Barrel, to the electron momentum p, measured in the tracker before (red points) and after (green points) corrections to ECAL crystal response variations due to transparency loss. A stable energy scale is achieved throughout 2016 run after applying laser corrections: in the EB (EE), the average signal loss is about 10% (20%), with an RMS stability after corrections better than 0.1% (0.5%).

3.8 Time-dependent energy scale corrections

Since 2016, a drift of the ECAL response in time was observed in the EB. It was first observed as a block pattern, compatible with the regions of the laser monitoring system sharing the same PN diode, in the η-ϕ map of the ratio of IC computed with data collected in different periods of the data-taking. An example is provided in the map of figure 3.15, where for each channel, the ratio between the IC derived with the π^0 method with data collected at the beginning, and at the end of the 2016 data-taking, is reported.

This effect was investigated at a deeper level with high energy electrons, employing the same concepts and techniques described in section 3.7. The monitoring of the ECAL

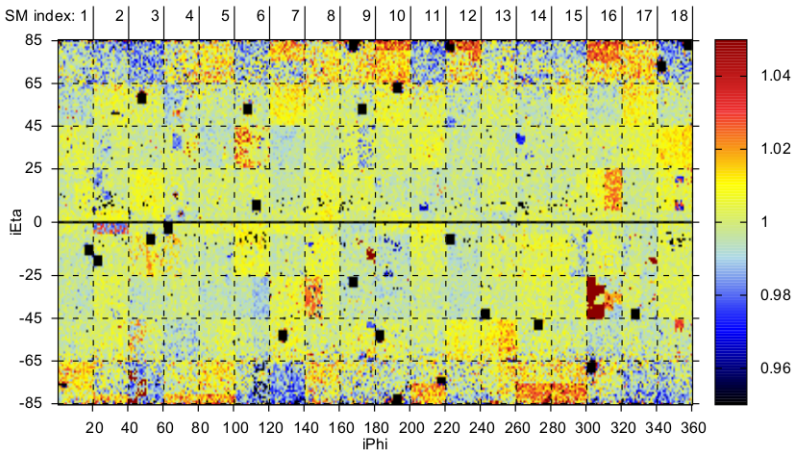

Figure 3.15: η-ϕ map in the EB of the ratio of the IC constants measured with using $\pi^0 \to \gamma\gamma$ decays with data at the end and at the beginning of the data-taking. Block arising in the map indicate a drift of the energy scale of the ECAL.

response was performed with a granularity equal to the size of a PN monitoring region. Each electron is assigned to a particular region, according to the position of the seed of its supercluster. The evolution in time of the E_{SC}/p_{trk} peak position was fitted to a straight line, whose angular coefficient represents the ECAL response loss per second (a slope of $3{\cdot}10^{-9}$ implies a drift of the energy scale of about 1 % in a month). An example of fit in a PN region located in the central EB is reported in figure 3.16.

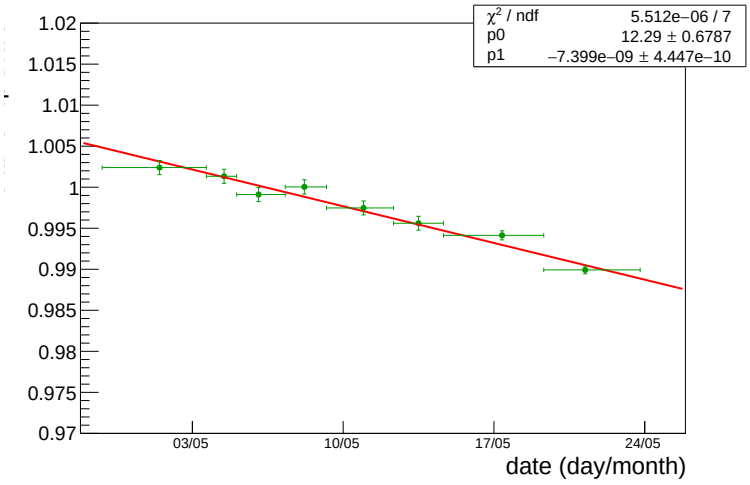

Figure 3.16: Example of the E/p peak position history, measured as explained in 3.7, in a single harness region located in the inner part of the EB. The fit to a straight line used to derive effective energy corrections is superimposed.

The map of figure 3.17 shows the slopes obtained in the 2016 for each PN region. The pattern in the slope map is in agreement with the one observed in the IC maps, confirming the drift in time of the ECAL response, whose behavior is compatible with an almost linear drift of the response of the monitoring system.

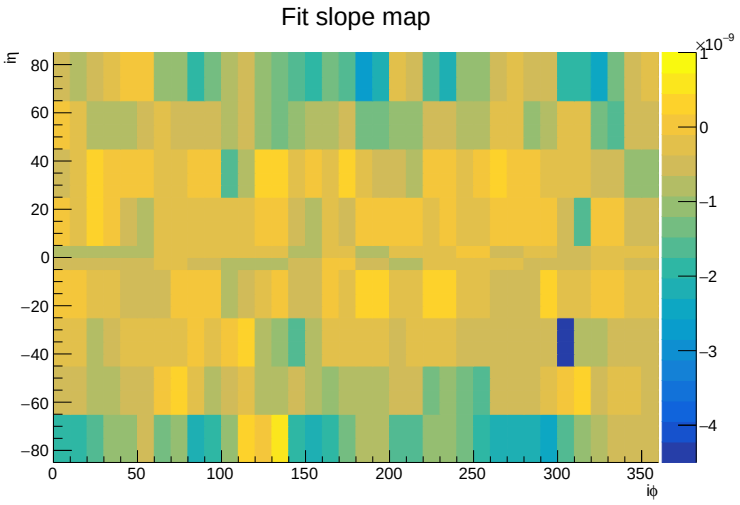

Figure 3.17: η-ϕ map of the slopes measured in the EB with the procedure described in the text.

Moreover, the response loss seems to correlate with η, suggesting that radiation damagemight be one of the possible roots of the response loss.

The slopes extracted from the fit were used to compute an effective correction, different for each region of the EB, to compensate the drift of the energy scale. The effect of the application of these corrections was tested with the $\pi^0 \rightarrow \gamma\gamma$ method to monitor the ECAL response. As it is possible to see in the plot of figure 3.18, it succeeded in restoring the ECAL scale stability in the barrel region at the level of the 0.3%.

The restored ECAL energy response stability also positively impacts the stability of variables related to the shape of the energy cluster reconstructed in the ECAL when photons and electrons traverse the detector. Such variables are crucial to discriminate between genuine photon and electron superclusters and jets with a huge electromagnetic component, and are exploited in analysis with photons in the final state, such as the H$\rightarrow \gamma\gamma$ analysis. The R9 variable can be considered as a candle to monitor the stability of variables describing the shape of energy clusters in the ECAL. Its evolution with time is monitored using Z$\rightarrow$ee events, binned according to the time the event was recorded. For each bin, the mode of the R9 distribution of electron and positron is taken as reference. The results of the R9 monitoring in 2017 after applying the energy scale corrections described in this section is showed in figure 3.19. Stability about the 0.3% is achieved after applying the corrections described in this paragraph and the channel intercalibration.

3.9 Energy correction factors for photon reconstruction

In order to achieve ultimate performance in terms of electrons and photons energy resolution, a further refinement of the energy estimation was developed initially in the context of the H$\rightarrow \gamma\gamma$ analysis and currently used also in the standard electron and photon reconstruction at CMS [50]. A multivariate regression technique is used to estimate the

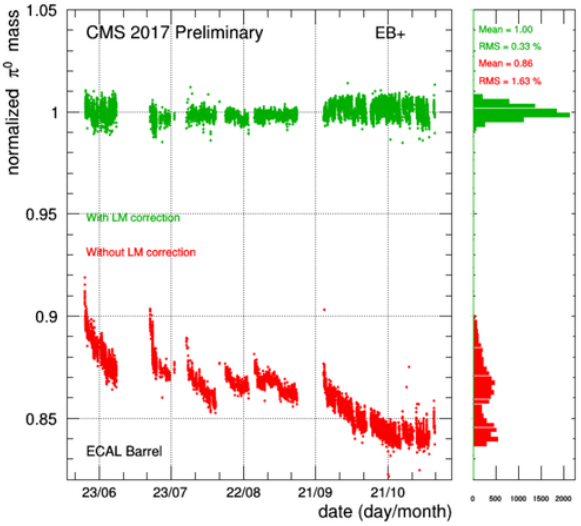

Figure 3.18: Stability of the relative energy scale measured from the invariant mass distribution of $\pi^0 \to \gamma\gamma$ decays in the ECAL barrel (positive z, EB+). The energy scale is measured by fitting the invariant mass distribution of approximatively 500k photon pairs in the mass range of the π^0 meson. Each point is obtained from a fit to approximatively 5 minutes of data taking. The energy scale is plotted as a function of time over the 2017 data taking period. The plots show the data with (green points) and without (red points) light monitoring (LM) corrections applied. The right-hand panel shows the projected relative energy scales.

correction factors entering the measurement of the energy of the superclusters. They account either for global effects, especially in the case of converted photons or local containment of the electromagnetic shower, affected by the energy loss in gaps and cracks between the ECAL modules. The input variables are the η and ϕ coordinate of the supercluster, quantities describing its shape in the ECAL, such as the R9 variable, the energy-weighted extension in terms of crystals of the supercluster along η and ϕ, and the energy reconstructed in the HCAL in the region behind the supercluster in ECAL. In the endcap, the ratio between the energy measured in the preshower detector and the supercluster energy is also used. These variables provide information about the probability and the location of a photon conversion, and the subsequent loss of energy due to showering in the material upstream of the ECAL. Additional information concerning the seed cluster and the seed crystal is provided, such as the relative energy and position of the seed cluster, ratios between the energy reconstructed in matrices with various sizes containing the seed channel and the total supercluster energy, and the position and energy of the seed crystal. This information is exploited to estimate the quantity of energy loss due to modules gaps, compensating local non-containment of the electromagnetic shower. The number of reconstructed vertices and the density of energy in the event are employed in the regression to account for the contribution of pileup events to the energy measurement.

A semi-parametric regression method is used [61]. The training is performed using MC simulation samples of true photons, with the target of the regression being the ratio E_{true}/E_{raw}, where E_{true} is taken from the MC simulation truth information and E_{raw} is measured as the sum of the energy deposits in each crystal of the supercluster. The regression output is the probability density function (pdf) of the E_{true}/E_{raw} variable for each photon. The most probable value of the pdf is taken as the correction to use in the

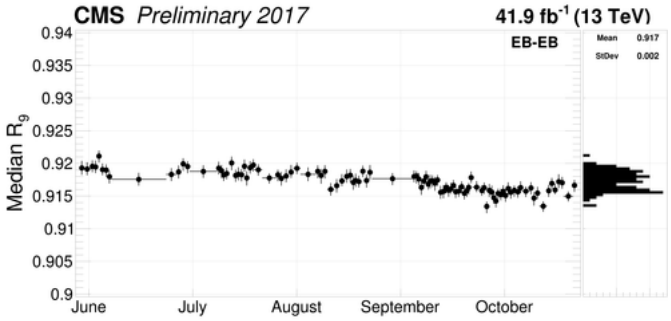

Figure 3.19: Stability of the shape of the electromagnetic deposits in the ECAL barrel for electrons from Z decays. A dedicated re-calibration using the full 2017 dataset was performed. The plot shows the stability of the variable R9, which is responsive to changes in pedestals and noise. Each point on the left panel is obtained by taking the median of the R9 histogram for the respective time bin. The histogram on the right shows the spread of median R9. The R9 variable is stable within 0.3% during the year.

energy measurement.

The improvement in energy resolution due to these correction factors are estimated in Z→ee events, with the electrons reconstructed using only ECAL-related information. In the plots of figure 3.20, the distributions of the m_{ee}, with both the electron and the positron in the EB (left plot) and EE (right plot) with data collected in 2015, using reconstruction algorithms including different levels of refinement in the energy estimation, is showed. In particular, they are: the energy reconstructed in a 5×5 matrix of crystals centered on the most energetic channel, the raw energy of the supercluster, computed summing the energies reconstructed in the channels belonging to the supercluster, and the same with the further application of the corrections provided by the regression.

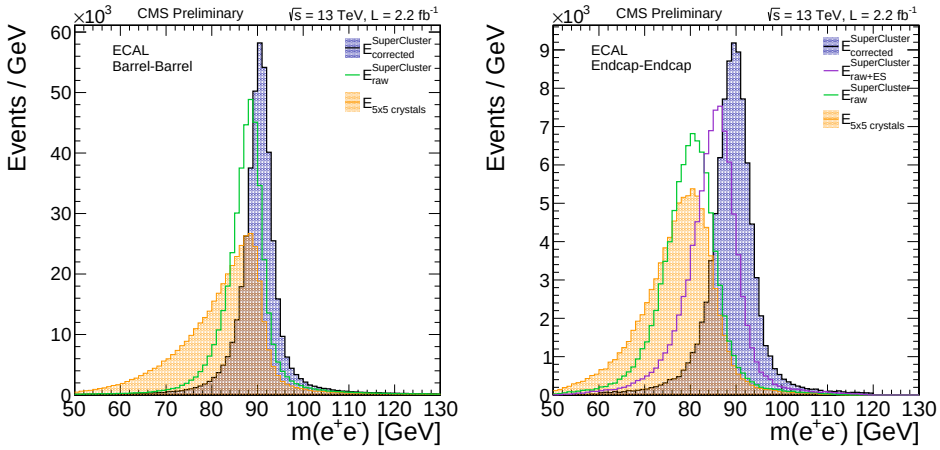

Figure 3.20: The two plots show the improvements to the Z→ee energy scale and resolution from the incorporation of more sophisticated clustering and cluster correction algorithms (energy sum over the seed 5×5 crystal matrix, bremsstrahlung recovery using supercluster, inclusion of preshower (ES) energy, energy correction using a multivariate algorithm). The invariant mass of the two electrons is reconstructed using all of the 2015 Run 2 ECAL data at B=3.8 T [62].

3.10 Conclusions

The main purpose of the ECAL is the precise measurement of the energy of photons and electrons, with a target resolution of 1 GeV for photons produced in the Higgs boson diphoton decay. Two of the main contributions to the ECAL energy resolution are due to the correction of the loss of transparency of the ECAL crystals as a result of irradiation and the channel intercalibration. Even if the ECAL is provided with a laser monitoring system able to measure and correct the loss of response of the ECAL crystals, a drift of the ECAL response was observed during 2016 and 2017 data-taking. A time-dependent correction was derived employing electrons from the decay of the W and Z bosons, restoring the ECAL response stability in time at the per mill level in the EB. Moreover, in 2017, a new intercalibration of the ECAL channels was performed. The improvement in energy resolution was assessed using the di-electron decay of the Z boson to be up to 10% in the EB and 15% in the EE. The benefits at analysis level are multiple and affect not only the reconstruction of photons and electrons but also of other physics objects such as jets and tau leptons, both in terms of energy resolution and data/MC agreement. In particular, for what concerns the H$\rightarrow \gamma\gamma$ analysis, the main benefit is the improvement in the diphoton invariant mass resolution that results in a higher signal-over-background ratio and sensitivity of the analysis. The gain in sensitivity of the analysis due to the intercalibration performed in 2017 can be estimated to be about 5%.

Chapter 4

The H$\rightarrow \gamma\gamma$ analysis with the 2016 dataset

In this chapter, the H$\rightarrow \gamma\gamma$ analysis of the data recorded in 2016 by the CMS experiment from the proton-proton collisions at a center-of-mass energy of 13 TeV provided by the LHC is presented. The analyzed amount of data corresponds to an integrated luminosity of 35.9 fb^{-1}.

4.1 Analysis strategy

The analysis relies on the excellent performance of the CMS, and in particular of the ECAL, in terms of identification and energy measurement of photons.

The main sources of background for the H$\rightarrow \gamma\gamma$ channel arise from diphoton production, which is an irreducible background for the analysis, and production of a single photon plus jets (pp$\rightarrow \gamma$+jets), and QCD production of multiple jets in the final state, with the jets being misidentified as photons, whose impact can be reduced thanks to a high-performance photon identification. At this purpose, a dedicated algorithm optimized for photons produced in the Higgs boson diphoton decay was developed.

To improve the sensitivity of the analysis, exclusive categories are defined to select events produced through specific mechanisms, according to the presence of additional particles in the final state, such leptons or jets identified as originated from a bottom quark decay, which allows to tag the event as arising from a specific production mechanism. If no other objects but the diphoton is reconstructed, the events are collected in "untagged" categories defined according to the event kinematics and per-event estimated diphoton invariant mass resolution through a multivariate classifier.

In the statistical analysis, the signal model is built from MC simulation of the four dominant production processed at the LHC (ggF, VBF, VH, ttH) with different values of the Higgs boson mass. The background parametrization is totally data-driven. The Higgs boson signal strength is extracted by means of likelihood fits to the diphoton invariant mass reconstructed in each category.

4.2 Data samples

The analyzed dataset corresponds to an integrated luminosity of 35.9 fb^{-1}, collected by the CMS experiment during the 2016. The analysis is performed in the invariant mass

region $100 < m_{\gamma\gamma} < 180$ GeV, keeping the signal region between 115 and 135 GeV blinded for the optimization of the analysis strategy and selections.

4.2.1 Trigger

At the L1 trigger, the data used in the analysis are selected requiring a combination of algorithms selecting events with a single or a pair of electromagnetic candidate in the calorimeter system, labeled as SingleEG and DoubleEG, respectively. Due to the limited bandwidth available at the L1 trigger, the threshold applied to the single object algorithm was set at 40 GeV. For what concerns the DoubleEG algorithms, lower thresholds can be used without saturating the available bandwidth. During 2016, the lowest transverse energy thresholds in DoubleEG paths, were 22 and 15 GeV. Since both the photons coming from the Higgs boson decay can meet the trigger requirements, the maximum efficiency is achieved when a combination of SingleEG and DoubleEG algorithms is used to seed the HLT.

At the HLT, a more refined event reconstruction is performed and stricter selections can be applied , improving the purity of the selected sample and reducing the bandwidth assigned to the path. The identification exploits variables related to shape of the cluster of energy reconstructed in the calorimeters, such as the R9 and the isolation variables. The transverse energy thresholds applied in the HLT path used to select diphoton events, are of 30 and 18 GeV, for the leading and subleading photon of the event, respectively. In addition, a minimum diphoton invariant mass of 90 GeV is required.

4.3 Simulated samples

Simulated events are generated with the MC method. A detailed simulation of the CMS detector [63] based on the GEANT4 [64] software is used to simulate the interactions of particles with the CMS subdetectors and the subsequent signal that are read out. The effect of additional interactions either in the same bunch crossing or in the adjacent ones, known as *pileup*, is taken into account by adding soft interactions to the hard-scattering process. The simulated events are weighted such that the distribution of the number of vertices in the MC samples, agrees with the one obtained in data.

4.3.1 Signal samples

Signal MC samples of the four main production processes are produced assuming different values for the Higgs boson mass, ranging from 120 to 130 GeV, and are used to build the signal model used in the statistical analysis. The MC samples related to the Higgs boson associated production with a pair of bottom quarks (bbH) and a single top quark (tH) are produced considering only a mass of the Higgs boson equal to 125 GeV. The signal samples of the four dominant production mechanism are generated with MADGRAPH5_aMC@NLO [65]. The parton level samples are interfaced to PYTHIA 8 [66] for parton showering and hadronization. The cross-sections and branching ratios suggested by the LHC Cross-Section Working Group [67] for each sample are used. The samples used in the analysis are listed in table 4.1.

4.3.2 Background samples

The main usage of the MC samples of the background processes is the validation of the MVA tools employed in the analysis, and the study and optimization of the selections

Production mode	Event Generator	Cross Section for m_H=125 GeV (pb)
ggH	MADGRAPH5_aMC@NLO	48.5800
VBF	MADGRAPH5_aMC@NLO	3.7820
VH	MADGRAPH5_aMC@NLO	2.2569
ttH	MADGRAPH5_aMC@NLO	0.5071
tHq	MADGRAPH5	0.0742
tHW	MADGRAPH5	0.0151
bbH	MADGRAPH5_aMC@NLO	0.5329

Table 4.1: List of MC simulation samples of signal events used in the analysis.

used in the analysis categories.

Two generators were used for the production of the simulated events: MADGRAPH [65], with PYTHIA 8 [66] for the hadronization stage, and SHERPA [68], which includes also the hadronization step of the simulation. The background MC samples used for the training and validation of the MVA techniques common to all the categories of the analysis are listed in table 4.2

Production mode	Event Generator	Cross Section for m_H=125 GeV (pb)
Diphoton + jets ($m_{\gamma\gamma}$ 40-80)	MADGRAPH5_aMC@NLO	48.5800
Diphoton + jets ($m_{\gamma\gamma}$ 80-Inf)	MADGRAPH5_aMC@NLO	3.7820
γ + Jets P_T 20-40 $m_{\gamma\gamma}$ 80-Inf	MADGRAPH5_aMC@NLO	2.2569
γ + Jets P_T 20-Inf $m_{\gamma\gamma}$ 40-80	MADGRAPH5_aMC@NLO	2.2569
γ + Jets P_T 40-Inf $m_{\gamma\gamma}$ 80-Inf	MADGRAPH5_aMC@NLO	2.2569
Dijet P_T 30-40	MADGRAPH5_aMC@NLO	0.5071
Dijet P_T 40-Inf	MADGRAPH5_aMC@NLO	0.5071
Drell-Yan with Z→ll + 0-2 Jets	MADGRAPH5_aMC@NLO	0.5071

Table 4.2: List of MC simulation samples of main background sources the for the analysis.

4.4 Vertex identification

The diphoton invariant mass ($m_{\gamma\gamma}$) is computed as:

$$m_{\gamma\gamma} = \sqrt{2E_1E_2(1-\cos\theta)} \tag{4.1}$$

where E_1 and E_2 are the energy of the photons and θ is the opening angle between their flight direction. The correct identification of the signal vertex affects the estimation of θ with an impact on the $m_{\gamma\gamma}$ resolution that can be as large as 1 GeV. For the events with the best energy resolution (two unconverted photons reconstructed in the ECAL barrel), 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. It was found to be negligible with respect to the contribution due to the energy resolution on photons, if the signal vertex is identified within 1 cm along the beam axis (z-axis) from the true one. The determination

of the Higgs boson decay vertex in the diphoton channel is not trivial, in particular in the case of unconverted photons, which can not be tied to any vertex exploiting hits produced in the tracker. Furthermore, the ECAL has no pointing capabilities, since it is not longitudinally segmented. However, the recoiling tracks associated to the vertices can be exploited to identify the primary vertex of the event. Obviously, if one of the photons convert, the additional information provided by the conversion tracks is used, improving the identification performance.

4.4.1 The vertex identification BDT

The algorithm developed for the vertex identification purpose is based on a boosted decision tree (BDT) that exploits the correlation between the diphoton system and the recoiling tracks. In the case of unconverted photons, three variables are used:

- sumpt2 = $\sum_i |\vec{p}_t^{\,i}|^2$
- ptbal = $-\sum_i p_T^i \cdot \frac{\vec{p}_T^{\,\gamma\gamma}}{|\vec{p}_T^{\,\gamma\gamma}|}$
- ptasym = $\frac{|\sum_i \vec{p}_T^{\,i}| - p_T^{\gamma\gamma}}{|\sum_i \vec{p}_T^{\,i}| + p_T^{\gamma\gamma}}$

where the i index runs over all the charged particles associated to the vertex, p_T^i is the transverse momentum of the i-th track associated to the vertex, and $p_T^{\gamma\gamma}$ is the transverse momentum of the diphoton pair. The sumpt2 variable is expected to have higher values for the signal vertex than for the other vertices of the event, reflecting the higher activity associated to the hard-scattering vertex. The ptbal variable, which corresponds to the negative sum of the projections of the transverse momentum of each track on the diphoton transverse momentum, tends to be positive in the signal vertex, since tracks recoil against the diphoton system. On the other hand, it is distributed around 0 for the wrong vertices, since there is not a preferred direction. The same applies to the ptasymm variable, which measures the asymmetry between the total momentum of the tracks associated to a vertex and the diphoton pair momentum. This variable peaks at -1 for a wrong vertex, while it takes higher values for the correct signal vertex.

When one of the photons, or even both, convert, the conversion tracks can be used to aid the reconstruction of the z-coordinate of the primary interaction vertex. Two methods were developed and used, with different performance depending on the distance between the conversion vertex and the ECAL surface. They both exploit the tracks of the electron-positron pair to determine the flight-direction of the converted photon, which is extrapolated back to the beam line to assess the z-coordinate of the vertex. In the first method, the angle between the momentum reconstructed from the conversion tracks and the beam axis (α_{conv}) is used to derive the z-coordinate of the signal vertex (z_{PV}^{conv}), through the equation:

$$z_{PV}^{conv} = z_{conv} - r_{conv} \cot(\alpha_{conv}) \tag{4.2}$$

where z_{conv} and r_{conv} are the z-coordinate and the distance from the beam axis of the conversion vertex.In the second method, the direction of the converted photon is determined employing a combination of the information related to the conversion vertex and the superclusters reconstructed in the ECAL. The z-coordinate of the primary vertex (z_{PV}^{SC}) is computed as:

$$z_{PV}^{SC} = z_{conv} - r_{conv} \frac{z_{SC} - z_{conv}}{r_{SC} - r_{conv}} \tag{4.3}$$

where z_{conv} and r_{conv} are the same as in the previous method and z_{SC} and r_{SC} are the z position and the distance from the beam axis of the supercluster in the ECAL. The two methods are complementary since if the conversion occurs early in the inner part of the tracker, the photon direction is well reconstructed from the electron-positron pair tracks. On the other hand, if the conversion occurs in the outer region of tracker, less hits are available for the track reconstruction, resulting in a poorer estimation of the photon direction, but the impact point of the photon in the ECAL is determined with higher precision. When at least one photon conversion occurs, two more variables are used in the vertex identification BDT: the number of converted photons, and the pull_{conv}, with:

$$pull_{conv} = \frac{|z_{PV} - z_{PV}^{conv}|}{\sigma_{conv}} \tag{4.4}$$

where z_{PV} is the z-coordinate of the tested vertex, σ_{conv} is the resolution on the vertex z-coordinate reconstructed with conversion tracks, according to the tracker region where the conversion occurs, and z_{PV}^{conv} the vertex location on the z-axis estimated by the method providing the best resolution on z_{conv} between the two described above. It is estimated according to the region of the tracker where the conversion vertex is reconstructed.

4.4.2 Validation of the vertex identification BDT

Since the algorithm is optimized on MC simulation, it is necessary to validate in on data. The validation of the vertex identification BDT in the case of unconverted photons is performed comparing the performance in a sample of $\rightarrow \mu\mu$ events selected in data and in a MC simulation sample of Drell-Yan events, with the Z boson decaying into a muon pair. The muon tracks are used to identify the right signal vertex and are subsequently removed from the event before running the vertex identification algorithm, in order to mimic a diphoton system topology. The efficiency as a function of the transverse momentum of the Z boson is showed the plot of figure 4.1.

In the case of at least a photon conversion, the BDT is validated using γ+jets events in data and simulated events, with the photon required to undergo a conversion. The tracks inside the jet provide a reference to tag the right vertex with high efficiency. Once the primary vertex is tagged, they are removed from the event to mimic the topology of the diphoton system.

Scale factors are derived from Z$\rightarrow \mu\mu$ events, in bins of Z boson p_T, to account for the observed discrepancy between the performance in the MC simulation events and data.

4.4.3 Correct vertex probability

The probability to locate the vertex with an accuracy of 1 cm is estimated on a per-event basis by means of a second BDT. The probability obtained by the MVA is subsequently provided as input of a BDT used to categorize the events, in order to identify those events for which the vertex is likely to have been correctly located, resulting in a negligible impact on the $m_{\gamma\gamma}$ resolution due to the diphoton opening angle reconstruction. A second BDT is trained to estimate this probability. The same MC simulation samples used for the vertex identification BDT are used for the training of this second BDT. The training input variables are:

- transverse momentum of the diphoton system.

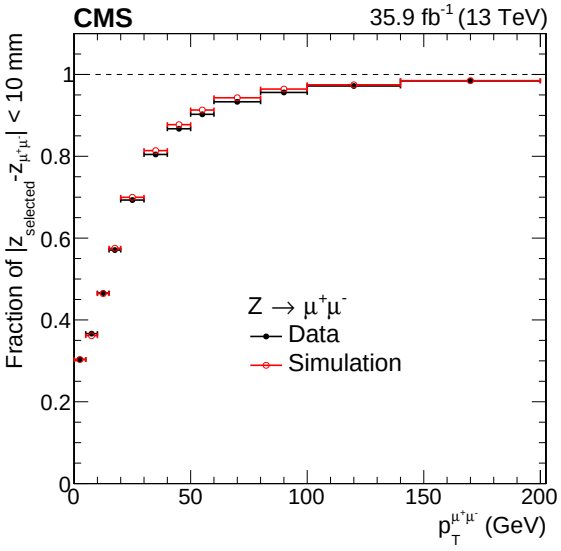

Figure 4.1: Validation of the diphoton vertex identification BDT performed on Z→ $\mu\mu$ events.

- the number of vertices in the event.
- BDT score for the three best ranked vertices obtained from the vertex identification BDT
- distance along the z-axis between the best-ranked vertex and the second and the third vertex in the ranking
- number of converted photons

The per-event probability to reconstruct the z-coordinate of the vertex within 1 cm of the true one is parametrized separately for unconverted and converted photons, as a function of the BDT score, with a 4-th order polynomial function with a constraint at BDT score = -1, where the probability is 1.

The algorithm was validated with Z→ $\mu\mu$ and photon plus jet events, for unconverted and converted photons, respectively. The procedure is the same as the one described in section 4.4.2.The algorithm was validated with Z→ $\mu\mu$ and photon plus jet events, for unconverted and converted photons, respectively. using the same procedure described in section 4.4.2.

4.4.4 Performance on H→ γγ simulated events

The performance of the vertex identification algorithm was tested on a MC simulation sample of H→ $\gamma\gamma$ events with m_H=125 GeV. The efficiency of choosing the vertex within 1 cm of the true one is showed in figure 4.2 as a function of the reconstructed Higgs boson transverse momentum (left plot) and as a function of number of vertices in the event (right plot). The overall efficiency is about 81%.

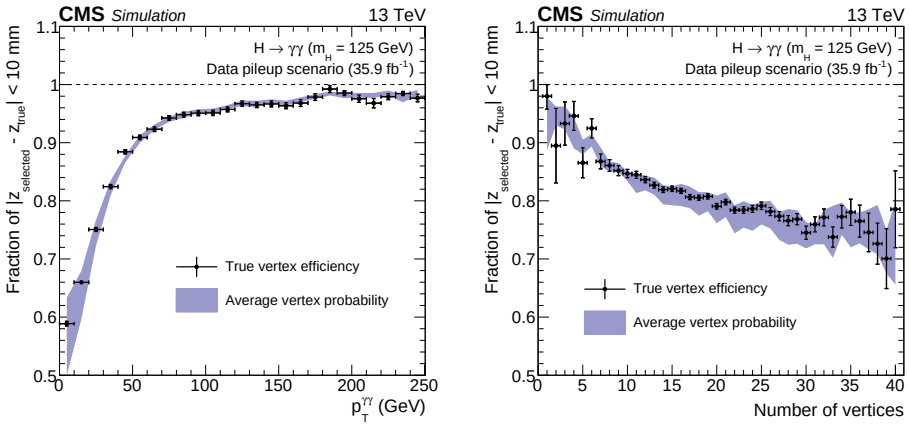

Figure 4.2: Comparison of the true vertex efficiency and the average vertex probability as a function of the reconstructed transverse momentum of the diphoton system and the number of vertices in the event in simulated H→ $\gamma\gamma$ events with m_H=125 GeV.

4.5 Further photon identification and corrections

As already mentioned, a huge contribution to the background of the analysis comes from reducible background due to γ+jets and QCD events, with the jets misidentified as photon. It occurs because of the production of high-energy neutral mesons in the jet fragmentation that can decay into a pair of close-by photons, which are reconstructed in the detector as a single, high-energy photon. The photon identification algorithm aims to reject this source of background exploiting the differences in the shape of the energy cluster in the ECAL between true and fake photons, and the isolation variable. A dedicated Boosted Decision Tree (BDT) optimized for photons produced in the decay of a Higgs boson was developed and used in the analysis.

The BDT training is performed on γ+Jets simulated events. Photons emitted by partons produced in the hard-scattering interaction are considered as prompt photons and are identified through the MC truth information by spatially matching the reconstructed candidate to the generator-level photon, while all the remaining reconstructed photons are considered as non-prompt photons. In order to obtain a training as independent as possible from the photon kinematics, the transverse momentum and pseudorapidity distribution of the supercluster associated to prompt photons were weighted to match the distributions of the non-prompt photons.

The analysis preselection are applied before training in order to improve the algorithm performance in the signal phase space. The input variables are related to the shape of the energy cluster reconstructed in the ECAL, on the isolation variable, and the position of the cluster in the detector. The usage of shower shape variables is motivated by the fact that even if collimated photon pairs produced in neutral mesons decays can be reconstructed a single photon, they tend to have a wider shower profile. The difference with respect to genuine photons is more prominent along η, since along the ϕ-coordinate, it is partially washed out by the effect of the magnetic field. The isolation variables provide further discrimination power since in the case of fake photons, energy deposits in the detector due to the particles produced in the jet are expected in the proximity of the cluster.

They input variables are:

- R9: the ratio between the energy reconstructed in a 3×3 matrix of crystals centered on the most energetic crystal (seed) of the supercluster and the energy of the

supercluster.

- $E_{2\times2}/E_{SC}$: the ratio between the energy reconstructed in a 2×2 matrix containing both the seed and the second most energetic channel in the supercluster, and the energy of the supercluster
- $\sigma_{i\eta i\eta}$: extension of the supercluster along η, measured in terms of crystals
- $cov_{i\eta i\phi}$: covariance along η and ϕ of the crystals within a 5×5 matrix centered on the seed of the supercluster
- Supercluster width along η
- Supercluster width along ϕ
- Preshower σ_{RR}: standard deviation of the shower spread in the preshower in the x-y plane (used only in the region covered by the preshower detector $1.65< |\eta| <2.8$)
- Charged isolation with respect to the vertex selected by the verted identification BDT: sum of the transverse momentum of charged particles associated to the vertex, lying in a cone of size R=0.3 around the photon direction
- Photon isolation: sum of the transverse energy of candidates identified as photons falling inside a cone of size R=0.3
- Charged isolation with respect to the worst vertex: sum of the transverse momentum of charged particles lying in a cone of size R=0.3 around the photon and associated to the vertex resulting in the largest isolation sum
- η of the supercluster
- Raw energy of the supercluster correspondent to the photon candidate
- ρ: average density of energy reconstructed in the detector. Futhermore, this variable is used to correct all the isolation variables in order to account for the contribution due to pileup.
- E_{ES}/E_{raw}: the ratio between the energy reconstructed in the preshower detector in correspondence of the photon-related supercluster, and the supercluster raw energy.

4.5.1 Photon identification performance

The performance of the photon identification algorithm was tested using simulated MC samples corresponding to the main background processes of the inclusive analysis (diphoton production, γ+jets, and multijet events), and the H→ $\gamma\gamma$ signal sample. The distributions of the photon identification BDT output of the lower-scoring photon in data and in simulated events are showed in figure 4.3 for events passing the analysis preselections and with $m_{\gamma\gamma}$ between 100 and 180 GeV.

In the analysis preselections, both the photons are required to have a BDT score higher than -0.9, corresponding to a signal efficiency of 99% and 49% background rejection. Further background rejections is achieved by including the BDT score of each photon among the input variables of a second MVA used to categorize diphoton events.

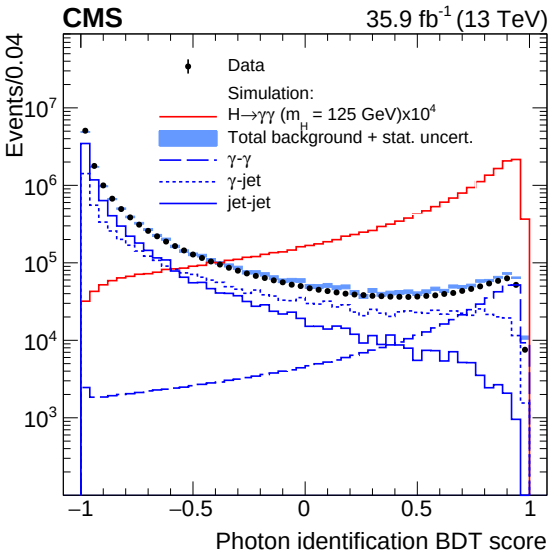

Figure 4.3: Distribution of the photon identification BDT score of the lowest-scoring photon of diphoton pairs with $100 < m_{\gamma\gamma} < 180$ GeV and passing the analysis preselections in the 2016 dataset (black dots) and simulated background events (blue histogram), including the distributions per background component are also showed, and simulated signal events with m_H=125 GeV (red histogram).

4.5.2 Validation of the photon identification algorithm

The performance of the BDT used for photon identification purpose is very sensitive to the accurate modeling of the electromagnetic shower development and detector response in the simulation. The level of agreement between data and simulation in the variables related to cluster shape and isolation, which are the input variables of the photon ID BDT, is probed by comparing them in Z→ee events selected in data and in simulated events, with electrons reconstructed as photons (no information from the electron track is used). Since the clustering algorithm used in the ECAL does not make any assumptions on the particle that originated the cluster, this process represents an excellent proxy for photons. Moreover, the reference provided by the Z boson mass peak enables to select a high purity sample. A HLT path imposing tight identification criteria and a minimum transverse momentum of 27 GeV of the electron is applied to both data and MC simulation. The standard analysis preselections are also applied but the electron veto, which is inverted.

A non-perfect agreement between data and MC simulation arose from the comparison. It was improved through an event-by-event correction derived from Z→ee event, applied to shower shape variables in the simulation, to remap them in such a way that the total distribution in MC simulation matches the distribution observed in data. A dedicated correction is applied to the isolation variables. It is derived by means of a technique called *stochastic isolation*, which consists in adding a quantity to the isolation variable obtained by randomly sampling a template distribution a number of times that depends on η of the photon and number of vertices in the event.

Finally, the variables corrected with the procedures described above are used in the training of the BDT used for photon identification. In figure 4.4, the comparison between the BDT score distribution in data and simulated events is showed. In order to account

for the residual discrepancy between data and MC simulation, a systematic uncertainty of ±0.03 is applied to the BDT output. Moreover, a further linearly increasing term is added to account for the bigger discrepancy in the tail at low values of the BDT output. The systematic uncertainty assigned to the output of the photon ID BDT represents one of the main contributions to the systematic uncertainty on the output of the BDT subsequently employed to categorize events (4.6), which is implemented as a source of migration of events among different categories in the signal model.

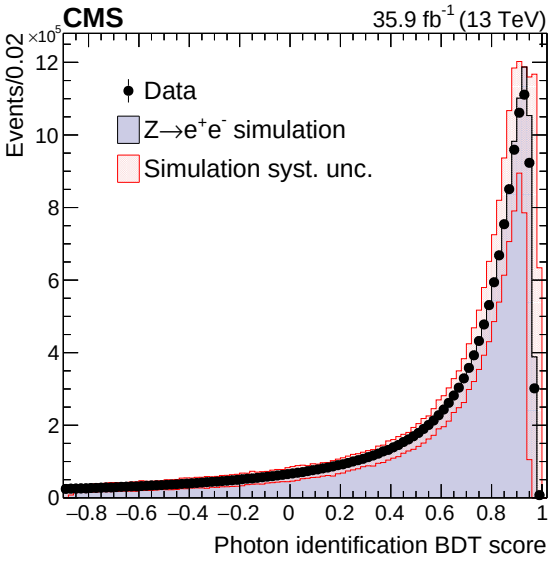

Figure 4.4: Distribution of the photon identification BDT score in Z→ee events, with electrons reconstructed as photons in data (black points) and simulated events (grey histogram), including the impact of the systematic uncertainty (pink region).

4.6 Diphoton reconstruction and classification

In the analysis preselection, photons are required to have a BDT score of the photon identification MVA higher than -0.9. In events where at least a pair of photons passes the analysis preselection, which will be described in detail in 4.7.1, a diphoton object is built combining the photon pair and the primary vertex selected by means of the vertex identification MVA presented in 4.4. If more than a pair of photons passes the preselections, the diphoton object with the highest transverse momentum is selected.

Diphoton objects are categorized according to the likelihood between kinematics in the event and the one expected in the signal, and the expected diphoton invariant mass resolution. A multivariate classifier using a boosted decision tree (BDT) is employed. The input variables are chosen in such a way that the classifier can not infer the diphoton invariant mass, in order to avoid any bias in the BDT output due to the mass of the Higgs boson in the samples used in the training. They input variables are:

- ratio between the transverse momentum of each photon of the pair and the diphoton invariant mass
- pseudorapidity of both photons

- cosine of the angle between the two photons in the transverse plane
- output of the BDT used for photon identification
- per-event estimate of the diphoton invariant mass resolution assuming a correct vertex assignment (σ_{rv})
- per-event estimate of the diphoton invariant mass resolution assuming a wrong vertex assignment (σ_{wv})
- per-event probability of correctly identifying the primary vertex

The invariant mass resolution in the case of correct vertex assignment is computed as:

$$\frac{\sigma^{rv}_{m_{\gamma\gamma}}}{m_{\gamma\gamma}} = \frac{1}{2}\sqrt{\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2} \tag{4.5}$$

The resolution on the photon energy measurement is extracted from the semi-parametric regression described in section 3.9. In the right vertex hypothesis, the impact on the diphoton invariant mass resolution due to accuracy in the determination of the opening angle between the photons, is negligible. The resolution on $m_{\gamma\gamma}$ under the wrong vertex assumption is calculated as:

$$\frac{\sigma^{wv}_{m_{\gamma\gamma}}}{m_{\gamma\gamma}} = \sqrt{\left(\frac{\sigma^{rv}_{m_{\gamma\gamma}}}{m_{\gamma\gamma}}\right)^2 + \left(\frac{\sigma_{vtx}}{m_{\gamma\gamma}}\right)^2} \tag{4.6}$$

where σ_{vtx} is the contribution to the diphoton invariant mass resolution due to a wrong identification of the diphoton vertex. This term can be analytically computed from the impact position of the photons reconstructed in the ECAL and the position of the selected vertex, assuming that the distance between the selected and the true vertex is distributed as a gaussian distribution with width equal to $\sqrt{2}\sigma_Z^{BS}$, where σ_Z^{BS} is the width of the beamspot along the beam axis. Its impact spans from less than 1 GeV, if the vertex is located with an accuracy of about 1 cm and can increase up to more than 3 GeV, dominating the total $m_{\gamma\gamma}$ resolution, if the distance along the z-axis between the true vertex and the selected one is larger than 2 cm. In the training of the BDT, the expected invariant mass resolution has to be manipulated to get an inverse proportionality with the signal-over-background ratio. For this purpose, the signal events are weighted with a weight w_{sig}:

$$w_{sig} = \frac{p_{vtx}}{\sigma^{rv}_{m_{\gamma\gamma}}} + \frac{1 - p_{vtx}}{\sigma^{wv}_{m_{\gamma\gamma}}} \tag{4.7}$$

where p_{vtx} is the correct vertex assignment probability. The effect of this weight is to push events with a better invariant mass resolution toward higher scores of the BDT.

MC simulation samples are used for the training. Simulated ggH, VBF, VH, and ttH events with mass of the Higgs boson equal to 124 GeV, weighted accordingly to the cross-section of each production process, is used as signal. MC simulation samples of diphoton, γ+Jets, and QCD events are used for the background information, each weighted to reflect the different cross-sections of the involved processes. A fraction equal to 1/4 of the total amount of MC simulated events are not used for the training in order to avoid any possible bias in the subsequent optimization procedure of the event categorization performed based on the diphoton BDT score.

The normalized distribution of the BDT score for signal and background, each normalized to unity, are showed in the plot of figure 4.5.

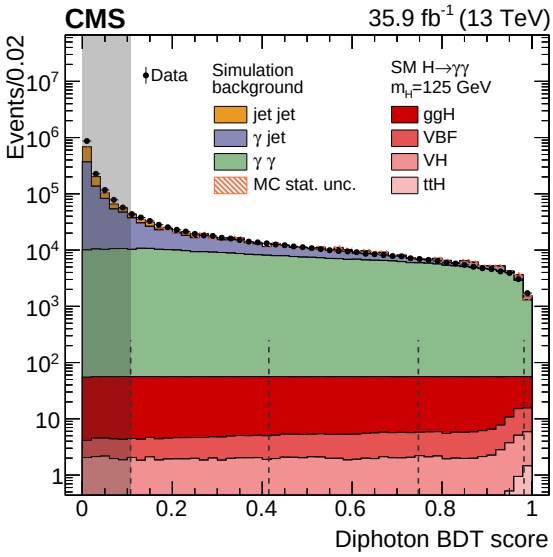

Figure 4.5: Distribution of the diphoton classifier BDT for events with photon pairs passing the analysis preselections in data (black points), simulated background events (colored histograms), and simulated signal events (red shade histograms). The output of the BDT was transformed to have a flat distribution for signal events. The hashed lines correspond to the boundaries of the untagged categories of the analysis, while the grey area corresponds to events that are discarded in the event.

The output of the BDT described in this section is used in a double way in the analysis. In the categories targeting specific production processes, it is used to enhance the background rejection. In the untagged categories, the output of the diphoton BDT is used to categorize events according to the expected diphoton invariant mass resolution and the compatibility with the kinematics in signal.

4.6.1 Diphoton BDT validation

Some of the input variables of the diphoton BDT, such as the output of the photon identification BDT and the estimated diphoton invariant mass resolution, are very sensitive to imperfect modeling in the MC simulation, resulting in a non perfect agreement between the diphoton BDT output in data and simulated events. These discrepancies affects the signal model, by potentially inducing a migration of events among categories. Z→ee events selected in data and a MC simulation sample of Drell-Yan events, with the Z boson decaying into an electron-positron pair, are used to assess the simulation accuracy and the level of agreement between BDT output distribution in data and MC simulation. Although electron from Z→ee events are considered a good proxy of photons from the Higgs boson diphoton decay, some intrinsic differences must be considered in the validation procedure of the output of the diphoton BDT. Firstly, the different spin of the Z and Higgs boson, 1 and 0, respectively, results in an intrinsically different kinematics. Furthermore, since the width of the Z boson resonance is comparable with the experimental resolution of the diphoton invariant mass, the BDT output is expected to be slightly less sensitive to not perfect modeling in the MC simulation than the H→ γγ decay.

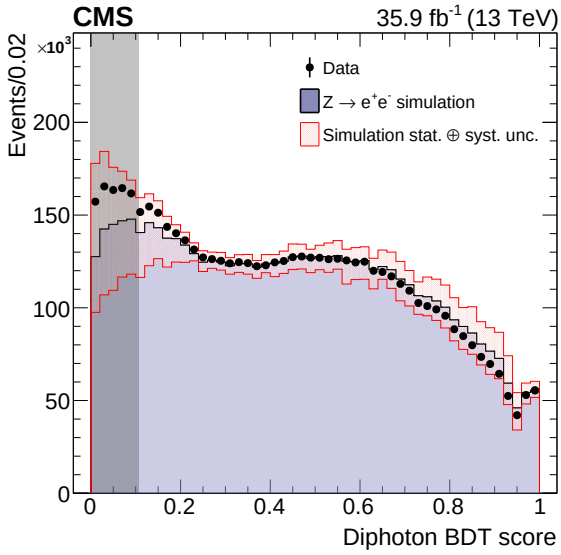

Figure 4.6: Distribution of the output of the diphoton classifier BDT in Z→ee events with electrons reconstructed as photons in data (points) and simulated events (grey histogram). The statistical and systematic uncertainties are represented by the pink back.

4.7 Event categorization

To improve the sensitivity of the analysis, the events passing the preselections(section 4.7.1) are categorized accordingly to the presence in the final state of physics objects which allow to tag the event as arising from a particular Higgs boson production mechanism. When no particles in addition to the photon pair is reconstructed in the final state, the output of the diphoton BDT described in section 4.6 is used to categorize the events.

The final state for VBF production is characterized by the presence of two high-energy jets, separated by a large distance in pseudorapidity. In the case of VH, the additional particles in the final state depend on the V, which can be either a Z or a W bosons, and its decay channel. Either the V boson decay to a pair of quark or the leptonic decays are considered. The ttH production is tagged by the presence in the final state of jets identified as originated by bottom quarks, since the top quarks decay into a bottom quark with a branching ratio almost equal to 1, in addition to the decay product of the W bosons, which can be either leptons or jets.

The categories targeting exclusive production mechanisms are listed below, ordered according to their priority:

- ttH leptonic: ttH events with one or two leptons produced in fully leptonic or semi-leptonic decays of the top quark
- ZH leptonic: associated production of the Higgs boson with a Z boson, decaying into an electron or muon pair
- WH leptonic: events with a lepton and huge amount of missing transverse energy due to the leptonic decay of the W boson
- VHLeptonic Loose: tag aiming at collecting both ZH and WZ leptonic events, which do not satisfy the tight requirements of the two previous tag

- ttH hadronic: ttH events with multiple jet final state due to top quarks hadronic decays
- VBF categories: 3 tags target events with two jets in the forward region
- VH MET: events with large amount of missing transverse energy
- VH hadronic: jet pairs resonating on the mass of the W or Z boson are used to tag the VH production with the V hadronically decaying

Events with no physics objects in addition to a photon pair in the final state are collected in *untagged* categories, which are mainly populated by ggH events. Such events are categorized on the base of the output of the diphoton identification BDT described in section 4.6, in order to aggregate events with high diphoton invariant mass resolution in high-performance categories with high signal over background ratio.

In addition to the improvement in the analysis sensitivity, the usage of exclusive categories to tag specific production processes, significantly enhance the accuracy in the measurement of the couplings of the Higgs boson. In particular, the tagging of VBF and VH mechanisms, improve the measurement of the Higgs boson coupling to vector bosons. Analogously, the tag of ttH production, provides a way to directly probe the Higgs boson coupling to the top quark.

4.7.1 Preselection

A set of preselections is applied to photons within the geometrical acceptance of the detector in both data and MC simulation events. The requirements are slightly tighter than the selections applied at the HLT level, resulting in a common phase space between data and simulated events, where the trigger is not required.

The selections concern both the minimum amount of energy of the photons and several variables related to the shape of the photon clusters and their isolation. The requirements are different according to the photon candidate R9 and its location in the ECAL. They are listed here and summarized in table 4.3:

- The leading and subleading photon of the event are required to have p_T >30 and 20 GeV, respectively
- both the photons must pass the conversion-safe electron veto
- H/E (ratio between the energy reconstructed in the HCAL in correspondence of the photon supercluster in the ECAL) < 0.08
- in the case of R9 less than 0.85 in the EB or 0.9 in the EE, $\sigma_{i\eta i\eta}$ is required to be less than 0.015
- the PF Photon isolation, which is the sum of the transverse momentum of all the photons reconstructed by the PF algorithm, lying in a cone of R=0.3 centered on the photon direction, must be less than 4.0 GeV. A correction accounting for pileup contribution in the isolation cone is applied.
- tracker isolation, which is the sum of the transverse momentum of the tracks associated to the primary vertex and reconstructed within a cone with radius R=0.3 around the photon direction, must be less than 4.0 GeV.

	R_9	H/E	$\sigma_{\eta\eta}$	$\mathcal{I}_{\text{ph}}$ (GeV)	$\mathcal{I}_{\text{tk}}$ (GeV)
Barrel	[0.5, 0.85]	<0.08	<0.015	<4.0	<6.0
	>0.85	<0.08	—	—	—
Endcaps	[0.8, 0.90]	<0.08	<0.035	<4.0	<6.0
	>0.90	<0.08	—	—	—

Table 4.3: Photon requirements applied in the preselections.

- R9 greater than 0.5 and 0.8 for photons reconstructed in the barrel and endcap region, respectively.

The efficiency of the preselection was tested in data using Z→ee events, using the tag-and-probe method. The efficiencies measured in data (ϵ_{data}) and MC simulation (ϵ_{MC}) are compared, and scale factors are derived to recover a good agreement. Even if electrons are good proxies for photons, a non negligible difference between the R9 distribution for electrons and photons is present. This discrepancy is addressed by weighting the R9 distribution for electrons to match the photon one.

However, it is impossible to measure the efficiency of the conversion-safe electron veto with the same technique. A different procedure based on Z→ $\mu\mu\gamma$ events is employed. In this process, one of the muons produced in the decay of a Z boson, emits a photon as final-state radiation. Thanks to the cleanliness of the physics objects in the final state and the reference provided by the Z boson mass, a high-purity sample of Z→ $\mu\mu\gamma$ events can be selected in data, by requiring the presence of two opposite charge, isolated muons with a transverse momentum exceeding 10 GeV and an invariant mass greater than 35 GeV, in association with an isolated photons with transverse energy greater than 20 GeV. The photon must pass all the analysis preselections but the conversion-safe electron veto. In addition, to enhance the purity of the sample, the distance in the η-ϕ plane between the photon and the closest muon, computed as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is required to be less than 0.8. The same selections are applied to a sample of simulated Drell-Yan events to assess the efficiency in MC simulation, and derive the data-MC scale factors. The values are reported in table 4.4. The preselection efficiencies for four categories of photons are reported in table 4.4.

Preselection category	ϵ_{data} (%)	ϵ_{MC} (%)	$\epsilon_{\text{data}}/\epsilon_{\text{MC}}$
Barrel; $R_9 > 0.85$	94.2 ± 0.9	94.7 ± 0.9	0.995 ± 0.001
Barrel; $R_9 < 0.85$	82.5 ± 0.7	82.5 ± 0.7	1.000 ± 0.003
Endcap; $R_9 > 0.90$	90.1 ± 0.2	91.3 ± 0.1	0.987 ± 0.005
Endcap; $R_9 < 0.90$	49.7 ± 1.4	53.8 ± 1.5	0.923 ± 0.010

Table 4.4: Photon preselection efficiencies measured in four photon categories, obtained with tag-and-probe techniques in Z→ee and Z→ $\mu\mu$. The quoted uncertainties include the statistical and systematic components.

4.7.2 Untagged event categorization

The output of the diphoton MVA presented in section 4.6 can be used to categorize events that are not tagged as arising from a specific Higgs boson production mechanism. The categories are defined according to the score of the diphoton BDT, resulting in different expected signal over background ratio and diphoton invariant mass resolution, improving the overall sensitivity of the analysis.

Both the number of categories and their boundaries were optimized for an integrated luminosity correspondent to the analyzed dataset of 35.9 fb^{-1}. For the optimization procedure, the signal model is built from MC simulation and fitted with the sum of two gaussian distributions, one accounting for events with a correct identification of the vertex position along the beam axis, and the other one accounting for the wrong vertex identification case. The background model is built from an exponential fit of the diphoton invariant mass spectrum . The p-value is extracted from the fit of Asimov datasets created from the combined distributions of signal and background.

Four categories were used in the analysis, with inferior boundaries in the diphoton MVA output: -0.405, 0.204, 0.564, 0.864.

4.7.3 The ttH categories

The measurement of the ttH process cross-section enables to directly measure the coupling of the Higgs boson with the top quark. This measurement is particularly interesting in the study of the electroweak symmetry breaking mechanism, given the strong Yukawa coupling of the top quark, about one, considerably higher than the couplings of the other fermions in the SM.

The experimental signature in the detector of the ttH production mechanism is characterized by the presence of two jets originating by bottom quarks produced in the top quark pair decay, along with additional jets or leptons, depending on the top quark decay channel.

Two categories are defined in the context of the H→ $\gamma\gamma$ analysis according to the decay channel of the top quark: a leptonic and a hadronic tag.

The ttH Leptonic Tag

The ttH leptonic tag aims at collecting events with the top quark pair decaying either fully leptonically or semi-leptonically ($t\bar{t}$ →blνbqq' and $t\bar{t}$ →bl$\nu_l$bl'$\nu_{l'}$). The presence in the signal final state of a lepton with high p_T provide an extremely effective handle to reduce the background events contamination in this category.

The events are required to fulfill the following criteria:

- Ratio of the leading photon transverse momentum over diphoton invariant mass ($p_T/m_{\gamma\gamma}$)>0.2
- Ratio of the subleading photon transverse momentum over diphoton invariant mass ($p_T/m_{\gamma\gamma}$)>0.4
- at least one lepton which passes the identification requirements listed in section 2.3.2 and with $p_T > 20$ GeV
- in the electron channel, $|m_{e\gamma} - m_Z| > 5$ GeV, in order to reject events with a Z boson decaying into an electron-positron pair, with one of them misidentified as a photon

- at least 2 jets with $|\eta| < 2.4$ and p_T >25 GeV
- at least one jet with the characteristics of the previous point must be b-tagged according to the medium working point of the b-jet discriminator (ϵ_{sign} ∼62%)

Given the very low number of background events expected pass the selections of the ttH leptonic category, a loose selection on the diphoton identification corresponding to an efficiency on the signal of about 96% was chosen.

The ttH hadronic category

The ttH hadronic category targets event with a fully hadronic decay of the top quark pair. For the identification of signal events, a discrimination based on the output of a Boosted Decision Tree (BDT) is used. The input variables are:

- number of jets in the $|\eta|$ <2.4 region with p_T > 25 GeV
- the maximum output of the b-tagging discriminator among the jets satisfying the previous requirement
- the second maximum output of the b-tagging discriminator among the jets satisfying the previous requirement
- p_T of jet with the highest transverse momentum in the event

The BDT was trained using MC simulation events. Simulated ttH events, with the top quark pair decaying into hadrons was used as signal, while a sample of photons plus jets events was used as background, since this process is the main source of background for this category. For validation and subsequent optimization step, a control sample is built in data by inverting the selection based on the output of one the photon identification MVA. In order to improve the description of the kinematics in data, the control sample is reweighted according to η and p_T, to match the kinematics in data. In figure 4.7, the comparison of the BDT output for events with at least three jets, one of which b-tagged according to a loose working point, is showed. The agreement between simulation and data is very good, moreover, the distribution of signal and background events are well separated, highlighting the discrimination power of the BDT output.

The selections based on the output of the diphoton identification MVA and the dedicated MVA for ttH hadronic events are simultaneously optimized through a bidimensional likelihood scan to maximize the expected sensitivity to the ttH production. Three different approaches for the modeling of the background were considered and for each of them the optimization procedure was independently performed. The background model was derived from sidebands in $m_{\gamma\gamma}$ distribution (100<$m_{\gamma\gamma}$ <120 GeV and 130<$m_{\gamma\gamma}$ <180 GeV) in data, MC simulation samples, or the control sample described above. The results obtained with the three different estimation of the background are in agreement. The optimal pair of selections was found to be for ttH MVA output >0.75 and diphoton identification BDT score >0.4.

The final set of selection used in the analysis is:

- Ratio of the leading photon transverse momentum over the diphoton invariant mass ($p_T/m_{\gamma\gamma}$) >0.375)

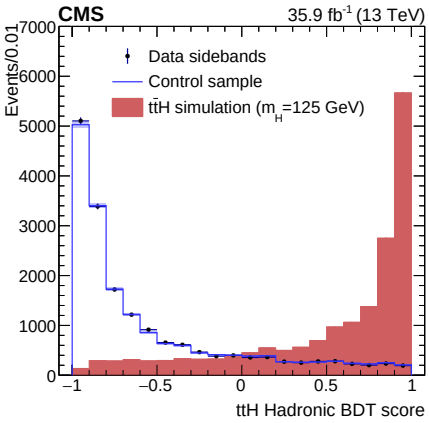

Figure 4.7: Comparison between the distribution of the output of the ttH hadronic MVA in data sidebands, control sample, and MC simulation signal events.

- Ratio of the subleading photon transverse momentum over the diphoton invariant mass ($p_T/m_{\gamma\gamma}$) >0.25)
- at least three jets with $|\eta|$ <2.5 and p_T >25 GeV
- at least one jet with satisfying the previous point requirements and b-tagged according to a medium working point of the b-jet discriminator (ϵ_{sign} ∼62%)
- no leptons, defined as in the ttH leptonic tag, in the event
- output of the ttH hadronic BDT >0.75
- output of the diphoton classification BDT > 0.4

4.7.4 The VBF categories

The VBF production process has a cross section one order of magnitude lower than the ggH mechanism. However, thanks to the experimental signature in the detector, it is possible to define categories with a high signal-over-background ratio. The final state is made up of two jets, mainly in the forward region of the detector and separated by a large gap in pseudorapidity, in addition to the photon pair produced in the Higgs boson decay. Moreover, additional QCD activity in the central part of the detector is suppressed. The tagging of VBF events relies on the reconstruction and identification of the jets in the forward region of the CMS.

The VBF events categorization is based on the output of two boosted decision trees: one is the diphoton categorization BDT presented in section 4.6, the second one is a dedicated MVA, called kinematic dijet MVA, that exploits the peculiar features of the genuine jets produced in VBF events, to discriminate the signal from background process with the same final state, either with a real diphoton object plus additional jets or with multiple jets in the final state, which mimic the diphoton system. Moreover, this second MVA is also very effective in reducing the contamination from ggH production.

The kinematic dijet MVA

The goal of the kinematic dijet MVA is the discrimination between reconstructed dijet systems formed by jets produced in VBF events from jets arising from other background

processes or pileup. Also ggH events are considered as background and the BDT is designed to reject them, resulting in a more pure category, enabling a more accurate measurement of the VBF process cross-section and of the coupling of the Higgs boson with vector bosons. The training is performed using MC simulation events. A sample of simulated VBF events, with mass of the Higgs boson equal to 125 GeV, is used as signal. MC simulation samples of diphoton, γ+jets, and multijet production events, along with a MC simulation sample of ggH events with mass of the Higgs boson equal to 125 GeV are used as background in the MVA training. A boosted decision tree using a gradient boosting (BDTG) algorithm implemented in the scikit-learn framework [69] is used. A set of preselections is applied before training, in order to perform it in a more signal-like phase space. The ratio between the transverse energy of the leading and subleading photons and the diphoton mass are required to exceed 0.25 and 0.20, respectively. Two jets with p_T above 30 and 20 GeV, respectively, with a minimum dijet invariant mass of 100 GeV are required to be present in the event. Furthermore, events with $m_{\gamma\gamma}$ between 100 and 180 GeV are used.

The input variables of the BDTG are:

- ratio of the transverse momentum of the leading photon and the diphoton invariant mass
- ratio of the transverse momentum of the subleading photon and the diphoton invariant mass
- transverse momentum of the leading jet
- transverse momentum of the subleading jet
- dijet invariant mass
- gap in pseudorapidty between the two jets
- difference in the azimuthal angle between the diphoton and the dijet system
- centrality variable, which is defined as:

$$C_{\gamma\gamma} = exp\left(-\frac{4}{(\eta_1-\eta_2)^2}\left(\eta_{\gamma\gamma}-\frac{\eta_1+\eta_2}{2}\right)^2\right) \tag{4.8}$$

  where η_1, η_2, and $\eta_{\gamma\gamma}$ are the pseudorapidities of the two jets, and of the diphoton
- difference in the azimuthal angle between the two jets
- the minimum distance among the distances between each photon and each jet

The combined MVA

The combined MVA employs a boosted decision tree with a gradient boost, which takes as input variables the output of the kinematic dijet MVA, the diphoton identification MVA, and the ratio between the transverse momentum of the diphoton system and the diphoton invariant mass. This MVA is meant to provide discrimination between the background and the VBF signal. For training, a simulated sample of VBF events, with mass of the Higgs boson equal to 125 GeV is used, while MC simulation samples of diphoton, γ+jets, and multijets events are given as background.

Three categories are defined based on the output if this BDTG. An optimization procedure was performed to define the number of categories, and the boundaries of each category in order to achieve the best sensitivity. The number of categories was chosen to be equal to three, with lower boundaries:

- VBF-0: 0.957
- VBF-1: 0.902
- VBF-2: 0.553

The whole procedure was validate using Z→ee events which also contain two jets, with electrons reconstructed as photons. These events were selected requiring the transverse momentum of the leading electron or positron to exceed 40 GeV, and the two jets to have p_T above 40 and 30 GeV, which transverse mass higher than 250 GeV. The comparison of the output distribution obtained with simulated events and data is showed in figure 4.8. The agreement is fair and well contained within the uncertainty band.

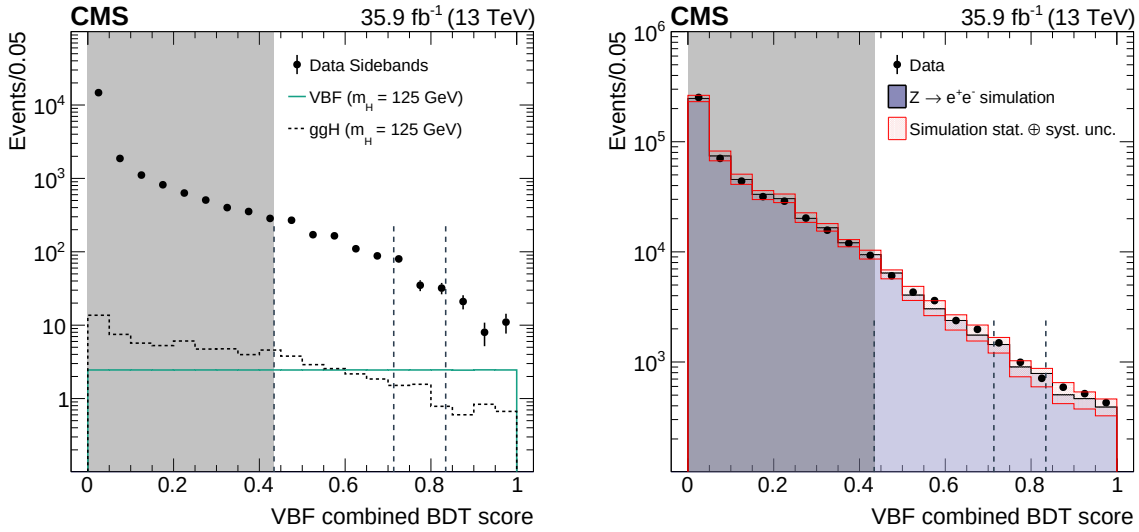

Figure 4.8: *Left*: comparison of the distribution of the combined MVA score after a transformation to get it flat for VBF events. The distributions for ggH and VBF signal events, and background events in the mass sidebands region are showed. *Right*: comparison of the distributions of the combined MVA with Z→ee events, with electrons reconstructed as photons, in data (points) and simulated events (grey histogram), including statistical and systematic uncertainties (pink region). The hashed lines correspond to the boundaries of the VBF categories defined in the analysis.

4.8 Statistical analysis

The events are categorized in a total of fourteen exclusive tags, which are defined on the basis of the four dominant production mechanism of the Higgs boson at the LHC, and the expected signal over background ratio. Three categories are build to collect events where the Higgs boson is produced via VBF, called VBFTag0, VBFTag 1, VBFTag 2.

The ttH mechanism is targeted by two tags, based on the decay of the top quark pair: TTHLeptonicTag and TTHHadronicTag. Three categories are defined to collect event with the Higgs boson produced via VH, with the vector boson decaying into leptons: ZHLeptonicTag, WHLeptonicTag, VHLeptonicLooseTag. ZH events with the Z boson decaying into a pair of neutrinos and WH events with the W boson decaying into a lepton and a neutrino, with the lepton not reconstructed, are collected into a dedicated tag, called VHMetTag. VH events with the V boson (either W or Z boson) decaying into a pair of quark are included in the VHHadronicTag. The remaining event are categorized on the base of their diphoton MVA score, in four untagged categories: UntaggedTag0, UntaggedTag1, UntaggedTag2, UntaggedTag3.

In order to statistically interpret the data, an estimation of the number of Higgs bosons produced within the detector geometrical acceptance, along with the efficiency of reconstruction is needed. Furthermore, the expected shape of the diphoton invariant mass for signal events, and a parametrization of the background distribution are necessary for the statistical analysis of data. The signal model is derived from MC simulation samples with the procedure described in 4.8.1, while the background parametrization is performed through a data-driven procedure(4.8.2). The statistical analysis is performed independently in each category.

4.8.1 Signal Model

The shape of the signal diphoton invariant mass distribution is modeled in each category, using MC simulation samples. As already mentioned, when the vertex selected in the analysis lies within 1 cm from the true one, the resolution on the diphoton invariant mass is dominated by the resolution on the measurement of photon energy in the ECAL. Otherwise, it is further smeared because of a not accurate estimation of the opening angle between the photons. In order to account for these effects, the signal modeling is performed separately for the two scenarios. MC simulation samples with different values of the Higgs boson mass are employed to construct the signal model, since the value of the Higgs boson mass is not exactly known. The method employed to build the signal model is the simultaneous signal fit (SSF). A simultaneous fit of all the simulated samples with different mass of the Higgs boson is performed, with the individual parameters of the functional form, being functions of the Higgs boson mass. In this procedure, the floating parameters of the fit are the coefficient of these polynomials.

The final signal model in each category is given by the combination of the $m_{\gamma\gamma}$ shapes obtained in the right, and wrong vertex location scenarios. Their relative fraction is equal to the correct vertex location efficiency assessed in the MC simulation. Also the diphoton vertex identification efficiency can change with the mass of the Higgs boson. For this reason, in the signal model, it is interpolated between the different Higgs boson mass scenarios. The contributions of the different Higgs boson production mechanisms to the expected $m_{\gamma\gamma}$ shape in each category are weighted according to the ratio of the process cross-sections predicted by the SM. Figure 4.9 shows the signal model corresponding to $m_H = 125$ GeV for all the analysis categories combined together, weighted by the S/(S+B) ratio, where S is the number of signal events, and B the number of background events in a window around the signal peak.

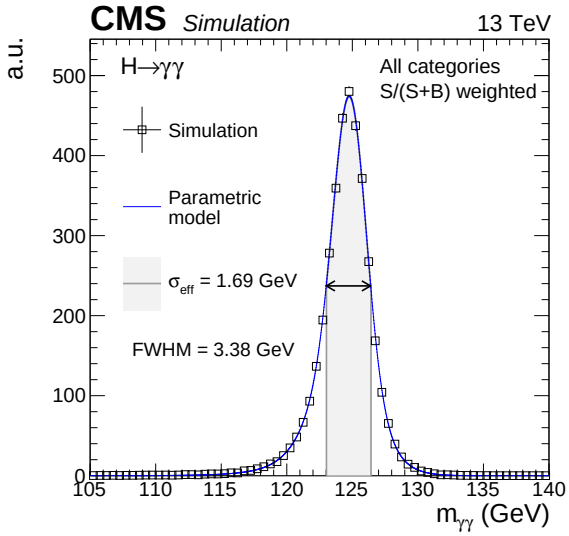

Figure 4.9: Signal shape for all the analysis categories combined together as described in the text, for simulated H→ γγ signal events with $m_H = 125$ GeV. The open squares represent weighted simulated events and the blue lines are the corresponding models. Also the σ_{eff} (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) is showed as a grey band, while the interval corresponding to the full width at half maximum (FWHM) is indicated by a double arrow.

4.8.2 Background Model

In the analysis, the background parametrization is fully data-driven. The diphoton invariant spectrum is fitted in the range within 100 and 180 GeV. In principle, the knowledge of the true functional form of the diphoton invariant mass spectrum in the background-only case would result in the most accurate outcome of the statistical analysis. The usage of a wrong parametrization of the background distribution can lead to negative effects such as biases in the results of the analysis. In the case of the H→ γγ analysis, this approach can not be pursued. Therefore, an alternative method, called discrete profiling, or *envelope* method [70], is used. It aims at identifying the functional form which better described the background distribution, by quantifying the uncertainty associated to the choice of a certain functional form for the background parametrization, and treating it as a discrete parameter in the likelihood fit. In order this method to be valid, it is necessary to test a complete set of function must be tested. In the background modeling procedure, four families of functions are considered:

- sum of exponential functions:

$$f_N(x) = \sum_{i=1}^{N} p_{2i} e^{p_{2i+1} x} \tag{4.9}$$

- sum of polynomials in the Bernstein basis:

$$f_N(x) = \sum_{i=0}^{N} p_i b_{(i,N)} \tag{4.10}$$

where $b_{(i,N)} := \binom{N}{k} x^i (1-x)^{N-i}$

- series of Laurent:

$$f_N(x) = \sum_{i=0}^{N} p_i x^{-4+\sum_{j=1}^{i}(-1)^j(j-1)} \tag{4.11}$$

- sum of power-law functions:

$$f_N(x) = \sum_{i=0}^{N} p_{2i} x^{-2p_{2i+1}} \tag{4.12}$$

where for all of these functional forms, i, p_i are a set of floating parameters in the fits. In figure 4.10, an example of the fit of the background in the "Untagged0" category is showed.

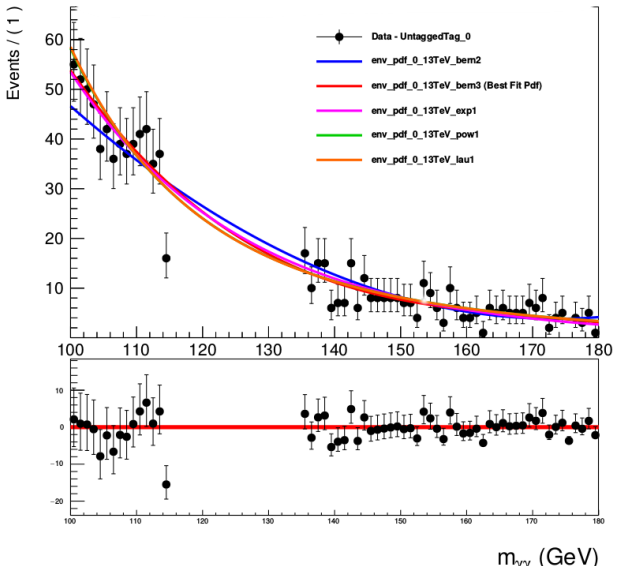

Figure 4.10: Fit of the background distribution in the Untagged0 category with the families of functions described in the text-

The order of the particular function among the families defined above that best describes the data is chosen by means of a procedure aiming at quantifying the improvement in the fit quality by passing from the order N to the order N+1. When fitting the diphoton invariant mass background distribution with the function of the order previously defined, twice the negative logarithm of the likelihood fit (-2Logλ) is minimized. A penalty factor is added to account for the different number of floating parameters in the fit of the different functions. When performing the measurement of a parameter of interest, the profiling method builds the envelope of the lowest values of -2Logλ, including the penalty factors, profiled as a function of the parameter of interest. The 1σ uncertainty band is obtained from the width of the 68% range, or equivalently from the point for which -2Logλ=1.

4.8.3 Systematic Uncertainties

The systematic uncertainties affect only the signal model, since the background is derived using a data-driven technique. The uncertainty due to the choice of the functional form chosen for the background parametrization is accounted for by the profiling method. The

systematic uncertainties are assumed to impact the signal model in three ways: modifying the $m_{\gamma\gamma}$ shape, changing the overall yield per process, and inducing event migration among categories. Source of systematic uncertainty of the first type are directly accounted for in the signal modeling procedure by considering them as nuisance parameters. In the case where, the shape of the $m_{\gamma\gamma}$ is largely unchanged, the systematic uncertainties impact the signal yields following a log-normal distribution. Where the systematic uncertainty has an effect on the inputs of the MVAs employed in the analysis, the variation takes the form of correlated log-normal uncertainty on the category yield, resulting in event migration between categories.

The source of systematic uncertainty considered in the analysis are:

- parton density function (PDF) uncertainties: the impact of the choice of the parton density function used in MC simulation is estimated after re-weighting the simulated events according to the PDF4LHC combined PDF set and NNPDF30 using the MC2hessian procedure. The impact on the overall signal yield and migration among categories are estimated separately. Less than 1% of the events were estimated to migrate. The overall signal yield is taken from [67].

- α_s: the uncertainty on the value of the strong coupling constant (α_s) is propagated in the analysis following the PDF4LHC prescription. The effect on the overall signal yield normalization and migration of events among categories are handled separately. The overall change in the relative event yield variation due to uncertainty on α_s was found to be 2.6%.

- QCD scale uncertainty: it is related to the choice of the scale in the renormalization and factorization scale in the QCD theory. The impact on the signal yield is estimated separately for the four main Higgs boson production processes at the LHC (ggH, VBF, VH, ttH). The migration of events among categories is assessed by means of MC simulation samples with modified values of the QCD scale.

- Underlying event and parton shower uncertainty: MC simulation samples with different choices and tuning of the model used to simulate underlying event and parton shower are used to assess the event migration due to this uncertainty. The resulting migration evaluated on VBF and untagged categories were found to be of 7%.

- Theoretical uncertainty on the H→ $\gamma\gamma$ branching fraction: about 2%

- Gluon-gluon fusion contamination in VBF and ttH categories: the theoretical prediction for the ggH mechanism, when the Higgs boson is produced in association with several jets, is not accurate. The uncertainty on the yield of ggH events in the VBF categories was estimated by means of the Stewart-Tackmann procedure, following the prescriptions of the LHC Higgs Cross Section Working Group [**?**]. The variation of the overall signal yield was estimated to be 29%, while the event migration among the VBF categories were at most 7%. In the estimation of the impact on the ttH categories, several sources were taken into account, such as the limited size of the simulated sample, resulting in 10% uncertainty, uncertainty form the choice of the parton showering, and gluon splitting models. The contribution from the choice of the parton shower model was estimated from the difference in the jet multiplicity distribution in simulation and data, and estimated to be about 15% in the ttH leptonic category and 35% for the ttH hadronic category. The uncertainty on the gluon splitting model was assessed by scaling the fraction of gluon-gluon fusion events with real jets arising from b quarks, by the observed difference between

data and MC simulation in the ratio $\sigma(t\bar{t}bb)/\sigma(t\bar{t}jj)$ [71]. The uncertainty on the ggH yield in the ttH categories was assessed to be about 50%.

- Integrated luminosity: it is measured from data. The related systematic uncertainty on the signal yield is estimated to be 2.5%

- Trigger efficiency: it is measured in Z→ee events using the tag-and-probe technique. The uncertainty is calculated in bins of pseudorapidity, transverse energy, and R9 of the photon. It results at most in a 0.1% variation in the signal yield.

- Photon preselection: the uncertainty due to the photon preselection is computed as the ratio of the preselction efficiency measured in data and MC simulation. It results in an event yield variation ranging from 0.2% to 0.5%, depending on the category.

- Vertex identification efficiency: it is calculated from the ratio of the vertex identification efficiency in data and MC simulation, using Z→ $\mu\mu$ events. Also the uncertainty due to the underlying event modeling and the uncertainty on the transverse momentum of the Higgs boson contribute to uncertainty on the vertex identification efficiency. It is accounted for in the signal model construction, by considering it as a nuisance parameter, allowing the fraction of events in the right and wrong vertex scenarios to vary. The uncertainty was assessed to be about 2%.

- Energy scale and resolution: energy scale and resolution is studied with electron from Z→ee events. The main source of systematic uncertainty arises from the different interaction of electrons and photons in the material upstream of the ECAL and their shower propagation inside the ECAL. The uncertainties are estimated by varying the R9 distribution, using electrons for the training of the MVAs, and the selection in the procedure to derive the corrections employed in the analysis. The uncertainties are assessed in bins of pseudorapidity and R9 to the Higgs boson signal. These sources of systematic uncertainty are accounted for, including nuisance parameters in the signal model construction. The systematic uncertainty amount to less than 0.5%, depending on the photon category. The impact on the inclusive signal strength was assessed to be of about 2.5%.

- Photon identification BDT output: a systematic uncertainty is assigned to the photon identification BDT output to cover the discrepancy between the distribution obtained in data and MC simulation. The uncertainty on the total signal yield is about 3%.

- Per-photon energy resolution estimation: it is parametrized as a rescaling of the resolution estimate of plus or minus 5% around its nominal value, covering the discrepancy between data and MC simulation samples, in the estimator output distribution

- Jet energy scale and smearing corrections: this uncertainty is implemented as migration between VBF, ttH, and untagged categories. Jet energy scale corrections result in 8% to 18% event migration within VBF tags, 11% of events from VBF to untagged categories, and 5% in the ttH categories. The event migration due to jet energy resolution is below 3%.

- b-tagging efficiency: it is estimated by varying the ratio between the measured b-tagging efficiency in data and MC simulation, within their uncertainty. For the ttH

hadronic category, where the output of the b-tagging algorithm is given as input to a BDT, the systematic uncertainty is assessed modifying the shape of the output of the b-tagging discriminant. The uncertainty on the signal yields are about 2% and 5% in the ttH lepton and ttH hadronic categories, respectively.

- Lepton identification: the systematic uncertainty due to lepton identification is estimated by varying the ratio of the efficiency measured in data and MC simulation by its uncertainty. The induced differences in the selection efficiency are less than 1%.

- Background modeling: the uncertainty due to the choice of the function for the parametrization of the background diphoton invariant mass spectrum is estimated by means of the discrete profiling method described in section 4.8.2.

The following sources of systematic uncertainty have a low impact in the analysis but are important for the precision measurement of the Higgs boson mass that, however, is beyond the scope of the analysis presented here:

- Uncertainty on the material upstream of the ECAL: it arises from the not perfect modeling of the real amount of material in front of the ECAL in the CMS MC simulation. It is estimated from dedicated MC samples with different material budget scenarios, in bins of pseudorapidity and R9. The impact on the energy scale uncertainty was found to be lower than 0.24%.

- Non uniformity of the light collection: it is due to the non uniformity of the light collection inside the crystal. The impact on the photon energy scale uncertainty was assessed to be below 0.1%.

- Shower shape modeling: it is related to account for a non perfect simulation of the development of the electromagnetic shower shape inside the ECAL. It is estimated in bins of pseudorapidity and R9 of the photon. The contribution to the systematic uncertainty on the photon energy scale was estimated to be below 0.05%.

4.9 The VH categories

The associated production of a Higgs boson with a vector boson, also called *higgs-strahlung*, can be tagged by requiring the presence of the decay products of the vector boson.

The VH mechanism cross section is an order of magnitude lower than the dominant ggH production at the LHC. However, the presence in the final state of the decay product of the vector boson is a very effective handle for background rejection, especially in the leptonic decay channels of the V. The VH categories contribute to the measurement of the Higgs boson coupling to vector bosons.

In the H→ $\gamma\gamma$ analysis, five categories are defined, according to the different decay channels of the vector boson produced in association with the Higgs boson:

- ZH leptonic category, targeting ZH events in the Z→ll decay channel;
- WH leptonic category, targeting WH events in the W→lν decay channel. Also ZH events with the Z boson decaying into a lepton pair fall in this category, if one of the two leptons is out of the geometrical acceptance of the detector or is not reconstructed due to detector inefficiency;
- VH leptonic loose category, aiming at collecting residual events with a lepton in the final state, which are not collected in the previous tight leptonic categories;
- missing E_T category, targeting WH events with the W boson leptonically decaying fall in this category, if the lepton is out of the geometrical acceptance of the detector or is not reconstructed because of detector inefficiency. Also ZH events in the Z→ $\nu\nu$ decay channel can fall in this category;
- VH hadronic category, targeting either ZH or WH events, with the vector boson decaying in a pair of quark;

The VH Missing E_T tag

The VH missing E_T category collects events with a large amount of missing energy in the final state, which is exploited to tag the event and reject background, resulting in a category with a high signal-over-background ratio. The high E_T can be due either to the lepton from the W boson decay that has not been reconstructed either because it is out of the geometrical acceptance of the detector or because of detector inefficiency.

Events are included in this category if they satisfy the following requirements:

- ratio between the transverse momentum of the leading photon and the diphoton invariant mass higher than 0.375
- ratio between the transverse momentum of the subleading photon and the diphoton invariant mass higher than 0.25
- missing E_T higher than 85 GeV
- difference in the azimuthal angle between the diphoton system and the missing momentum >2.4
- output of the diphoton categorization BDT higher than 0.6

The ZH leptonic tag

The ZH leptonic tag targets events where a Z boson is produced in association with the Higgs boson, and decays into a pair of leptons. Thus, the experimental signature in the detector of this process is made up of two isolated and high-energy leptons, with same flavor and opposite charge, resonating in the mass of the Z boson. Thanks to the presence of the two leptons in the final state, the background due to multijet events is almost totally suppressed. The high signal-over-background ratio of this category compensates the low branching ratio of the dilepton decay of the Z boson ($\sim$7%). The identification of ZH leptonic events is based on a cut-based selection, whose requirements are:

- ratio between the transverse energy of the leading photon and the diphoton invariant mass higher than 0.375
- ratio between the transverse energy of the subleading photon and the diphoton invariant mass higher than 0.25
- two leptons of the same flavor and opposite charge with transverse momentum exceeding 20 GeV. Both electrons and muons are required to fulfill the identification criteria described in section 2.3
- dilepton invariant mass within 70 and 110 GeV
- ΔR between the photons and the electrons larger than 1. In events with muons, the threshold is lowered to 0.5. This selection aims at rejecting events where one of the photon in the diphoton object is actually emitted by the lepton
- score of the diphoton MVA higher than -0.405. Given the high purity of this category, a loose selection is sufficient to achieve a high signal over background ratio.

The VH Leptonic Loose tag

This category aims at collecting events with a lepton in the final state which are not collected in the other VH leptonic categories. The set of selections applied to events to be collected in this category are the same as the WH leptonic category but the one involving missing transverse energy, which must be lower than 45 GeV. This difference results in a huger contamination from background with respect the WH leptonic tag.

4.10 The VH Hadronic Category

In this section, the category dedicated to the exclusive selection of events where a Higgs boson, decaying to two photons, is produced in association with a vector boson that decays in a pair of quarks: $H \to \gamma\gamma + V \to jj$ (with V = W/Z).

4.10.1 The VH hadronic channel

The decay of a vector boson into a pair of quarks is the most probable one, with a branching ratio of about 67% and 70%, for the W and Z boson, respectively. The experimental signature in the detector of VH events in the hadronic decay channel is characterized by a pair of photons arising from the decay of the Higgs boson and two jets with high transverse momentum, resonating in the invariant mass of the V boson. From the experimental point of view, the identification of this process at the experimental environment of the LHC is affected by the copious production of jets. The dominant sources of background events are events with a pair of prompt photons, in association with two jets in the final state, which can originate from the underlying event activity or pileup, and photons plus jets events.

The identification of signal events exploits the harder spectrum of physics objects in the final state in signal than the ones arising from background processes since the energy scale of the event is set by the masses of the decaying bosons (H and V). This difference can be observed in figure 4.11, where the transverse momentum of the diphoton system (left) and the subleading jet (right) in signal events are compared to the distribution expected in background events, and gluon-gluon fusion production of a Higgs boson. Also,

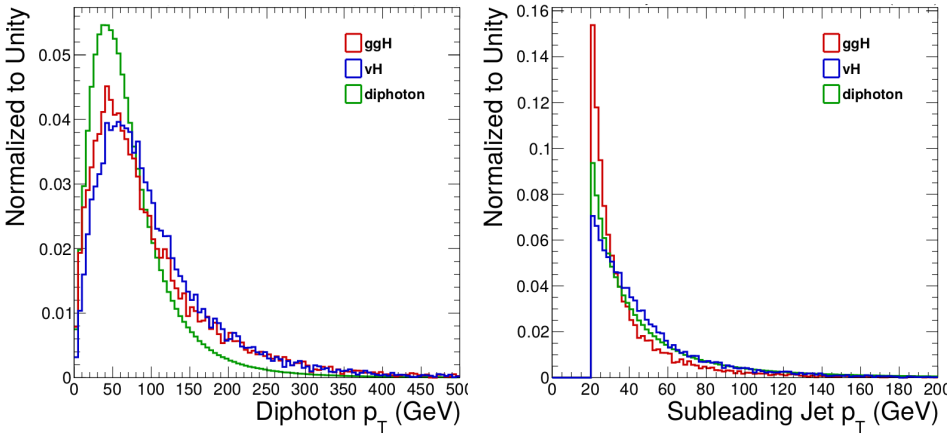

Figure 4.11: Diphoton transverse momentum (left) and subleading jet transverse momentum (right) distributions. The expected shape for VH events (blue line) is compared to the shape of gluon fusion Higgs boson production (red line) and the major background: diphoton continuous production (green line). All distributions are normalized to unity.

since the two jets arise from the decay of a massive boson, their invariant mass resonates in the vector boson mass, while it is a continuous distribution for the background events since the two jets are uncorrelated. The dijet invariant mass distributions in VH (V→qq), gluon fusion Higgs-boson production and background events are showed in the left plot of figure 4.12 Since both W and Z hadronic decays contribute to the peak in the dijet invariant mass spectrum, its width is larger than the one expected for a single hadronic resonance at a similar mass. Moreover, given the resolution on the measurement of the

energy of jets at CMS, the two peaks are merged, and it is impossible to distinguish the vector boson produced in association with the Higgs boson.

Furthermore, the complete reconstruction of both the Higgs boson and the vector boson decays enables to perform a refined angular analysis of the final state reconstructed in the detector. As signal events stem from a two-body decay, the direction of the diphoton and the dijet system are correlated. This is not true in background events since they do not arise from the decay of a single particle. Based on this consideration, the variable θ^*, defined as the angle between the direction-of-flight of the V^* in the laboratory frame, and its decay products in the V^* rest frame, was found to provide a good discrimination power. The expected distribution for the signal is a flat distribution, while in background events, the boost to the V^* rest frame is artificial and results in a preference toward $\theta^* = 0$ (the boost direction). This difference can be noticed in the $\theta*$ distributions reported in the right plot of figure 4.12, for the signal, ggH, and background events.

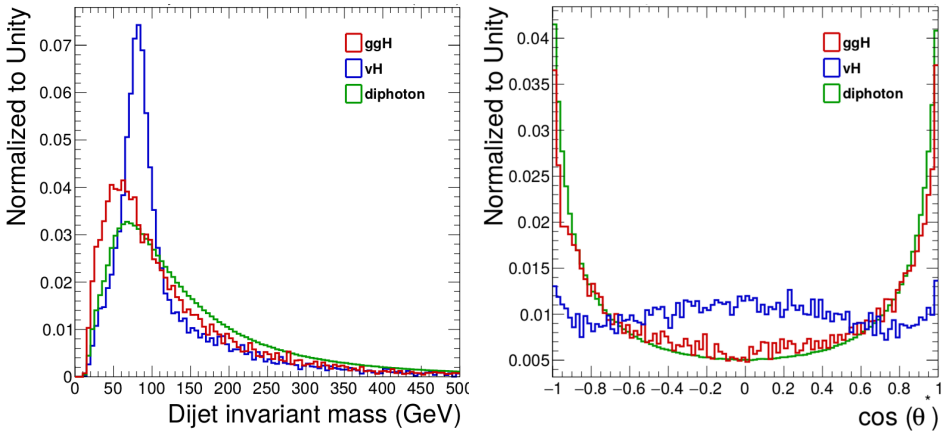

Figure 4.12: Dijet invariant mass (left) and $\cos\theta^*$ (right) distributions. The expected shape of VH events (blue line) is compared to the shape of gluon fusion Higgs boson production (red line) and the major background: diphoton continuous production (green line). All distributions are normalized to unity.

4.10.2 Selection and categorization

The categorization follows a cut-based approach. The selections specific to this category are based on the presence of two jets from the decay of a W or Z boson, and on kinematic requirements, such as $\cos\theta^*$ and $p_T^{\gamma\gamma}/m_{\gamma\gamma}$, introduced earlier in the text. The same variables as in the the analysis of the LHC Run 1 dataset, which led to the discovery of the Higgs boson in 2012 [72], are exploited for signal from background discrimination:

- ratio between the photons transverse energy and the diphoton invariant mass: it is required to be above 0.5 and 0.25 for the leading and the subleading photon, respectively. The threshold on the leading photon is tighter than the one used in preselections for the majority of the other categories of the analysis to exploit the higher boost of the photons produced via VH and achieve a more effective background rejection.
- dijet invariant mass (m_{jj}): $60<m_{jj}<120$ GeV
- at least two jets with transverse momentum above 40 GeV

- $\cos\theta^*$
- $\mathrm{p_T^{\gamma\gamma}/m_{\gamma\gamma}}$.

In the analysis of data collected during 2016, the discrimination provided by these variables was reviewed in order to spot possible effects due to the different experimental conditions of the collisions during the LHC Run 2, mainly the different center-of-mass energy of the collisions, increased from 7 (in 2011) and 8 (in 2012) TeV to 13 TeV. The selection based on the $\cos\theta^*$ and $\mathrm{p_T^{\gamma\gamma}/m_{\gamma\gamma}}$ variables, whose impact is expected to be significant for this category, were studied in more detail. In particular, some differences are expected for what concerns the $\mathrm{p_T^{\gamma\gamma}/m_{\gamma\gamma}}$, since at Run 2, particles produced by pileup interactions are expected to be more energetic on average. The precision on the measurement of the Higgs-boson coupling to vector bosons (σ_{RV}) was considered as the figure of merit of the study. A cut and count experiment was performed to compute σ_{RV} applying all the other selection of the category and varying the value of the variable to analyze in steps of 0.1. The number of signal events was estimated from simulated H$\rightarrow\gamma\gamma$ events, including all the four main production processes (ggH, VBF, VH, ttH). The number of background events was assessed through a control sample (CS) defined in data, orthogonal to the signal sample (SS). It was obtained requiring one of the photon to fail the identification preselection (BDT score > -0.9). Since the efficiency of the analysis preselecions can be slightly different in CS and SS and depends on the η and $\mathrm{p_T}$ of the photon, the events in the CS were re-weighted in bins of $\mathrm{p_T}$ and η of the photons, in order the CS to accurately reproduce the kinematics of the SS. In the plots of figure 4.13, the comparison of the distribution of some variables representative of the event kinematics, and the analyzed variables, in data sidebands and the CS are showed. The agreement is fair, ensuring that the control sample is well describing the kinematics of the signal sample.

The diphoton invariant mass distribution obtained in the CS was fitted to an exponential function, as depicted in figure 4.15, and the number of background events was obtained from the integral of the fitted function in the mass window between 123.5 and 126.5 GeV. The plots of figure 4.16 show the σ_{RV} obtained with the procedure described above as a function of the selection value, for $\mathrm{p_T/m_{\gamma\gamma}}$ (left plot) and $\cos\theta^*$ (right plot). The vertical, red lines identify the selection used in the Run 1 analysis. For what concerns the selection on $\cos\theta^*$, the value used in the analysis of the Run 1 dataset is very close to the optimal value indicated by this study and was used also in this analysis. The threshold on $\mathrm{p_T/m_{\gamma\gamma}}$ was lowered to 1 (slightly lower than the threshold of 13/12 used in the Run 1 analysis). This variable is correlated to the score of the diphoton classifier BDT, which was subject of a more significant change between Run 1 and Run 2. It was found that the combination of a looser selection on $\mathrm{p_T/m_{\gamma\gamma}}$ and a tighter selection on the diphoton BDT score provided better performance.

4.10.3 Diphoton MVA selection optimization

The selection based on the output of the diphoton BDT(4.6) was optimized to maximize the significance of the category. The figure of merit of the optimization was the precision on the measurement of the coupling of the Higgs boson to vector bosons (σ_{RV}). The

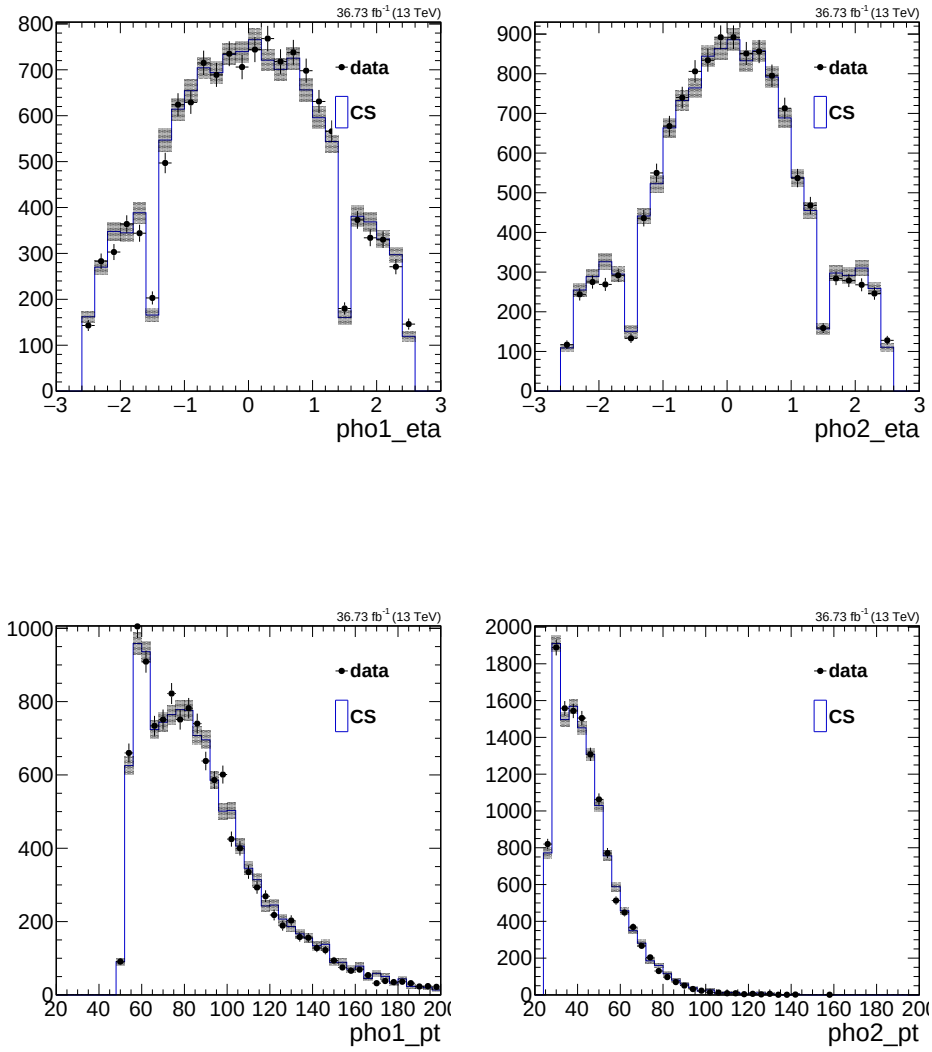

Figure 4.13: Distribution of η (top plots) and p_T (bottom plots) of the leading (left plots) and subleading (right plots) photon for the signal and the control sample (CS). The analysis preselection are applied and the events in the CS are reweighted as described in text.

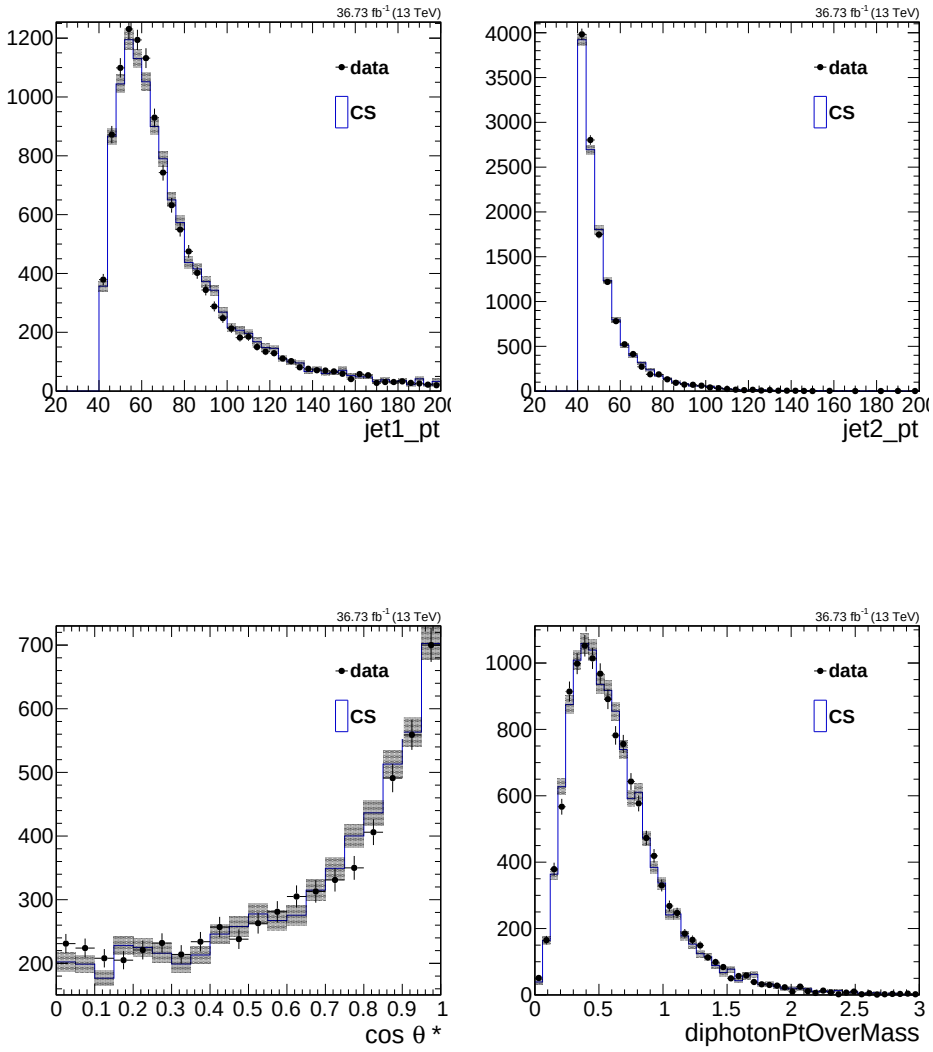

Figure 4.14: Distribution of the leading (top left) and subleading (top right) jet p_T, $\cos\theta^*$ (bottom left), and $p_T^{\gamma\gamma}/m_{\gamma\gamma}$ (bottom right) for the signal and the control sample (CS). The analysis preselection are applied and the events in the CS are reweighted as described in text.

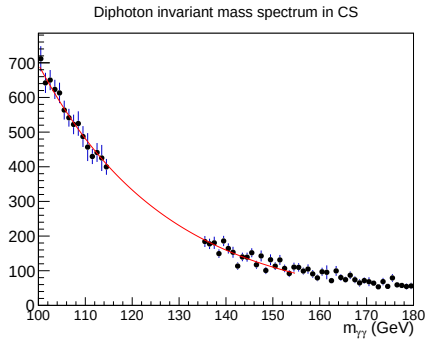

Figure 4.15: Diphoton invariant mass distribution shape comparison between CS and SS. The fit to the exponential function used to estimate the background events is superimposed.

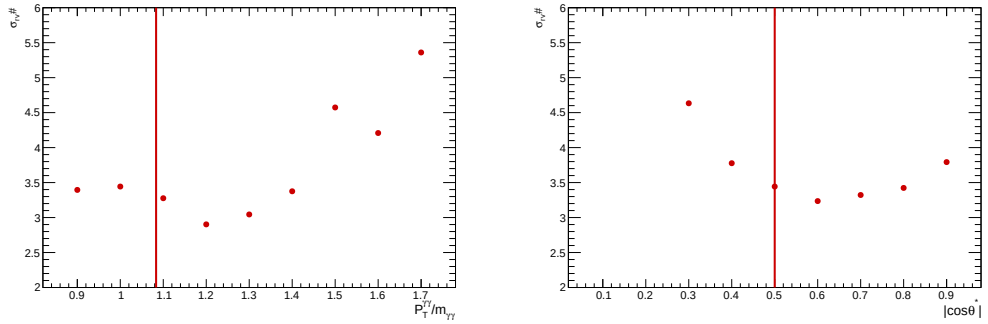

Figure 4.16: Expected precision on the Higgs-boson coupling to vector bosons σ_{RV} as a function of $\mathrm{p_T}(\gamma\gamma)/\mathrm{m}_{\gamma\gamma}$ (left) and $|\cos\theta^*|$ (right). The red lines indicate the cut used in Run 1 analysis.

number of signal events was estimated from VH MC simulation samples, with $\mathrm{m}_H = 125$ GeV, considering only those ones with the V boson decaying into a pair of quarks. The number of background events was assessed through a control sample defined in data, orthogonal to the signal sample (SS) that was obtained applying all the standard selections of the category but the one on $|\cos\theta^*|$, which was inverted ($|\cos\theta^*| > 0.5$). The events in CS were weighted in bins of p_T and η of the photons, to improve the agreement between the kinematics in the CS and the SS. The diphoton invariant mass spectrum obtained in the CS was rescaled to have the same number of events in the sideband of the SS spectrum and fitted to an exponential function. The number of background events was estimated from the integral of the fit function between 123.5 and 126.5 GeV. In figure 4.18, the distributions of some kinematic variables of the photons in the CS and the SS are compared. The agreement is fair, hence the CS can be safely used to estimate the number of background events expected in the SS.

For each value of the diphoton MVA selection, a cut and count experiment was performed to compute σ_{RV}. The result of the optimization procedure is shown in Figure 4.19.

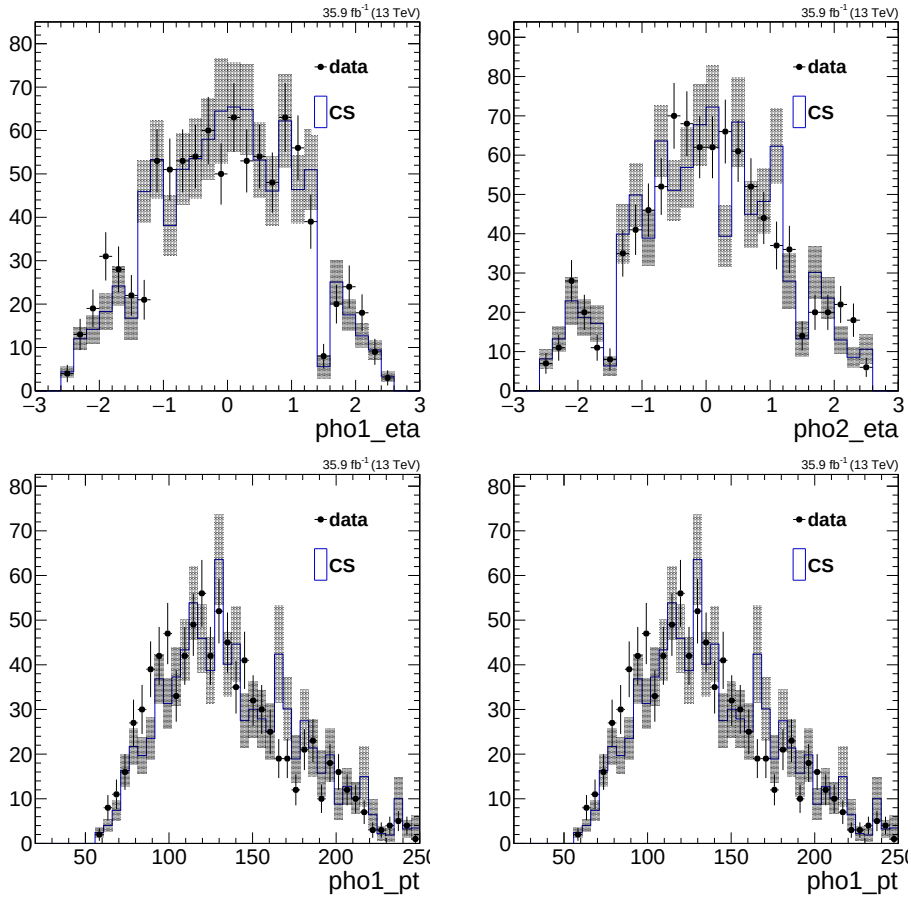

Figure 4.17: Distribution of η (top plots) and p_T (bottom plots) of the leading (left plots) and subleading (right plots) photon for the signal and the control sample (CS). All the category selection but the one on diphoton BDT are applied and the events in the CS are reweighted as described in text.

Selecting events with an output of the diphoton BDT >0.7 was found to maximize the significance in the category. Such a different value is motivated by a substantial difference in the diphoton classification BDT output between Run 1 and Run 2. This can be noticed by comparing the shape of the distributions in the left and right plot of figure 4.20, where the normalized distributions of the diphoton BDT output in MC simulation samples of ggH, VBF, VH, and ttH events, in Run 1 (left plot) and Run 2 (right plot) are showed.

4.10.4 Conclusions

The category described in this section was included in the analysis of the data recorded during 2016 by the CMS experiment, corresponding to an integrated luminosity of 35.9 fb^{-1}. The final set of selections used for the event categorization are:

- ratio of the transverse momentum of the leading photon and the diphoton invariant mass > 0.5;
- ratio of the transverse momentum of the subleading photon and the diphoton invariant mass > 0.375;

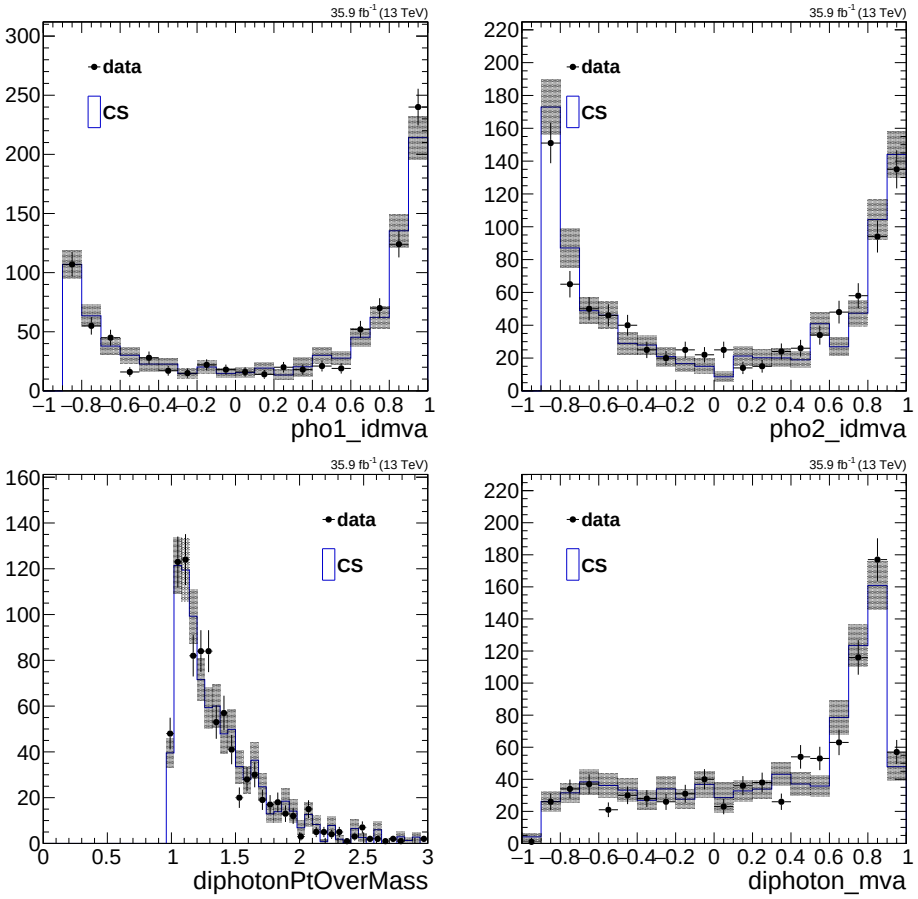

Figure 4.18: Distribution of the photon identification BDT score of the leading (left plots) and subleading (right plots) photon. All the category selection but the one on diphoton BDT are applied and the events in the CS are reweighted as described in text.

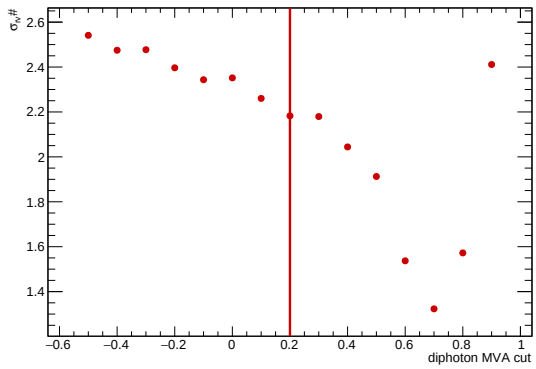

Figure 4.19: Expected precision on the Higgs-boson coupling σ_{RV} as a function of the selection of cut diphoton MVA cut. The red line indicates the cut used in Run 1 analysis.

- ratio of the diphoton momentum over the diphoton invariant mass > 1;
- dijet invariant mass between 60 and 120 GeV;

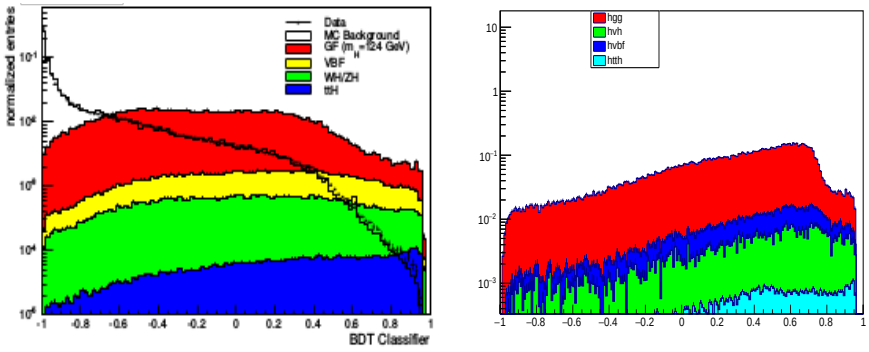

Figure 4.20: *Left*: normalized distributions of the diphoton categorization BDT used in the analysis of the 2012 dataset, for MC samples simulation samples of ggH (red histogram), VBF (yellow histogram), VH (green histogram), and ttH (yellow histogram) events. The distribution related to simulated background events (transparent histogram) and data (black dots) for events with 100 $<m_{\gamma\gamma}$ <120 GeV and 130$<m_{\gamma\gamma}$ <180 GeV. *Right*: normalized distributions of the diphoton categorization BDT score used in the Run 2 analysis described in this thesis, for MC simulation samples of ggH (red histogram), VBF (blue histogram), VH (green histogram), and ttH (azure histogram) events.

Categories	SM 125GeV Higgs boson expected signal													Bkg (GeV^{-1})
	Total	ggH	vbf	ttH	bbH	tHq	tHW	WHLep	ZHLep	WHHad	ZHHad	σ_{eff}	σ_{HM}	
VH Had	7.9	47.6%	4.5%	4.4%	0.4%	1.7%	0.3%	0.2%	0.5%	**25.2%**	**15.1%**	1.38	1.30	7.2

Table 4.5: Expected number of signal events in the VH hadronic category with the percentage breakdown per production mode. The σ_{eff} and σ_{HM} are also reported.

- $|\cos\theta^*|$ <0.5;
- diphoton MVA output >0.7;

The expected number of signal events after the selections is estimated from MC simulation. In table 4.5, the breaking down of the number of events collected in the VH hadronic category for each of the dominant Higgs boson production process and the number of background events per GeV, are reported. Also the smallest interval containing the 68% of the events of the distribution (σ_{eff}) and the width of the distribution at half of its maximum (σ_{HM}) divided by 2.35 are showed.

The expected significance (σ) of the category was estimated to be about 0.58, computed as:

$$\sigma = \frac{S}{\sqrt{B}} \tag{4.13}$$

where S is the number of VH events with the V boson, either a Z or a W boson, decaying into a pair of jets, and B is the number of background events in a window of $\pm 3\,\sigma_{eff}$ around 125 GeV. In figure 4.21, the diphoton invariant mass distribution for events collected in the VHHadronic category is showed, together with signal and best background fit, with the $\pm 1\,\sigma$ and $\pm\,2\,\sigma$ bands.

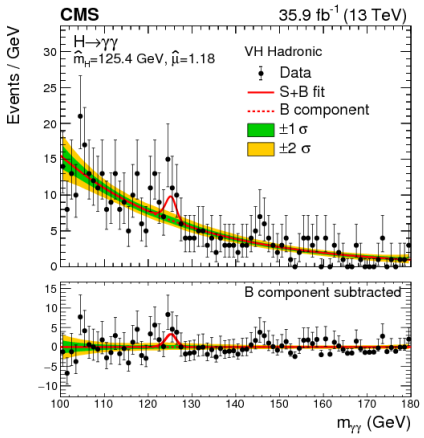

Figure 4.21: Diphoton invariant mass distribution in the VH hadronic category. Also, the best background fit, with its 1σ and 2σ uncertainty bands, and the signal model are showed.

4.11 The WH Leptonic Category

The branching ratio for a W boson decaying into a lepton and a neutrino is about 30%. The experimental signature in the detector of WH events, in the Higgs boson diphoton decay channel, is characterized by the presence of a lepton with high transverse momentum, large missing transverse energy, and a pair of photons with high transverse energy. Thanks to the presence of a lepton in the final state, the background due the copious production of jets at the LHC collisions is heavily suppressed. In this section, the categorization used to select WH events in the analysis of the data collected during 2016 is presented. Moreover, the study of possible improvements in view of the analysis of the entire dataset collected during the LHC Run 2 thanks to a multivariate analysis approach is showed.

4.11.1 Specific WH Leptonic background processes

The background contribution due to QCD processes resulting in multijets events is massively reduced by the requirement of a high-energy lepton in the final state and large transverse missing energy. However, because of the very large production of jets at the LHC, a little, residual contamination is expected. Moreover, other processes can mimic the signal final state:

- events with a prompt pair of photons, or a single photon, in addition to jets. Their impact is heavily suppressed by the request of a lepton in the final state;
- Drell-Yan events with the Z boson decaying into a lepton pair, in addition to jets. In particular in the Z→ee channel, since if the track of the electron (or positron) is not reconstructed because of tracking inefficiency, it can be misidentified as a photon, that together with a fake photon due to jets produced by the underlying event activity or pileup can give raise to a fake diphoton object, plus a lepton, mimicking the signal signature;
- Production of top quark pair and single top quark, in association with either jets or photons ($t\bar{t}$+Jets, $t\bar{t}$+γ, $t\bar{t}$+$\gamma\gamma$). The top quark decays before undergoing hadronization, with a branching fraction almost 1 into a b quark and a W boson, which, in

turn, can decay into a lepton and a neutrino, as in the leptonic sector of the signal final state. In events without real photons arising from the hard scattering or emitted via final state radiation, a fake photons can be due to the misidentification of either jets with a large electromagnetic fraction or electrons (positrons), whose track is not reconstructed. This source of background can be rejected through the identification of the jets arising from the bottom quarks hadronization, thanks to the b-tagging algorithms used at the CMS [48]. Furthermore, stringent veto for photons compatible with reconstructed tracks can be used for further suppression of this background contribution, when a photon is faked by an electron;

- W boson in association with a photon (W+γ): this process is very difficult to reject, since the final state reconstructed in the detector is identical to the signal. However, different kinematics correlations between the diphoton system and the lepton can be exploited to separate the signal from these source of background;
- Other SM electroweak production of diboson events: WW, WZ; Zγ, ZZ;

4.11.2 The category in 2016 analysis

The WH leptonic category relies on the presence in the final state of an isolated lepton with high transverse momentum, and a large amount of missing transverse energy. The combined request for a lepton and missing energy results in a huge background rejection, in particular from QCD processes. The main sources of background are SM events with diboson production (WW, ZZ), in particular Wγ and Zγ events. The categorization strategy follows a cut-based approach.

To be included in this category, events must pass the following criteria:

- ratio between the transverse energy of the leading photon and the diphoton invariant mass higher than 0.375
- ratio between the transverse energy of the subleading photon and the diphoton invariant mass higher than 0.25
- at least one lepton (electron or muon) with transverse momentum higher than 20 GeV, which fulfill the identification criteria presented in section 2.3
- ΔR between the photons and the electrons larger than 1. In events with muons, the threshold is lowered to 0.5. This selection aims at rejecting events where the photon is emitted by the lepton
- $|m_{e\gamma} - m_Z| > 10$ GeV: this selection is applied only in the electron channel. In particular, it aims to reject Zγ, with Z$\rightarrow$ee events, where one of the two electrons is misidentified as a photon because of failure in reconstruction of the electron track.
- Missing transverse energy exceeding 45 GeV
- output of the diphoton MVA higher than 0. This working point results in 90% efficiency on signal and 50% background rejection
- less than two jets reconstructed within $|\eta|$ 2.4, with transverse momentum exceeding 20 GeV. This selection is meant to reject background events due to top quark pairs, decaying leptonically, which otherwise would result in a final state very similar to the signal one. For the same reason, it is also effective in reducing the contamination from ttH events.

4.11.3 MVA approach to the WH leptonic category

The program for the publication of the results of the H→ $\gamma\gamma$ analysis does not envisage the analysis of the 2017 dataset only. Rather, the analysis of the entire dataset recorded by the CMS during the LHC Run 2 is targeted. In this context, a new approach to the categorization aiming to select WH leptonic events targeting the analysis of the entire run 2 analysis has been studied. A multivariate analysis technique is employed to enhance the background rejection thanks to a more fruitful exploitation of the peculiar features of the signal topology. The selection of the events is based on the output of a Boosted Decision Tree (BDT), using the Ada boosting algorithm [73], implemented in the TMVA framework [74].

Training of the BDT

In order to perform the training in the proper phase space, the standard preselection of the analysis, listed in section 4.7.1 were applied. Moreover, additional very loose requirements are applied in order to select a sub-sample more representative of the signal: at least one lepton with p_T >15 GeV and $\not{E}_T$ > 15 GeV.

Simulated WH events with mass of Higgs boson equal to 124 GeV were used as signal. Only events with the W boson decaying into a lepton and a neutrino, identified using the MC truth information, were used in the training. The MC simulation samples related to the other Higgs boson production mechanisms were not included in the training as signals, in order to avoid to dilute the discrimination power provided by the features of the lepton and missing energy in the WH (W→lν) mechanism, and their correlation with the diphoton system. The MC simulation samples of the main background processes are used as background in the training. Every MC sample was re-weighted after the preselection step in the training procedure to reproduced the relative fraction of background events passing the preselections. The list of MC samples used in the training, together with the number of entries after applying the preselections used for the training procedure, are summarized in table 4.6.

Sample	N. of Events
Signal	
VH, m_H=124 GeV	6000
Background	
Z→ll+Jets	659
Diphoton	14035
Multijets	5
γ+Jets	1858
WW	4
ZZ	482
WZ	349
Wγ	167
$t\bar{t}$+Jets	57
$t\bar{t}\gamma$+Jets	1119
$t\bar{t}\gamma\gamma$	3427
tγ+Jets	20
Total N. of entries	26005

Table 4.6: Table summarizing the list of MC simulation samples used to train the BDT

The input variables provided to the BDT can be divided in three types, according to the physics object in the final state they refer to, among the diphoton system, the lepton, and global quantities of the event:

Photon variables:

- Ratio of the transverse energy and the diphoton invariant mass of both the leading photon and the subleading photon in the event
- Ratio between the transverse momentum and the invariant mass of the diphoton system. Given the topology of the VH production mechanism, it is expected to be higher than in background events.
- Output of the photon identification BDT of both the leading and the subleading photon
- Pixel seed veto of both photons: this veto probes the compatibility of the photon supercluster reconstructed in the ECAL with a track seed reconstructed in the pixel detector of the CMS, extrapolated to the ECAL surface. This is expected to be particularly effective in rejecting background events due to top quark pair production with an electron arising from a top quark decay faking a photon.
- pseudorapidity of both photons. It is useful to account for possible differences in the performance of photon reconstruction and identification in different pseudorapidity regions of the detector.

W boson-related variables:

- Transverse momentum of the lepton with the highest transverse momentum in the event
- Distance in the η-ϕ plane, computed as $\Delta R(\eta, \phi) = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, between the lepton with the highest p_T in the event and each photon of the selected diphoton object
- The angle between the direction of the lepton and the direction of the diphoton system (θ). In the signal, the Higgs boson and the real W boson in the final state stem from a two-body decay of a virtual W boson, implying that a correlation between the flight direction of the Higgs boson and the real W boson exists. This correlation is exploited in the VH hadronic category, to discriminate signal-like from background-like events. In the WH leptonic final state, the presence of a neutrino, which escape the detector without being reconstructed, prevents from completely reconstructing the final state and fully exploit this correlation. However, the direction of the lepton and the direction of the Higgs boson, reconstructed from the photons in the final state, can present a residual correlation, which can aid the discrimination between signal and background
- Missing transverse energy: amount of missing transverse energy in the event. It is interpreted as the experimental signature of the neutrino arising from the W boson decay
- Transverse mass resulting from the lepton with the highest p_T in the event and $\not{E}_T$. In the signal, it presents an end-point at the value of the W boson mass.

Global and jet-related variables

- Number of reconstructed electrons and muons in the event.
- Number of jets with $p_T > 20$ GeV with $|\eta| < 2.5$
- Transverse momentum of the jet of the most energetic jet in the event. In the signal, jets in the final state are mainly due to the underlying event or pileup interactions. On the other hand, in particular in background events involving the production of top quark pairs, the energy spectrum is expected to be tighter.
- Output of the b-tagging algorithm of the two jets with the highest p_T in the event. These variables are meant to enhance the identification of background events due to top quarks pairs production.

The distributions of each of the variables described above are reported in figure 4.22 and 4.23, for the signal (blue histogram) and the background events (red histogram). Each histogram is normalized to unity, in order to highlight the differences in the shape of signal and background distributions.

4.11.4 BDT output and performance

The distributions normalized to unity of the BDT score obtained in simulated signal events with m_H=125 GeV, and in the sidebands of the diphoton invariant mass distribution ($100<m_{\gamma\gamma}<120$ GeV and $130<m_{\gamma\gamma}<180$ GeV) in data are showed in figure 4.24, for electrons and muons, separately.

The BDT score distribution in events with the W boson decaying into an electron and in a muon are slightly different: it is narrower and with a higher mean value for muons than for electrons, both in signal and background. This difference can be explained by the cleaner reconstruction of muons at CMS, and the higher impact of background in the electron channel. As explained in section 2.3.3, muons are reconstructed exploiting the combined information of the CMS tracker and muon stations, placed outside the return yoke of the CMS magnet. Besides very high energy jets that can traverse all the material placed in front of the muon subdetectors and induce spurious signals in them, muons are the only particle expected to reach the muon stations. Therefore, they are reconstructed and identified with very high efficiency and purity, and their momentum is measured with an excellent resolution. On the other hand, electron reconstruction is more affected by pileup and noise in the detector. Furthermore, since the signal final state includes a photon pair from the Higgs boson decay, there is a higher probability that a background process involving electrons can mimic the experimental signature of the signal.

Given the different BDT output, two distinct working points are set for electrons and muons. To check the potential performance of a selection based on the BDT output, the significance (σ) of the category in the hypothesis of no WH production was computed as in equation 4.14 as a function of the threshold on the BDT score. In addition, the same preselections as the one used in the BDT training were used.

$$\sigma = \sqrt{2 * (s+b) \ln\left(1 + \frac{s}{b}\right) - s} \tag{4.14}$$

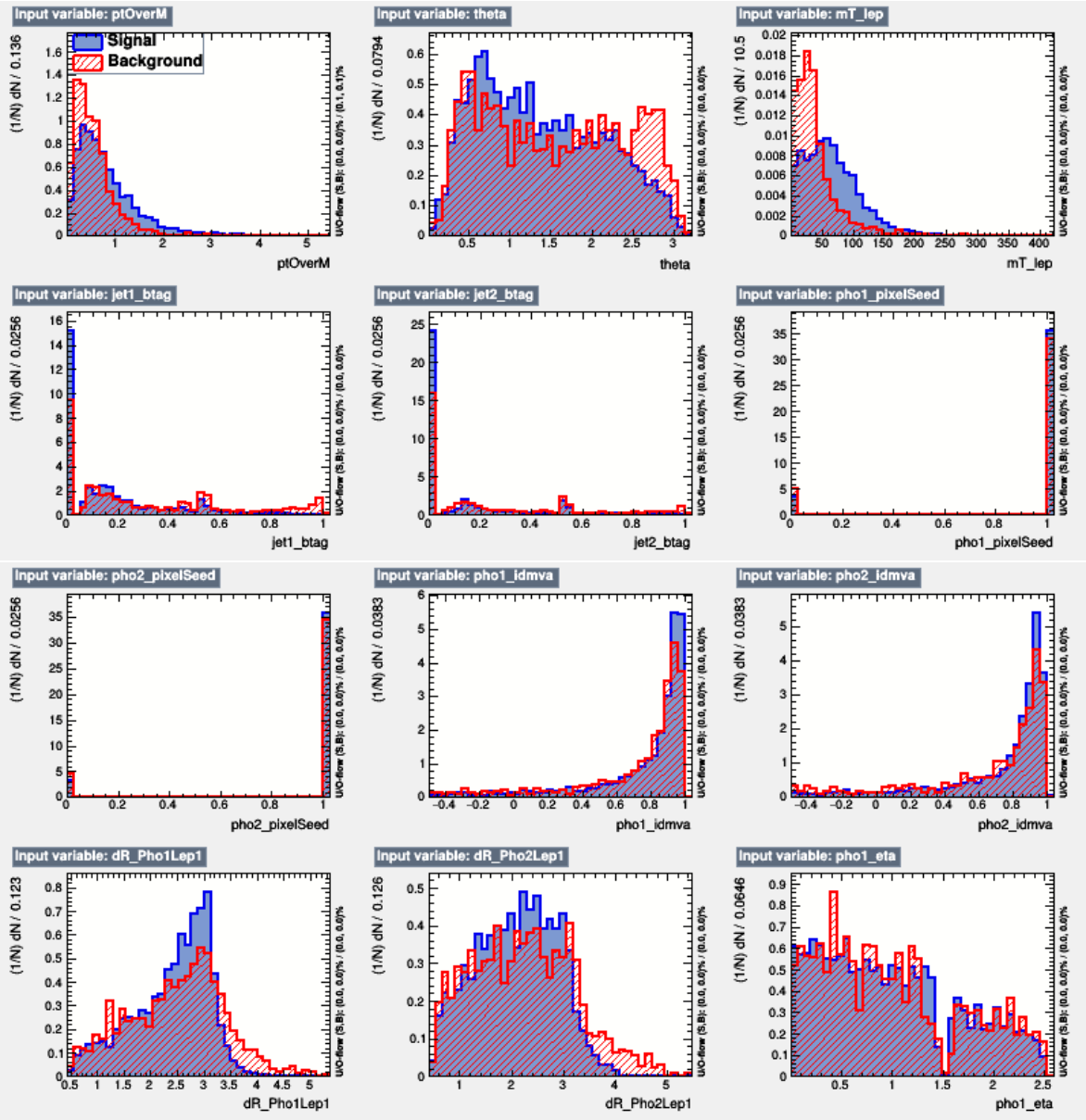

Figure 4.22: Distribution of the variables used in the BDT, in signal (blue) and background (red) processes.

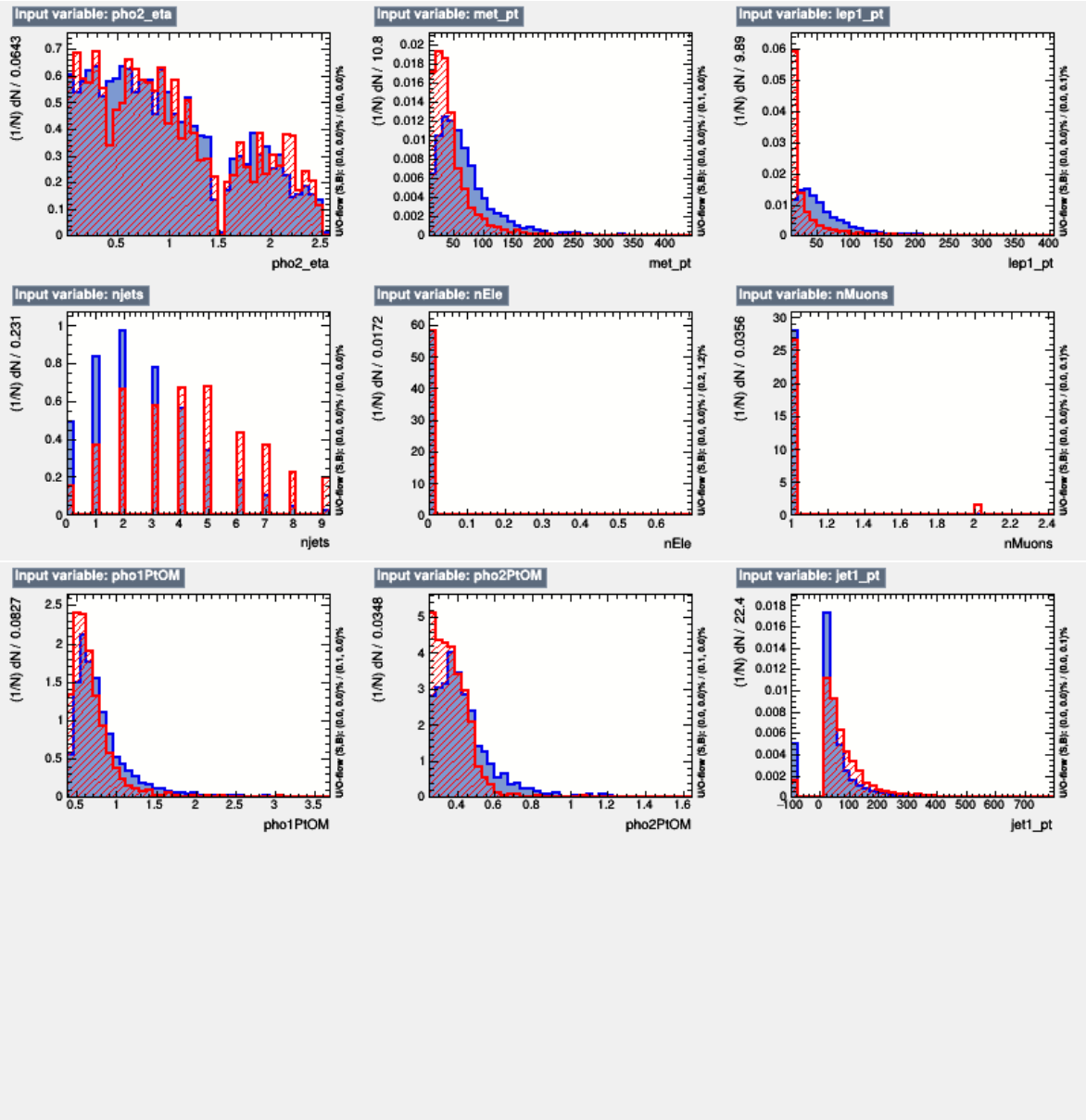

Figure 4.23: Distribution of the variables used in the BDT, in signal (blue) and background (red) processes.

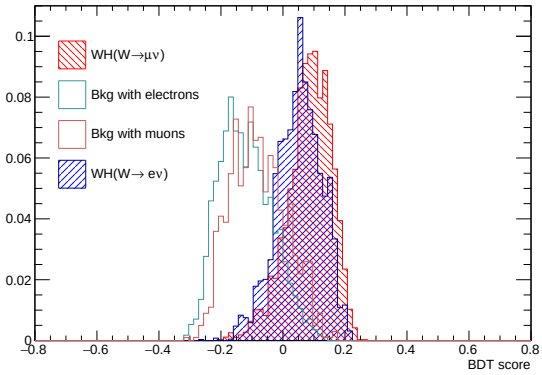

Figure 4.24: Distribution of the BDT output obtained in data sidebands (red) and simulated WH, in the W→lν decay channel (blue).

In this procedure, the number of signal events (s) was estimated from a VH MC simulated sample with m_H=125 GeV, considering only W→lν events. The number of background events was estimated from data. The sidebands of the diphoton invariant mass were fitted to an exponential function, and the number of background events was set to the integral of the fitted function between 123.5 and 126.5 GeV. The significance was evaluated increasing the threshold of the BDT score in steps of 0.1. The resulting σ as a function of the minimum BDT score is plotted in figure 4.25, for electrons (blue dots) and muons (red dots), along with the correspondent selection efficiency on the signal with respect to the number of events passing the preselections used in the BDT training.

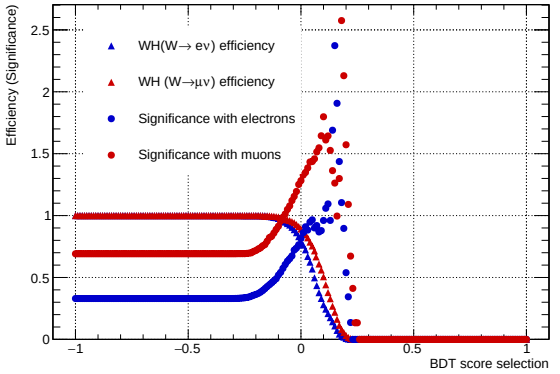

Figure 4.25: Significance of the WH Leptonic category as a function of the requirement on the minimum BDT score of the event.

The same scan was performed also on MC simulation samples with the number of background events equal to the number of events in the diphoton invariant mass spectrum between 123.5 and 126.5 GeV. The same MC simulation samples used in the BDT training and listed in table 4.6 were used. Nevertheless, because of the lack of events left in the samples after applying the analysis preselection, it could not be used for a full scan of the BDT output range. In the plot of figure 4.26, the σ computed estimating the number of background events from simulated events (azure dots) and from data sidebands (green

dots) as a function of the selection on the BDT score are compared. For BDT > 0 almost no events are present in the MC simulation samples. However, it can be noticed that where the number of events in the MC simulation samples are sufficiently high, the significance obtained with the two methods are in fair agreement.

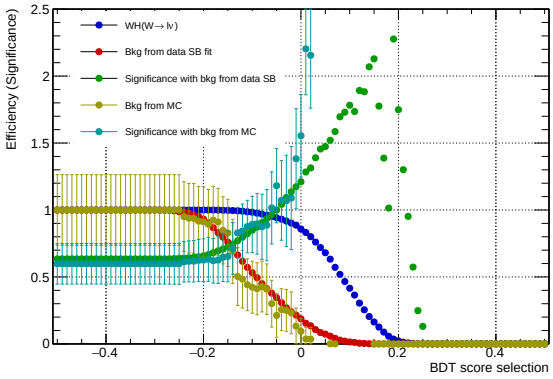

Figure 4.26: Significance of the WH leptonic category as a function of the requirement on the minimum BDT score of the event, estimating the number of background events from simulated events.

In order to prevent to introduce a bias due to an optimization performed on the same distribution ($m_{\gamma\gamma}$ spectrum) used to extract the final analysis results, the BDT output selection was set to have a number of signal events as in the 2016 analysis (2.97 signal events). The related threshold are 0.11 for muons and 0.007 for electrons. To assess the possible improvement due to this new strategy, the σ of this category with the strategy described in this section and the one used for the analysis of the 2016 dataset were compared. In table 4.7, the number of signal and background events, estimated through a fit of diphoton invariant mass distribution to an exponential function, and the significance computed as 4.14 are reported. This test indicated that an improvement on

	Analysis 2016	MVA analysis
Signal events	2.97	3.0
Background events	5.11	2.7
σ	1.16	1.58

Table 4.7: Number of signal and background events with BDT output higher thann x for muons and y for electrons. Also the significance of the category for the same selection is reported.

σ of about 35% can be achieved. Obviously, an improved background rejection would positively impact also the measurement of the WH cross-section, which is part of the Simplified Template Cross-Section (STXS) [75] framework that will be employed for the interpretation of the analysis results of the Run 2 dataset.

4.11.5 Validation of the BDT

The BDT was validated using WZ events, with the W boson decaying into a lepton and a neutrino (W→lν), and the Z boson decaying into an electron-positron pair (Z→ee),

Sample
WZ (W→lν, Z→ee), m$_H$=124 GeV
Z→ll+Jets
Diphoton
QCD
γ+Jets
W+Jets
WW
ZZ
WZ
Wγ
$t\bar{t}$+Jets
$t\bar{t}\gamma$+Jets
$t\bar{t}\gamma\gamma$
tγ+Jets

Table 4.8: List of the background MC simulation samples used in the WZ validation of the BDT

that are reconstructed as photons (not using any information from the tracker system). Therefore, the final state is identical to the final state of the WH process with the W boson decaying leptonically. The same set of preselections used in the H→ $\gamma\gamma$ analysis was applied, but the electron veto, which was inverted in order to select the electron-positron pair from the Z boson decay. The background contribution was estimated from the MC simulation samples listed in table 4.8. In addition, at least one lepton with p$_T$ >20 GeV and $\not{E}_T > 20$ GeV are required. Finally, only events with electron-positron invariant mass within 15 GeV of the nominal mass of the Z boson were used for the validation.

The comparison between the distributions of the BDT input variables in data and simulated events is reported in figures 4.27 and 4.28. The MC histograms were rescaled to match the number of event observed in data.The agreement is fair within the statistical uncertainty, whose main contribution is due to Z→ll+Jets events.

4.12 Results

The expected number of signal events after the selections is estimated from MC simulation. Table 4.9 shows the breaking down of the number of events collected in the VH categories of the analysis, per production process. The VH events are splitted according to the decay channel of the V boson, in ZH and WH events with the W or Z boson leptonically decaying (labelled as WH lep and ZH lep, respectively), and events with the W and Z boson decaying into quarks (labelled as WH had and ZH had, respectively). The percentage of events related to the production mechanism targeted by each category, is highlighted in bold. Also the smallest interval containing the 68% of the entries of the distribution (σ_{eff}) and the width of the distribution at half of its maximum, divided

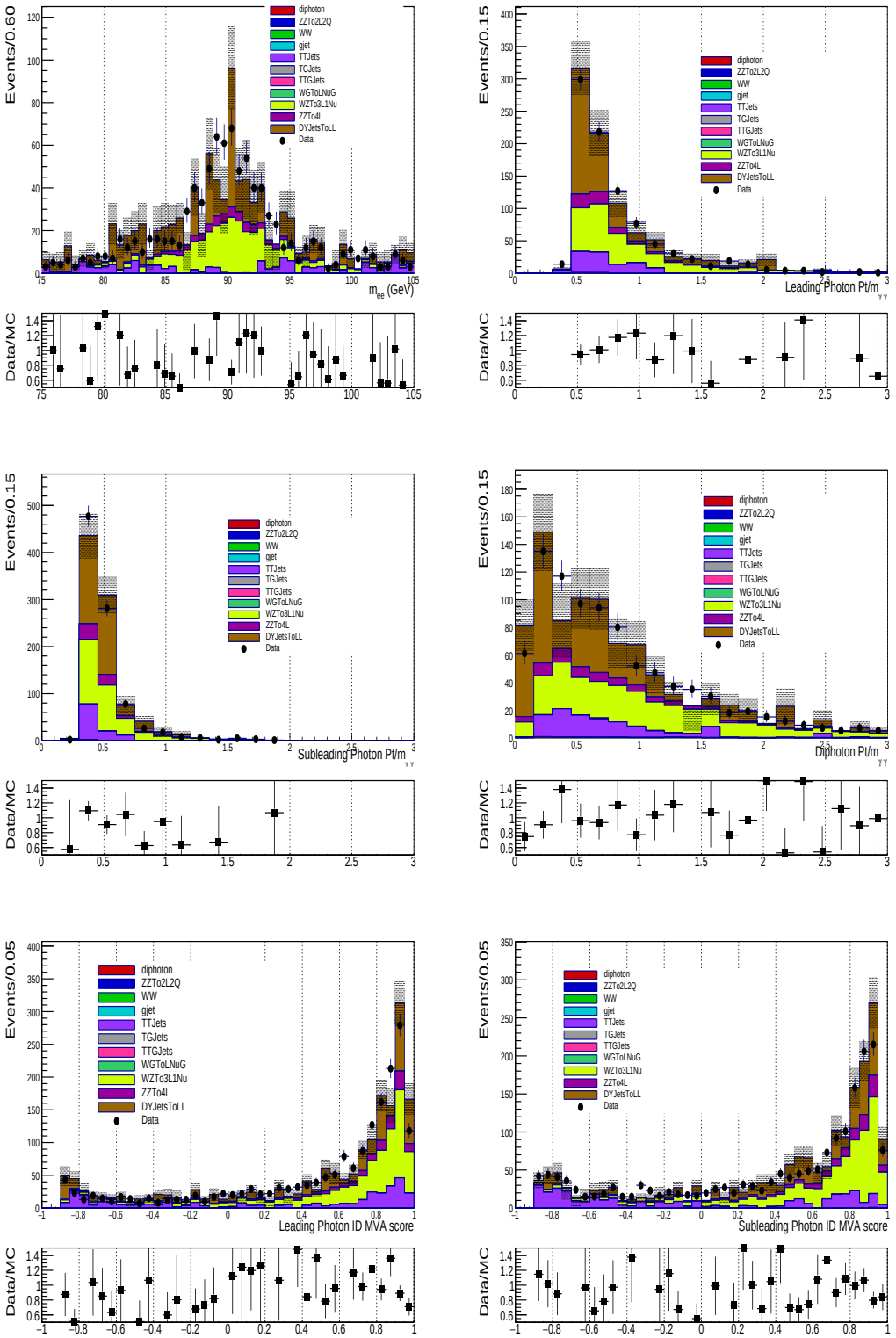

Figure 4.27: Comparison of the BDT input distributions in data and simulated events, selecting WZ events with the W boson decaying leptonically and the Z boson decaying into an electron-positron pair, reconstructed as photons.

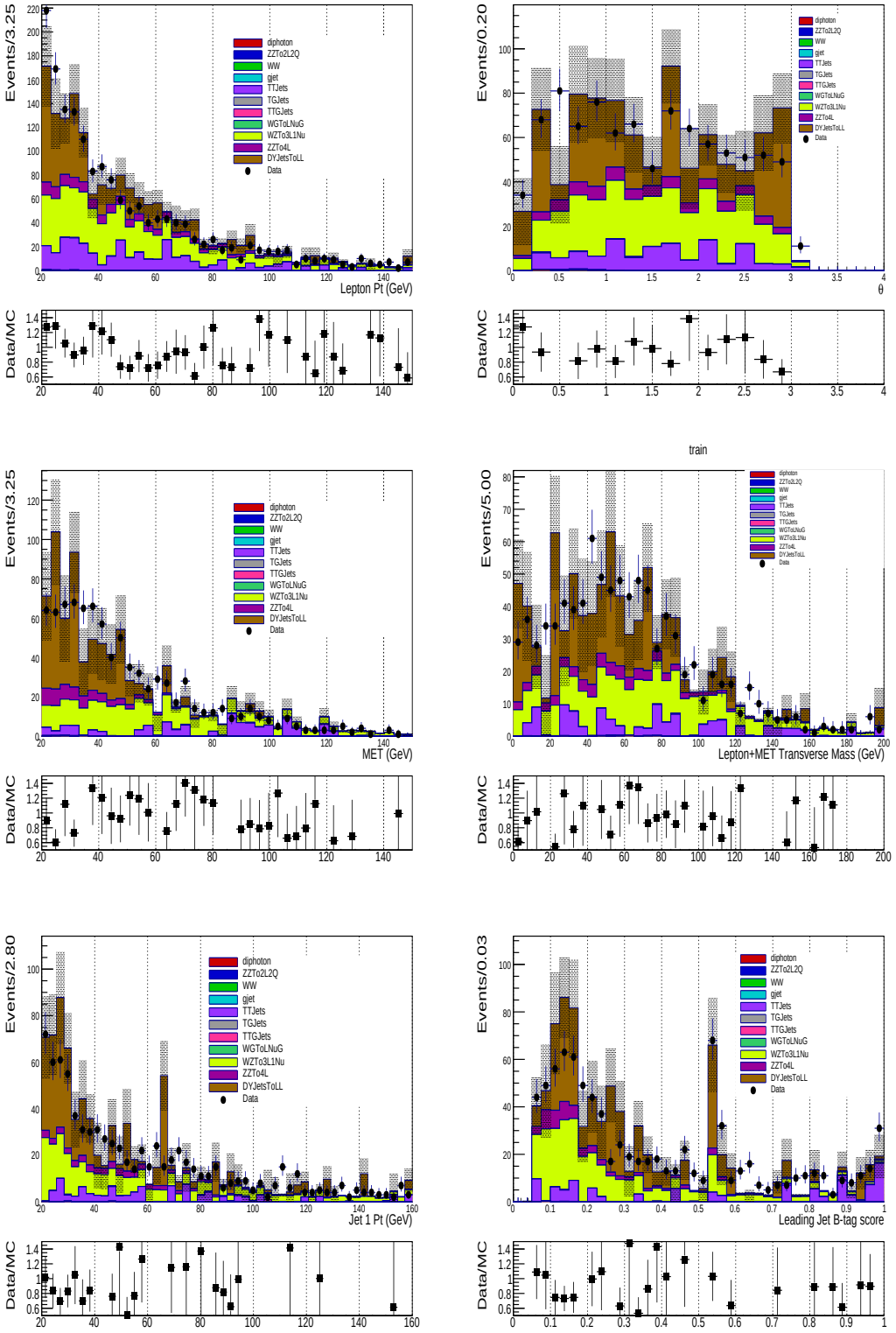

Figure 4.28: Comparison of the BDT input distributions in data and simulated events, selecting WZ events with the W boson decaying leptonically and the Z boson decaying into an electron-positron pair, reconstructed as photons.

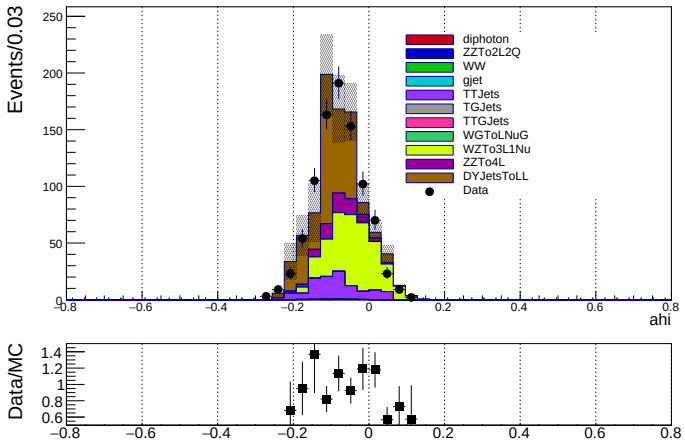

Figure 4.29: Comparison of the BDT output in data and simulated events, selecting WZ events with the W boson decaying leptonically and the Z boson decaying into an electron-positron pair, reconstructed as photons.

by 2.35 (σ_H), are showed. The expected number of background events per GeV, in the corresponding σ_{eff} interval, estimated from the fit to the $m_{\gamma\gamma}$ distribution, is also included in the table.

Categories	SM 125GeV Higgs boson expected signal													Bkg (GeV^{-1})
	Total	ggH	vbf	ttH	bbH	tHq	tHW	WHLep	ZHLep	WHHad	ZHHad	σ_{eff}	σ_{HM}	
Untagged 0	32.5	72.0%	16.6%	2.6%	0.6%	0.7%	0.3%	0.6%	0.3%	4.2%	2.2%	1.32	1.26	21.8
Untagged 1	469.3	86.5%	7.9%	0.6%	1.2%	0.1%	<0.05%	0.5%	0.3%	1.9%	1.1%	1.46	1.32	925.1
Untagged 2	678.3	89.9%	5.4%	0.4%	1.2%	0.1%	<0.05%	0.5%	0.3%	1.4%	0.8%	1.93	1.67	2391.7
Untagged 3	624.3	91.3%	4.4%	0.5%	1.0%	0.1%	<0.05%	0.5%	0.3%	1.2%	0.7%	2.61	2.27	4855.1
VBF 0	9.3	15.5%	**83.2%**	0.4%	0.4%	0.3%	<0.05%	<0.05%	<0.05%	0.2%	<0.05%	1.52	1.31	1.6
VBF 1	8.0	28.4%	**69.7%**	0.4%	0.6%	0.4%	<0.05%	0.1%	<0.05%	0.3%	0.1%	1.66	1.38	3.3
VBF 2	25.2	45.1%	**51.2%**	0.9%	0.8%	0.6%	0.1%	0.2%	0.1%	0.8%	0.3%	1.64	1.37	18.9
ttH Hadronic	5.6	7.0%	0.7%	**81.1%**	2.1%	4.3%	2.1%	0.1%	0.1%	0.7%	1.9%	1.48	1.30	2.4
ttH Leptonic	3.8	1.5%	<0.05%	**87.8%**	0.1%	4.7%	3.1%	1.5%	1.2%	<0.05%	<0.05%	1.60	1.35	1.5
ZH Leptonic	0.5	<0.05%	<0.05%	2.6%	<0.05%	<0.05%	0.1%	<0.05%	**97.3%**	<0.05%	<0.05%	1.65	1.43	0.1
WH Lep	3.6	1.3%	0.6%	5.2%	0.2%	3.0%	0.7%	**84.5%**	4.3%	0.1%	0.1%	1.64	1.43	2.1
VH LepLoose	2.7	8.1%	2.7%	2.4%	0.6%	1.8%	0.1%	**64.4%**	**19.1%**	0.6%	0.2%	1.67	1.56	3.5
VH Had	7.9	47.6%	4.5%	4.4%	0.4%	1.7%	0.3%	0.2%	0.5%	**25.2%**	**15.1%**	1.38	1.30	7.2
VH Met	4.0	18.7%	2.6%	15.4%	0.4%	2.1%	1.2%	**26.8%**	**30.4%**	1.4%	0.9%	1.56	1.39	3.5
Total	1875.0	86.9%	7.1%	1.0%	1.1%	0.2%	<0.05%	0.8%	0.4%	1.6%	0.9%	1.96	1.62	8237.8

Table 4.9: Expected number of signal events per category and the percentage breakdown per production mode. The σ_{eff} is also reported as an estimation of the diphoton invariant mass resolution in each category.

In figure 4.30, the same information is provided, together with a figure of merit of the signal-over-background ratio, expressed as the ratio of the number of expected signal events and the sum of the expected number of signal and background events.

The diphoton invariant mass spectra in data, with the signal plus background model fit, with the 1 σ and 2σ bands associated to the uncertainty in background component of the fit are showed in figures 4.31, 4.32,4.33, 4.34, for each category of the analysis.

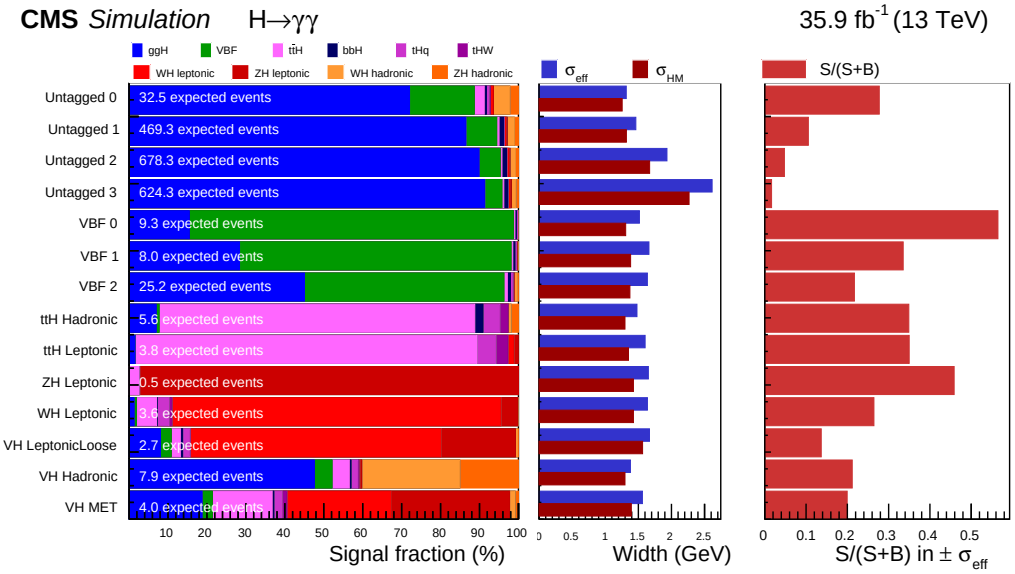

Figure 4.30: The expected number of signal events per category and the percentage breakdown per production mode are showed. The σ_{eff} is also provided as an estimate of the $m_{\gamma\gamma}$ resolution in each category and compared to the full-width at half maximum. The ratio of the number of expected signal events (S) to the sum of the number of signal background events (S+B) is showed on the right panel.

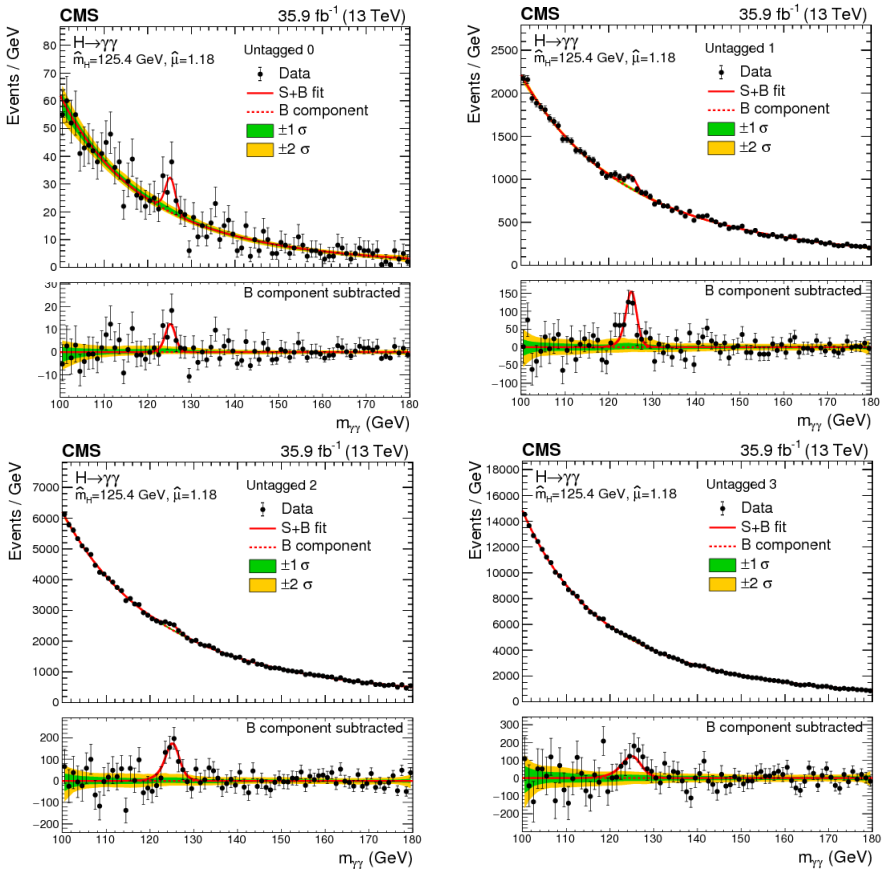

Figure 4.31: Diphoton invariant mass distribution in each analysis category. Also the best background fit, with its 1σ and 2σ uncertainty bands, and the signal model are showed.

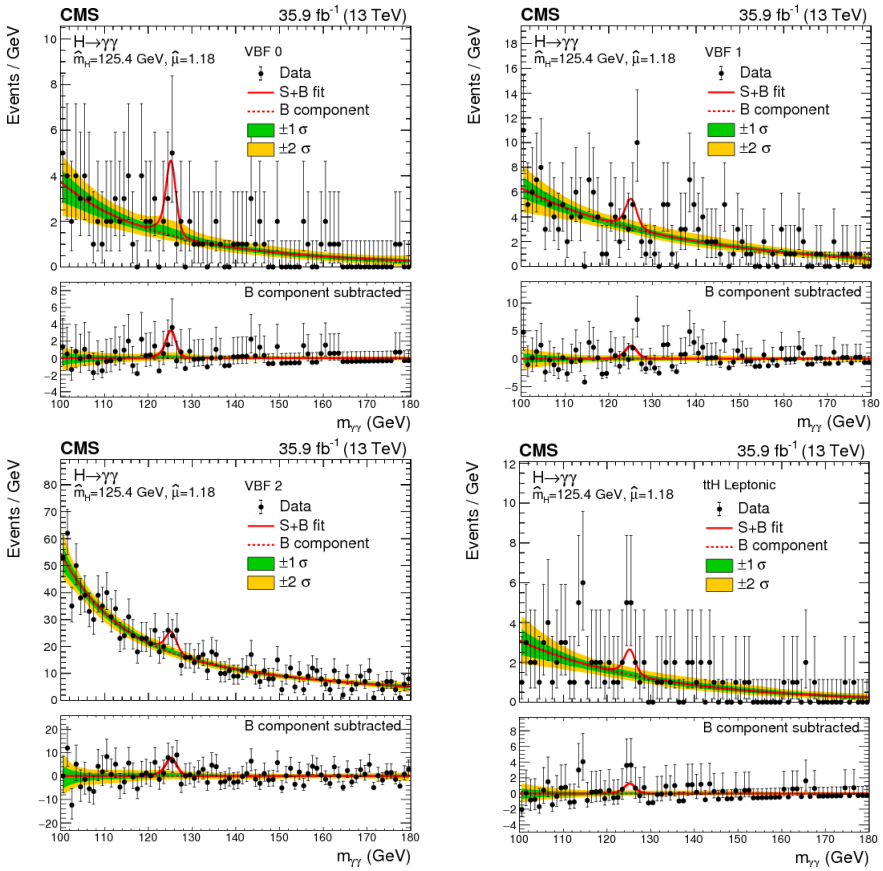

Figure 4.32: Diphoton invariant mass distribution in each analysis category. Also the best background fit, with its 1σ and 2σ uncertainty bands, and the signal model are showed.

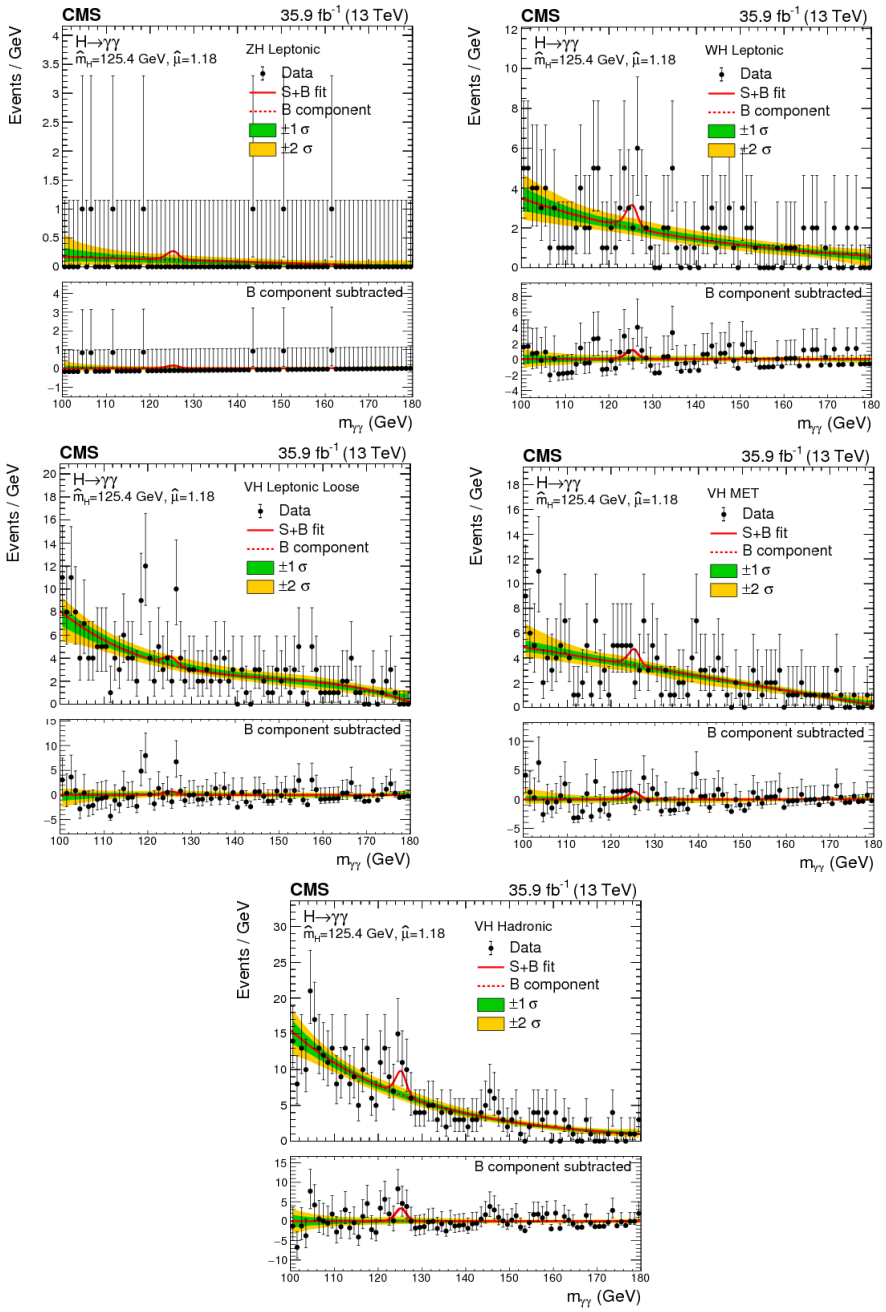

Figure 4.33: Diphoton invariant mass distribution in each analysis category. Also the best background fit, with its 1σ and 2σ uncertainty bands, and the signal model are showed.

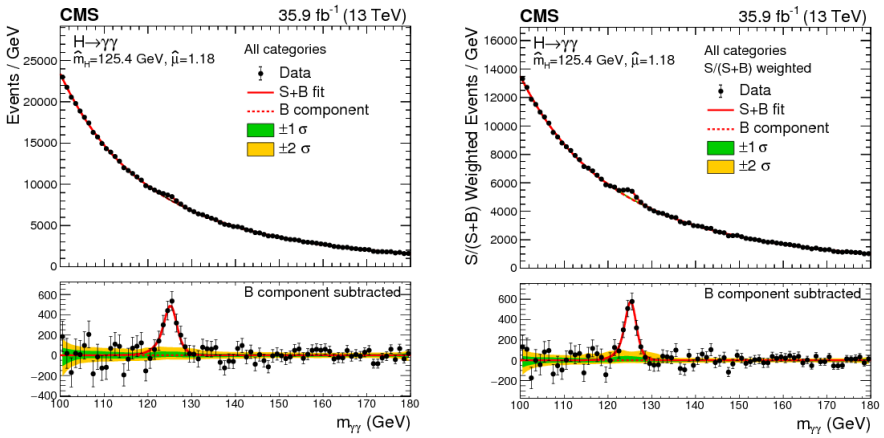

Figure 4.34: Diphoton invariant mass distribution in each analysis category. Also the best background fit, with its 1σ and 2σ uncertainty bands, and the signal model are showed.

A set of different results is extracted from the data collected in each category. The *signal strength* (μ), which quantifies the agreement between the measurement and the SM prediction, is defined as the ratio between the measured Higgs boson cross-section and the prediction of the SM. The overall best fit signal strength was measured to be $\mu=1.18^{+0.17}_{-0.14}=1.18^{+0.12}_{-0.11}(\text{stat.})^{+0.09}_{-0.07}(\text{syst.})^{+0.07}_{-0.06}(\text{theoretical})$. It is measured through a likelihood scan, visible in figure 4.35, with all the other parameters of the model left floating. Systematic uncertainties are included as nuisance parameters.

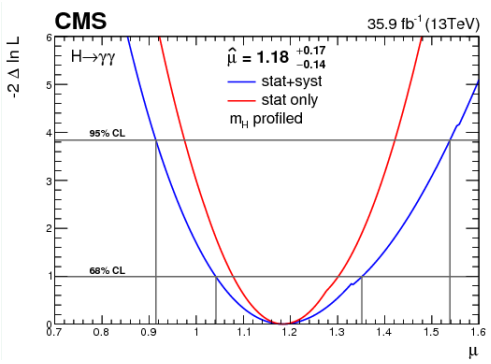

Figure 4.35: The likelihood scan of the signal strength, with all nuisance parameters, including also the Higgs boson mass, profiled in the fit.

The signal strength for each of the four main production processes, defined in the same way as described above, were measured. The inclusion in the analysis of the VH dedicated categories enabled to measure μ_{VH} in the Higgs boson diphoton decay channel for the first time with data collected during the LHC Run 2: $\mu_V H = 2.4^{+1.1}_{-1.0}$. The previous

measurement of $\mu_V H$, derived from the analysis of the whole LHC Run 1 dataset was: μ_{VH}=-0.16$^{+1.16}_{-0.79}$ [72]. The significance of the observed and expected excesses with respect to the absence of the VH production mechanism are 2.4 σ and 1.2 σ, respectively. The results obtained for each mechanism are summarized in figure 4.36.

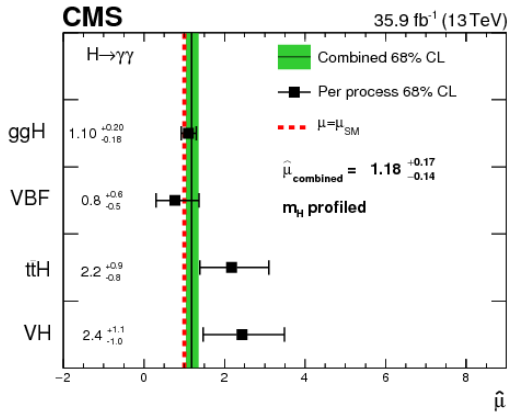

Figure 4.36: Measured signal strength of each Higgs boson production process. The green band represents the measurement of the overall signal strength, while the red dashed line corresponds to the SM prediction

A two-dimensional likelihood scan of the signal strengths related to the fermionic production processes, ggH (μ_{ggH}) and ttH (μ_{ttH}), and the bosonic production modes, VBF (μ_{VBF}) and VH (μ_{VH}), with the m$_H$ profiled in the fit was performed. The bets fit values, μ_{ggH}, μ_{ttH}=1.19$^{+0.22}_{-0.18}$ and μ_{VBF}, μ_{VH}=1.21$^{+0.58}_{-0.51}$, along with the 1 σ and 2 σ contours are reported in figure 4.37.

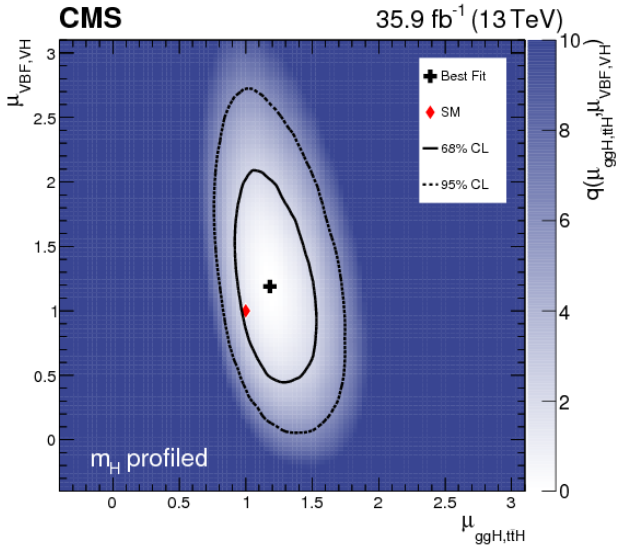

Figure 4.37: Two-dimensional best fit (cross) of the signal strength modifier of the fermionic (μ_{ggH}, μ_{ttH}) and bosonic (μ_{VBF}, μ_{VH}) production mechanisms.

Coupling measurement in the k-framework

The couplings of the Higgs boson to SM particles can be derived from the measured signal strengths. A theoretical framework, called k-framework [67], was developed to quantify the consistency of the characteristics of Higgs boson discovered in 2012, with the SM expectation. A set of coupling modifiers, labeled by the letter k, is used to parametrize deviations of the measured values from the SM predictions. For each production process or decay mode, a coupling modifier is defined as $k_i^2 = \sigma_i/\sigma_{SM}$ or $k_i^2 = \Gamma_i^2/\Gamma_{SM}$, respectively. The SM expectations correspond to values of the k modifiers equal to 1. Coupling modifiers for each tree-level Higgs boson coupling to fermions and bosons are defined, as well as effective modifier accounting for the ggH production and H→ γγ decay, which occur through loops and may be modified by contribution to the loop from particles predicted by beyond standard model particles. The result of the H→ γγ analysis were interpreted also within the k-framework. Two-dimensional likelihood scans of the modifier of the Higgs boson coupling to fermions (k_f) versus the modifier for vector bosons (k_V), and for the effective coupling of the Higgs boson with gluons (k_g) and photons (k_γ) was performed. The value of the test statistics, and the best fit values with the 1 σ and 2 σ contours are reported in the plots of figure 4.38. The point (k_V, k_f) = (1, − 1) has an observed (expected) q of 35.2 (53.7), inconsistent with the best fit to an observed (expected) level of 5.8 σ (7.0 σ). In both cases, the measured values are in agreement with the SM values, within the uncertainty.

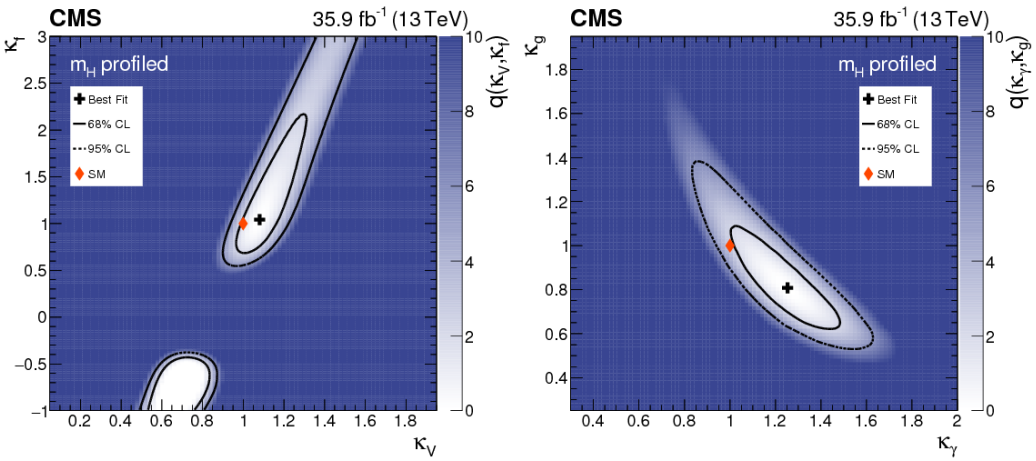

Figure 4.38: 2-dimensions likelihood scan of the Higgs boson coupling modifier to fermions and vector bosons (left), and effective modifier of the coupling to gluons and photons (right). The red diamond corresponds to the SM values, while the black cross indicates the best-fit value. The solid line and dashed line trace the 1σ and 2σ contours.

Interpretation in the Simplified Template cross-section framework

The measurement of the Higgs boson properties performed so far in terms of signal strengths and modifiers of couplings, in the k-framework. The simplified template cross section framework (STXS) [75] represents an evolution in the study of the Higgs boson properties. It aims at reducing the theory dependence of the measurements, resulting both in a reduction of the uncertainty due to theory and the underlying physics model.

Moreover, it enables to investigate the properties of the Higgs boson with a finer granularity. The core of the framework is the measurement of the Higgs boson production cross-sections in mutually exclusive portions of the phase-space (called bins). The bins are defined in such a way that the dependence on theoretical assumptions is minimized, and the possible effect arising from beyond standard model particles or interactions is isolated and enhanced. A schematic overview of the stage 0 categorization is depicted in figure 4.39.

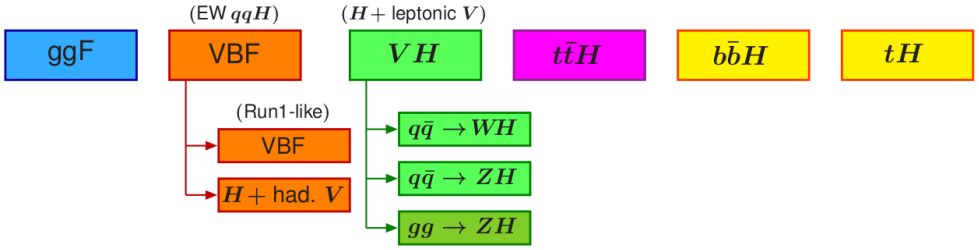

Figure 4.39: Overview of the bin definition of the Stage 0 Simplified template cross-section.

The ratio in the VH hadronic bin was measured to be $5.1^{+2.5}_{-2.3}$. The results of the H$\to \gamma\gamma$ analysis in the STXS stage 0 interpretation are summarized in figure 4.40.

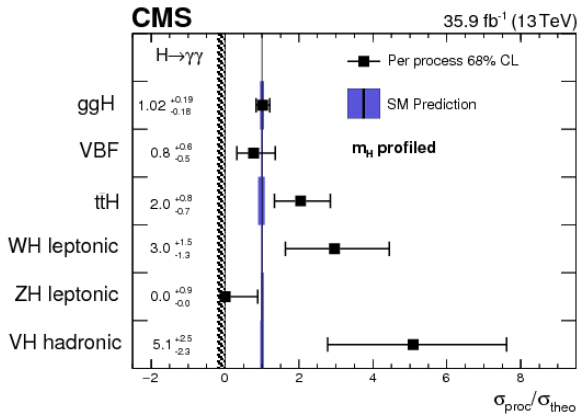

Figure 4.40: Measurement of the Higgs boson production cross sections ratios to the SM predictions, in the STXS stage 0 bins.

Chapter 5

Conclusions

The research activity during my Ph.D. activity was performed within the CMS experiment collaboration. The primary focus of my work was the study of the associated production of the Higgs boson with a vector boson, in the Higgs boson diphoton decay channel(H$\rightarrow \gamma\gamma$), in proton-proton collisions, using data collected at a center-of-mass energy of 13 TeV with the CMS experiment in 2016 for an integrated luminosity of 35.9 fb^{-1}.

The electromagnetic calorimeter (ECAL) plays a key role in the H$\rightarrow \gamma\gamma$ analysis and must provide an excellent resolution on the measurement of the photon energy. In the range of energy relevant for photons produced in the Higgs boson decay, two main contributions to the energy resolution arise from an accurate channel intercalibration and response stability throughout the datataking period. My main contribution to the ECAL activity regarded the channel intercalibration and the monitoring and stabilization of the ECAL energy scale using high-energy electrons from W and Z bosons.

The overall best fit signal strength, defined as the ratio between the measured cross-section and the SM prediction, was measured to be μ=$1.18^{+0.17}_{-0.14}$=$1.18^{+0.12}_{-0.11}$(stat.)$^{+0.09}_{-0.07}$(syst.)$^{+0.07}_{-0.06}$(theoretical). The signal strength for each of the four main Higgs boson production processes at the LHC, the gluon-gluon fusion (ggH), vector boson fusion (VBF), associated production with a vector boson (VH), and associated production with a top quark pair (ttH), were measured as well. In particular, a two-dimensional likelihood scan of the signal strength for fermionic production modes (ggH and ttH) and for vector boson production modes (VBF, VH) was performed. The measured signal strength for the vector boson production mechanisms is $\mu_{VBF,VH}$=$1.21^{+0.58}_{-0.51}$. The inclusion in the analysis of the dedicated VH categories enabled to measure the signal strength related to this production process for the first time with data collected during the LHC Run 2, with $\mu_{VH} = 2.4^{+1.1}_{-1.0}$. It was previously measured with the data collected during the whole LHC Run 1 to be μ_{VH}=-$0.16^{+1.16}_{-0.79}$. The significance of the observed and expected excesses with respect to the absence of the VH production mechanism are 2.4 σ and 1.2 σ, respectively. The coupling of the Higgs boson to vector bosons, which is involved in the VH mechanism, was measured within the k-framework and no deviations from the SM prediction was found. The results of the analysis were also interpreted in the Simplified Template Cross-Section framework, where the ratio between the observed and the SM cross-section in mutually exclusive portions of the phase-space are measured. The ratio in the VH hadronic bin was measured to be $5.1^{+2.5}_{-2.3}$. All the measurement are in agreement with the prediction of the standard model of particles and interactions.

The WH leptonic tag was subject of a longer-term study targeting the analysis of the entire LHC Run 2 dataset. The improvement in background rejection thanks to a multivariate technique analysis approach was estimated considering the significance of

the category as a figure of merit. It was estimated to be approximately 35%, resulting a significance with a dataset corresponding to 100 fb^{-1} integrated luminosity of about 2.48σ.

Bibliography

[1] Lyndon Evans and Philip Bryant. Lhc machine. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001.

[2] Albert Einstein. The Foundation of the General Theory of Relativity. *Annalen Phys.*, 49(7):769–822, 1916. doi: 10.1002/andp.200590044,10.1002/andp.19163540702. [,65(1916)].

[3] E. Fermi. Zur quantelung des idealen einatomigen gases. *Zeitschrift für Physik*, 36 (11):902–912, Nov 1926. ISSN 0044-3328. doi: 10.1007/BF01400221. URL `https://doi.org/10.1007/BF01400221`.

[4] On the theory of quantum mechanics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 112(762):661–677, 1926. ISSN 0950-1207. doi: 10.1098/rspa.1926.0133. URL `http://rspa.royalsocietypublishing.org/content/112/762/661`.

[5] M. Tanabashi et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL `https://link.aps.org/doi/10.1103/PhysRevD.98.030001`.

[6] C N. Yang and R L. Mills. Conservation of isotopic spin and isotopic gauge invariance. 96:191–195, 10 1954.

[7] E. Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918:235–257, 1918. URL `http://eudml.org/doc/59024`.

[8] David J. Gross and Frank Wilczek. Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, 30:1343–1346, Jun 1973. doi: 10.1103/PhysRevLett.30.1343. URL `https://link.aps.org/doi/10.1103/PhysRevLett.30.1343`.

[9] The ALEPH, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the SLD Electroweak and Heavy Flavour Groups. Precision Electroweak Measurements on the Z Resonance. *Phys. Rept.*, 427:257, 2006.

[10] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964. doi: 10.1103/PhysRevLett.13.321. URL `https://link.aps.org/doi/10.1103/PhysRevLett.13.321`.

[11] Peter W. Higgs. Spontaneous symmetry breakdown without massless bosons. *Phys. Rev.*, 145:1156–1163, May 1966. doi: 10.1103/PhysRev.145.1156. URL `https://link.aps.org/doi/10.1103/PhysRev.145.1156`.

[12] Luis Álvarez-Gaumé and John Ellis. Eyes on a prize particle. *Nature Physics*, 7:2 EP –, Dec 2010. URL http://dx.doi.org/10.1038/nphys1874.

[13] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. Broken symmetries. *Phys. Rev.*, 127:965–970, Aug 1962. doi: 10.1103/PhysRev.127.965. URL https://link.aps.org/doi/10.1103/PhysRev.127.965.

[14] 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. doi: 10.1103/PhysRevLett.114.191803.

[15] Apollinari G., Béjar Alonso I., Brüning O., Fessia P., Lamont M., Rossi L., and Tavian L. *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1*. CERN Yellow Reports: Monographs. CERN, Geneva, 2017. URL https://cds.cern.ch/record/2284929.

[16] Bruce Mellado Garcia, Pasquale Musella, Massimiliano Grazzini, and Robert Harlander. CERN Report 4: Part I Standard Model Predictions. May 2016. URL https://cds.cern.ch/record/2150771.

[17] J Butler, D Contardo, M Klute, J Mans, L Silvestris, and Collaboration on behalf of the CMS. CMS Phase II Upgrade Scope Document. Technical Report CERN-LHCC-2015-019. LHCC-G-165, CERN, Geneva, Sep 2015. URL https://cds.cern.ch/record/2055167.

[18] Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV. Technical Report CMS-PAS-HIG-15-002. ATLAS-CONF-2015-044, CERN, Geneva, 2015. URL https://cds.cern.ch/record/2053103.

[19] Morad Aaboud et al. Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector. *Phys. Lett.*, B786:59–86, 2018. doi: 10.1016/j.physletb.2018.09.013.

[20] A. M. Sirunyan et al. Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121(12):121801, 2018. doi: 10.1103/PhysRevLett.121.121801.

[21] Combined measurements of the Higgs boson's couplings at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-HIG-17-031, CERN, Geneva, 2018. URL http://cds.cern.ch/record/2308127.

[22] The Large Electron-Positron Collider. Jul 2012. URL https://cds.cern.ch/record/1997351.

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

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

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

[26] *LHCb : Technical Proposal*. Tech. Proposal. CERN, Geneva, 1998. URL https://cds.cern.ch/record/622031.

[27] Esma Mobs. The CERN accelerator complex. Complexe des accélérateurs du CERN. Jul 2016. URL https://cds.cern.ch/record/2197559. General Photo.

[28] Esma Mobs. The CERN accelerator complex. Complexe des accélérateurs du CERN. Jul 2016. URL https://cds.cern.ch/record/2197559. General Photo.

[29] Tai Sakuma and Thomas McCauley. Detector and event visualization with sketchup at the cms experiment. *Journal of Physics: Conference Series*, 513(2):022032, 2014. URL http://stacks.iop.org/1742-6596/513/i=2/a=022032.

[30] V Karimaki, M Mannelli, P Siegrist, H Breuker, A Caner, R Castaldi, K Freudenreich, G Hall, R Horisberger, M Huhtinen, and A Cattai. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/368412.

[31] CMS Collaboration. ECAL Technical Design Report (TDR) Figures from Chapter 1. CMS Collection., Dec 1997. URL http://cds.cern.ch/record/1327662.

[32] *The CMS hadron calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/357153.

[33] Min Suk Kim. CMS reconstruction improvement for the muon tracking by the RPC chambers. *PoS*, RPC2012:045, 2012. doi: 10.1088/1748-0221/8/03/T03001. [JINST8,T03001(2013)].

[34] Vardan Khachatryan et al. Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s} = 8$ TeV. *JINST*, 10(02):P02006, 2015. doi: 10.1088/1748-0221/10/02/P02006.

[35] CMS Collaboration. Particle-flow reconstruction and global event description with the cms detector. *Journal of Instrumentation*, 12(10):P10003, 2017. URL http://stacks.iop.org/1748-0221/12/i=10/a=P10003.

[36] Performance of electron reconstruction and selection with the cms detector in proton-proton collisions at$\sqrt{}$ s = 8 tev. *Journal of Instrumentation*, 10(06):P06005, 2015. URL http://stacks.iop.org/1748-0221/10/i=06/a=P06005.

[37] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the cms tracker. *Journal of Instrumentation*, 9(10):P10009, 2014. URL http://stacks.iop.org/1748-0221/9/i=10/a=P10009.

[38] R. Fruhwirth. Application of kalman filtering to track and vertex fitting. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 262(2):444 – 450, 1987. ISSN 0168-9002. doi: https://doi.org/10.1016/0168-9002(87)90887-4. URL http://www.sciencedirect.com/science/article/pii/0168900287908874.

[39] W Adam, R Frühwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the cms tracker at the lhc. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9, 2005. URL http://stacks.iop.org/0954-3899/31/i=9/a=N01.

[40] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. Fastjet user manual. *The European Physical Journal C*, 72(3):1896, Mar 2012. ISSN 1434-6052. doi: 10.1140/epjc/s10052-012-1896-2. URL https://doi.org/10.1140/epjc/s10052-012-1896-2.

[41] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The catchment area of jets. *Journal of High Energy Physics*, 2008(04):005, 2008. URL http://stacks.iop.org/1126-6708/2008/i=04/a=005.

[42] Matteo Cacciari and Gavin P. Salam. Pileup subtraction using jet areas. *Physics Letters B*, 659(1):119 – 126, 2008. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2007.09.077. URL http://www.sciencedirect.com/science/article/pii/S0370269307011094.

[43] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti- k t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008. URL http://stacks.iop.org/1126-6708/2008/i=04/a=063.

[44] Vardan Khachatryan et al. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *JINST*, 12(02):P02014, 2017. doi: 10.1088/1748-0221/12/02/P02014.

[45] Henning Kirschenmann. Jet performance in CMS. Technical Report CMS-CR-2013-325, CERN, Geneva, Oct 2013. URL https://cds.cern.ch/record/1627818.

[46] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup Per Particle Identification. *JHEP*, 10:059, 2014. doi: 10.1007/JHEP10(2014)059.

[47] CMS Collaboration. Pileup Jet Identification. 2013.

[48] CMS Collaboration. Identification of b quark jets at the CMS Experiment in the LHC Run 2. 2016.

[49] Vardan Khachatryan et al. Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s} = 8$ TeV. *JINST*, 10(02):P02006, 2015. doi: 10.1088/1748-0221/10/02/P02006.

[50] Performance of photon reconstruction and identification with the cms detector in proton-proton collisions at $\sqrt{}$ s = 8 tev. *Journal of Instrumentation*, 10(08):P08010, 2015. URL http://stacks.iop.org/1748-0221/10/i=08/a=P08010.

[51] Emanuele Di Marco. CMS electromagnetic calorimeter calibration and timing performance during LHC Run I and future prospects. Technical Report CMS-CR-2014-410, CERN, Geneva, Nov 2014. URL http://cds.cern.ch/record/1975982.

[52] Emanuele Di Marco. CMS electromagnetic calorimeter calibration and timing performance during LHC Run I and future prospects. Technical Report CMS-CR-2014-410, CERN, Geneva, Nov 2014. URL http://cds.cern.ch/record/1975982.

[53] The CMS Electromagnetic Calorimeter Group. Radiation hardness qualification of pbwo 4 scintillation crystals for the cms electromagnetic calorimeter. *Journal of Instrumentation*, 5(03):P03010, 2010. URL http://stacks.iop.org/1748-0221/5/i=03/a=P03010.

[54] T J Orimoto. The CMS ECAL Laser Monitoring System. page Pages, 2008. URL https://cds.cern.ch/record/1742291.

[55] A Van Lysebetten and P Verrecchia. Performance and Measurements of the Light Monitoring System for CMS-ECAL from 2002 Test Beam Data. Technical Report CMS-RN-2004-001, CERN, Geneva, Feb 2004. URL `https://cds.cern.ch/record/787485`.

[56] Alessio Ghezzi, Martina Malberti, Stefano Ragazzi, Chiara Rovelli, and Tommaso Tabarelli de Fatis. Analysis of the response evolution of the CMS electromegnetic calorimeter under electron and pion irradiation. Technical Report CMS-NOTE-2006-038, CERN, Geneva, Feb 2006. URL `https://cds.cern.ch/record/934066`.

[57] Petar Adzic and Others. Energy resolution performance of the cms electromagnetic calorimeter. Technical Report CMS-NOTE-2006-140, CERN, Geneva, Sep 2006. URL `http://cds.cern.ch/record/1000388`.

[58] M. Oreglia. *A Study of the Reactions* $\psi' \to \gamma\gamma\psi$. PhD thesis, SLAC, 1980. URL `http://www-public.slac.stanford.edu/sciDoc/docMeta.aspx?slacPubNumber=slac-r-236.html`.

[59] Technical proposal: L3. Technical Report CERN-LEPC-83-5. LEPC-P-4, 1983. URL `https://cds.cern.ch/record/297266`.

[60] G. Breit and E. Wigner. Capture of slow neutrons. *Phys. Rev.*, 49:519–531, Apr 1936. doi: 10.1103/PhysRev.49.519. URL `https://link.aps.org/doi/10.1103/PhysRev.49.519`.

[61] David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. doi: 10.1017/CBO9780511755453.

[62] 2015 ECAL detector performance plots. Dec 2015. URL `https://cds.cern.ch/record/2114735`.

[63] S Banerjee. Cms simulation software. *Journal of Physics: Conference Series*, 396 (2):022003, 2012. URL `http://stacks.iop.org/1742-6596/396/i=2/a=022003`.

[64] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506: 250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8.

[65] 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. doi: 10.1007/JHEP07(2014)079.

[66] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to pythia 8.1. *Computer Physics Communications*, 178(11):852 – 867, 2008. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2008.01.036. URL `http://www.sciencedirect.com/science/article/pii/S0010465508000441`.

[67] J R Andersen et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 2013. doi: 10.5170/CERN-2013-004.

[68] 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. doi: 10.1088/1126-6708/2009/02/007.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[70] P D. Dauncey, M Kenzie, N Wardle, and G J. Davies. Handling uncertainties in background shapes: The discrete profiling method. *Journal of Instrumentation*, 10, 08 2014. doi: 10.1088/1748-0221/10/04/P04015.

[71] A.M. Sirunyan et al. Measurements of $t\bar{t}$ cross sections in association with b jets and inclusive jets and their ratio using dilepton final states in pp collisions at s=13tev. *Physics Letters B*, 776:355 – 378, 2018. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2017.11.043. URL `http://www.sciencedirect.com/science/article/pii/S0370269317309358`.

[72] Observation of the diphoton decay of the Higgs boson and measurement of its properties. *Eur. Phys. J. C*, 74:3076, 2014. doi: 10.1140/epjc/s10052-014-3076-z.

[73] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL `http://www.sciencedirect.com/science/article/pii/S002200009791504X`.

[74] Jan Therhaag. TMVA Toolkit for multivariate data analysis in ROOT. *PoS*, ICHEP2010:510, 2010. doi: 10.22323/1.120.0510.

[75] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016. doi: 10.23731/CYRM-2017-002.