SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Fisica G. Occhialini

Dottorato di Ricerca in / PhD program: Physics and astronomy Ciclo / Cycle XXXII

MEASUREMENT OF THE ASSOCIATED PRODUCTION OF TOP QUARK PAIRS WITH A HIGGS BOSON IN THE DIPHOTON FINAL STATE WITH THE CMS DETECTOR

Cognome / Surname Beschi Nome / Name Andrea

Matricola / Registration number 746069

Tutore / Tutor: Prof. Tommaso Tabarelli de Fatis

Coordinatore / Coordinator: Prof.ssa Marta Calvi

ANNO ACCADEMICO / ACADEMIC YEAR 2018/2019

Live as if you were to die tomorrow.
Learn as if you were to live forever.

Abstract

This thesis presents the analysis of the data collected by the CMS detector at the Large Hadron Collider to perform a precise measurement of the cross section of the Higgs boson production in association with a pair of top-antitop quarks ($t\bar{t}H$). The diphoton decay channel of the Higgs boson is exploited to perform the measurement because it is one of the most sensitive channels. The measurement is based on 35.9 fb^{-1} and 41.5 fb^{-1} of data collected in 2016 and 2017 at a centre-of-mass energy of 13 TeV.
The precise determination of the cross section of the $t\bar{t}H$ process is one of the major targets of the CMS experimental program as it could provide a stringent consistency test for the Standard Model of particle interactions (SM). The $t\bar{t}H$ process offers the unique possibility of a direct measurement of the top quark Yukawa coupling (y_t). The value of y_t can be inferred from the production cross section of processes involving virtual contributions (loops) from massive particles, where the top quark gives major contributions, as the gluon fusion Higgs boson production or the decay of the Higgs boson to a pair of photons. The derivation of y_t from loops is possible only assuming no contributions from unobserved particles, while the measurement from the $t\bar{t}H$ process would live aside this assumption. The direct measurement of y_t would provide a direct comparison with the indirect constraint, offering the unique opportunity to explore the inner structure of the loops predicted by the SM.
The choice to exploit events with a diphoton decay of the Higgs boson is due to the rather clean experimental environment provided by this topology and to the fully reconstructed final state. The experimental signature of a Higgs boson decaying to a pair of photons ($H \rightarrow \gamma\gamma$) is the presence of a narrow resonant peak, only smeared by the experimental resolution, arising in the invariant mass distribution of the photon pairs. The $t\bar{t}H$ production can be exclusively identified by the presence in the final state of the decay products of the top quarks. Events with similar topology but arising from different processes, namely backgrounds, challenge the identification of the $t\bar{t}H$ process. The main backgrounds are due to non-resonant photon production or jet fragments misidentified as photons in association with jets or leptons, mimicking the $t\bar{t}$ quarks, or to real $t\bar{t}$ pairs accompanied by photons or jets mis-reconstructed as photons.
The identification of the $t\bar{t}H$ process starts from the identification of the photon candidates. Photons are reconstructed as isolated energy deposits in the electromagnetic calorimeter (ECAL) not linked to any reconstructed track. In addition, the energy distribution in the ECAL cells must be compatible with the one expected from a photon shower. Events with two photon candidates are selected for the analysis.
Further selections are applied, which exploit the presence of top quarks decay products in the final state. As the top quark decays with nearly 100% probability to a W boson and a b quark, distinct experimental signatures to tag the presence of a top quark pair are

provided by the decays of the W bosons. If both the W bosons decay hadronically, the resulting final state features six jets, two of which originating from b quarks (b jets). If one of the W bosons decays in a lepton and one hadronically, the final state has a charged lepton, a neutrino, and four jets, two of which are b jets. Finally, when both the W bosons decay leptonically, the resulting final state has two charged leptons, two neutrinos, and two b jets. The definition of the selections to exclusively identify the $t\bar{t}H$ production has been the area of contribution of this work.

The analysis of the 2016 data is performed splitting the events into two categories, one targeting hadronic decays of the top quarks and the second one collecting events with at least one lepton. The background rejection is performed with a Boosted Decision Tree (BDT) in the hadronic category, exploiting variables related to jets and b jets, and by a cut-based selection in the leptonic category, based on requirements on jets and high-momentum leptons. The signal-to-background ratio is enhanced by a second BDT, common to the categories, which aims at the selection of high-energy and well-reconstructed photons compatible with originating from the decay of the Higgs boson.

The analysis of the 2017 data featured several improvements, with a sensitivity increase of about 50%. Two BDTs are trained, one for the hadronic and one for the leptonic topology, exploiting variables related to kinematics of the event and to the quality of the reconstruction of photons, jets, b jets and leptons. The BDTs are capable to exploit the correlation among the input variables, providing a more efficient rejection of the background. The hadronic category is further split into three subcategories, based on the BDT output score, while the leptonic one is split in two. The BDTs are validated exploiting a sample of $t\bar{t}Z$ events as a proxy of the $t\bar{t}H$ process.

The $t\bar{t}H$ production rate is extracted from a fit to the invariant mass spectrum of the photon pairs. The branching ratio of the Higgs boson to photons is constrained to the SM prediction. The fit function adopts a signal model, built from simulation of the Higgs boson production processes, and a background one, derived directly from fitting the data. The fit is performed with floating the signal strength $\mu_{t\bar{t}H}$, defined as the ratio between the measured $t\bar{t}H$ production cross section and the expectation from the SM. The analysis of the 2016 data resulted in an observed value of $\hat{\mu}_{t\bar{t}H} = 2.2^{+0.9}_{-0.8}$, corresponding to a rejection of the background-only hypothesis at the level of 3.2 standard deviations, where 1.5 is expected for a SM Higgs boson. The 2017 analysis measured a signal strength of $\hat{\mu}_{t\bar{t}H} = 1.3^{+0.7}_{-0.5}$, rejecting the background-only hypothesis at the level of 3.1 standard deviations, where 2.2 are expected for a SM Higgs boson. The combination of the two analyses resulted in an observed signal strength of $\hat{\mu}_{t\bar{t}H} = 1.7^{+0.6}_{-0.5}$, corresponding to a significance of 4.1 standard deviations.

The analysis of the 2016 data, together with the analyses targeting final states with the Higgs boson decaying to b quarks, vector bosons, and τ leptons, allowed the first experimental observation of the $t\bar{t}H$ process. This result exploited the analysis of the data collected in 2016 at centre-of-mass energy of 13 TeV, as well as the data collected in 2011 and 2012 at a centre-of-mass energy of 7 and 8 TeV, respectively. The best-fit value of the combination of the different channels is $\hat{\mu}_{t\bar{t}H} = 1.26^{+0.31}_{-0.26}$, in agreement with the SM expectation. The background-only hypothesis is rejected at the level of 5.1 standard deviations. This result proves for the first time the tree-level coupling of the Higgs boson with the top quarks and, hence, with an up-type quark.

The whole document is divided into six parts. At first, Chapter 1 gives an overview of the SM and of the knowledge of the Higgs boson at the beginning of this work. The

reasons that lead to the investigation of the $t\bar{t}H$ production are also described. After a short description of the LHC accelerator and of the CMS experiment in Chapter 2, the contributions to the maintenance and to the refinement of the detector performance are described in Chapter 3. The analysis of the data conducted to achieve a measurement of the $t\bar{t}H$ process cross section is described in Chapter 4. Chapter 5 describes the observation of the $t\bar{t}H$ process and gives an overview of the future prospects for the measurements. Finally, Chapter 6 presents a summary of the results achieved in this thesis.

Contents

Chapter 1

Introduction

La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto.

Galieo Galilei

This work presents the analysis of the data collected by the CMS experiment during 2016 and 2017 in proton-proton collisions at a centre-of-mass energy of 13 TeV to achieve a precise measurement of the cross section of associated production of Higgs bosons and top-antitop quark pairs ($t\bar{t}H$). The $t\bar{t}H$ process, within the standard model of particle interactions (SM), enables the direct measurement of the Yukawa coupling of the top quark (y_t), allowing a stringent consistency test of the SM. The diphoton Higgs boson decay channel is exploited to perform the measurement, as among the most experimentally sensitive channels. After a short introduction on the SM, this chapter presents the motivations suggesting possible extensions of the SM and of the Higgs boson sector, before reviewing in detail the experimental knowledge of the Higgs boson. Finally the motivation which make the $t\bar{t}H$ a key process in the context of the SM and its experimental topology are discussed.

1.1 Overview of the standard model of particle interactions

The impressive structure of the standard model of particle interactions [1–4] towers above the theoretical and experimental physics landscape of the last sixty years. Despite incapable to include all the observed phenomena, its unprecedented predicting capability over several orders of magnitude drove the research in particle physics, until its final corroboration with the discovery of the Higgs boson by the ATLAS and CMS Collaborations in 2012 [5, 6]. The SM is a gauge field theory based on a symmetry group SU(3) $\times$ SU(2) $\times$ U(1), where the SU(3) group comes from the Quantum Chromodynamics (QCD) and the SU(2) $\times$ U(1) group describes to the electroweak interaction (EW). The SM features three families of fermions, which differ only for their masses, and the corresponding antifermions; interactions are mediated by gauge bosons. A summary of properties of fermions and bosons can be found in Tables 1.1 and 1.2. The model is entirely described by its Lagrangian, which can be split in a QCD term and an EW term:

		Mass	Electric charge	Colour charge	Mean lifetime (s)
1^{st} gen	u	2.2 MeV	2/3	RGB	-
	d	4.7 MeV	-1/3	RGB	-
	e	511 keV	-1	0	∞
	ν_e	0	0	0	∞
2^{nd} gen	c	1.27 GeV	2/3	RGB	-
	s	95 MeV	-1/3	RGB	-
	μ	106 MeV	-1	0	2.2×10^{-6}
	ν_μ	0	0	0	∞
3^{rd} gen	t	173 GeV	2/3	RGB	-
	b	4.18 GeV	-1/3	RGB	-
	τ	1.78 GeV	-1	0	290×10^{-15}
	ν_τ	0	0	0	∞

Table 1.1: List of known fermions, grouped in the three generations. The mass, the electric charge (in units of positron electric charge), the colour charge and the mean lifetime of each fermion are reported. All values and the convention for quark masses are taken from Ref. [7]. Neutrinos are assumed massless. No lifetime is reported for quarks, since the hadronisation makes the concept ill defined.

	Mass	Electric charge	Spin
γ	0	0	1
$W^\pm$	80.37 GeV	± 1	1
Z	91.18 GeV	0	1
g	0	0	1
H	125.18 GeV	0	0

Table 1.2: List of known bosons. The mass, the electric charge (in units of positron electric charge) and the spin of each boson are reported. All values are taken from Ref. [7].

$$\mathcal{L}^{\mathrm{SM}} = \mathcal{L}^{\mathrm{QCD}} + \mathcal{L}^{\mathrm{EW}}. \tag{1.1}$$

1.1.1 The QCD Lagrangian

The QCD Lagrangian reads:

$$\begin{aligned}\mathcal{L}^{\mathrm{QCD}} = &\sum_q \bar{\psi}_{q,a}\left(i\gamma^\mu\partial_\mu\delta_{ab} - g_s\gamma^\mu t^C_{ab}\mathcal{A}^C_\mu - m_q\delta_{ab}\right)\psi_{q,b} \\ &- \frac{1}{4}F^A_{\mu\nu}F^{A\ \mu\nu} + \theta\frac{g_s^2}{64\pi^2}F^A_{\mu\nu}\epsilon_{\mu\nu\sigma\rho}F^{A\ \mu\nu},\end{aligned} \tag{1.2}$$

where repeated indexes are summed over. The γ^μ are the Dirac matrixes. The $\psi_{q,a}$ are the spinors for a quark of flavour q and colour a, colour index running over R, G, B (three colours exist). Quarks are the fundamental representation of the SU(3) group while colours can be seen as the QCD charge, thus colourless particles, such as leptons, do not experience strong interaction. The interaction term consists of eight matrixes $\mathcal{A}^C_\mu$ corresponding to the gluon fields, with C running from 1 to $\mathrm{N_C^2} - 1 = 8$ (the theory has eight gluons), where $\mathrm{N_C} = 3$ is the number of colours. The t^C_{ab} correspond to the generators of the SU(3) group, representing rotations in the colour space. The quantity g_s is the QCD coupling constant, related to the widely used α_s as α_s= $g_\mathrm{s}^2/4\pi$. The field term is given by:

$$F^A_{\mu\nu} = \partial_\mu\mathcal{A}^A_\nu - \partial_\nu\mathcal{A}^A_\mu - g_\mathrm{s}f_{ABC}\mathcal{A}^B_\mu\mathcal{A}^C_\nu \qquad \left[t^A, t^B\right] = if_{ABC}t^C, \tag{1.3}$$

with f_{ABC} the structure constants of the SU(3) group. The last term can accommodate a CP-violating interaction in the Lagrangian. The parameter θ is a free parameter of the Lagrangian and $\epsilon_{\mu\nu\sigma\rho}$ is the fully-asymmetric Levi-Civita tensor. Present experimental limits on θ constrain the parameter to $|\theta| \leq 10^{-10}$ [8].
Quarks and gluons are not observed as free particles, but only combined in colourless particles, called hadrons. This confinement is an effect of the running of the α_s coupling constant discussed below. Most of the QCD predictions exploited in collider physics rely on perturbative QCD. Observables are expressed as power series of the coupling constant $\alpha_\mathrm{s}(\mu^2_\mathrm{R})$, function of an unphysical renormalisation scale μ_R. If the scale μ_R is computed at the momentum Q transferred during the interaction, α_s(Q) is an estimate of the strength of the QCD interaction at that scale. The coupling itself is not a physical observable, but rather a parameter defined in the context of the perturbation theory which affects the prediction of physical quantities, such as cross sections. Multiple experiments measured α_s at different energy scale exploiting a variety of processes: hadronic τ leptons decay, heavy quarkonia resonances, deep inelastic scattering of electron on protons, electron collisions and proton-proton collisions. Lattice QCD computations are also exploited to derive the value of α_s. The present world average at a scale equal to the Z boson mass m_Z is [7]:

$$\alpha_\mathrm{s}\left(m_\mathrm{Z}\right) = 0.1181 \pm 0.0011. \tag{1.4}$$

Figure 1.1 shows the value of α_s as a function of the energy scale Q. The dependence on the energy scale of QCD predictions is exactly cancelled if the perturbation series is computed up to infinite order. However, for real world application, a dependence on the arbitrarily chosen scale affects all the QCD predictions. This effect is kept into account by an uncertainty on the prediction which is typically assigned by varying the renormalisation

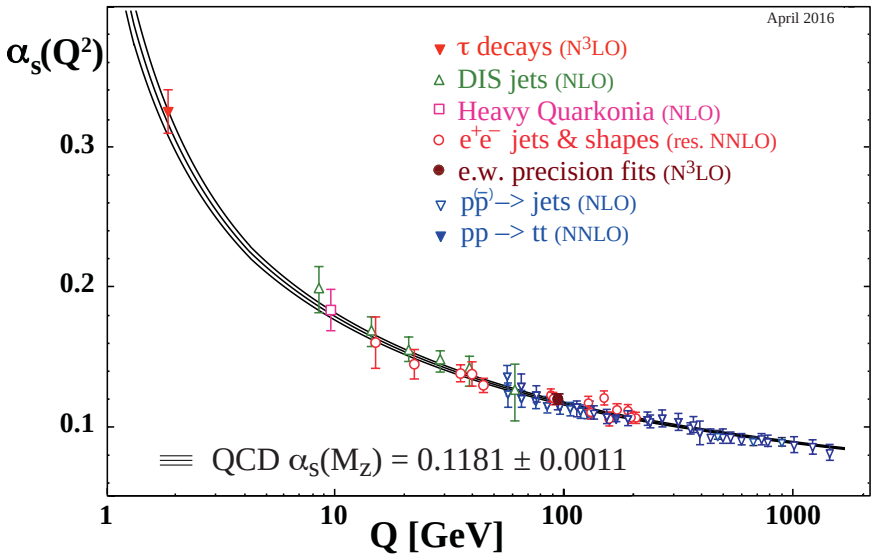


Figure 1.1: Dependence of the strong coupling constant α_s from the energy scale Q [7]. Points represent different experimental measurements. The degree of perturbation theory used to extract the value of α_s is indicated in brackets in the caption.

scale up and down by a factor of 2. Two notable features of the α_s scale dependence are the asymptotic freedom, which causes the interaction to vanish at extremely high transferred momentum, and the divergence at low scale, which is responsible for the non-perturbative behaviour of QCD at low energy; the energy scale Λ at which the QCD is no longer perturbative is $\mathcal{O}(200$ MeV). The divergence of α_s at low scale causes the confinement of quarks and gluons: the five lighter quarks hardonise in a timescale which is of order $\approx 1/\Lambda$, while the top quark decays before it can hadronise, due to its considerable mass. Predictions from perturbative QCD concern interaction of quarks and gluons; to apply the QCD domain to collider physics, the partonic structure of protons should be determined. Probability density functions (PDFs) describe the inner structure of the proton in terms of probability of interacting with a quark of a given flavour (or gluon) within the proton structure. The PDFs are non-perturbative and are derived from experimental data, using e, μ and ν collision on protons. The cross section for a hadronic collision $h_1\ h_2 \to X$ can thus be written as:

$$\sigma_{h_1 h_2 \to X} = \sum_{n=0}^{\infty} \alpha_s^n \left(\mu_R^2\right) \sum_{i,j} \int dx_1 dx_2 f_{i/h_1}\left(x_1, \mu_F^2\right) f_{j/h_2}\left(x_2, \mu_F^2\right) \hat{\sigma}_{ij \to X}\left(x_1 x_2 s, \mu_R^2, \mu_F^2\right), \tag{1.5}$$

where $f_{i/h}(x)$ is the PDF, the probability to find a quark (or a gluon) of type i which carries a fraction x of the longitudinal momentum of the hadron inside the hadron h. The $\hat{\sigma}_{ij \to X}\left(x_1 x_2 s, \mu_R^2, \mu_F^2\right)$ is the partonic cross section as from perturbative QCD.

A factorisation scale μ_{F} appears: a quark before interaction can emit a gluon which modify the quark momentum; most of the gluon emission will be soft and collinear. The factorisation scale μ_{F} can be interpreted as the minimum transverse momentum for a gluon to be included in the partonic cross section; gluons with transverse momentum below μ_{F} are kept into account by the PDF $f_{i/\mathrm{h}}(x)$. The dependence of the PDFs from the factorisation scale is known from the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) [9–12] equation which at leading order (LO) is:

$$\mu_{\mathrm{F}}^2 \frac{\partial f_{i/\mathrm{h}}\left(x, \mu_{\mathrm{F}}^2\right)}{\partial \mu_{\mathrm{F}}^2} = \sum_j \frac{\alpha_{\mathrm{s}}\left(\mu_{\mathrm{F}}^2\right)}{2\pi} \int_x^1 \frac{dz}{z} P_{i \leftarrow j}^{(1)}(z) f_{i/\mathrm{h}}\left(\frac{x}{z}, \mu_{\mathrm{F}}^2\right), \tag{1.6}$$

where $P_{i \leftarrow j}^{(1)}(z)$ are the LO splitting functions, while z is the ratio between the momentum before and after the gluon emission from a quark. The choice of the factorisation scale, as well as of the renormalisation one, is arbitrary, but when adding infinite orders in perturbation theory, any dependence is cancelled out. Similarly to the renormalisation scale, the QCD prediction are affected by an uncertainty due to the choice of the factorisation scale, which is customarily estimated from varying the scale up and down by a factor two.
The final ingredients to obtain a QCD prediction suitable for collider physics are parton showering and hadronisation. Perturbative QCD can not describe phenomena at low energy, therefore the parton-to-hadron transition is addressed differently. Fragmentation-functions are the final-state equivalent of PDF to address the non-perturbative modelling of the final state.
The simulation of a collision with two protons interacting is therefore a complex task, with multiples steps. Full QCD prediction starts from the random generation of the hard scattering process, according to the computed cross section convolved with the PDFs. The second step is the parton showering, based on the successive emission of gluons and gluon splitting into quarks. The showering is stopped at some energy scale, usually around 1 GeV. At this point, hadronisation take place and quarks and gluons are merged into final state hadrons. If the collision originated from protons, the final ingredient is the description of the interaction of the partons which did not take part in the hard scattering, generally referred to as underlying event. It is usually handled as an additional scattering and it is tuned to match the experimental observations [13, 14].
The uncertainty on a QCD prediction based on a perturbative calculation to a fixed order is expected to be of the order of the leading neglected term. It is usually estimated from varying independently the renormalisation and factorisation scale. An additional uncertainty, of order Λ/Q, arises due to non-perturbative effects of the theory. It is estimated from the difference in the observables at partonic level and after hadronisation.

1.1.2 The EW Lagrangian

The EW interaction corresponds to the SU(2) × U(1) part of the Lagrangian. It features three gauge bosons W_μ^i $(i = 1, 2, 3)$ for the SU(2) part, and a boson B_μ for the U(1), with couplings g and g', respectively. The EW couplings are universal as they do not depend on the fermion. Left-handed fermions transform as doublets of the SU(2), while right-handed fermions are singlets:

$$\Psi_1 = \begin{pmatrix} \mathrm{u} \\ \mathrm{d}' \end{pmatrix}_{\mathrm{L}}, \quad \Psi_2 = \mathrm{u}_{\mathrm{R}}, \quad \Psi_3 = \mathrm{d}_{\mathrm{R}}. \tag{1.7}$$

Here u stands for any up-type quark, while d' are the down-type quarks after the Cabibbo-Kobayashi-Maskawa (CKM) mixing:

$$\mathrm{d}'_i = \sum_j V_{ij}\ \mathrm{d}_j. \tag{1.8}$$

The sum runs on all the down-type quarks, while V is the CKM matrix [15, 16]. The mixing implies that mass eigenstates are not eigenstates under the SU(2) group. The CKM matrix is almost diagonal, thus interaction of quarks of the same generation are favoured. Similarly, left-handed leptons and neutrinos form SU(2) doublets, while the right-handed counterparts are singlets:

$$\Psi_1 = \begin{pmatrix} \nu_\ell \\ \ell \end{pmatrix}_{\mathrm{L}}, \quad \Psi_2 = \ell_{\mathrm{R}}, \quad \Psi_3 = \nu_{\mathrm{R}}. \tag{1.9}$$

A weak isospin T is defined to identify the members of the SU(2) doublet, with its third components $T_3 = \pm 1/2$ that is conserved by EW interactions. The positive eigenstate is associated with up-type quarks and neutrinos, while the negative one with down-type quarks and leptons. For right-handed fermions holds $T = 0$. The Lagrangian describing EW phenomena is:

$$\begin{aligned} \mathcal{L}_0^{\mathrm{EW}} &= \sum_i \bar{\Psi}_i i\gamma^\mu \partial_\mu \Psi_i \\ &- \frac{g}{2\sqrt{2}} \sum_i \bar{\Psi}_i \gamma^\mu \left(1 - \gamma^5\right) \left(T^+ W_\mu^+ + T^- W_\mu^-\right) \\ &- e \sum_i Q_i \bar{\Psi}_i \gamma^\mu \Psi_i A_\mu \\ &- \frac{g}{2\cos\theta_W} \sum_i \bar{\Psi}_i \gamma^\mu \left(g_V^i - g_A^i \gamma^5\right) \Psi_i Z_\mu \\ &- \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W_{\mu\nu}^i W_i^{\mu\nu}. \end{aligned} \tag{1.10}$$

The first line is the kinetic term, the sum running on all the fermions. The second line represents the charged currents, where $W^\pm$ are the physical W boson fields, linked to the W^i of the SU(2) group as $W^\pm = \left(W^1 \mp W^2\right)/\sqrt{2}$. The operators $T^\pm$ are the weak isospin raising and lowering operators. The third line models the electromagnetic interaction, where $e = g \sin\theta_{\mathrm{W}}$ is the positron electric charge, θ_{W} is the Weinberg angle defined below, Q_i is the electric charge of the fermion and A is the photon field. The fourth line is the weak neutral current mediated by the Z boson, where g_V^i and g_A^i are the vector and axial coupling for the i-th fermion:

$$g_V^i = T_3^i - 2Q^i \sin^2\theta_W, \qquad g_A^i = T_3^i. \tag{1.11}$$

The photon and Z boson fields are linear combinations of the W^3 and B fields, the mixing depending on the Weinberg angle $\theta_{\mathrm{W}} = \tan^{-1}\left(g'/g\right)$:

$$\begin{pmatrix} W^3 \\ B \end{pmatrix} = \begin{pmatrix} \cos\theta_{\mathrm{W}} & \sin\theta_{\mathrm{W}} \\ -\sin\theta_{\mathrm{W}} & \cos\theta_{\mathrm{W}} \end{pmatrix} \begin{pmatrix} Z \\ A \end{pmatrix}. \tag{1.12}$$

The present world average for the measurement of θ_{W} is [7]:

$$\sin^2\theta_{\mathrm{W}} = 0.22343 \pm 0.00007. \tag{1.13}$$

Finally, the last line of Eq. 1.10 encodes the field term for the W^i and B bosons, analogue to the one of Eq. 1.3, when the form factors of the SU(2) and U(1) groups are exploited. The Lagrangian of Eq. 1.10 does not accommodate a mass term for bosons or fermions, as a term of the form $m\bar{\Psi}\Psi$ would explicitly break the SU(2) symmetry, mixing of right and left-handed fermions. The problem of a massless theory is overcome by the mechanism of spontaneous symmetry breaking [17–22]. The Lagrangian is extended with a potential of the form:

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda\left(\phi^\dagger\phi\right)^2, \qquad \phi = \begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix}, \tag{1.14}$$

where ϕ is the Higgs scalar doublet and $\phi^{+,0}$ are complex fields. For $\mu^2 < 0$, ϕ develops a vacuum expectation value $v/\sqrt{2} = \mu/\sqrt{2\lambda}$, thus an infinite number of degenerate states present minimum energy. When a ground state is chosen as:

$$V(\phi) = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v + H(x)\end{pmatrix}, \tag{1.15}$$

part of the EW symmetry group is spontaneously broken while the $\mathrm{U(1)_{QED}}$ groups remains unbroken, to preserve the photon massless. When the potential of Eq. 1.14 is added to the Lagrangian, by requiring V to be locally gauge invariant and after choosing the vacuum state as in Eq. 1.15, the Lagrangian acquires additional terms:

$$\begin{aligned}\mathcal{L}_1^{\mathrm{EW}} = \mathcal{L}_0^{\mathrm{EW}} &+ \frac{1}{2}\partial_\mu H\partial^\mu H + \frac{1}{4}\lambda v^4 - \frac{1}{2}m_{\mathrm{H}}^2H^2 - \frac{m_H^2}{2v}H^3 - \frac{m_H^2}{8v^2}H^4\\ &+ m_{\mathrm{W}}^2W_\mu^+W^{-\mu}\left(1 + \frac{2}{v}H + \frac{H^2}{v^2}\right) + m_{\mathrm{Z}}^2Z_\mu^+Z^{-\mu}\left(1 + \frac{2}{v}H + \frac{H^2}{v^2}\right).\end{aligned} \tag{1.16}$$

The first line describes the behaviour of the Higgs field H. A new particle is originated by the additional degrees of freedom included in the Lagrangian. Three more massless Goldstone bosons [23] are originated by the Higgs doublet, whose excitation are unphysical thanks to the gauge symmetry, and thus do not enter the Lagrangian. A quadratic term giving mass m_{H} to the Higgs boson appears, together with a cubic and quartic term which are responsible for the Higgs boson self-interaction. The second line of Eq. 1.16 describes the interaction of gauge bosons and the Higgs boson. A mass term for the W and Z bosons is included, providing mass m_{W} and m_{Z}, respectively, while fermions are still massless. Boson masses, at tree level, can be expressed as:

$$\begin{aligned}m_{\mathrm{H}} &= \sqrt{2\lambda}v,\\ m_{\mathrm{W}} &= \frac{1}{2}gv = \frac{ev}{2\sin\theta_{\mathrm{W}}},\\ m_{\mathrm{Z}} &= \frac{1}{2}v\sqrt{g^2+g'^2} = \frac{ev}{2\sin\theta_{\mathrm{W}}\cos\theta_{\mathrm{W}}} = \frac{M_{\mathrm{W}}}{\cos\theta_{\mathrm{W}}},\\ m_\gamma &= 0.\end{aligned} \tag{1.17}$$

The final ingredient to obtain massive fermions is the addition of the Yukawa interaction to the Lagrangian:

$$\mathcal{L}^{\mathrm{EW}} = \mathcal{L}_1^{\mathrm{EW}} - \sum_i \bar{\Psi}_i \left(v + H\right) y_i \Psi_i, \tag{1.18}$$

where the y_i are the Yukawa coupling of the i-th fermion, related to its mass as:

$$y_i = \frac{\sqrt{2} m_i}{v}. \tag{1.19}$$

The Yukawa interaction, finally, provides mass terms for the fermions as well a new interaction between the Higgs boson and the fermion, proportional to fermion masses, without explicitly breaking the gauge symmetry. Neutrinos in the SM are explicitly massless.

1.1.3 The SM free parameters

The SM Lagrangian, given by the sum of EW and QCD terms as in Eq. 1.1, has 19 free parameters to be determined experimentally:

- three parameters for the SU(2) × U(1) gauge sector:
  - the EW couplings g and g';
  - the Higgs boson vacuum expectation value v;
- two parameters for the SU(3) gauge sector:
  - the strong coupling g_{s};
  - the CP violating term coefficient of QCD θ;
- four parameters for the CMK matrix (3 mixing angles and 1 phase);
- one parameter for the Higgs boson potential (λ or the Higgs boson mass);
- nine parameters for the Yukawa interaction.

To accommodate non-zero mass neutrinos, 3 more Yukawa couplings (or masses) and 4 parameters to describe neutrino oscillations with a mechanism analogue to the CKM are needed.
There is some freedom in the choice of the parameters, as an example the EW couplings can be replaced by the Weinberg angle and the electric charge of the positron. Generally the parameters with the lowest experimental uncertainty are exploited as input to the SM: the Z boson mass, the Fermi constant G_{F} and the fine structure constant $\alpha = e^2/4\pi$ substitute the EW couplings and v. Interestingly enough, most of the free parameters of the SM comes from the Higgs boson sector; it's not difficult to identify the Higgs boson as the less known (both experimentally and theoretically) part of the SM.
Once the parameters are defined, the SM predicts the cross sections, the decay amplitudes and several other observables of the interaction processes described by its Lagrangian and involving the particles included in the model.
A full and extensive review of the SM, with a detailed discussion of the experimental measurement of the free parameters and the success of its predictions can be found in Ref. [7].

1.1.4 The success of the SM

The SM has survived with a spectacular precision to several experimental tests over the last 50 years. The self-consistency of the SM has been intensively verified with impressive precision over several orders of magnitude without any significant failure in its predictions. Figures 1.2 and 1.3 show a comparison between the SM predictions and experimental data from the ATLAS and CMS Collaborations, respectively, in proton-proton collisions at centre-of-mass energy of 7, 8 and 13 TeV. Multiple processes are investigated, involving several phenomena predicted by the SM Lagrangian and all of them are in agreement with the theory predictions. The spontaneous symmetry breaking mechanism, a feature key of the SM, has been proved by the experimental observation of the Higgs boson in 2012 by the ATLAS and CMS Collaborations [5,6] providing an additional confirmation of the SM consistency and predicting capability.

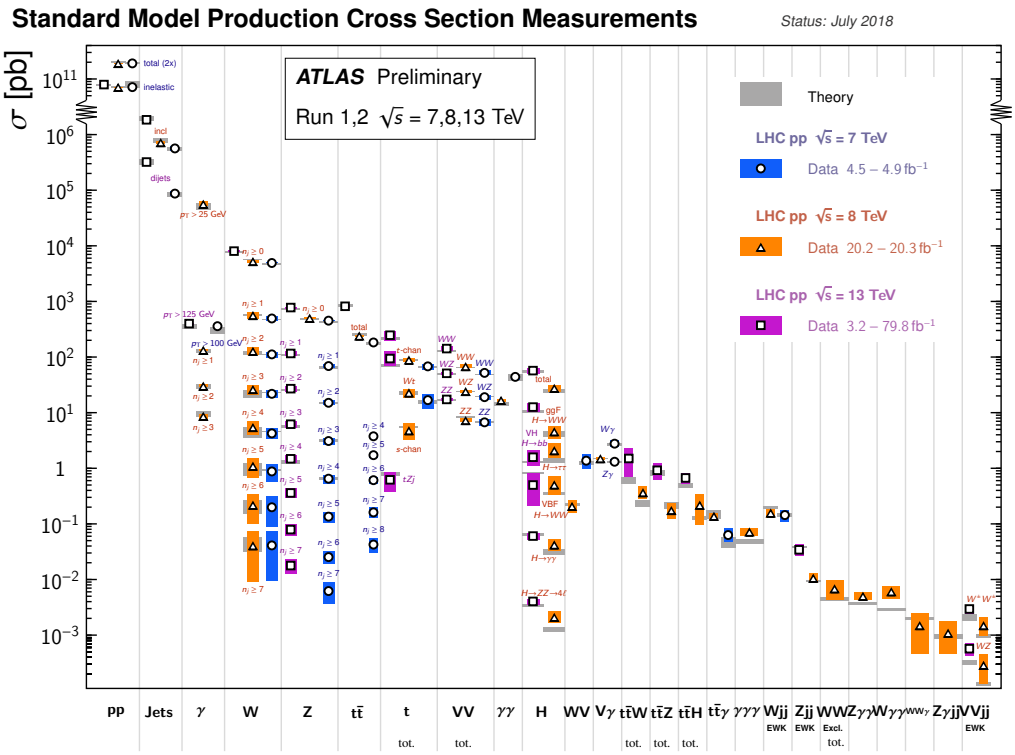


Figure 1.2: Comparison between the theoretical predictions of the SM and the experimental data collected by the ATLAS Collaboration in different processes.

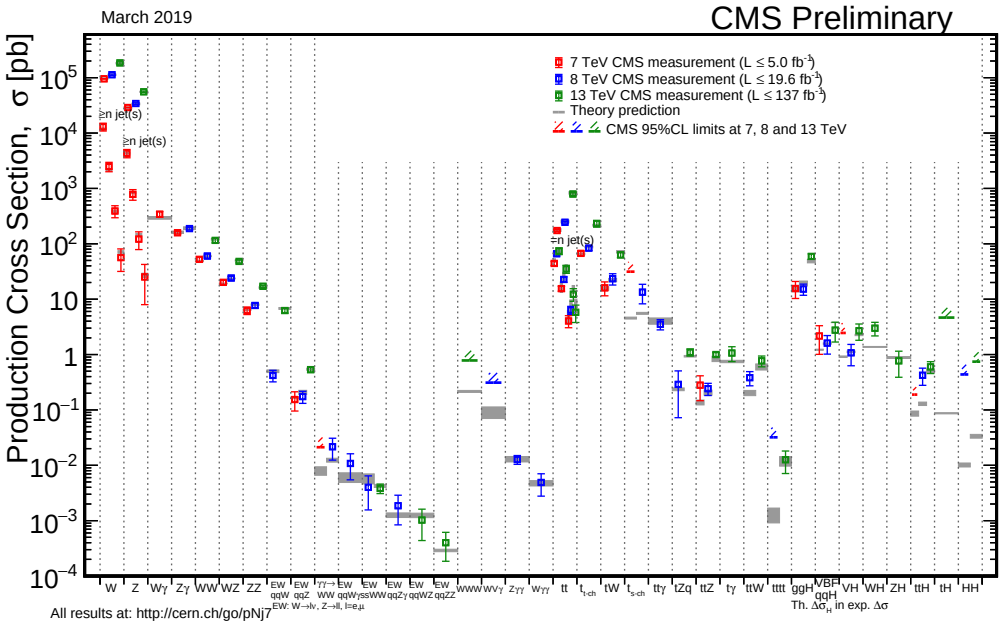


Figure 1.3: Comparison between the theoretical predictions of the SM and the experimental data collected by the CMS Collaborations in different processes.

1.2 Need for further tests of the SM

The SM proved to be a self-consistent theory up to unprecedented degree of precision. At present, no significant failure in any of the SM prediction has been observed in collider experiments. After its recent discovery, also the Higgs boson sector is coherent with the SM expectations, within the experimental uncertainties inevitably large at this stage.
Yet not all the the experimental data can be described by the SM Lagrangian. As examples, the asymmetry between matter and antimatter predicted by the SM is not enough to justify the existence of the universe, the gravity can not be included in the SM Lagrangian and neutrino masses and oscillations are not described by the SM.
An additional problem of the SM is the lack of the so-called 'naturalness' [24]. A 'natural' theory would have only few free parameters of the same order of magnitude. Instead, the SM features 19 free parameters to be determined experimentally (26 to introduce neutrino masses) spread over several orders of magnitude.
In the Higgs boson sector, the naturalness involves the value of the Higgs boson mass (m_{H}). The value of m_{H} at tree level is given by Eq. 1.17 and it is function of λ and v. When higher order in perturbation theory are added, the Higgs boson mass receives quantum corrections proportional to energy scale up to which the SM is assumed to be valid, which could be as high as the Plank scale. If the cut-off is at more than 500 GeV the corrections are bigger than the value of the mass itself [25], implying a large fine-tuning of the parameters of the theory to match the experimental observation of m_{H}, causing in an 'un-natural' behaviour of the theory parameters.
Based of the naturalness argument, two broad classes of models have been proposed as possible extensions of the SM. Supersymmetric models [26–30] solve the naturalness of the Higgs boson sector by adding a class of particles symmetric to the known ones which introduce an automatic cancellation of the fine-tuning required in the SM. The composite

Higgs boson models [31–33] describe the Higgs boson as a bound state of a new interaction and its mass is not a parameter of the theory but emerges with a mechanism analogue to the proton mass. Those models feature desirable naturalness of the theory parameters and are capable to describe the full SM phenomenology, but no experimental observation has ever substantiated any of them.
Experimentally two viable ways to investigate the SM consistency and to test its possible extensions can be identified. The first one is the direct search of unpredicted phenomena, such as high-mass particles or new sources of CP violation. The second one, which is the research path followed in this thesis, is to improve the precision in measuring quantities predicted by the SM, to search for small deviations induced by quantum corrections due to unobserved particles.
The Higgs boson sector offers a unique possibility to pursue precision tests of the SM as the Higgs boson is responsible for more than a half of the free parameters of the SM. In particular, the study of the coupling of the Higgs boson to the SM particles offer the possibility to investigate both the mechanism of the spontaneous symmetry breaking and the nature of the Yukawa interaction, by measuring the Yukawa couplings of the fermions. Following its relative recent discovery, the Higgs boson have been intensively investigated, but the uncertainty on different measurements is large enough to accommodate both the SM and several of its extensions. The complete characterisation of the Higgs boson properties is a major goal of the LHC experimental program to assert the consistency of this unexplored sector of the SM.

1.3 Higgs boson properties

An overview of the present experimental knowledge of the Higgs boson is presented in this section. At the beginning of this work (Autumn 2016), the most up-to-date result of the ATLAS and CMS Collaborations was Ref. [34], relative to the analysis of the LCH Run I data, collected by the experiments from 2010 to 2012 in proton-proton collisions at centre-of-mass energy of 7 and 8 TeV. At the time of writing (Summer 2019), the analysis of the LHC Run II data, collected in the period 2016-2018 at a centre-of-mass energy of 13 TeV, is ongoing. A much higher accuracy will be reached compared to Ref. [34], thanks to the higher centre-of-mass energy and to the larger amount of data. The measurements reported in this section generally refers to Ref. [34]; whenever a single experiment has already reached a precision higher than the Run I combination by analysing a part of the Run II dataset, the result is quoted as well. A notable exception is the discussion about the couplings: updated results with Run II data are already available from single experiments [35, 36]. Since this work is one of the input used for the measurement of the Higgs boson couplings, the up-to-date knowledge of the couplings will be discussed in detail later in this document.

1.3.1 Couplings of the Higgs boson

The Higgs boson mass and the vacuum expectation value of the Higgs boson field are unbound by the theory and must be determined experimentally, since they input the SM predictions of the Higgs boson couplings. The value of v as tree level is linked to the Fermi constant by the relation $v = \left(\sqrt{2}G_{\mathrm{F}}\right)^{1/2} \approx 246$ GeV. The Fermi constant is measured with a precision of 0.6 part per million from muon decays by the MuLan experiment [37].

The Higgs boson mass $m_{\mathrm{H}} \approx 125$ GeV has been measured by the ATLAS and CMS Collaborations [38]. More details on the mass measurement are given below.
The Higgs boson does not couple universally to fundamental particles, but according to their masses, establishing a new kind of interaction. The coupling to fermions is linear with their masses, while it is quadratic for the bosons. The dominant mechanisms of production and decay thus involve couplings with Z or W bosons or third generation fermions. The coupling to gluons and photons at tree level is zero, given their massless nature. The interaction of the Higgs boson with gluons proceeds through virtual contributions (loops) dominated by top quarks; similarly the interaction with photons happens thanks to a loop of W bosons, with minor contributions from the top quarks. The Higgs boson couplings are fixed once the mass of the particles and the vacuum expectation value are known: the measurement of the Higgs boson couplings represents a powerful tool to ascertain the consistency of the SM. Figure 1.4 summarises the status of the measurements of the Higgs boson couplings as a function of the particle masses at the beginning of this work; all the results are compatible, within the experimental uncertainties, with the SM expectations.

1.3.2 Production of the Higgs boson at the LHC

At the LHC, a Higgs boson can be produced through four main production mechanisms, with three additional subdominant processes. The Feynman diagrams of the seven processes are shown in Fig. 1.5, while Fig. 1.6 (left) displays the predicted cross sections for the various production modes as a function of the centre-of-mass energy of the colliding protons. For sake of clarity, the same information is provided in Table 1.3 for the centre-of-mass energies relevant for the LHC operations.
The gluon-fusion (ggH) process (Fig. 1.5a) has the largest cross section, of about 50 pb, at a centre-of-mass energy of 13 TeV. It proceeds through a loop mediated by the exchange of virtual top quarks; contribution from other quarks are suppressed proportional to the mass of the quark squared. The second leading process is the vector boson fusion (VBF), $\mathrm{qq} \to \mathrm{qqH}$, with a cross section of about 4 pb at centre-of-mass energy of 13 TeV. The process (Fig. 1.5b) involves the scattering of two quarks exchanging Z or W bosons, the Higgs boson being radiated from the virtual propagator. The final state can be identified by the presence of two jets with notable angular separation in addition to the Higgs boson. Once proper selections are applied, the VBF production offers a relative clean experimental environment. The associated production with vector a boson (VH, V being either the W or the Z boson) is the third leading process (Fig. 1.5c), with a cross section of about 2.3 pb (1.4 for WH and 0.9 pb for ZH) at centre-of-mass energy of 13 TeV. The VH process provides direct access to the Higgs boson coupling to vector bosons. Finally the $\mathrm{t\bar{t}H}$ process, $\mathrm{gg} \to \mathrm{t\bar{t}H}$ (Fig. 1.5d), has a cross section of about 0.5 pb at centre-of-mass energy of 13 TeV. The $\mathrm{t\bar{t}H}$ process directly probes y_{t} and it presents a complex final state according to the top quarks pair decay.
Among the subdominant processes, the $\mathrm{b\bar{b}H}$ production (Fig. 1.5d) is the analogue of $\mathrm{t\bar{t}H}$ with b-type quarks. Despite the cross section comparable to $\mathrm{t\bar{t}H}$, $\mathrm{b\bar{b}H}$ offers a final state which is difficult to identify in the LHC. The single-top associated production (Fig. 1.5e), tHq, has a cross section about one order of magnitude lower than the $\mathrm{t\bar{t}H}$, but can can be strongly enhanced in case of a CP violating coupling to the top quark [39]. Finally the tHW process (Fig. 1.5f) is an extremely rare process, out of reach for the

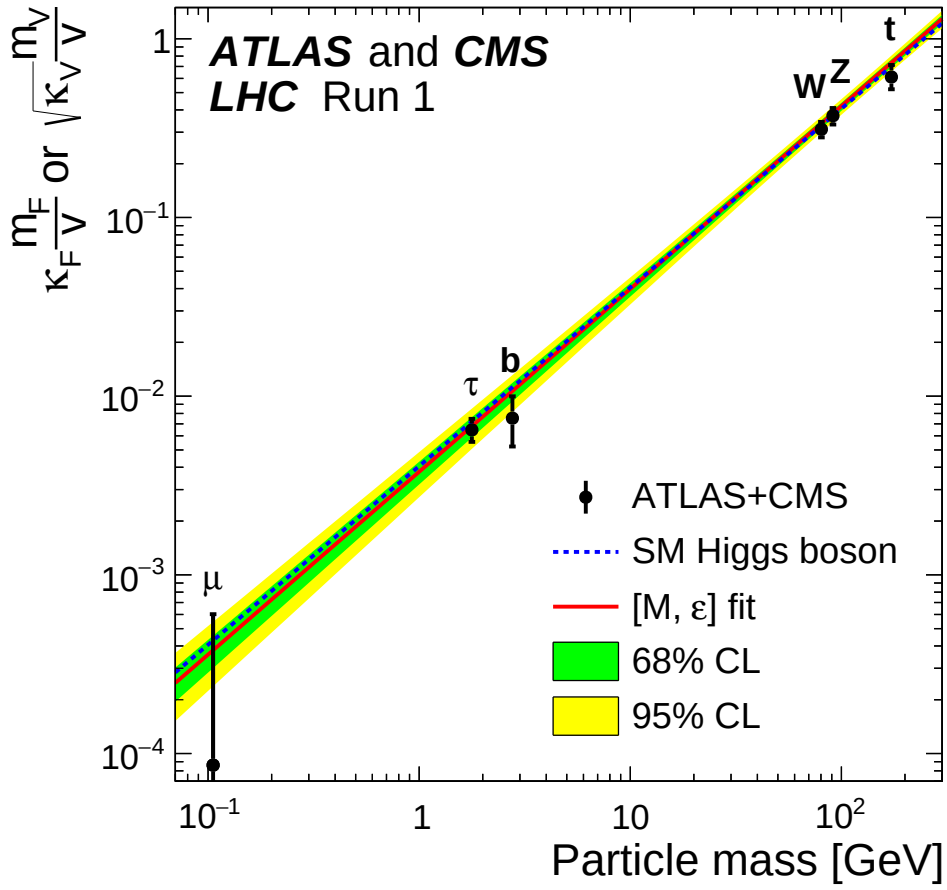


Figure 1.4: Best-fit values of the Higgs boson couplings as a function of the particle mass. The y axis reports $\kappa_{\mathrm{F}} \cdot m_{\mathrm{F}}/v$ for fermions and $\sqrt{\kappa_{\mathrm{V}}} \cdot m_{\mathrm{V}}/v$ for bosons, where m is the mass of the particle, v is the Higgs field vacuum expectation value and κ is the coupling modifier of the κ-framework defined in the text. The dotted line shows the SM prediction. All results are compatible with the SM within the experimental uncertainties.

$\sqrt{s}$ (TeV)	Production cross section (pb) for m_{H}=125 GeV					
	ggH	VBF	WH	ZH	$\mathrm{t\bar{t}H}$	Total
7	$16.9^{+5\%}_{-5\%}$	$1.24^{+2\%}_{-2\%}$	$0.58^{+3\%}_{-3\%}$	$0.34^{+4\%}_{-4\%}$	$0.09^{+8\%}_{-14\%}$	19.1
8	$21.4^{+5\%}_{-5\%}$	$1.60^{+2\%}_{-2\%}$	$0.70^{+3\%}_{-3\%}$	$0.42^{+5\%}_{-5\%}$	$0.13^{+8\%}_{-13\%}$	24.2
13	$48.6^{+5\%}_{-5\%}$	$3.78^{+2\%}_{-2\%}$	$1.37^{+2\%}_{-2\%}$	$0.88^{+5\%}_{-5\%}$	$0.50^{+9\%}_{-13\%}$	55.1
14	$54.7^{+5\%}_{-5\%}$	$4.28^{+2\%}_{-2\%}$	$1.51^{+2\%}_{-2\%}$	$0.99^{+5\%}_{-5\%}$	$0.60^{+9\%}_{-13\%}$	62.1

Table 1.3: Predicted cross sections for the most relevant Higgs boson processes, in proton-proton collisions at centre-of-mass energies corresponding to the LHC past and foreseen operations. The uncertainty on the prediction is also indicated. The last column reports the total cross section for the production of a Higgs boson. Values are taken from Ref. [40].

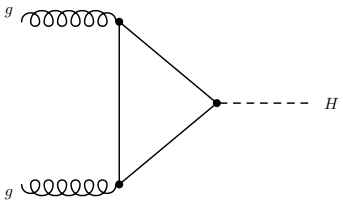
present sensitivity of the LHC experiments.

1.3.3 Decay channels of the Higgs boson

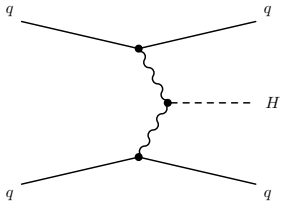
A Higgs boson with a mass of about 125 GeV features decays to multiple SM particles, allowing a deep and complete exploration of its couplings both with bosons and fermions. Since the coupling is proportional to the mass, direct decays to the most massive particles are favoured, if not limited by phase-space constraints. The branching ratios $\mathcal{B}$ of the different decay channels are reported in Fig. 1.6 (right) for a Higgs boson with mass around 125 GeV, while Table 1.4 reports the branching ratios of the most relevant channels for the LHC experiments. Figure 1.7 shows the leading order Feynman diagrams for the decays to fermions and bosons (1.7a), as well as the loops involved in the decay to a pair of photons (1.7b). The decay to a pair of b-type quarks is the most probable decay, with $\mathcal{B} = 58\%$. The second leading decay is $\mathrm{H} \to \mathrm{WW}^*$, with $\mathcal{B} = 21\%$ followed by the decay to gluons, forbidden at tree level, with $\mathcal{B} = 9\%$. About 6% of the decays is in pairs of τ leptons and about 2% in pairs of c-type quarks. The decay to pairs of Z bosons has a branching fraction of about 2% and, in the final state with four leptons following the decay of the Z bosons, is among the most sensitive final states, despite the requirement of a fully leptonic final state lowers the branching fraction by about a factor 200. The diphoton channel, with $\mathcal{B} = 0.2\%$, offers a very clean environment to study the Higgs boson. Finally the decays to $\mathrm{Z}\gamma$ and to pair of muons are extremely rare decays, with very low branching ratios and they are not yet established experimentally. A detail discussion on the motivations that lead to the choice of the diphoton decay channel is provided in Section 1.5.1.

1.3.4 Experimental knowledge of the Higgs boson couplings

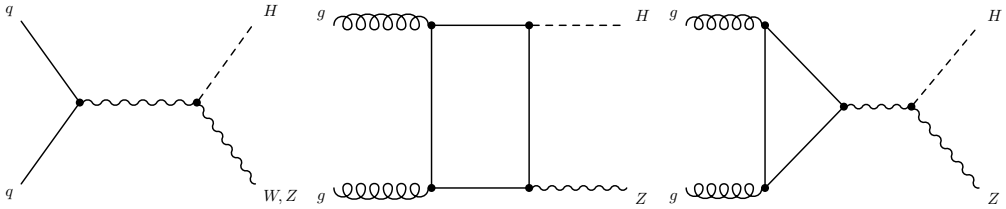
The combined study of the production and decay channels allows, at the present experimental sensitivity, the characterisation of the Higgs boson couplings to vector bosons and to third generation fermions. Four production and five decay channels have been



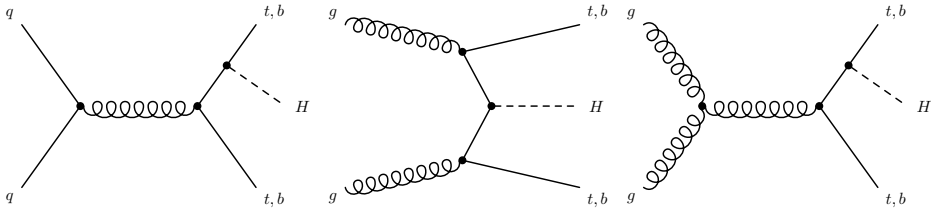
(a) Gluon fusion.



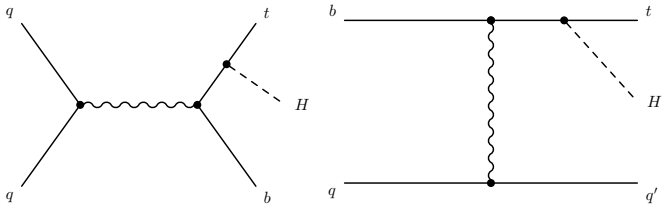
(b) Vector bosons fusion.



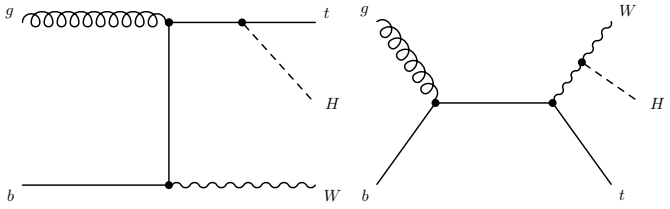
(c) Associated production with a vector boson.



(d) Associated production with a top-antitop or bottom-antibottom quark pair.



(e) Associated production a single top quark.



(f) Associated production with a single top quark and a W boson.

Figure 1.5: Leading order Feynman diagrams representing the Higgs boson production mechanisms relevant at the LHC.

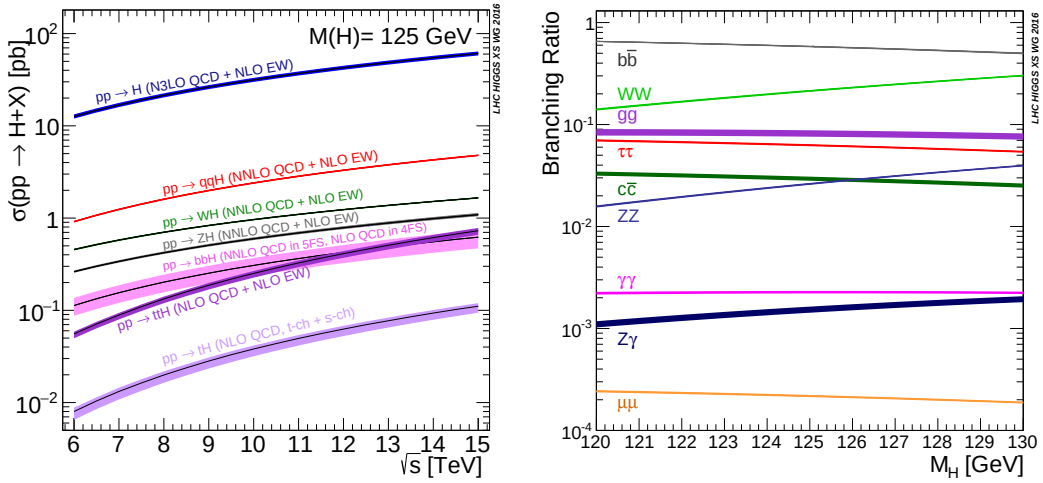
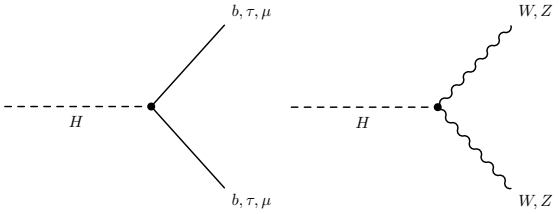


Figure 1.6: Left: cross section for the production of a Higgs boson in proton-proton collisions as a function of the centre-of-mass energy of the collision. Different colours represent different production mechanisms. The width of each line is the uncertainty on the prediction. Right: Higgs boson decay branching ratios as a function of the Higgs boson mass around m_{H}=125 GeV. Different colours represents different decay channels. The width of each line is the uncertainty on the prediction [40].

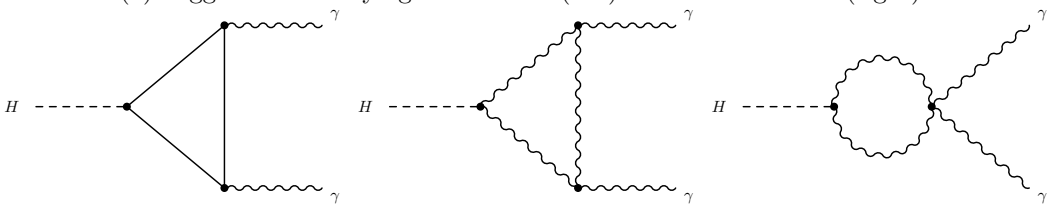
Decay channel	Branching ratio $\mathcal{B}$	Rel. uncertainty
$\mathrm{H} \to \mathrm{b\bar{b}}$	58.4%	$^{+3.2\%}_{-3.3\%}$
$\mathrm{H} \to \mathrm{W^+W^-}$	21.4%	$^{+4.3\%}_{-4.2\%}$
$\mathrm{H} \to \tau^+\tau^-$	6.27%	$^{+5.7\%}_{-5.7\%}$
$\mathrm{H} \to \mathrm{ZZ}$	2.62%	$^{+4.3\%}_{-4.1\%}$
$\mathrm{H} \to \gamma\gamma$	0.23%	$^{+5.0\%}_{-4.9\%}$
$\mathrm{H} \to \mathrm{Z}\gamma$	0.15%	$^{+9.0\%}_{-8.9\%}$
$\mathrm{H} \to \mu^+\mu^-$	0.02%	$^{+6.0\%}_{-5.9\%}$

Table 1.4: Most relevant branching ratios $\mathcal{B}$ for a Higgs boson with mass of 125 GeV. Values are taken from Ref. [40].

experimentally established; the ggH and VBF production were already observed after the Run I combination, as well as the decay to bosons (WW^*, ZZ^*, $\gamma\gamma$) and to τ leptons. The $t\bar{t}H$ process has been observed from the combination of the 2016 and the Run I data, using this work as one of the inputs, and will be discussed in greater detail later in this document. The decay to b-type quarks and the VH production have recently been observed [41, 42]. Beside the observation of a process, a precise measurement of the couplings is required for comparison with theory expectations. Multiple frameworks are adopted to provide an immediate interpretation of the results, as well as a simple exchange of information between theory and experiments. The easiest possible parametrisation is the introduction



(a) Higgs boson decaying to fermions (left) and vector bosons (right).



(b) Higgs boson decaying to photons.

Figure 1.7: Leading order Feynman diagrams representing the Higgs boson decays relevant at the LHC.

of signal strength modifiers $\mu = \sigma/\sigma_{\mathrm{SM}}$, defined as the ratio between the measured cross section and the SM prediction in a given production or decay channel. Figure 1.8 shows the signal strength modifiers for the production and decay mechanisms accessible at the LHC from the ATLAS and CMS Run I combination. A good agreement is found, with a mild deviation for the $\mathrm{t\bar{t}H}$ production of about 2 standard deviations from the SM expectation.

The κ-framework [43] provides an immediate way to access informations on the couplings. Instead of searching a deviation in production or decay modes, a modifier is assigned to each coupling: κ_{Z}, κ_{W}, κ_{t}, κ_{τ}, κ_{b}, κ_{g}, κ_{γ}. An additional parameter $\mathrm{B_{BSM}}$ accounts for invisible decays of the Higgs boson, coming from decays to invisible particle such as Beyond the Standard Model (BSM) particles. If the SM holds, $\kappa = 1$ and $\mathrm{B_{BSM}} = 0$. The coupling to photons and gluons, null at tree level in the SM, is generated radiatively by SM particles loops, so it is sensitive to particles still unobserved and involved in the loops. Finally, κ_{t} is mainly constrained from the gluon fusion and diphoton decay loops and its sign is still unconstrained. Figure 1.9 reports the best-fit value of the Run I combination for the κ-framework. Two different parameterisations are assumed: in the first one the BSM contribution is free to float and $\kappa_V \leq 1$ is required, where V is either W or Z because unitarity is imposed. Instead, in the second parameterisation the BSM is fixed to zero. A good agreement with the SM is found. More tests have been performed with this framework, for example varying all the couplings to fermions and bosons coherently or fixing all the couplings to the SM values except κ_{g} and κ_{γ} to investigate any distortion of the loops, without any significant deviation from the SM.

The Simplified Template Cross Sections (STXS) framework, described in detail in Ref. [40], aims at minimising the theoretical uncertainties, maximising the experimental sensitivity and isolating possible BSM effects. The cross section predictions for the Higgs boson are generally computed integrated over all the solid angle, regardless the direction of emission of the Higgs boson. Experimentally, the cross section can only be measured within the detector acceptance. The extrapolation from the experimental accessible volume to the full phase-space introduces large theoretical uncertainties, setting strong limits in the precision

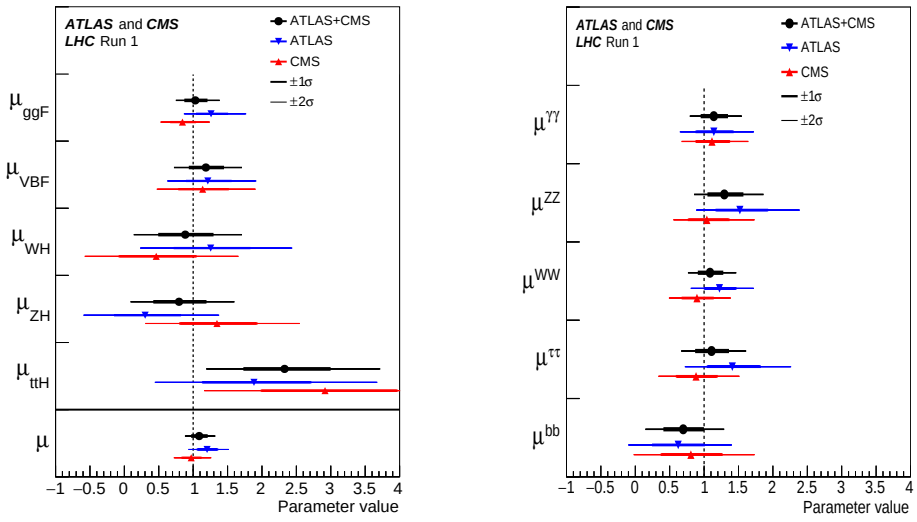


Figure 1.8: Signal strength modifiers $\mu = \sigma/\sigma_{SM}$ for the different production (left) and decay (right) mechanisms accessible at the LHC from the ATLAS and CMS Run I combination. The one and two standard deviations uncertainties are shown. The dotted line represents the SM expectation [34].

of the measurements. Within the STXS framework, a fiducial volume, close enough to the experimental acceptance, is defined. The ratio between the theory prediction and the experimental measurement of the cross section, analogue to the signal strength modifier, is computed within the fiducial volume, avoiding the extrapolation to the full solid angle. A more granular comparison is performed dividing the theory prediction in bins, defined according to the properties of the Higgs boson (and not of its decay products) and of associated objects common to all channels, such as jets. The ratio between the predicted and the observed cross section is extracted in each bin to compare the theory prediction with the experimental observation. The number of bins and their definition accounts both for the experimental sensitivity and the theoretical uncertainties. The Stage-0 STXS defines four bins, corresponding to the four production mechanisms, with an additional request on the Higgs boson rapidity $y_H \leq 2.5$. The CMS experiment has published a combined Stage-0 result in Ref. [35]. Stage-1 STXS are designed for a measurement with the full Run II data and feature additional bins for each process, defined in terms of the transverse momentum of the Higgs boson $p_T{}^H$ and number of jets. The $t\bar{t}H$ process is not split even at the Stage-1. Results form the Stage-1 STXS framework are available only for individual final states, since a combination is not yet available even from single experiments. None of the published results has shown any significative deviation from the SM expectations.

1.3.5 Higgs boson mass, width and self-coupling

The Higgs boson mass can be measured from the high resolution $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels, where the final state is completely reconstructed with excellent energy resolution. In those channels, a clear invariant mass peak arises over spurious events. The combined best-fit value from Run I is $m_H = 125.09 \pm 0.24$ GeV. Figure 1.10 summarises

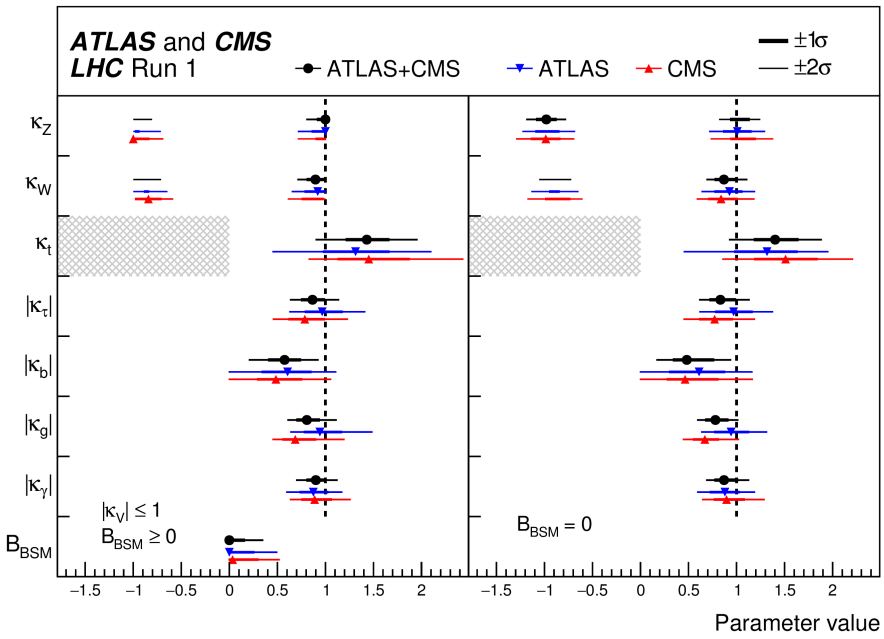


Figure 1.9: Best-fit values for the κ-framework from the ATLAS and CMS Run I combination. In the left panel the BSM contribution is free to float, while $\kappa_V \leq 1$, where V is either W or Z, because unitarity is imposed. In the right panel the BSM is fixed to zero. The best-fit value and the corresponding one and two standard deviations uncertainties are reported for each experiment, as well as for the combination. The κ_t parameter is constrained positive in the fit [34].

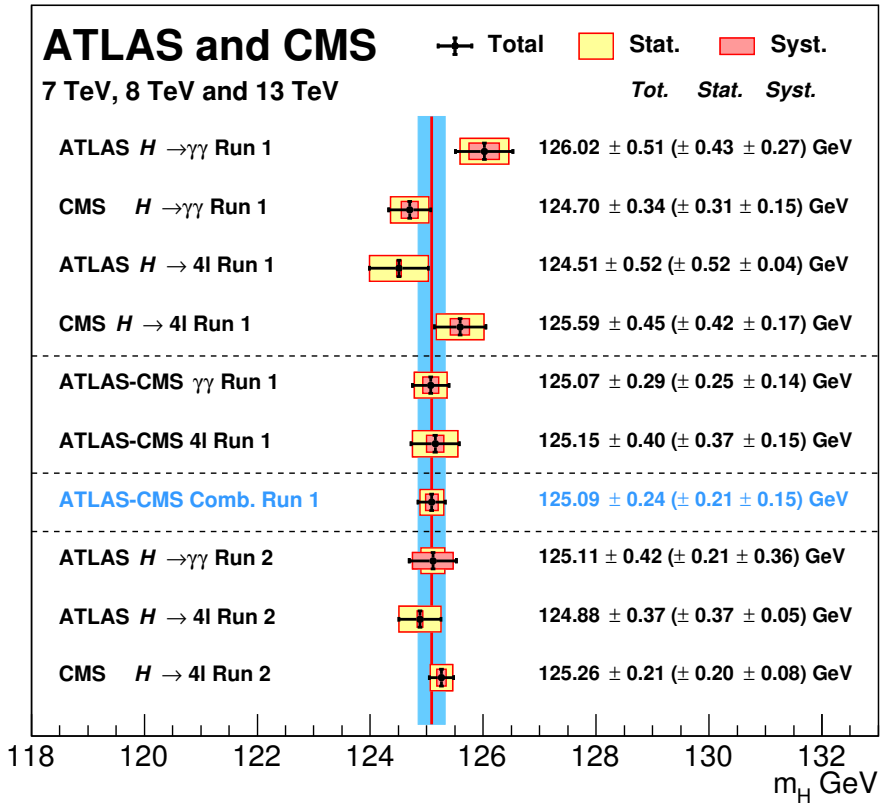


Figure 1.10: Summary of the Higgs boson mass measurement performed by the ATLAS and CMS Collaborations in the ZZ* and γγ final states. The vertical line represents the average of the Run I combination, while the azure band shows its uncertainty [7].

all the mass measurements performed by the two experiments. Figure 1.11 shows the invariant mass distribution of the photon pairs (left) and of the four leptons in the ZZ* channel (right) for two analysis on 2016 data from ATLAS and CMS experiments respectively. At present, the most precise available measurement of the Higgs boson mass is $m_{\mathrm{H}} = 125.26 \pm 0.21$ GeV [44], reaching alone a precision better than the Run I combination.

The total width of the Higgs boson is predicted in $\Gamma_{\mathrm{H}} = 4.07 \times 10^{-3}$ GeV, with about 4% of uncertainty [7], far below the experimental resolution in the invariant mass reconstruction of the Higgs boson candidates. The width can be derived from the interference between the off-shell and on-shell Higgs boson production in the four leptons final state. This methodology provides the highest sensitivity but assumes the knowledge of the Higgs boson off-shell production cross section, and thus it is not completely model independent. The Higgs boson width is constrained to $3.2^{+2.8}_{-2.2}$ MeV [46], in agreement with the SM expectations.

Once the Higgs boson mass and v are known, the values of the λ parameter of the Higgs field is completely determined. A direct measurement of λ would therefore once more test

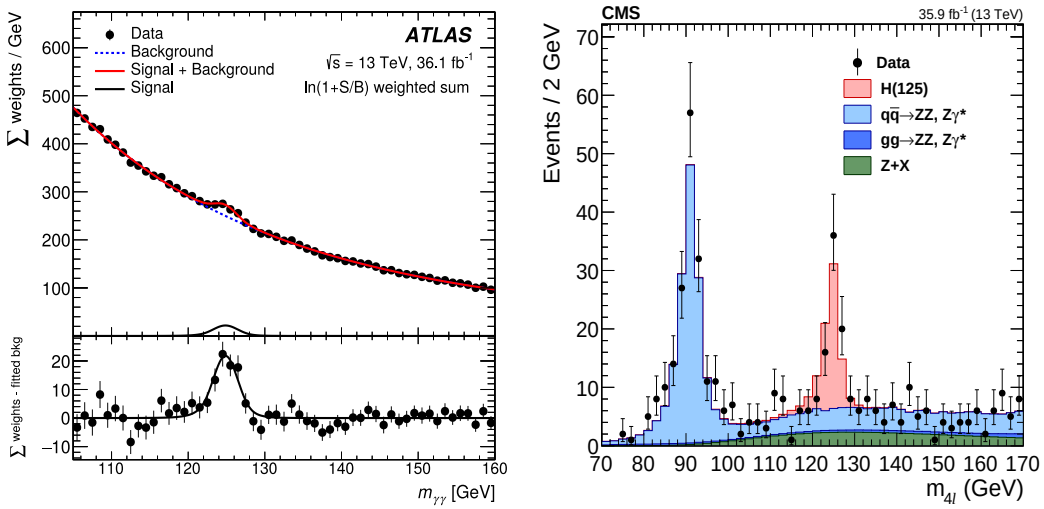


Figure 1.11: Left: Invariant mass distribution of the selected photon pairs from the ATLAS 2016 $H \to \gamma\gamma$ analysis, overlaid with the result of the fit (solid red line). Each event is weighted for the expected sensitivity. The bottom panel shows the residual after the subtraction of the non-resonant contribution [45]. Right: Invariant mass distribution of the four leptons for events selected in the CMS 2016 $H \to ZZ^* \to 4\ell$ analysis [44].

the consistency of the SM, probing the shape of the Higgs field potential. A direct test of λ requires a cubic (or quartic) Higgs boson vertex. The di-Higgs boson production, featuring the cubic vertex, is an extremely rare process, with a cross section of about 31 fb at centre-of-mass energy of 13 TeV. Present direct measurements set limits on λ and constrain this parameter to about 10 times the SM model expectations [47, 48]. The observation of the di-Higgs boson production is one of the major challenges for the LHC experimental program, and if the SM hold, the expected uncertainty on λ would be at least of 100% even after the end of the LHC experimental program. Preliminary studies suggest that a direct measurement of the quartic coupling would be out of reach even for a 100 TeV hadron collider.

1.4 The $t\bar{t}H$ production

Among the processes experimentally accessible at the LHC, the $t\bar{t}H$ production has been chosen as the object of investigation of this thesis for several reasons, summarised in this section.

At the beginning of this work, the $t\bar{t}H$ process was still an unobserved phenomenon of the SM, and establishing its existence was a first step for a consistency test for the SM. Moreover, after the LHC Run I, all the measurements of production and decay rates of the Higgs boson were in agreement with the SM expectation, except for a small tension of two standard deviations in the $t\bar{t}H$ production cross section, coherently observed by both ATLAS and CMS Collaborations (see Fig. 1.8). A more precise measurement of the $t\bar{t}H$ production rate was necessary to understand the nature of this mild departure from the SM prediction and to ascertain whether it was due to statistical fluctuations or to some

unpredicted feature. Some supersymmetric models predict heavy particle with cascade decays which, according to the mass of the heavy particles, could result in final states involving Higgs bosons, top and bottom quarks [49–52], mimicking the $t\bar{t}H$ signature and enhancing the observed $t\bar{t}H$ event rate.
The associate production of Higgs bosons and top quark(s) could be exploited to perform a stringent consistency test of the SM. The $t\bar{t}H$ and tHq processes are the only ones within the SM which allow a direct measurement of y_t, the tHq process being disfavoured by a factor of 10 in the cross section. Figures 1.4 and 1.9 present a measurement of the top quark Yukawa coupling; the experimental constrain mainly proceed through loops in the ggH process or in the diphoton decay channel. The value of y_t can be inferred from those processes under the assumption of no contribution of unobserved particles in the loops. The presence of unpredicted phenomena could affects the loops and mitigate a possible deviation of y_t from the SM prediction. The $t\bar{t}H$ process would live aside this assumption, thanks to the direct measurement of y_t. The direct constrain from $t\bar{t}H$ allows a direct comparison with the SM prediction of Eq. 1.19. The combination of the two measurements is a unique opportunity to explore the inner nature of the loops, thus supplying a strong consistency test for the whole SM.
The large value of y_t compared to all the other couplings and its value close to unity (see Eq. 1.19) may indicate that y_t or the top quark have a still-unknown special role in the mechanism of the electroweak symmetry breaking, making the direct measurement of y_t an interesting possibility to investigate the symmetry breaking mechanism.
Finally the $t\bar{t}H$ production rate is sensitive to CP-violation in the Higgs sector [53] since the production rate could be enhanced in case of CP-violating couplings. The Higgs boson could have both scalar and pseudo-scalar behaviour in presence of CP-violation with couplings modifiers κ_t and $\tilde{\kappa_t}$, respectively. The Lagrangian depicting the interaction between the Higgs boson and the top quark would therefore be modified as

$$\mathcal{L}_t = -\frac{m_t}{v}\left(\kappa_t \bar{t}t + i\tilde{\kappa_t}\bar{t}\gamma_5 t\right), \tag{1.20}$$

with a CP-violation phase given by:

$$\xi_t = \arctan\frac{\tilde{\kappa_t}}{\kappa_t}. \tag{1.21}$$

Figure 1.12 shows the ratio between the $t\bar{t}H$ cross section in case of CP violation and the SM one as a function of the two coupling modifiers κ_t and $\tilde{\kappa_t}$. The $t\bar{t}H$ production rate can be modified by up to a factor 2 in case of CP-violation and a measurement of the $t\bar{t}H$ cross section with an accuracy of about 20% would be enough to constrain $\xi_t = 0 \pm 30^o$.
A complete characterisation of the $t\bar{t}H$ process is thus a primary goal of the LHC experimental program, which should proceed through the experimental observation of the process followed by a precise measurement of its cross section exploiting the data collected in the next two decades.

1.5 The topology of $t\bar{t}H$ at the LHC

The $t\bar{t}H$ process can be experimentally identified from the presence in the final state of the decay products of the Higgs boson and of the top quark pair. The topology thus depends on the decay channel of the Higgs boson and of the top quarks. The choice of the diphoton

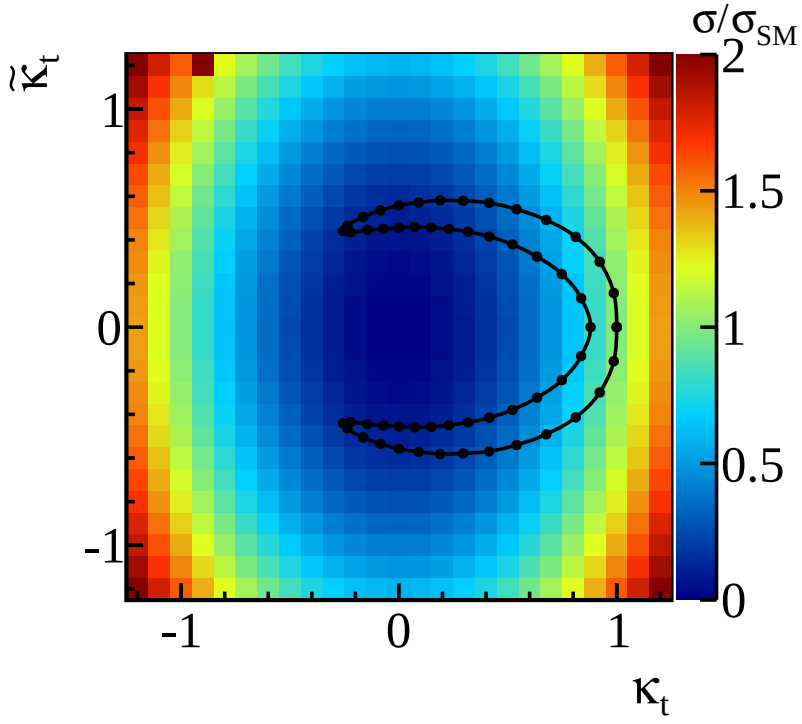


Figure 1.12: Ratio of the $t\bar{t}H$ cross section to the SM prediction as a function of the coupling modifiers κ_t and $\tilde{\kappa_t}$. The black line represents the region allowed from the Run I data. The black points are the points used for the simulation [53].

decay channel is motivated in Section 1.5.1 while the possible final states following the decay of the top quarks are illustrated in Section 1.5.2.

1.5.1 The diphoton decay channel

The study of the Higgs boson at the LHC is a real challenge: the production cross section is 10^9 times smaller than the proton-proton total cross section (see Fig. 1.2). The isolation of a Higgs boson signal has to deal with other processes mimicking the experimental signature of the Higgs boson itself, generally referred to as backgrounds. After the inelastic proton-proton interaction, the process with the highest cross section is the production of jets mediated by strong interaction, simply abbreviated as QCD or multi-jet production. Since the QCD production cross section is four orders of magnitude bigger than the Higgs boson one, the choice of final states with jets is extremely challenging, while final states with lepton are generally cleaner. Consequently, some decay channels of the Higgs boson, such as the $H \to gg$ and $H \to cc$, are overwhelmed by the background, and thus unaccessible (at least for the moment) at the LHC. Instead, some clean channels as $H \to \mu\mu$ and $H \to Z\gamma \to \ell\ell\gamma$ generally present low level of background but the low branching ratio

makes the final state difficult to study with the present amount of data.
A second element relevant to address the experimental sensitivity of a given final state is the invariant mass resolution. The invariant mass is computed from the decay products of the Higgs boson, and, for a fully reconstructed final state in an ideal detector, a peak with a width $\Gamma = \Gamma_H$ arises over a continuous background due to non-resonant events. In a real detector, given the narrow width of the Higgs boson $\Gamma_H \approx 4$ MeV, the width of the peak is driven by the experimental resolution. The mass of the Higgs boson can be inferred from the position of the peak, while its area, after properly subtracting the background, represents the signal yield. A better mass resolution allows the invariant mass peak to be narrower and thus to be more easily discernible from the background, enhancing the sensitivity, while a broad peak makes the search more difficult (see Fig. 3.1 for an example). Channels involving leptons and photons in the final state generally means good mass resolution. Topologies with neutrinos do not allow a fully reconstruction of the peak, while final states with jets generally presents poor mass resolution. In that case, *ad hoc* built variables are exploited to extract the signal yield [41, 42, 54, 55].
Table 1.4 shows the Higgs boson branching ratios for the most relevant decays at the LHC. Despite the large branching ratio, the $H \to bb$ channel has limited sensitivity for the overwhelming multi-jet background. Similarly, the $H \to \tau\tau$ channel is challenging in the final state with hadronic decays of the τ leptons, while it is easier to isolate in topologies involving leptons. The $H \to WW$ channel has a good signal-to-background ratio in the final states involving leptonic decays of the W bosons, but the presence of neutrinos does not allow a full reconstruction of the Higgs boson, reducing, in turn, the sensitivity of this channel. The $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels allow a full reconstruction of the event with an excellent invariant mass resolution and good signal-to-background ratio. The $ZZ^* \to 4\ell$ has a lower branching ratio but a better signal-to-background ratio and mass resolution compared to the $\gamma\gamma$ channel.
When combining the low $t\bar{t}H$ production cross section with the different Higgs boson decay channels, the $H \to \gamma\gamma$ turns out to be one among the most sensitive channels. The $H \to ZZ^* \to 4\ell$ has a lower background and a better mass resolution, but the branching ratio is about 10 times less than the $H \to \gamma\gamma$ channel, and, at present, the low expected signal yield makes the sensitivity of this channel subdominant with respect to the diphoton one [44]. Other final states, such as the bb and the final states involving leptons coming from $\tau\tau$, WW or $ZZ^* \to 2\ell 2$jets decays of the Higgs boson, have a sensitivity similar or slightly better than the diphoton channel but are already dominated by the systematic uncertainty [56–58]. Therefore, the $H \to \gamma\gamma$ channel provides an optimal channel to study the $t\bar{t}H$ production and it is expected to lead the precision of the measurement in the near future.

1.5.2 The top quark

A short review of the properties of the top quark is presented in this section, which summarises Ref. [7]. The top quark phenomenology is widely by its large mass. The mass has been measured both at Tevatron and at the LHC exploiting different final states, single and double top quark production, and the dependence of its production cross section from the mass. The two most precise measurements available are the ATLAS and CMS measurements, combining different techniques and exploiting the full Run I dataset. The measured values are 172.69 ± 0.48 GeV from ATLAS [59] and 172.44 ± 0.48 GeV from

CMS [60].
The top quark is the only quark heavier than the W boson, and thus the only one that can decay in an on-shell W boson and a down-type quark. As a consequence, its lifetime is extremely short, shorter than the typical time-scale of the hadronisation process $\mathcal{O}(1/\Lambda) \approx 10^{-24}$ s, and it decays weakly before hadronisation can occur. The top quark decays approximately with $\mathcal{B} = 100\%$ in a W boson and a bottom quark. The top quark branching fraction in b-type quark, $\mathcal{R} = \mathcal{B}(\mathrm{t} \to \mathrm{bW})/\mathcal{B}(\mathrm{t} \to \mathrm{qW})$, is measured $\mathcal{R} = 1.014 \pm 0.003$ (stat.) ± 0.032 (syst.), with $\mathcal{R} > 0.955$ at 95% confidence level [61].
The W boson further decays in a lepton and a neutrino or in a pair of quarks (obviously the decay in a top quark is kinematically forbidden). Since the weak coupling is universal, each possible final state has the same probability to occur (except for small differences coming from the different masses of the decay products). The branching ratio of the W boson is about 1/3 in leptons (11% in each leptonic flavour) and 2/3 in jets (2 jets families in 3 different colours). If a τ lepton and the corresponding neutrino are produced, the final state features hadrons, with $\mathcal{B}_{\mathrm{h}} \approx 67\%$, or electron and muons, with $\mathcal{B}_{\mathrm{e},\mu} \approx 17\%$ following the decay of the τ lepton. Hadronically decaying τ leptons can be experimentally identified as collimated jets composed of few tracks.
For a pair of top-antitop quarks, as in the $\mathrm{t\bar{t}H}$ process, three groups of final states can be identified:

- fully hadronic decays: $\mathrm{t\bar{t}} \to \mathrm{b\bar{b}\ W^+W^-} \to \mathrm{b\bar{b}\ q\bar{q}\ q'\bar{q}'}$;
- semi-leptonic decays: $\mathrm{t\bar{t}} \to \mathrm{b\bar{b}\ W^+W^-} \to \mathrm{b\bar{b}\ q\bar{q}}\ \ell\bar{\nu}_\ell$;
- fully leptonic decays: $\mathrm{t\bar{t}} \to \mathrm{b\bar{b}\ W^+W^-} \to \mathrm{b\bar{b}}\ \ell\bar{\nu}_\ell\ \bar{\ell}'\nu_{\ell'}$.

The possible final states following the decay of a pair of top quarks and their branching fractions are reported in Table 1.5. Quarks in the final state go through hadronisation and experimentally are observed as a collimated jet of hadrons.
The bottom quark is the lightest quark of the third generation and it can decay only through CKM-suppressed generation-changing decays. When a b-type quark is produced, it forms b-mesons with sizeable lifetime which, thanks to the relativistic boost originated in the collision, can travel for $\mathcal{O}(mm)$ before decaying. Jets originating from bottom type quarks, generally referred to as b jets, can be identified thanks to the presence of hadrons following the decay of the b meson which are displaced from the proton-proton interaction point. Additionally, the sizeable mass of the b-type quark allows the presence of leptons within the jets with a sizeable transverse momentum with respect to the jet axis. The presence of displaced hadrons or of leptons within a jet is exploited for the identification of the b jets; a description of the algorithm used by the CMS experiment will follow in Section 4.3.6.

Category	Final State	Branching ratio
Fully leptonic	$t\bar{t} \to b\bar{b}$ ee $\nu_e\nu_e$	1.1%
	$t\bar{t} \to b\bar{b}$ $\mu\mu$ $\nu_\mu\nu_\mu$	1.1%
	$t\bar{t} \to b\bar{b}$ $\tau\tau$ $\nu_\tau\nu_\tau$	1.2%
	$t\bar{t} \to b\bar{b}$ eμ $\nu_e\nu_\mu$	2.2%
	$t\bar{t} \to b\bar{b}$ eτ $\nu_e\nu_\tau$	2.4%
	$t\bar{t} \to b\bar{b}$ $\mu\tau$ $\nu_\mu\nu_\tau$	2.4%
Semi-leptonic	$t\bar{t} \to b\bar{b}$ $q\bar{q}$ eν_e	14.4%
	$t\bar{t} \to b\bar{b}$ $q\bar{q}$ $\mu\nu_\mu$	14.4%
	$t\bar{t} \to b\bar{b}$ $q\bar{q}$ $\tau_h\nu_\tau$	15.2%
Fully hadronic	$t\bar{t} \to b\bar{b}$ $q\bar{q}$ $q'\bar{q}'$	45.6%

Table 1.5: Branching ratios $\mathcal{B}$ of the possible final states following the decay of a pair of top-antitop quarks.

Chapter 2

Experimental apparatus

Non aetate verum ingenio
apiscitur sapientia.

Titus Maccius Plautus

The European Organisation for Nuclear Research (CERN) is a world-leading particle physics laboratory, hosted in Geneva, Switzerland. Its rich experimental program covers most of the topics of interest in modern particle physics, from high energy physics to neutrinos, from nuclear interactions to matter-antimatter symmetries. The Large Hadron Collider (LHC) is a proton-proton collider of about 27 km of circumference, which accelerates two beams of protons up to a centre-of-mass energy of 14 TeV. The beams are collided in four different interaction points where the ALICE, ATLAS, CMS and LHCb experiments records the outcome of the collisions to mine the underlying law of nature.
The Compact Muon Solenoid (CMS) detector is a general purpose detector which collects and records the outcome of proton-proton and ion collisions delivered by the LHC. The CMS Collaboration gathers more than 4000 physicists, engineers, computer scientists and technicians working in about 200 research institutes and universities from more than 50 different countries around the world. The experiment is taking data since 2010 and in 2012, along with the ALTAS Collaboration, contributed to the Higgs boson discovery [5, 6].
After a description of the most relevant parameters of the LHC, this chapter gives an overview of the CMS experiment, from design to performance in particle reconstruction and identification.

2.1 The Large Hadron Collider

The physics motivation underlying the construction of the LHC was to ascertain the existence of the Higgs boson exploring the mass range allowed by the SM, up to about 1 TeV. The SM without the Higgs boson with $m_{\mathrm{H}} \lesssim 1$ TeV shows mathematical inconsistencies which would led to the divergence of the cross section of some processes with increasing the energy. In case of no discovery of the Higgs boson, the LHC would have been able to prove the inconsistency of the SM and to investigate on mechanisms alternative to the spontaneous symmetry breaking to recover the SM consistency. The LHC reached is main goal after only three years of operations.
The machine has been designed to collide proton beams with a centre-of-mass energy

up to 14 TeV and lead ions with a centre-of-mass energy up to 2.76 TeV per nucleon. The first beam was injected in the machine in 2008, but a major failure of the system caused sever damages to the accelerator and the start of the operations was delayed by one year. The first collision happened in 2009 and the LHC has kept delivering collisions and continuously exceeding its best performance up to now.

2.1.1 The LHC design

The LHC [62] is a 26.7 km long proton and ion accelerator hosted in the underground tunnel build to accommodate the Large Electron-Positron Collider. The tunnel runs across the Swiss-France border in the region surrounding Geneva at a depth between 50 and 175 m underground.
As a proton accelerator, the LHC accelerates two separated and counter-rotating beams of protons up to an energy of 7 TeV per beam. The orbit is defined by 1232 superconducting Niobium-Titanium dipole magnets, each one long about 15 m and capable of providing a magnetic field up to 8.33 T thanks to a current of about 11 kA which flows through the magnets. The superconductivity is granted by a cooling system which exploits liquid Helium-4 at 1.9 K. To ensure smooth running of the operations, beam pipes are kept at a pressure of 10^{-10} mbar, a level of vacuum comparable with the Moon atmosphere. The beam dynamics is controlled by 392 quadrupole magnets positioned throughout the arc. Before each interaction point, triplets of quadrupole magnets squeeze the beams to increase the proton density, dramatically enhancing the probability for protons to interact. The beams are accelerated by 8 superconducting radio-frequency cavities working at a frequency of 400 MHz, providing an energy increase of 485 keV per turn and ensuring the ramp-up from injection to collision energy in about 20 minutes. Each beam consists of at most 2808 bunches, with a spacing of 25 ns per bunch; each bunch features 1.15×10^{11} protons. The total energy stored in the beam is 362 MJ and each beam performs 11245 revolutions per second.
The LHC is the last step of an accelerator chain, starting from hydrogen atoms ending up with ultra-relativistic proton-proton collisions. A pictorial view of the full accelerator chain is depicted in Fig. 2.1. The hydrogen atoms are ionised with a plasma source and the protons are injected in the Linear Accelerator (LINAC2) for the first acceleration to 50 MeV. The protons are then injected in the Proton Synchrotron Booster and in the Proton Synchrotron (PS), where they reach an energy of 1.4 and 25 GeV, respectively. Finally the Super Proton Synchrotron (SPS) enhances the energy to 450 GeV before the injection in the LHC.
Alongside with the energy, the luminosity $\mathcal{L}$ is the LHC parameter which has bigger implications from the experimental point of view. The luminosity is defined as the proportional coefficient between the event rate of a given process dN/dt and its cross section σ:

$$\frac{dN}{dt} = \mathcal{L} \cdot \sigma. \tag{2.1}$$

The luminosity depends only on the beam parameters and, for Gaussian beam distribution and two identical beams, it can be written as:

$$\mathcal{L} = \frac{N_{\mathrm{b}}^2 n_{\mathrm{b}} f_{\mathrm{rev}} \gamma_{\mathrm{r}}}{4\pi \epsilon_{\mathrm{n}} \beta^*} F, \tag{2.2}$$

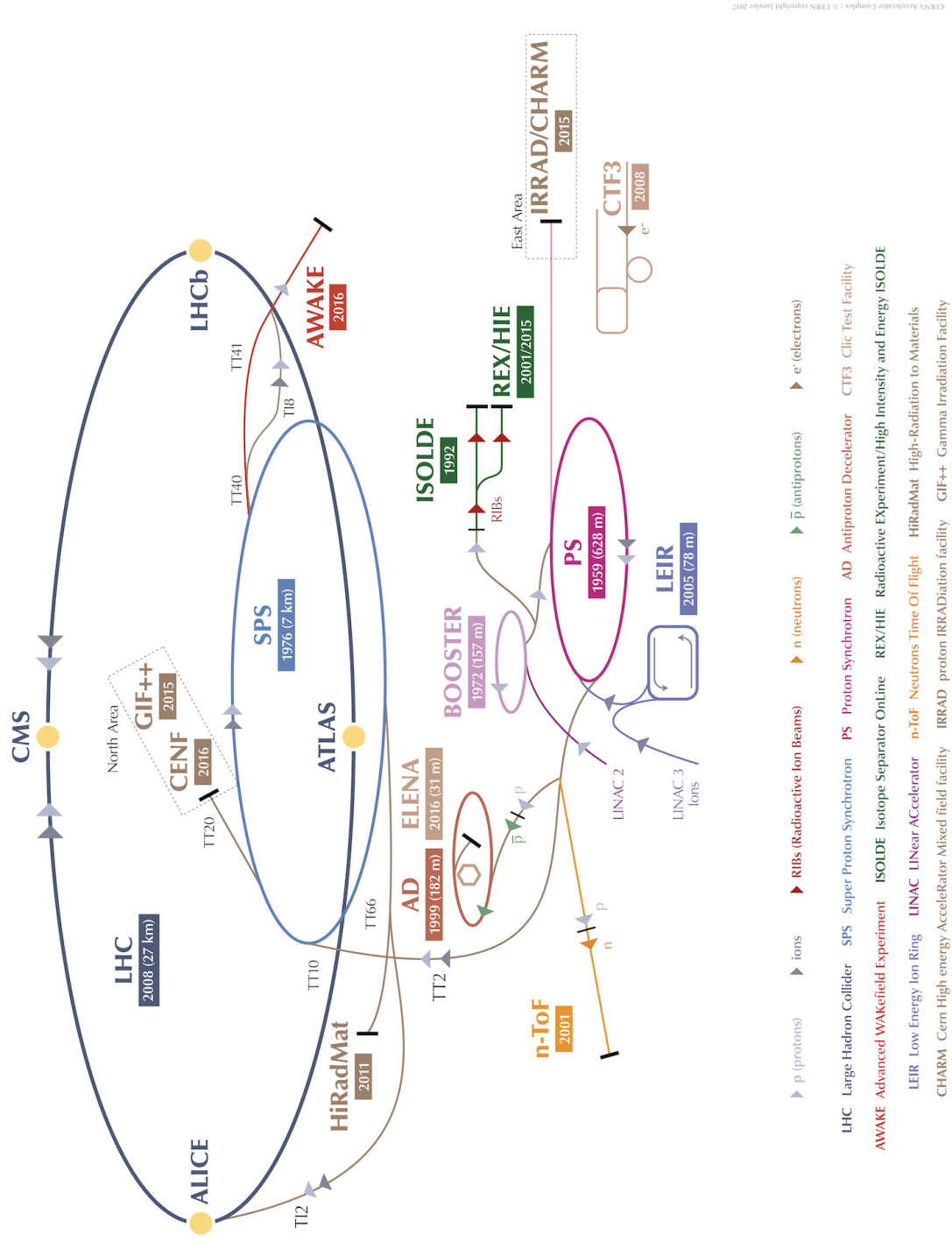


Figure 2.1: A pictorial view of the CERN accelerator complex [63].

where N_{b} is the number of particles per bunch, n_{b} is the number of bunches in the machine, f_{rev} is the revolution frequency and γ_{r} is the relativistic factor of the circulating particles. The transverse emittance ϵ measures the average spread of the beam in a space-momentum phase space in a plane perpendicular to the direction of motion. As the beam keeps accelerating, the emittance decreases as the beam size is reduced; the normalised emittance ϵ_{n} does not depends on the beam energy and is function only of the beam dynamic. The β^* function at the interaction point is related to the transverse beam size, while F is a reduction factor which keeps into account the crossing angle at the interaction points θ_{C} and depends on the r.m.s. bunch length σ_{z} and on the r.m.s. beam size at the interaction point σ_{xy}:

$$F = \left(1 + \left(\frac{\theta_{\mathrm{C}}\sigma_{\mathrm{z}}}{2\sigma_{\mathrm{xy}}}\right)^2\right)^{-1/2}. \tag{2.3}$$

For heads-on collisions the factor F is equal to unity. The design parameter of the LHC are reported in Table 2.1. The design peak luminosity of 10^{34} cm^{-2}s^{-1} was outperformed by a factor more than 2 during the Run II operations, reaching a value of 2.2×10^{34} cm^{-2}s^{-1}.

$\sqrt{s}$	Centre-of-mass energy of the collisions	14 TeV
n_{b}	Number of bunches per beam	2808
Δt_{b}	Spacing between adjacent bunches	25 ns
N_{b}.	Protons per bunch	1.15×10^{11}
f_{rev}	Revolution frequency	11.2 kHz
γ_{r}	Relativistic factor	7461
ϵ_{n}	Normalised transverse emittance	3.75 μm rad
β^*	β^* function	0.55 m
σ_{z}	Bunch length (r.m.s.)	7.55 cm
σ_{xy}	Bunch size (r.m.s.)	16.7 μm
F	Luminosity geometrical reduction factor	0.836
$\mathcal{L}$	Peak luminosity	10^{34} cm^{-2}s^{-1}

Table 2.1: Design parameters for the LHC as a proton-proton collider.

The integrated luminosity $\mathcal{L}^{\mathrm{int}}$ is the integral of the luminosity over a period of time T:

$$\mathcal{L}^{\mathrm{int}} = \int_T \mathcal{L} dt. \tag{2.4}$$

The total number of observed events is proportional to the integrated luminosity, therefore higher integrated luminosity means higher probability to observe rare events: high integrated luminosity is the key to fulfil the ambitious physics program of the LHC experiments. To get higher integrated luminosity, either the luminosity is increased, or longer operations are planned. The running schedule of the LHC keeps into account the maintenance necessary to the LHC and to the detectors, so the only viable way to increase the running time is to reduce the dead time of the accelerator. During Run II, the LHC achieved the impressive fraction of 70% of time dedicated to operations, excluding the winter shout down and the machine commissioning. The second way to go is to increase the luminosity. Higher $\mathcal{L}$ implies higher probability of multiple proton-proton interactions

for each bunch crossing. The ATLAS and CMS experiments, at the peak luminosity of Run II recorded data with up to 70 concurrent collision per bunch crossing, generally referred to as pileup (PU), causing a serious challenge in the event reconstruction.
The LHC features four interaction points; the CMS and ATALS (A Toroidal Apparatus) experiments are located at the opposite sides of the ring, where maximum luminosity can be provided. They are general purpose, hermetic detectors surrounding the interaction point. The LHCb (LHC bottom) experiment is a spectrometer designed specifically to study the phenomenology of the bottom quark and runs at a luminosity $\mathcal{L} = 10^{32}$ cm^{-2}s^{-1} to achieve an average pileup $\langle \text{PU} \rangle = 1$. Finally, the ALICE (A Large Ion Collider Experiment) experiment is designed to study ion collisions and its target luminosity is $\mathcal{L} = 10^{27}$ cm^{-2}s^{-1} during Pb-Pb collisions.
The complete description of the LHC design can be found in Ref. [64], while a good synthesis is provided by Ref. [65].

2.1.2 Operations of the LHC

Figure 2.2 provides a sketch of the schedule of the LHC operations, starting from 2010 up to the end of its experimental program beyond 2035.

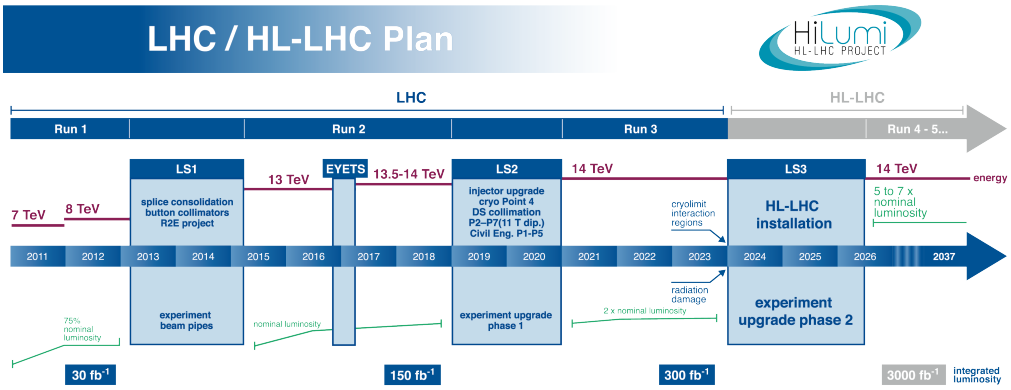


Figure 2.2: Schematic view of the LHC baseline schedule [66].

The beginning of the LHC Run I happened in 2010: the accelerator delivered collisions to the experiments with a bunch spacing of 50 ns for three consecutive years. To reduce the mechanical stress on the LHC magnets, the centre-of-mass energy of the collisions was limited to 7 and 8 TeV in 2010-2011 and 2012, respectively. The delivered integrated luminosity corresponded to 45 pb^{-1}, 6 fb^{-1} and 23 fb^{-1} for the three years. Thanks to the Run I data, the Higgs boson was discovered by the ATLAS and CMS experiments, reaching the main target of the LHC after just three years of operations.
At the end of 2012, the LHC entered the Long Shutdown 1 (LS1), a period of maintenance and renewal of both accelerator and detectors. The collisions come back in 2015, with the beginning of the LHC Run II. The preparation work of the LS1 allowed the LHC to get closer to its nominal parameters. The magnets were trained to 6.5 TeV, and, after a short period at 50 ns, the bunch spacing was reduced to its design of 25 ns. The Run II conditions allowed the experiments to start the precision era for the Higgs boson and to explore the TeV scale searching for unpredicted phenomena. Since 2015, mainly devoted to

the machine and experiments commissioning after the shutdown, three years of data taking provided a total luminosity of about 163 fb^{-1} at a centre-of-mass energy almost doubled with respect to Run I. The large expertise developed with the LHC beam dynamic allowed the study of new solutions to outperform the design luminosity: new injection schema and beam transports enabled a sizeable reduction in the beams emittance, with the record luminosity of 2.2×10^{34} $cm^{-2}s^{-1}$. This value represents the superior limit achievable at the present LHC due to the cooling of the triplet magnets; higher luminosity would induce a heat loss unbearable by the current system. A summary of the total integrated luminosity since the beginning of operations and of the peak luminosity achieved in the different years is presented in Fig. 2.3.

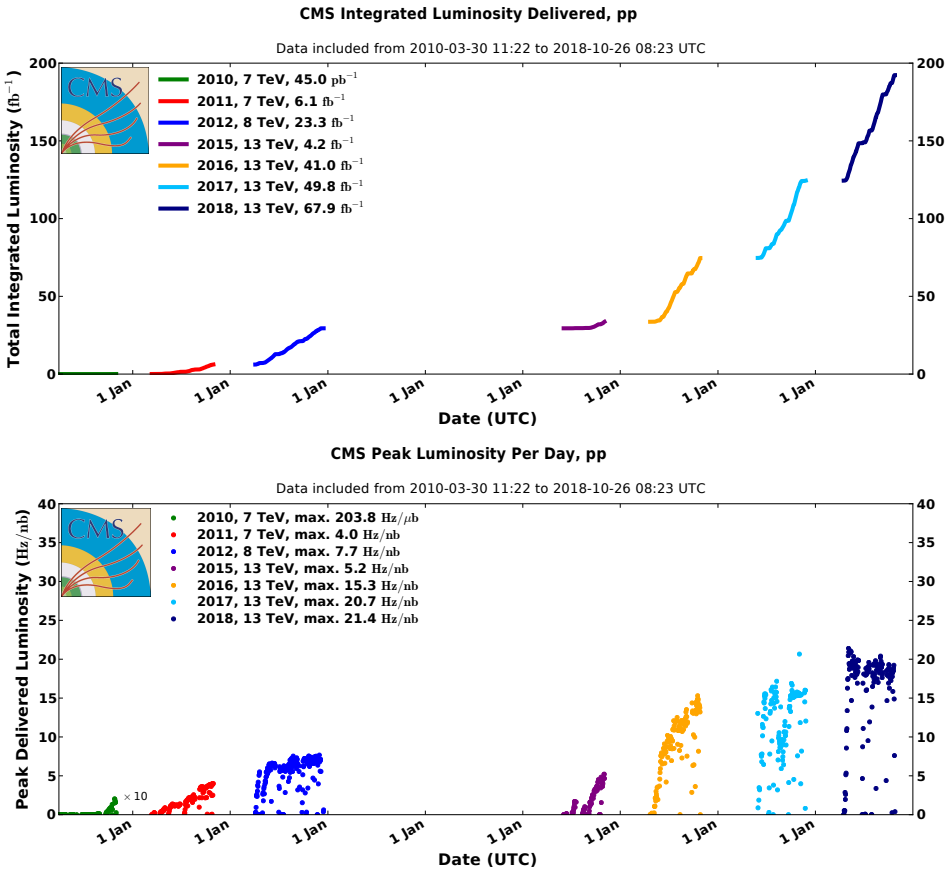


Figure 2.3: Top: integrated luminosity delivered by the LHC from the beginning of the operations as a function of the time. Bottom: peak luminosity delivered by the LHC as a function of the time.

With the end of 2018, the LHC entered the Long Shutdown 2 (LS2), a second period of maintenance and further improvement of the accelerator performance. The end of LS2 is foreseen in 2021, when the Run III will start and up to 300 $fb^{-1}$of integrated luminosity are expected in three years. The LS2 will be exploited to train the magnets to 7 TeV so to reach the nominal energy of the LHC. Additionally, the machine will be virtually able to deliver a luminosity up to 4 times the design one, where virtually means that the cooling limitation will prevent the actuation of such a high luminosity. This expertise will

be exploited for the luminosity levelling, with a constant luminosity of 2×10^{34} cm^{-2}s^{-1} hold for long periods of time.
Beyond the LHC Run III, a third Long Shutdown (LS3) is planned before entering the High Luminosity LHC (HL-LHC) [67]. The triplet magnets are expected to be severely damaged by 2025 by radiation. The magnets will be replaced with Ni_3Sn superconducting magnets, capable to provide a magnetic field of about 12 T. The higher magnetic field will provide a smaller beam profile and, in turn, a higher luminosity. Compact superconducting cavities will be used to precisely rotate the beam and reduce the crossing angle, enhancing the factor F of Eq. 2.2. The luminosity will experience a value from 5 to 7 times higher than the original LHC design, with the impressive target of delivering 3000 fb^{-1} in 10 years of operations.

2.2 The CMS experiment

The CMS detector [68] has been designed to understand the EW symmetry breaking mechanism, for which the Higgs boson was deemed (and actually is) responsible. The detector has consequently been design to explore the TeV scale and to potentially discover he Higgs boson wherever in the allowed mass range. Since the decay branching ratios of the Higgs boson change drastically with its mass, unknown before the LHC, the CMS detector has been designed as a general-purpose cylindrical and hermetical detector surrounding the interaction point capable to optimally identify, reconstruct and measure different kind of particles. A detailed description of the most relevant aspects of the CMS detector can be found in Ref. [69].
In addition to the stringent requirements on energy and momentum resolution for different kinds of particles, the experimental design had to keep into account the high particle multiplicity in the final state of the LHC collisions. The accelerator was designed to deliver 20 proton-proton collisions every 25 ns, and during Run II operations, it reached 70 pileup collisions. As a consequence, a few thousands charged particles cross the detector at every interaction, challenging the event reconstruction and providing severe radiation dose. To cope with the expected level of pileup and radiation damage, the detector has been designed with radiation-hard high-granularity sensors; more than 80 millions sensors are exploited to obtain a coherent picture of beam interactions.
An additional challenge is given by the collision rate of 40 MHz provided by the LHC. The maximum rate at which events can be recorded on disk is about 1 kHz, so online selection of the most interesting events should happen. A two-staged trigger system has been implemented which provide the rejection factor of about 10^5 necessary to the operations. After a brief description of the coordinate system used in CMS (Section 2.2.1), Section 2.2.2 provides a description of all the subsystems of CMS and Section 2.3 describes the algorithm exploited to optimally merge the informations coming from the different sensors in a coherent picture of the interaction. Finally Sec 2.4 describes the data taking of CMS during Run II.

2.2.1 Coordinate system

The CMS detector is described by a right-handed coordinate system with the origin of the axis in the nominal interaction point; the x axis points towards the LHC centre, the y axis in the upward direction and the z one in the direction of the counterclockwise proton

beam.
It is convenient to define a cylindrical coordinate system, which reflects the geometry of the detector. The azimuthal angle φ is measured in the (x, y) plane ('transverse plane') as the angle formed between the positive axis x and the polar coordinate r. The polar angle θ is defined in the (r, z) plane as the angle between r and the positive direction of the z axis.
The polar angle is usually converted in units of pseudorapidity $\eta = -\log\tan(\theta/2)$. The usage pseudorapidity is a natural choice in a hadron collider; the main interaction (and main background in most of the channels) is QCD production, whose cross section is uniform as a function of η. The description (and segmentation) of the detector in terms of pseudorapidity ensures, on average, the same amount of particles for each unit in η. In addition the difference in pseudorapidity are invariant under Lorentz transformations for boost along the beam axis. Since η is strongly non-linear with the angle θ, the forward region of the detector has the highest occupancy and radiation levels.
The separation of two particles can be expressed in terms of a distance parameter ΔR, defined as:

$$\Delta R = \sqrt{(\Delta\varphi)^2 + (\Delta\eta)^2}. \tag{2.5}$$

The projection of the momentum in the transverse plane is generally referred to as transverse momentum p_{T}. Transverse momentum is independent of the Lorentz boost generated along the z axis in the collision. Its magnitude is referred to as transverse energy E_{T}.

2.2.2 Design of CMS

The CMS detector is a 14000 tons, 15 m high and 21 m long detector. Its central feature is a 13 m long, 6 m wide superconducting solenoid capable of providing a magnetic filed up to 4 T. Currently, to reduce the mechanical stress on the magnet itself, the magnetic field is operated at 3.8 T. The magnet provides high bending power necessary to curve the trajectories of charged particles in order to get a precise determination of their momentum. The return magnetic flux is large enough to saturate 1.5 m of iron, allowing four muon stations to be integrated between the iron yokes. The magnet is wide enough to host inside the tracking system and the calorimetry.
A sketch of CMS can be seen in Fig. 2.4. The detector is shaped as a cylindrical 'barrel' with two 'endcaps' covering the two extremities. The boundaries between the two regions vary for each of the CMS subsystem. The inner detector is a silicon tracker, which provides the reconstruction of the track of charged particles, exploited to determine the momentum of the particles and the coordinates of the interaction points ('interaction vertices'). The electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL) measure the energy of electromagnetic and hadronic showers, respectively. Finally the muon stations beyond the magnet ensure good muon identification capabilities and complement the tracker measurement of their momentum. A detail description of each subsystem follows in the next sections. The two-staged trigger system used in online event selection is also described.

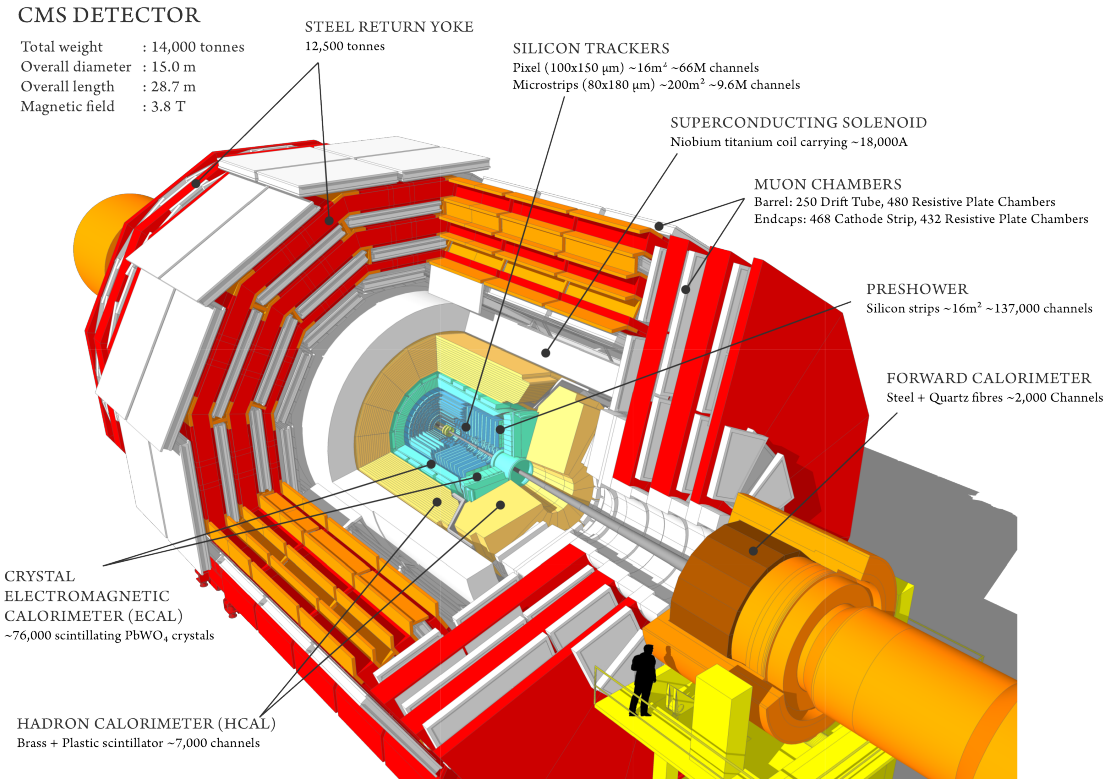


Figure 2.4: A 3D view of the CMS detector [70].

Inner tracker system

The tracking system of CMS is a 5.8 m long and 2.5 m wide detector surrounding the interaction point. The system is designed to deal with the large charged particle density and the induced radiation damage. Achieving excellent performance on momentum resolution is required to reach CMS goals. The technological choice fall on an extremely granular tracker made only by silicon sensors, equipped with fast readout on-board electronics. Several layers of silicon sensors surround the interaction point. Charged particles crossing the tracker deposit energy in the sensors ('hit') of the each layer and different hits are combined to form tracks. Each track provides information on the trajectory of a charged particle and thus on its momentum. The track *ensamble* allows a precise identification of the interaction vertices as well as of secondary vertices from decay of long-living particles such as b-flavoured mesons.
The particle density decreases as $1/r^2$ moving outward from the interaction point, therefore the detector granularity can be reduced in the outer layers without increasing the sensors occupancy. Inner layers feature pixel sensors with cell size 100×150 μm^2, with average detector occupancy of 10^{-7} per bunch crossing. The outer layers, where the particle density is lower, are equipped with silicon strips of different pitches, according to the distance from the interaction point. Figure 2.5 illustrates the composition of the tracker system.
The inner Pixel detector features 3 layers in the barrel at 4, 7 and 11 cm from the interaction points and 2 disks in each of the endcap at $|z|$ of 35 and 47 cm. The spatial

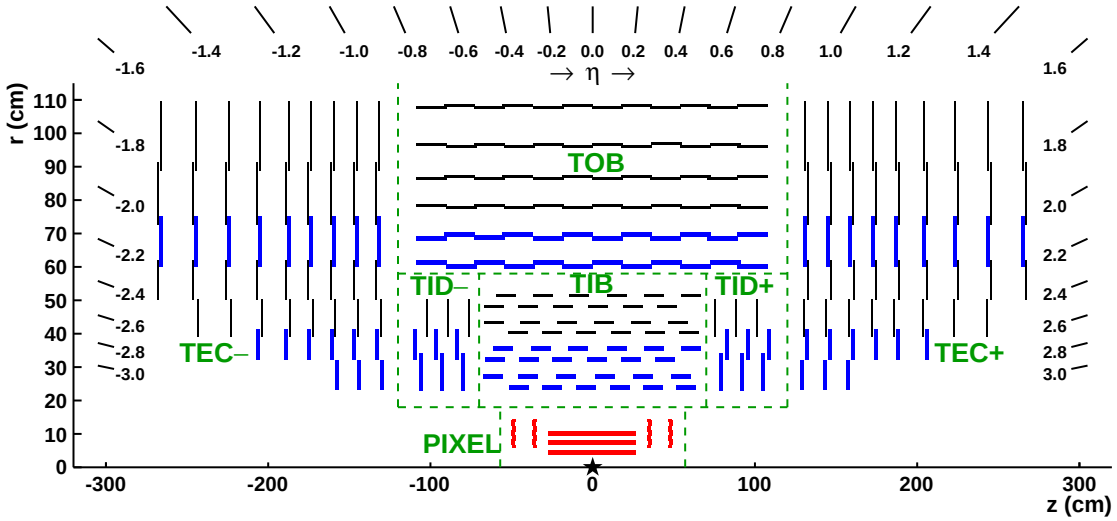


Figure 2.5: Layout of the CMS tracker system. The interaction point (black star) is surrounded by 3 layers of pixels in the barrel and two disks in each of the endcap (red). The modules equipped with strip sensors are shown by blue and black lines; blue lines represent the pair of modules with stereo vision, allowing 3D reconstruction of the hits. Green lines help in understanding the division of the sensors in the different regions of the tracker [71].

resolution of the single sensor is about 10 μm^2 in the (r, φ) plane and 20 μm^2 in the z direction. The Pixel detector is equipped with a total of 66 millions sensors covering a surface of about 1 m^2. The CMS Pixel detector has been replaced during the winter shutdown between 2016 and 2017 [72]. A comparison between the structure of the old and the upgraded pixel is shown in Fig. 2.6. The upgraded detector features one more layer in the barrel and one more disk in each of the endcap, with the innermost layer of the barrel at only 3 cm from the interaction point. The overall material budget has been reduced from 40 to 80%, depending on η, and performance in track identification and momentum resolution are better than with the previous Pixel detector. The largest gain is achieved in the identification of secondary vertices, thanks to the smaller distance from the interaction point and to the additional layer of sensors. As a consequences data collected in 2016 presents slightly different tracking than data collected in 2017 and 2018.
The silicon strip detector covers the region $20 < r < 110$ cm. The barrel is divided in two parts, the Tracker Inner Barrel (TIB), covering the region $r < 65$ cm, and the Tracker Outer Barrel (TOB) in the remaining space. The TIB consists of four layers paved with sensors with a thickness of 320 μm, length of 10 cm and a strip pitch between 80 and 120 μm. The average occupancy is 2-3% per bunch crossing. The first two layers are made with stereo modules which allow a 3D reconstruction of the hits (in blue in Fig. 2.5). The single point resolution is of about 23 μm in both the (r, φ) plane and in the z direction. The TOB, where the radiation levels are significantly lower, features strips with a thickness of 500 μm, ensuring good signal to noise ratio for larger pitch strips (up to 180 μm). The TOB is composed of 6 layers, 2 of which with stereo measurement; the single point resolution varies from 34 to 52 μm in the (r, φ) plane and it is 52 μm in the z direction. The TIB is shorter than the TOB to avoid excessive crossing angles

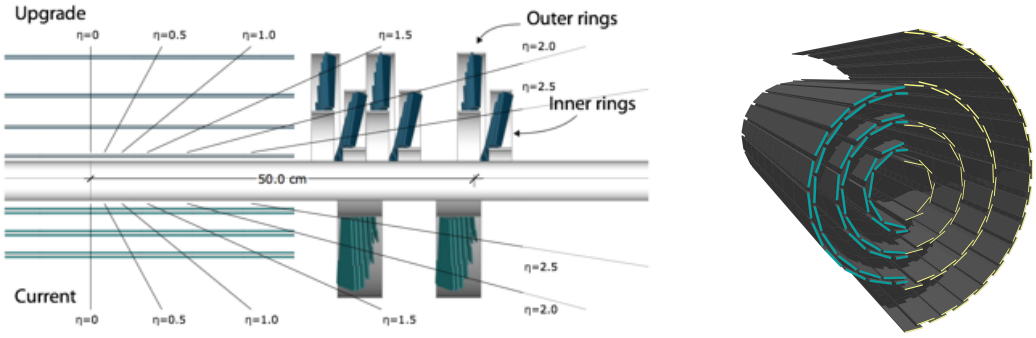


Figure 2.6: Left: longitudinal view of the upgraded pixel detector (top) compared with the old one (bottom). Right: Transverse view comparing the two pixel detectors [72].

of tracks. The transition region between TIB and Tracker End Cap (TEC) is equipped with the Tracker Inner Disk (TID), composed by three small disks of strips. The TEC has nine disks extending the coverage to the region $120 < z < 280$ cm. Both TID and TEC are arranged in rings centred on the beam line made of silicon strips of variable pitches, between 320 and 500 μm. The first two disks of both the regions are equipped with stereo modules. The full strip detector consists of about 9.6 millions silicon strips.
The full tracker system is operated at -20^oC, to minimise the impact of the radiation damage. An efficient cooling system ensures the dispersion of the heating generated by the on-board electronics.

Electromagnetic calorimeter

The ECAL provides a high resolution measurement of the energy of electrons and photons, as well as of the electromagnetic component of hadronic showers. It is a hermetic and homogeneous calorimeter made of lead tungstate ($PbWO_4$) scintillating crystals coupled with photodetectors. Incoming electrons and photons are converted into electromagnetic showers which, interacting with the crystals, produce scintillation light read by the photodetectors. The low material budget in front of the ECAL ensures low energy loss before the calorimeter, therefore a homogeneous calorimeter, where the active material coincides with the absorber, can be fully exploited to reach optimal performance in terms of energy resolution.
The choice of $PbWO_4$ as scintillating material was driven by the density (8.28 g/cm^3), the short interaction length ($X_0 = 0.89$), the low Moliere radius (2.2 cm), the fast scintillation signals (80% of the light is emitted in 25 ns) and by its radiation hardness. These parameters ensure optimal containment of the electromagnetic shower, alongside with the fast response and the durability necessary to withstand the LHC environment. The $PbWO_4$ drawback is the low light yield of about 30 photons per deposited MeV, which forced the choice of photodetectors with internal gain capable to operate embedded in a strong magnetic field. The ECAL Barrel (EB) is equipped with Avalanche PhotoDiodes (APDs) while the ECAL Endacps (EE) are instrumented with Vacuum PhotoTriodes (VPTs), less sensitive to the radiation damage.
Figure 2.7 illustrates the geometry of the detector. The EB features 61200 scintillating crystals arranged in a (η, φ) grid, each one equipped with two APDs. The crystals are

grouped in 36 identical 'supermodules', covering half of the barrel length ($0 < |\eta| < 1.479$) and 20^o along φ. Supermodules are further subdivided in five 'modules' and in 'Trigger Towers' (TTs), structures of 5×5 crystals which share the front-end electronics. The section of the crystals is 2.2×2.2 cm^2 or 0.0174 in units of $\Delta\varphi$ (1^o) and $\Delta\eta$. The depth of the crystals is 23 cm, corresponding to 25.8 radiation lengths. The crystal geometry is quasi-projective, with the crystals tilted by 3^o with respect to the nominal interaction point, to prevent photons from escaping in the crystal gaps.
The EE, equipped with 7324 crystals per side, is placed at 314 cm from the interaction point, extending the coverage of the ECAL up to $|\eta| = 3.0$. The crystals, with section of 28.6×28.6 cm^2 and a length of 22 cm (24.7 radiation lengths), are disposed according to an x-y grid. Each endcap is organised in 9 modules further subdivided in TTs. The ECAL Preshower (ES) detector is placed in front of each endcap, to improve the discrimination of photons from $\pi^0 \to \gamma\gamma$ decays. The active elements are two layers of silicon strips with pitch of 1.9 mm, laying beyond lead absorbers at a depth of 2 and 3 radiation lengths.

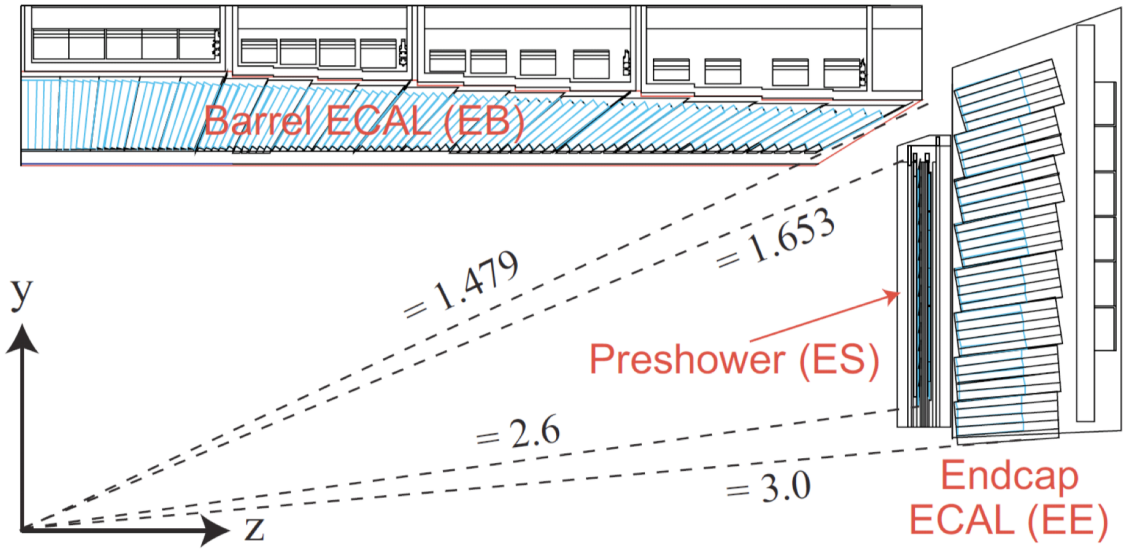


Figure 2.7: Longitudinal view of a quarter of the ECAL.

The ECAL is readout at 40 MHz by a multi-gain preamplifier. Three gains of 12, 6 and 1 are available; for each signal the highest gain which does not saturate the dynamic range of the electronics is exploited for the reconstruction. Each crystal is digitalised by a 12 bit analogue-to-digital converter (ADC) in 10 consecutive digitalisations at 25 ns distance, for a total readout window of 250 ns.
Both the light yield of the crystals and the gain of the photodetectors are sensitive to the temperature, with a loss of $2\%/^o$C for the light yield and of $2.3\%/^o$C for the sensors. To ensure constant detector conditions over time, the full ECAL is kept at 18^oC, with a stability of 0.05^oC in the EB and 0.1^oC in the EE. More details on the ECAL can be found in Refs. [73] and [74].
The energy resolution of a calorimeter can be parametrised as:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2. \tag{2.6}$$

Three effects concur to define the energy resolution. The stochastic term S keeps into account the randomness of the energy deposition inside the crystals. The number of elementary carriers of information (in the ECAL case photons from the scintillation) follow a Poisson distribution, with a standard deviation $\sqrt{n}$, where n is the average number of carriers, proportional to the energy deposited. As a consequence, the randomness in the number of carries induces an energy spread proportional to $\sqrt{E}$. The noise term N depends on the parameter of the electronics, which does not depend on the energy of the incoming particle. Finally, the constant term C accounts for detector inhomogeneities, resulting in a term which affect a constant fraction of the energy deposited. The values of the three terms were measured in $S = 2.8\%$, $N = 12\%$ and $C = 0.3\%$ on a test beam with incident electrons before the starting of the data taking [75].
The performance of the ECAL is crucial in a search targeting photons as a final state; consequently a consistent fraction of the work in the context of this thesis has been devoted to ensure high performance and excellent energy resolution of the calorimeter. Section 3.1 provides more details about the functioning of the ECAL and its performance during the Run II data taking, with special emphasis on the calibration procedure necessary to ensure uniform response across the whole calorimeter.

Hadronic calorimeter

The HCAL provides a measurement of the energy of hadronic showers. As the hadronic interaction length is much longer than the electromagnetic one, hadronic showers are not stopped by the ECAL. In the general case, a shower starts its development in the ECAL and gets absorbed by the HCAL. The intrinsic energy resolution of an hadronic shower is much worse than its electromagnetic counterpart. Nuclear and hadronic interactions cause non-Poissonian effects in the shower development, with undetectable particles created within the shower. The production of π^0 results in an electromagnetic component of the shower, which has a different response from the hadronic one and can be fully absorbed in the ECAL. Additionally, fitting HCAL in the magnet volume forced the usage of diamagnetic absorbers and limited the length of the calorimeter. As a result, the energy resolution of the HCAL is quite modest. Despite its limited performance, the HCAL is mandatory for an efficient jet reconstruction and to ensure an hermetic detector, capable of providing an accurate estimate of the transverse momentum imbalance of the collision. Within the limited space available, the design of the HCAL targeted to maximise the material budget of the absorber (in terms of interaction lengths) confining, in turn, the active material to the smallest possible volume. The choice of brass as absorber granted a diamagnetic material with short interaction length and an easy manufacturing. The active medium is given by plastic scintillators tiles read out with embedded wavelength shifters optical fibres. The light produced in the scintillators is carried on the fibres to hybrid photodiodes detectors.
The structure of the HCAL is depicted in Fig. 2.8. The Hadron Barrel (HB) covers the region $|\eta| < 1.4$. It features 2304 towers with a segmentation $\Delta\eta \times \Delta\varphi = 0.087 \times 0.087$, the same of the ECAL TTs, and it is read out as a single longitudinal sampling calorimeter. Particles leaving the ECAL, after crossing a 9 mm thick scintillator, go across 15 brass plates of about 5 cm of thickness with 3.7 mm scintillators in between. The first scintillator is optimised to provide 1.5 times more light than the others. The Hadronic Outer (HO)

calorimeter has been designed to increase the containment of the showers. It consists of 10 mm scintillators covering the region $|\eta| < 1.26$ collocated outside the magnetic coil. The HO serves as a tail-catcher for penetrating showers that leak through the rear of the calorimeter, increasing the effective thickness of the HCAL to more than 10 interaction lengths. The Hadronic Endcap (HE) covers the region $1.4 < |\eta| < 3$. The HE has the same technology of the HB, with a total of 2304 towers with a segmentation of 5^o in the φ direction and varying from 0.087 to 0.35 in η. The coverage is extended up to $|\eta| = 5$ thanks to the Hadronic Forward (HF) calorimeter. The HF is a steel and quartz fibre sampling calorimeter starting 11.2 m away from the interaction point, with an absorber depth of 1.65 m. This technology ensures ultimate radiation hardness, necessary to withstand the high level of radiation in the region. The signal originates from Cherenkov light emitted in the quartz fibres, which channel it to photomultipliers for the readout. Two lengths of fibres are exploited to obtain a longitudinal segmentation of the sampling, with depth optimised to get a readout of the electromagnetic and the hadronic component of the shower.

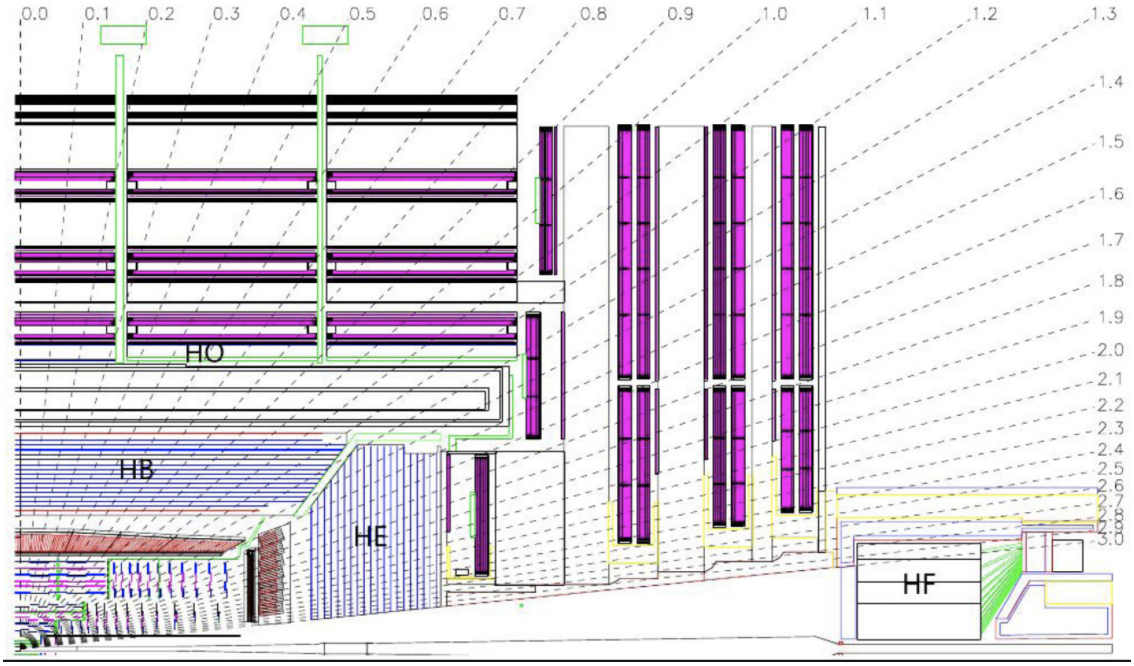


Figure 2.8: Longitudinal view of a quarter of CMS where the different HCAL regions are shown.

During the Run II, the HCAL was upgraded to improve its performance and to replace some components damaged by the radiation [76]. All the sensors were replaced; more radiation hard photomultipliers have been exploited for the HF, while the hybrid photodiodes have been replaced with more reliable silicon photomultipliers. The new read out system allow the possibility to longitudinally sample the shower, but for Run II data this feature has not been exploited yet.
The HCAL resolution is mainly limited by the imperfect containment of the shower; test beam performed before the beginning of the data taking measures the stochastic term in 110% and the constant term in 9% [77].

Muon system

Muons are extremely penetrating particles. They do not experience strong interaction and *bremsstrahlung* loss is negligible for momenta below 1 TeV [68]; they behave as minimum ionising particles, crossing the calorimetry, therefore they can be measured in the muon stations beyond the magnet.
The muon system is a highly redundant combination of gaseous detectors providing a measurement of muon momentum to complement the tracker information. For low p_{T} muons the tracker measurement in by one order of magnitude more precise than the muon system while, beyond $p_{\mathrm{T}} \approx 200$ GeV, the muon system starts to dominate; the combination of the two ensures optimal performance in muon reconstruction. In addition to complementing the tracker, the muon system provides to CMS unique capability of muon identification and muon trigger, since tracker information is not available during the first step of the trigger.
The muon chambers are hosted in the iron yokes, where the return field of the solenoid ensures the bending power necessary for precise momentum determination. Four stations of detectors instrument the region, to ensure a redundant system capable to identify muons with high efficiency. The technological choice of gaseous detectors is essentially driven by the large area to be instrumented, of about 25000 m^2.
The full muon system is depicted in Fig. 2.9. In the barrel, the stations are disposed in cylinders interleaved with the iron yokes, fragmented in five section along the beam direction. In the endcaps the stations are disks perpendicular to the beam axis, further divided in three concentric rings in the innermost station and two in the others.
Three different detector technologies are exploited in the system, to ensure redundancy of the information and good operations even in the high background regions.
The Muon Barrel (MB) region ($|\eta| < 1.2$), where the neutron induced background is small and the muon rate is low, Drift Tubes (DT) chambers are used. The DT in the different stations are staggered to ensure the crossing of at least three out of four stations for high p_{T} muons. The chambers are arranged to provide both a measurement in the transverse plane and in the z direction. The spatial resolution of each chamber ranges between 80 and 120 μm [78].
The Muon Endcap (ME) is equipped with Cathode Strips Chambers (CSCs), capable of sustaining the higher muon rate, the larger neutron induced background and the stronger magnetic field in this region. The coverage is up to $|\eta| < 2.4$. The gas ionisation in the CSCs, and the following electron avalanche, induces a fast signal on a wire while the ions migrate on the strips paving the side of the chambers. The fast signal of the wire provide a coarse estimate of the position and it is exploited for the trigger, while a better position resolution is obtained by the charge balance on the strips. Each CSC chamber has a position resolution of 40-150 μm [78].
Both MB and ME are equipped with Resistive Plate Chambers (RPCs), operated in avalanche mode to ensure good operations even at high rates (up to 10 $\mathrm{kHz/cm}^2$). The RPCs provide a fast response with good timing resolution of about 3 ns, at the price of a coarse spatial resolution of 0.8-1.2 cm [78]. The different kind of detectors are highly complementary, providing a redundant and highly efficient trigger and identification for muons. The overall efficiency of the system in muon reconstruction is between 95 and 98%, varying on the position, and of 96% at trigger level [78].

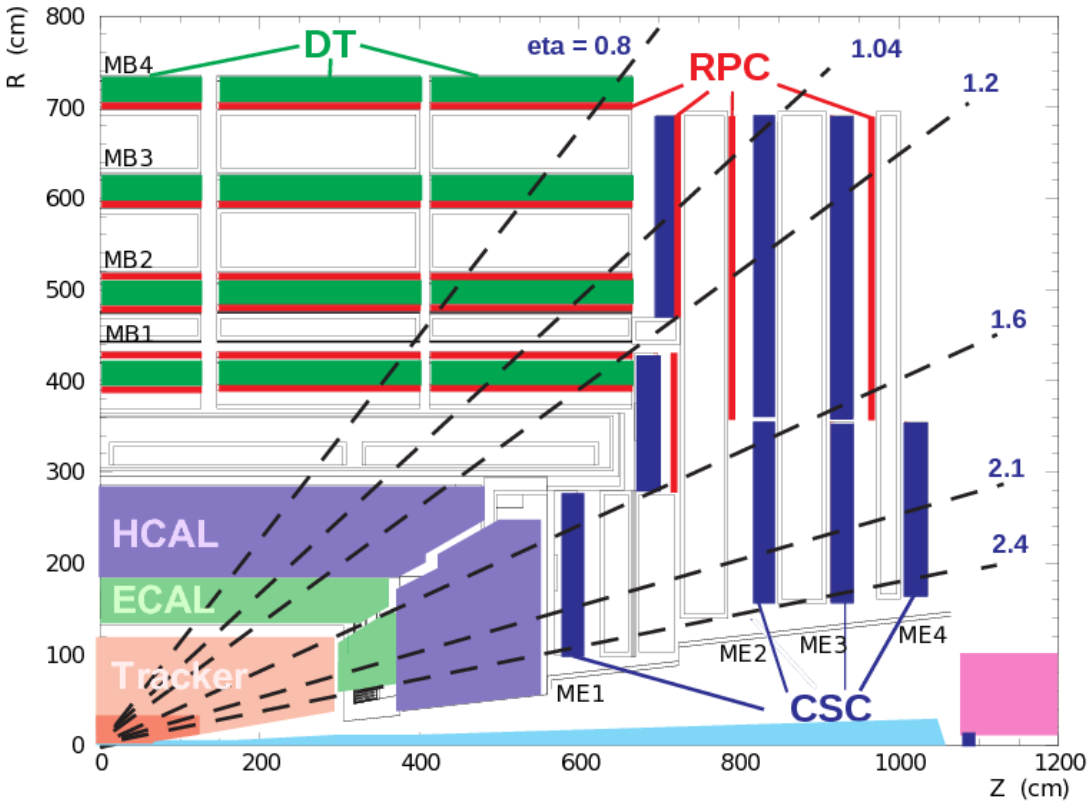


Figure 2.9: Longitudinal view of a quarter of CMS where the different regions and detectors of muon system are shown [79].

Trigger system

The information collected by the whole CMS detector can be stored in about a few MB per collision. Since the LHC is designed to deliver 40 millions collisions per second, a storage space of more than 100 TB/s would be necessary to record the outcome of every single event detected by CMS. At present, no technology exists to write, store and retrieve such an amount of data. The only viable way is to select the most interesting events for storage on disk, the interest being defined according to the physics program, and lose all the others. Figures 1.2 and 1.3 summarise the cross sections of most of the physical processes happening at the LHC. The total proton-proton cross section is about five orders of magnitude greater than the second leading process, which is the production of jets through strong interaction. The majority of the collisions will thus result in a low energy proton-proton interaction of poor interest in the context of the LHC physics program. The rejection of the vast majority of the events does not compromise the goal of CMS, as long as high efficiency is assured on the interesting physical processes. The trigger system takes care of the selection of the events to be recorded, reaching the rejection factor nearly 10^5 necessary to keep the writing rate below its maximum limit of 1 kHz. The trigger system is organised in two stages, the Level 1 (L1) hardware trigger and the software High Level Trigger (HLT).
The L1 trigger takes as inputs informations with reduced granularity from the front-end electronic and produces an output rate of at most 100 kHz. The lower granularity ensures

a fast processing of each event, necessary to withstand the LHC collision rate. The trigger is physically located in a different underground cavern with respect to the detector, thus a fixed latency of 3.2 μs per event is necessary to transport the information from the front-end electronic to the L1 hardware and back. During this latency, the full granularity information acquired by the detector is stored in buffers on the front-end electronics. Low granularity data are sent to the L1 trigger, which elaborates a decision and, in case of a positive answer, the front-end electronic in informed to transfer the event to the HLT. The latency is dominated by the travel time of the signals; only less than 1 μs is due to the trigger algorithm. The decision of the L1 trigger is taken on the base of 'trigger primitives': muon tracks from the muon system or energy deposits in the calorimeters. No information from the tracker is available at this stage since its processing would induce an unbearably high latency.
During the LS1, the L1 trigger underwent a major upgrade [80], to improve the trigger performance. Sophisticated algorithms can be implemented in the present L1 trigger system; more than 300 different L1 algorithms, generaaly referred to as 'seeds', have been developed during Run II to ensure good acceptance of data useful for physics measurement, calibration, monitoring and alignment of the detector.
As for the ECAL case, a fraction of this thesis was devoted to the development of efficient L1 seeds for electron and photons, thus the L1 trigger will be described in greater detail in Section 3.2.
The HLT is a computing farm running on over 50000 CPUs working in parallel and running the same trigger algorithm. After the recipient of a L1 trigger, the high-granularity data are transferred from front-end electronic to buffers accessible from the Data Acquisition (DAQ) system. From those buffers, data are sent to the first free CPU for processing.
The algorithm is as close as possible to the one exploited in the reconstruction of stored data (Section 2.3). The reconstruction starts around the L1 trigger primitives and proceed in subsequent steps, aiming at discarding non-interesting events as soon as possible. If the trigger primitive (form calorimetry or muon system) is identified as a potential signal, the full reconstruction of the calorimeters and of the muon system is performed, followed by the pixel and finally by the full tracking. The HLT reduces the output rate to less than 1 kHz.
Events selected by the HLT are sent to the CERN computing centre for full reconstruction, storage on disk, backup on tape and sharing among all the institution involved in the CMS Collaboration to be accessed for later analysis.

2.3 Event reconstruction

This section describes the Particle Flow (PF) algorithm [81], exploited in CMS to optimally merge the redundant information from all the subsystems. The typical signature that each different particle leaves in the detector is pictorially drawn in Fig. 2.10. A muon is identified as a track in the muon system connected to a track in the inner tracker. A hadron leaves an energy deposit in the HCAL and, if charged, a track is associated to it. Similarly, electrons and photons release energy in the ECAL, electrons being associated with tracks. The PF algorithm combines the informations coming from the different subdetectors to fully exploit the high resolution detectors and partially compensate for the limitations of each subsystem. The output of the PF algorithm is a collection of particles, such as muons, electrons, photons and hadrons (neutral or charged). Section 2.3.1 describes in

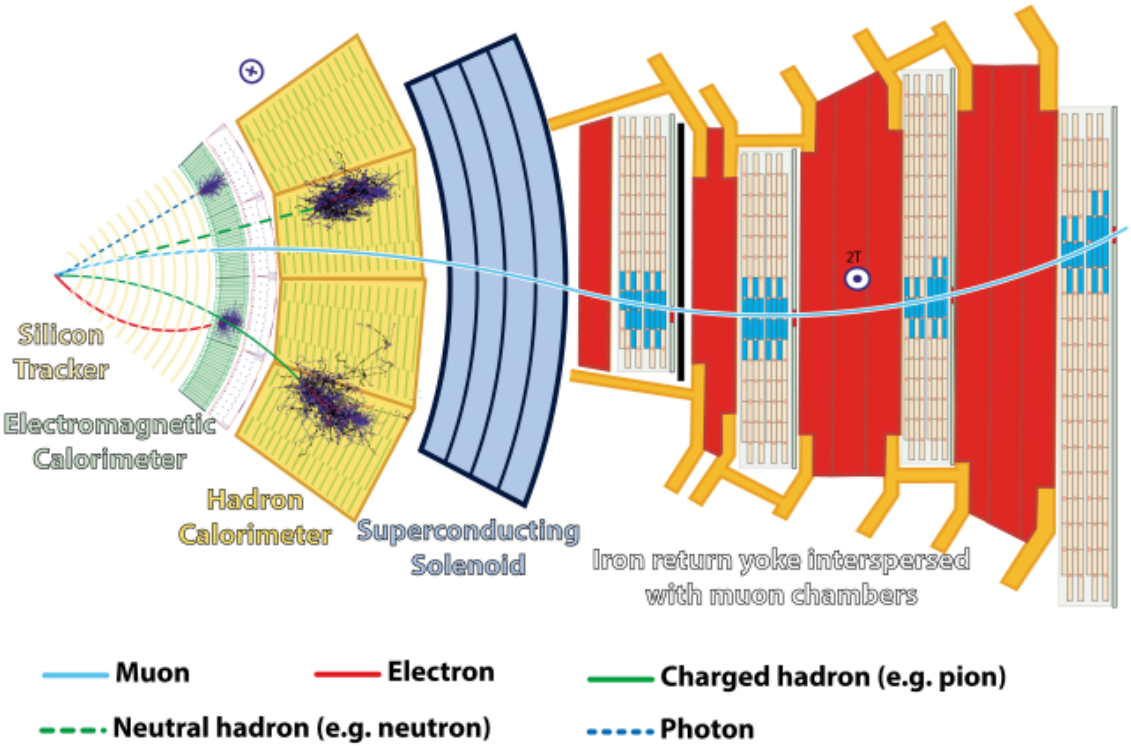


Figure 2.10: A sketch of the different particle interactions in a slice of CMS [81].

detail the principle underlying the PF algorithm, while the following sections explain how the different particles relevant for the $t\bar{t}H$ process are reconstructed.

2.3.1 Global event reconstruction

The aim of the PF algorithm is to provide a description of each collisions in terms of final state particles, combining information from the different detectors. Each detector provides useful information, with specific strengths and weaknesses; the combination of the different subsystems largely improves the energy resolution and the particles identification capability.
Before explaining the algorithm, as an example to better understand the idea underlying the PF algorithm, the reconstruction of a jet is described. A jet produces a mixture of electromagnetic and hadronic energy deposits in conjunction with a set of tracks. The traditional approach for its reconstruction would be to cluster the energy deposits in the calorimeters and to measure the energy of the jet from the sum of the clustered energy. Given the HCAL limited energy resolution, the jet energy would be poorly measured. On average, 65% of the jet energy is carried by charged hadrons, about 25% by photons following π^0 decays and only 10% by neutral hadrons. The high resolution of the ECAL provides a precise measurement of the electromagnetic fraction of the shower. If the tracker information is exploited for the charged component, the HCAL contribution is limited to 10% of the jet energy, therefore the energy resolution is largely improved. Figure 2.11 shows the reconstruction of a simulated event with two jets; the PF reconstructed jets are much more precise both in measuring the energy and direction of the jet than the traditional calorimetric jet.

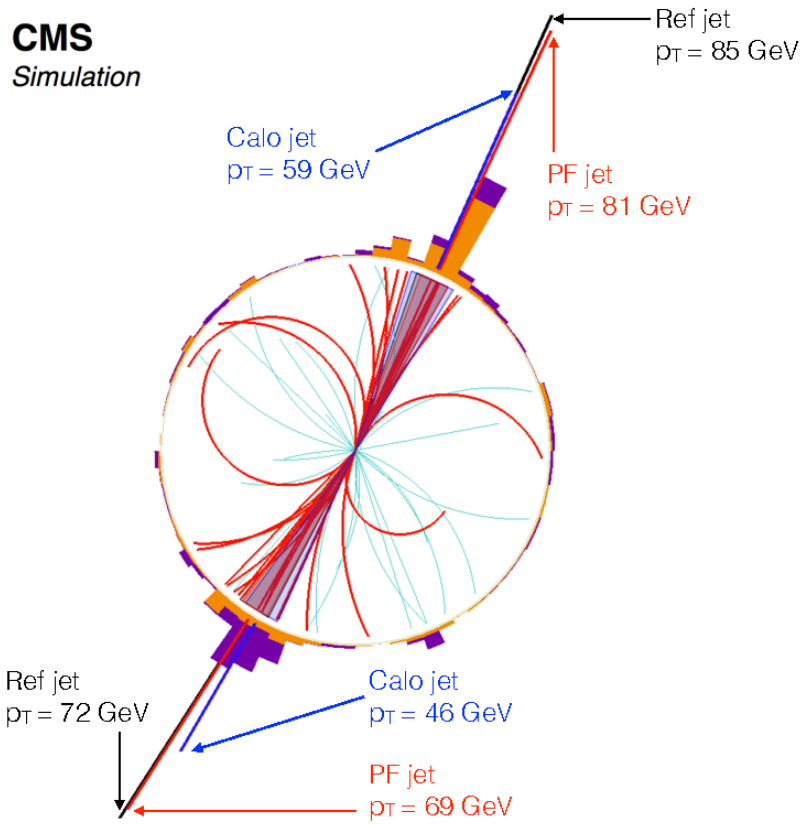


Figure 2.11: Reconstruction of a simulated event with two jets. The particles clustered in each jet are indicated by the red lines. For clarity, particles with $p_T < 1$ GeV are not depicted. The PF jet is compared with the corresponding generated jet (Ref) and with the traditional jet reconstruction based only on calorimeters (Calo) [81].

For the PF algorithm to succeed, high granularity of tracker and calorimeters, intense magnetic field capable to separate charged and neutral particles, hermeticity of the detector and excellent muon momentum resolution must be provided. The CMS detector is therefore an optimal environment to develop a PF reconstruction algorithm.
The PF algorithm proceeds in multiple steps. Initially, tracker hits are combined to form tracks; the same procedure is performed in the muon system. The energy deposits in the ECAL and in the HCAL are then clustered; at this stage a collection of tracks and clusters exists. The different objects are then linked together to create 'PF blocks', on top of which final state particles are built. Finally, global event quantities are computed, such as jets, interaction vertices and the missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, defined as the projection onto the transverse plane of the negative vector sum of the momenta of all reconstructed PF particles in each event.
The PF approach works only if high efficiency on track reconstruction is ensured. A complex iterative tracking algorithm [71] , based on the Kalman filter [82], grants the necessary efficiency with multiple reconstruction steps aiming at the reconstruction of tracks originated from different processes. After each iteration, the hits already associated with a track are removed from further processing and the following steps continue with

the remaining hits. The first steps target energetic isolated particles and decay products of displaced b-flavoured mesons, with stringent requirement on the track quality. At each step, quality criteria are released in order to increase the reconstruction efficiency, at the price of a modest increase of the fake rate. About 20% of the charged hadrons undergo a nuclear interaction within the tracker volume, originating a secondary vertex. On average, one secondary vertex from nuclear interaction is present in each $t\bar{t}$ event. Two steps of the tracking algorithm are explicitly designed to reconstruct extremely displaced tracks originated from nuclear interactions. In those steps no pixel information is used and tracks are built only from hits in inner or outer strip sensors. A dedicated treatment is reserved to vertices from nuclear interactions, aiming at the highest possible resolution for all the components originated in the nuclear interaction, including the neutral one. An *ad hoc* tracking is reserved for electrons, as described in Section 2.3.3.
Once tracking is completed, tracks segments are built in the muon system and clustering of calorimeters starts. The clustering is exploited to determine position and energy of deposits due to photons, neutral hadrons or charged particles whose tracks are not well reconstructed or not reconstructed at all. The clustering is performed separately in EB, EE, ES, HB and HE. At first, 'seed' cells are identified as calorimeter cells whose energy is a local maxima above a given threshold. Clusters are grown by topological clustering: cells neighbouring to a cell already in the cluster are added to the cluster itself if the energy deposit is above a given threshold, set according to the electronic noise. For each topological cluster, substructures are searched with a Gaussian-mixture model [83], to separate energy deposits close to each others, as the two photons of a $\pi^0 \to \gamma\gamma$ decay.
The following step of the PF is the 'link algorithm'. This step associates clusters and tracks which are compatible with originating from the same particle and a collection of PF blocks is produced. Linking is performed between elements spatially close to each other, using the distance parameter ΔR defined in Eq. 2.5. A track is linked to a cluster if its extrapolation falls inside an ECAL or HCAL cluster. An ECAL cluster can be linked to an HCAL or an ES, one if the position of the cluster in more granular calorimeter falls inside the envelope of the second one. A track in the inner tracker can be linked to a track in the muon system, if the two tracks are compatible with originating from the same particle. Finally tracks can be linked to each other by a common secondary vertex due to nuclear interaction.
The collection of final state particles is built upon the PF blocks. At first, muons are reconstructed and tracks associated with muons are removed from the track collection. Then, isolated electrons and photons are reconstructed and the tracks and the clusters associated with them are removed from the respective collections. The remaining elements are exploited to reconstruct hadrons and non-isolated photons which are clustered in jets or hadronically decaying τ leptons. Finally the missing transverse momentum is computed.

2.3.2 Muons reconstruction

Within CMS, muon reconstruction proceeds both in the muon systems and in the tracker. Hits in the DTs, CSCs, and RPCs are associated to form track segments in the muon chambers. The track in the inner tracker are reconstructed as well, therefore three possible reconstructions are possible for a muon:

- **standalone muon**: the muon track is reconstructed only from hits in the muon system;

- **tracker muon**: the muon track is reconstructed in the inner tracker. If at least one hit in the muon system lays on the track extrapolation, the track is interpreted as originated from a muon and it is referred to as a tracker muon;
- **global muon**: if a standalone muon track is matched to a track in the inner tracker, the muon is identified as a global muon. For large transverse momenta, $p_{\mathrm{T}} \gtrsim 200$ GeV, the addition of the muon system improves the resolution with respect to a tracker-only measurement. If a tracker muon shares the same inner trajectory of a global muon, the two candidates are merged together.

Global muon reconstruction is designed to be highly efficient for penetrating muons with high p_{T}. Since muons with low transverse momentum could fail this identification for the multiple scattering in the iron yokes, the tracker-only reconstruction extends the coverage to the low p_{T} region, ensuring a reconstruction efficiency of about 99% for muons produced within the geometrical acceptance of the detector. Data analyses involving isolated muons generally further select muons applying tighter identification criteria than the PF ones. The purpose of the selections is to reject charged hadrons escaping the HCAL and being misidentified as muons. The identification is based on track quality criteria and on the requirement of low energy deposits in the calorimeters cells surrounding the muon track. Standalone muons are rarely exploited in the data analyses due to the low momentum resolution and the large contamination of cosmic muons.
The muon momentum is measured from the curvature of the track. For muons with $p_{\mathrm{T}} < 200$ GeV, the momentum is measured only from the inner track, as the multiple scattering in the iron yokes would worsen the resolution. Instead, for muons with $p_{\mathrm{T}} > 200$ GeV the momentum is taken as the one from the track with the best fit among tracker only, muon system only or combined track. The momentum resolution is better than 6% for all the p_{T} ranges and all the η values within the muons system acceptance [84].

2.3.3 Electrons and isolated photons reconstruction

Both electrons and photons are reconstructed from an energy deposit in the ECAL; electrons are required to have a track associated to the ECAL cluster while for photons the track is vetoed. The reconstruction of electrons and photons is complicated by the high probability of *bremsstrahlung* emission and photon conversion in the the tracker material upstream ECAL. An incoming electron losses a considerable fraction of its energy by *bremsstrahlung* emission before reaching the ECAL. Good energy resolution can be achieved only if those energy losses (and only them) can be recovered. For this reason, ECAL clusters in a small window in η and an extended window in φ around the electron or photon direction are grouped into 'superclusters'. The extended window in φ keeps into account the azimuthal bending of the electrons in the magnetic field.
The *bremsstrahlung* emission challenges also the electron tracking. When the energy loss is low, the standard tracking based on Kalman filter performs nicely also for electrons. If energetic photons are radiated, the electron changes its original trajectory and the Kalman filter reconstruction capability drops. Tracks with poor quality fit are reclustered based on a Gaussian Sum Filter (GSF) [85], which describes better the electron patterns. The dedicated tracking allows a better measurement of the momentum and the reconstruction of low energy electrons, down to $p_{\mathrm{T}} \approx 2$ GeV, a phase space unaccessible to the ECAL. An electron can thus be seeded both by an ECAL cluster or a GSF track. The ECAL-seeded

electrons are generally isolated high-momenta electrons, while GSF seeded electrons are mainly exploited for low p_{T} electrons or for electrons within jets, whose energy deposit in the ECAL is overlapped with contributions from other hadrons.
Electrons seeded by an ECAL cluster are identified by requiring one track linked to the ECAL supercluster. The track is refitted with the GSF approach and compatibility in the measurement of ECAL energy and of the track momentum is required. Discrimination from jets is performed by requiring the ratio between the energy deposited in the HCAL cells beyond the ECAL cluster and the cluster itself (H/E) to be less than 0.1. As for the muons, additional selections are generally applied on electrons used in data analyses.
Photon candidates are identified as ECAL clusters isolated from any track and from other energy deposit. In addition, the ECAL cells energy distribution and the ratio between the HCAL and ECAL energies must be compatible with those expected from a photon shower. Even in this case, PF selections are generally looser than the ones exploited in the data analyses.
The energy resolution for electrons ranges from 1.7 to 4.5% depending on the electron pseudorapidity and from the *bremsstrahlung* energy loss. Photon energy resolution is better than 1% for unconverted photons in the central region of the EB and better than 3.5% in the whole calorimeter. More details on the electron and photon reconstruction, as well as on the performance achieved in terms of energy resolution and reconstruction efficiency, can be found in Refs. [86] and [87], respectively.

2.3.4 Jets Reconstruction

Once electrons and photons have been reconstructed, clusters and tracks associated with them are removed from the collection. What remains at this stage is used to build charged hadrons, neutral hadrons and non-isolated photons. Those elements are clustered to create jets (or hadronically decaying τ leptons).
Within the acceptance of the tracker ($|\eta| < 2.5$), all the ECAL clusters not linked to any track are interpreted as photons, while the HCAL clusters are interpreted as neutral hadrons. Instead, energy deposits in the HCAL linked to a track are considered as charged hadrons. Beyond the tracker acceptance, all the deposits are interpreted either as photons or neutral hadrons.
The resulting collection of particles is clustered in jets with the anti-k_{T} [88, 89] algorithm, with a distance parameter R=0.4. This algorithm clusters neighbouring PF candidates, creating approximately conic jets centred around the most energetic particle. The momentum of the jet is determined as the vectorial sum of all the PF particles clustered in the jet. Jet Energy Corrections (JEC) are derived to match the energy scale observed in data to the one in simulation. The corrections are derived in simulation from the relation between the reconstructed jet energy and the simulated energy of the generated particles and are validated directly on data. The main effects corrected by the JEC are the PU, the mis-modelling of the detector response and the residual difference between data and simulation used to derive the correction. The typical energy resolution for jets in the central region of the barrel is about 20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV [90].

2.3.5 Missing transverse momentum reconstruction

The missing transverse momentum is computed as the negative sum of all the PF objets identified in the previous iterations of the algorithm. Its magnitude is generally referred to as $p_{\mathrm{T}}^{\mathrm{miss}}$. It measures the momentum of undetected particles, such as neutrinos, or of not reconstructed particles. The $p_{\mathrm{T}}^{\mathrm{miss}}$ plays a major role in many searches of unobserved particles, which are expected not to interact with matters, such as dark matter candidates. It is also relevant in the $\mathrm{t\bar{t}H}$ search, since it is generated by neutrinos following the W boson decay or by jets outside the detector acceptance. The $p_{\mathrm{T}}^{\mathrm{miss}}$ is extremely sensitive to reconstruction errors: if a particle is not reconstructed, it contributes to the missing transverse momentum of the event. Similarly, if some hits of the detector are wrongly clustered in a particle, missing transverse momentum arises in the opposite direction. Even the electronic noise can contribute to induce fake missing transverse momentum. To prevent all those effects, a set of corrections is specifically designed to reduce events with large fake $p_{\mathrm{T}}^{\mathrm{miss}}$ [91]. The typical resolution on the $p_{\mathrm{T}}^{\mathrm{miss}}$ ranges from 15 to 30 GeV, depending on the number of reconstructed vertices [91].

2.4 Data taking during Run II

A whole set of procedures is set up to ensure a high quality data taking of the CMS detector. While LHC delivers collisions, the detector is constantly monitored to promptly react to any possible inconvenience. Inevitably, sometimes the detector can experience hardware issues, such as subsystems temporarely out of order or electronics failures. When part of the detector is not in condition to take data, and only in that case, the data are discarded. The result is a recording efficiency, with respect to the total LHC delivered luminosity, lower than one. The right panel of Fig. 2.12 shows the amount of data delivered by the LHC and collected by CMS during Run II as a function of the time. The recording efficiency, integrated on the whole Run II, has been as high as 92%.

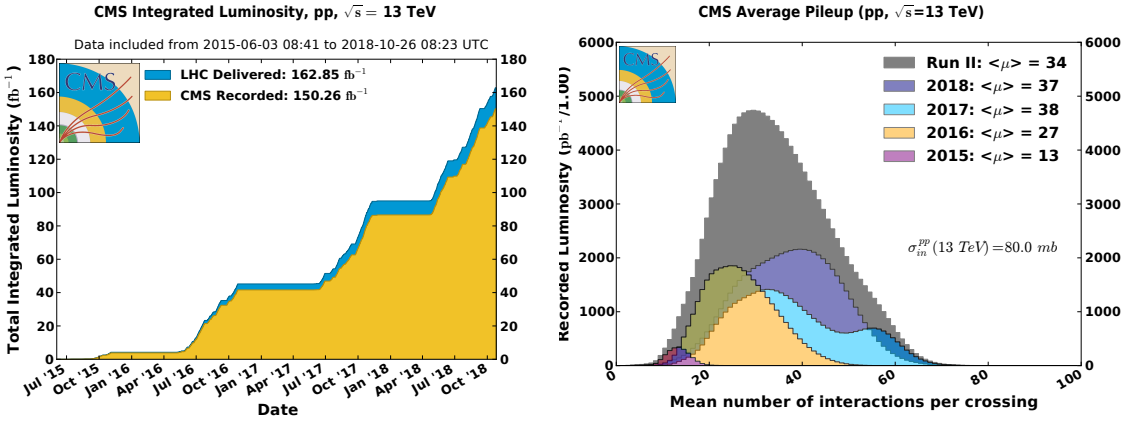


Figure 2.12: Left: luminosity delivered by the LHC and recorded by CMS during Run II as a function of time. The total delivered and recorded luminosity is also reported. Right: the distribution of the number of reconstructed interaction vertices per bunch crossing in the CMS experiment for each year of Run II. The grey histogram is the envelope of the full Run II. The top right corner displays the average pileup $\langle\mu\rangle$ for the full Run II and for each year of the data taking.

For operational reasons, the data taking during each year is further divided into different periods, summarised in Tables 2.2, 2.3 and 2.4. Once data are acquired, a prompt check is performed to ensure high quality, so that all the subsystems were active and that no issues are spotted in the data. The luminosity available for the physics analyses is therefore further reduced with respect to the recorded one. For Run II it corresponds to 35.9, 41.5 and 58.9 fb^{-1} in 2016, 2017 and 2018 respectively. The data collected in 2015, corresponding to 3.8 fb^{-1}, were mainly used for commissioning and are not exploited in this work.
As this thesis is going to completion, the analysis of 2018 data is still ongoing, therefore the results hereby presented involve only the 2016 and 2017 datasets.

Period	Beginning	Ending	Delivered $\mathcal{L}$ (fb^{-1})	Recorded $\mathcal{L}$ (fb^{-1})
2016B	28 Apr	20 Jun	6.3	5.8
2016C	24 Jun	4 Jul	3.3	2.9
2016D	4 Jul	14 Jul	4.7	4.4
2016E	15 Jul	25 Jul	4.7	4.3
2016F	30 Jul	13 Aug	3.5	3.2
2016G	14 Aug	9 Sep	8.8	8.0
2016H	16 Sep	26 Oct	10.3	9.5
2016	28 Apr	26 Oct	41.6	38.1

Table 2.2: CMS data taking periods during 2016 in proton-proton collisions. For each period the starting and ending dates, the total luminosity delivered by the LHC and the luminosity recorded by the CMS experiment are shown. The last line shows the values integrated over the whole year.

The LHC operational conditions largely varied during Run II, with the intent to keep increasing the machine performance. The result is a distribution of the number of pileup vertices different for each year, as shown in the right panel of Fig. 2.12. The 2016 data present an average number of 27 collisions per bunch crossing, which raises to 38 and 37 for 2017 and 2018 respectively. The 2017 data have a distribution with two peaks due to the reduction of the LHC bunches during the year, which forced to an increased pileup in order to avoid a reduction in the integrated luminosity.
The pileup affects the data reconstruction in two ways. The energy deposition due to the different collisions in the same bunch crossing is generally referred to as in-time pileup. Instead, the out-of-time (OOT) pileup is due to events in neighbouring bunch crossings and depends on the bunch spacing in the machine. The collision frequency is 25 ns and most of the electronics requires more than this amount of time to process a signal, for example a pulse in an ECAL crystal is reconstructed in 250 ns. The PU affects both the low-level reconstruction, where the amplitude of the electronic pulses is estimated, and the high-level event reconstruction performed by the PF. The best amplitude estimate performed on the signal pulse can receive a bias from the PU energy deposition while the PF algorithm must be robust against higher order effects induced by the PU.

Period	Beginning	Ending	Delivered $\mathcal{L}$ (fb^{-1})	Recorded $\mathcal{L}$ (fb^{-1})
2017B	16 Jun	18 Jul	6.5	5.5
2017C	18 Jul	30 Aug	12.4	10.8
2017D	30 Aug	17 Sep	5.0	4.6
2017E	21 Sep	12 Oct	10.5	9.8
2017F	13 Oct	11 Nov	15.7	14.4
2017	16 Jun	11 Nov	50.1	45.1

Table 2.3: CMS data taking periods during 2017 in proton-proton collisions. For each period the starting and ending dates, the total luminosity delivered by the LHC and the luminosity recorded by the CMS experiment are shown. The last line shows the values integrated over the whole year.

Period	Beginning	Ending	Delivered $\mathcal{L}$ (fb^{-1})	Recorded $\mathcal{L}$ (fb^{-1})
2018A	26 Apr	28 May	15.6	14.7
2018B	28 May	7 Jul	8.0	7.6
2018C	8 Jul	26 Jul	7.4	6.9
2018D	27 Jul	24 Oct	36.3	33.9
2018	26 Apr	24 Oct	67.3	63.1

Table 2.4: CMS data taking periods during 2018 in proton-proton collisions. For each period the starting and ending dates, the total luminosity delivered by the LHC and the luminosity recorded by the CMS experiment are shown. The last line shows the values integrated over the whole year.

Chapter 3

Detector performance

Quid lucri est homini
de universo labore suo, quo laborat sub sole?
Generatio praeterit, et generatio advenit,
terra autem in aeternum stat.

Liber Ecclesiastes

Excellent detector performance is the premise to achieve the ambitious experimental program of CMS. The highest standards in trigger efficiency, particle identification and reconstruction efficiency, robustness against pileup and energy resolution must be granted. Failure to do so would cause degraded physics results or, directly, the impossibility to perform the most challenging measurements. For these reasons, a considerable fraction of the work in the context of this thesis has been devoted to ensure optimal experimental conditions of CMS. Two main areas of contribution have been identified as the ones affecting more the $t\bar{t}H$ search, with $H \rightarrow \gamma\gamma$.
The first area is the calibration of the ECAL. As already mentioned in Section 1.5.1, a search in the diphoton decay channel of the Higgs boson is the search of a narrow peak in the diphoton invariant mass distribution over a continuum background due to non-resonant events. The sensitivity of the search is therefore directly related to the resolution of the invariant mass peak and, in turn, of the photon energy. Figure 3.1 shows the impact of the resolution on the most sensitive category of the 2017 $t\bar{t}H$, with $H \rightarrow \gamma\gamma$, data analysis. The signal-plus-background model is build by the sum of the signal model, derived from fitting the sum of two Gaussian functions on the $t\bar{t}H$ simulation, and a background model, an exponential function fitted on the data excluding the signal region $115 < m_{\gamma\gamma} < 135$ GeV. The left figure shows the signal plus background model as it is expected in the analysis, while the right one shows the same situation when the resolution of the photon is worsened by 10%. The signal peak in more difficult to discriminate from the background and the sensitivity of the analysis is lower, with the mean expected significance (defined in Section 4.5) reduced by 40%, from 1,4 to 0,85 standard deviations. The ECAL energy resolution is therefore a major actor in the diphoton channel and an effort in its calibration is mandatory to achieve a precise measurement of the Higgs boson properties.
The contribution to the ECAL calibration has been realised with the refinement and the application of the φ-symmetry method (described in Section 3.1.4) throughout the Run II data taking. The method has been adapted to work at the energy and the pileup achieved

during Run II and it has been refined to avoid loss in the precision of the calibration at the present ECAL conditions. Additionally, possible improvements of this method have been investigated targeting the LHC Run III.

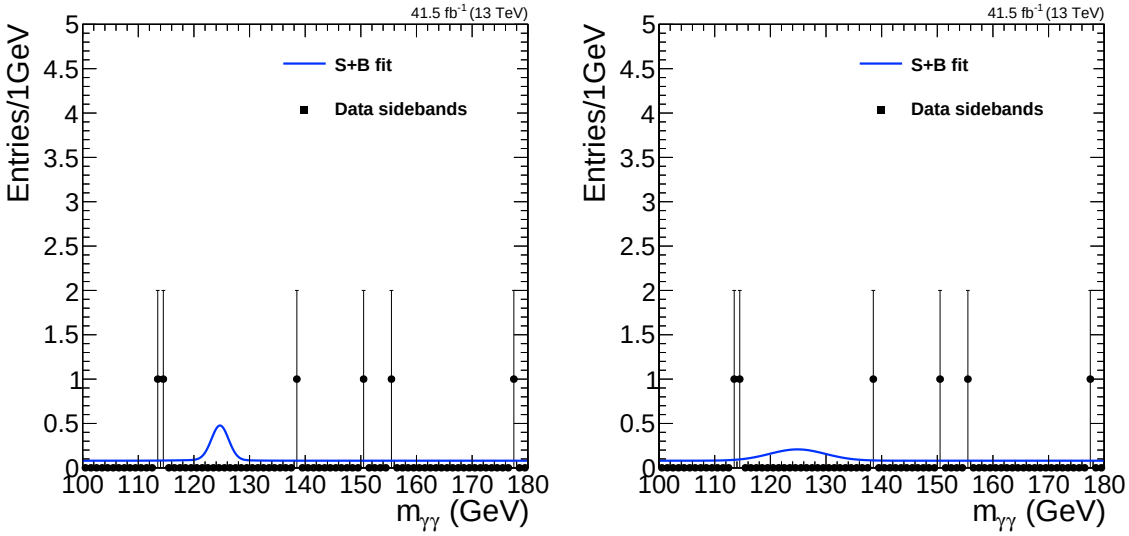


Figure 3.1: Impact of the photon energy resolution on the most sensitive category of the 2017 $t\bar{t}H$, with $H \rightarrow \gamma\gamma$, analysis. Black markers represents the data while the blue line is the signal plus background model. The left figure is realised with the energy resolution expected in the analysis, the right one with a photon energy resolution 10% worse.

The second area of interest has been the development of an efficient L1 trigger for electrons and photons. The upgrade of the trigger system performed during LS1 largely increased the trigger capabilities and the complexity of the algorithms running at this stage. The LHC ran with a peak luminosity of 10^{34} $cm^{-2}s^{-1}$ during 2016 and up to twice this value was expected (and achieved) for 2017. For the L1 trigger to sustain the doubled luminosity while avoiding a tremendous increase in the energy thresholds applied to trigger the events, a large effort was necessary to redesign the trigger algorithms and to fully exploit the possibilities of the upgraded trigger. The effort aimed at improving the trigger for electron and photons, exploited both for the $t\bar{t}H$ analysis and for the ECAL calibration.

3.1 The ECAL calibration

A description of the ECAL is provided in Section 2.2.2; this section illustrates the procedure exploited to ensure uniform response of the calorimeter. At first time evolution of the crystals response must be corrected. Radiation damage induces colour centres in the crystals, preventing light emission or trapping the emitted light and, in turn, reducing the transparency of the crystals. A laser system is specifically design to track the evolution of the transparency as a function of time. The second step is the equalisation of the about 75000 crystals of the calorimeter. The absolute energy scale is derived from the invariant mass peak of the $Z \rightarrow e^+e^-$, while four different methods are exploited to derive a per-crystal intercalibration constant (IC) that levels the response across the whole calorimeter.

The calibration of the detector directly affects the energy measurement of a crystal E_{ch}, which can be expressed as:

$$E_{\text{ch}} = IC_{\text{ch}} \cdot LM_{\text{ch}}(t) \cdot G \cdot A_{\text{ch}}, \tag{3.1}$$

where A_{ch} is the amplitude of the pulse reconstructed in the crystal expressed in counts of the ADC, G is the conversion factor from ADC counts to energy, $LM_{\text{ch}}(t)$ is the time dependent correction from the laser monitoring (LM) system and IC_{ch} is the intercalibration constant for that channel.
The precision achieved in the correction of the time dependent effect and of the crystals intercalibration directly influence the constant term of the resolution of the ECAL (see Eq. 2.6). For photons and electrons with $p_{\text{T}} \gtrsim 30$ GeV, as the photons following the decay of the Higgs boson, the energy-dependent terms of the resolution are subdominant with respect to the constant one. The effect of the ECAL calibration is therefore directly related with the resolution of the Higgs boson invariant mass peak, as shown in Fig. 3.2. The reconstructed invariant mass distribution in simulated $t\bar{t}H$, with $H \rightarrow \gamma\gamma$, events is shown. Events are required to have both the photons in EB, as this is the topology which provides the highest sensitivity to the analysis. The invariant mass is shown for events with the LM and IC corrections applied, with only the LM corrections and without any correction.The invariant mass resolution is degraded by 10% without the intercalibration constant and by a factor more than 5 by removing the LM correction. In addition the position of the peak is shifted by about 10% which is the average transparency loss of the EB.
The different terms of Eq. 3.1 are discussed in detail in the following sections; the energy reconstruction ($G \cdot A_{\text{ch}}$) in Section 3.1.1, the laser monitoring correction $LM_{\text{ch}}(t)$ in Section 3.1.2 and the intercalibrations IC_{ch} in Section 3.1.3. Section 3.1.4 describes the φ-symmetry calibration, which has been the method used and refined during this thesis.

3.1.1 Energy reconstruction

Electromagnetic showers crossing the ECAL produce scintillation light read by APDs in the barrel and VPTs in the endacp. The electrical signal of each sensor is amplified and digitalised by a multi-gain preamplifier coupled to three ADCs. The three gains, with gain factor 12, 6 and 1, ensure a good dynamic range to the electronic, with a coverage from 35 MeV to 1,7 TeV in EB and 2,8 TeV in EE. The pulse with the highest gain which does not saturate the dynamic range is acquired. Each pulse is digitalised in ten samples with a sampling frequency of 40 MHz, thus the signal is acquired in 250 ns, equivalent to ten LHC bunch crossings. The digitalisation is synchronised with the LHC such that the amplitude maximum for an energy deposit is on the fifth sample. The first three samples of each pulse are taken before a signal rise to provide a per-signal determination of the pedestal. The choice of the acquisition time is justified by the fast scintillation of the $PbWO_4$, with a decay time of 10 ns, and the 40 ns of the shaping time of the preamplifier. Therefore 250 ns of acquisition ensures the collection of the almost the whole pulse while granting a fast processing of the signal.
Since the reconstruction happens in the time of ten bunch crossings, the OOT pileup can affect the estimation of the energy deposited in the crystal. A *multifit* algorithm is specifically designed to suppress the contribution of the OOT pileup and to get an estimate of the in-time energy deposition. The amplitude is derived from a template

fit to the measured pulse, where the templates are the in-time signal and nine OOT signals. The OOT templates share the same exact shape but are shifted within the window [-5,+4] bunch crossings around the nominal interaction bunch, representing the energy deposited in each of the bunch crossings within the pulse reconstruction. A least-square fit is performed; the best fit determines how many templates are active and the relative contribution of each template to the measured pulse. The pileup contribution is subtracted from the pulse and the in-time component is assigned as the energy of the pulse. Figure 3.3 gives an example of pulse shape fits for two simulated pulses in EB and EE.
The template pulses are measured for each crystal directly on collision data, from randomly selected events saved by a dedicated trigger stream. To disentangle the in-time and OOT contributions, isolated bunches in the LHC are exploited. The LHC filling scheme features trains of bunches with 25 ns spacing and, additionally, few isolated bunches with much bigger spacing. During the collision of two isolated bunches, the OOT pileup contribution is zero, therefore the in-time pulse shape can be measured. The OOT templates have the same shape shifted by the appropriate number of bunch crossings. The templates are taken as the average pulse, weighted by its energy, of the selected events normalised to the same amplitude.
The *multifit* method has been introduced in 2015 and exploited during the full Run II data taking to cope with the increased pileup level and the reduced bunch spacing. The algorithm used in Run I, described in Ref. [92], was designed as a filter to optimally

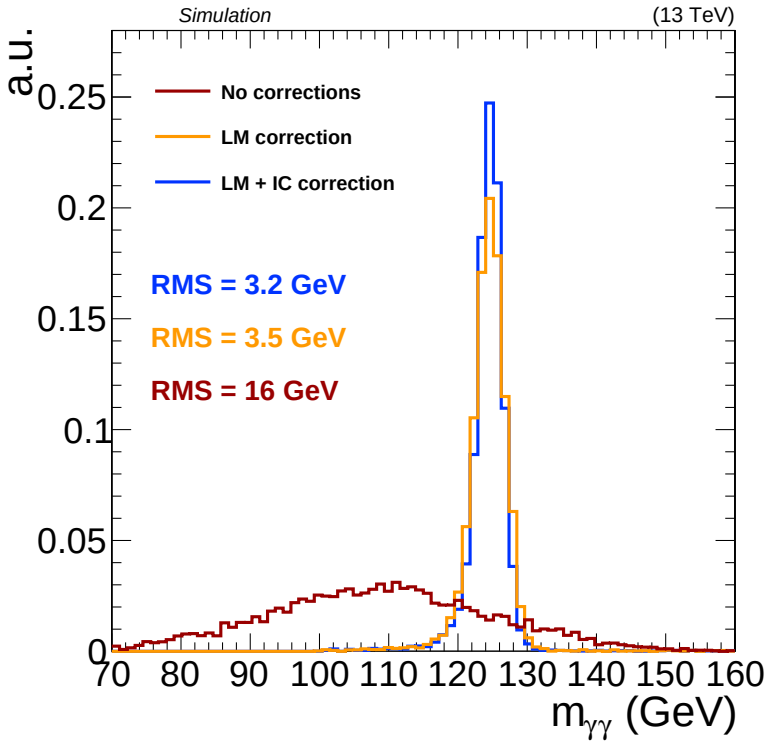


Figure 3.2: Reconstructed invariant mass peak in simulated $t\bar{t}H$, with $H \rightarrow \gamma\gamma$, events with both the photons in EB. The red histogram is obtained without any correction, the orange one with the LM corrections and the blue one with LM and IC corrections. The root mean square of the distributions is also indicated.

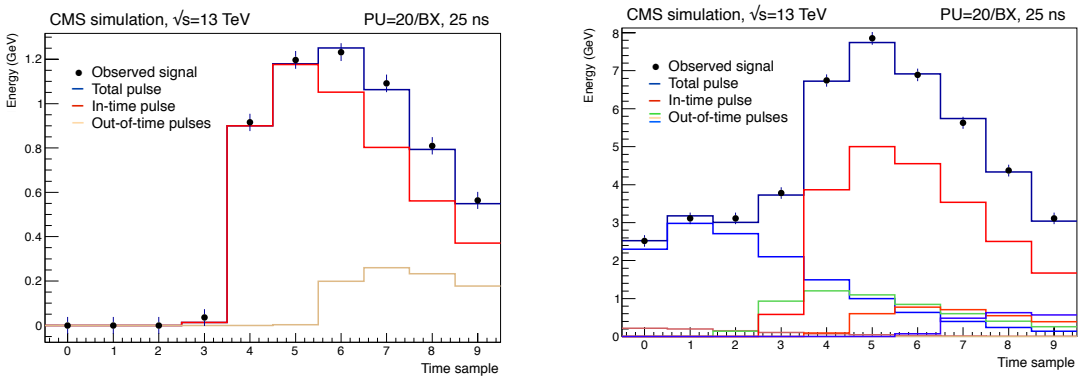


Figure 3.3: Example of the template fit performed by the *multifit* algorithm on a simulated pulse in the EB (left) and EE (right) with average PU=20 and 25 ns spacing between adjacent bunches. The digitalised signal is shown by black markers, the in-time contribution estimated by the *multifit* algorithm is shown in red, while the other lines represent the OOT templates. The dark blue line is the total reconstructed pulse given by the sum of the in-time and OOT contributions.

suppress the electronic noise. The improvement coming from the usage of the *multifit* algorithm is mainly on low energy pulses, where the impact of the pileup is more severe. A modest energy resolution improvement with respect the Run I method of 0.5% (1%) is observed in EB (EE) for electrons and photons with $p_{\mathrm{T}} > 25$ GeV, while the improvement is up to 10% for low p_{T} electrons and photons. The Run I algorithm is still exploited to reconstruct events digitalised with a gain one or six, since a small non-linearity of the gain-switch prevents from applying the *multifit*. For those kind of events, the OOT contribution is completely subdominant with respect to the in-time one, therefore no loss in the performance is observed.

3.1.2 Laser Monitoring system

As the LHC keeps running, CMS is subject to impressive levels of radiation dose. The ECAL crystals suffer from two kinds of radiation damage, the recoverable electromagnetic damage and the permanent damage due to hadronic interactions. Electromagnetic interactions move the electrons from the valence bands to the conduction one; the electrons tend to migrate back with a relaxation time typical of the crystal, causing a recovery of the induced damage [93]. The second kind of damage is due to hadrons interacting with the crystals. When nuclear reactions happen, the crystal is permanently modified and no recovery happens. The damages induce traps for the scintillation light, reducing in turn, the transparency of the crystals. If not corrected for, the signals from an APD for a fixed amount of energy deposited in a crystal get lower and lower as the damage increases.
A complex LM system has been designed to track the evolution of the crystal transparency in time [94]. Figure 3.4 shows the response to the laser monitoring system as a function of time along with to the luminosity delivered by the LHC. As the integrated luminosity increases, the response to the laser (or the crystal transparency) lowers. During LHC shutdowns, the transparency recovery is clearly visible. Since the radiation dose vary with the pseudorapidity, the damage of the crystals is not uniform across the calorimeter

but depends on the position of the crystals. At the end of 2018 the transparency loss, relative to a non irradiated crystal, amounted at about 10% in the EB, to 62% at $|\eta| = 2.5$, corresponding to the electron and photon acceptance.

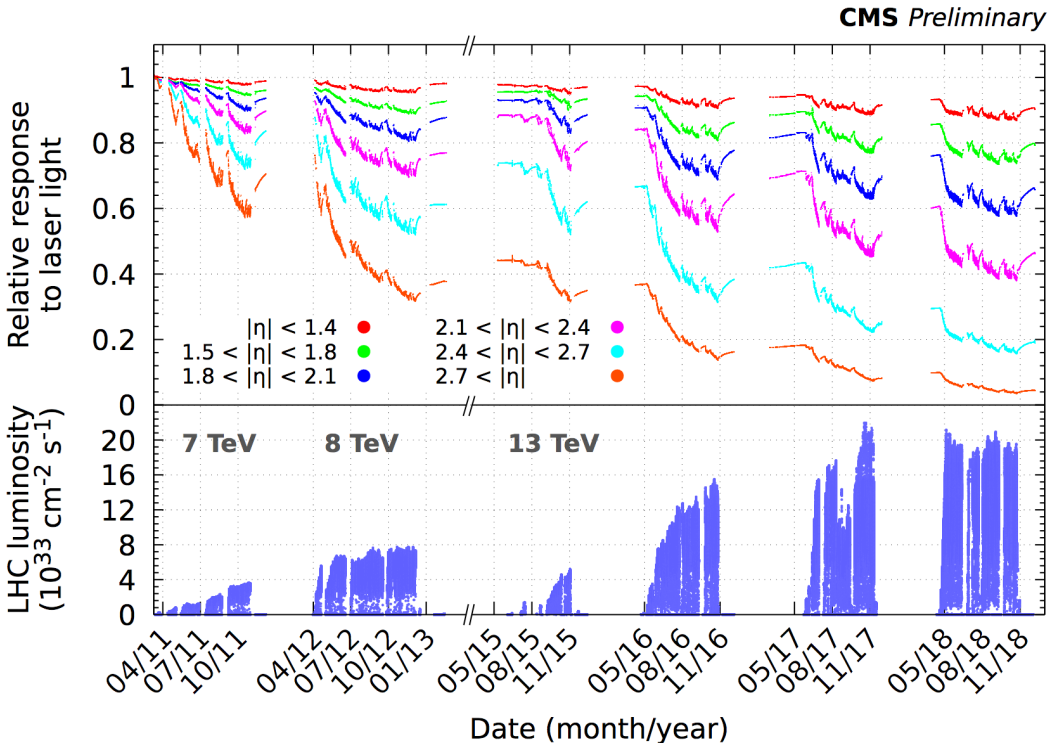


Figure 3.4: Relative response of the ECAL to the LM system since the beginning of the data taking in 2011. Each colour represents a different pseudorapidity region of the calorimeter. The bottom panel shows the peak luminosity reached by the LHC. The recovery of the radiation damage during LHC shutdowns is visible.

The working principle of the LM system is to measure the response to a known amount of light injected by a laser in each crystal at regular intervals in time. The laser injects light at 447 nm, a frequency close to the scintillation light, every 40 min. The light is channeled by a system of optical fibres directly in the crystals, in the front face in EB and in the rear of the crystals in EE. To prevent fluctuations in the amount of light injected by the laser to affect the calibration, the injected light is monitored by PN diodes. Crystals are grouped in 'harness regions', of 100 or 200 crystals, and each harness is monitored by the same PN diode. The laser light is injected in the PN diodes and, for each crystal, the response variation R is estimated from the ratio between the amplitude $\mathcal{A}$ of the APD signal and the amplitude of the signal of the PN diode:

$$R = \frac{\mathcal{A}\,(\mathrm{APD})}{\mathcal{A}\,(\mathrm{PN})} \tag{3.2}$$

The final ingredient to correct for the response of the crystals is the relation between the response to the laser light and to the scintillation one. The spectral composition and the path within the crystal are different for laser and scintillation, as the scintillation light is emitted isotropically while the laser light is injected from one edge. A correction factor

α relates the response to the scintillation light $S(t)$, normalised to the response at the beginning of the data taking S_0, to the response to the laser $R(t)$, again normalised at the response at the beginning of the data taking R_0 [95]:

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

The parameter α has been measured in tests beam before the data taking for the two configuration of fibres of EB and EE; typical value of α is 1.5 in EB and 1 in EE. Both measurements have been refined with data from the $Z \to e^+e^-$ peak.
The laser correction is validated directly with collision data exploiting the four methods described in Section 3.1.3. Figure 3.5 shows the stability of the invariant mass of the π^0 and of the Z peaks as a function of time in 2017 data. The stability of the Z invariant mass peak is also used as a figure of merit to fine tune the value of the α parameter.

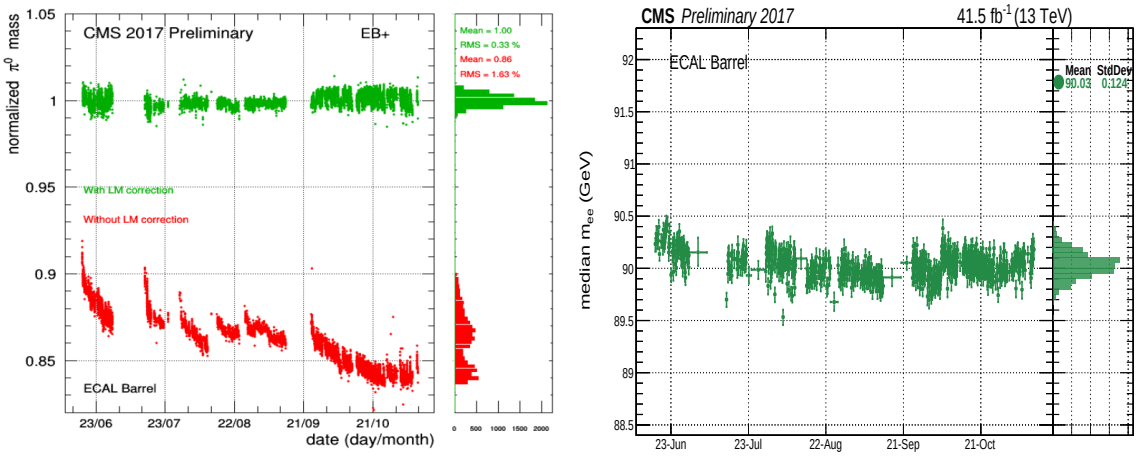


Figure 3.5: Left: Invariant mass distribution of the $\pi^0 \to \gamma\gamma$, normalised to unity, as a function of time in 2017 data. The red points are obtained without the LM correction, while the green ones with the LM correction applied. The right panel shows the normalised distribution of the mass peak. Right: median of the $Z \to e^+e^-$ invariant mass peak as a function of time in 2017 data. The right panel shows the distribution of the medians.

Since Run II a miscorrection of the LM system has been observed. The cause was identified in the damaging, due to radiation, of the reference diodes and 'harness corrections' were developed to take care of the problem. For each harness region, the energy scale is measured from the ratio of the energy to the momentum of candidate electrons in time bins corresponding to about one bin per LHC fill. The dependence of the scale as a function of time is linear as long as the LHC conditions are stable. The time dependence is modelled with a linear fit per LHC period and the outcome of the fit is used to correct the time dependence. The largest observed shift has been during the period 2017F, with a drift of the energy scale of about 1% in a month in EB.

3.1.3 Calibration of the ECAL

Once data are corrected for response variation with time, the derivation of the absolute energy scale and of the inter-crystal calibration starts. The IC constants are necessary

to equalise the crystal response over the whole detector. If the factorisation of the effects shown in Eq. 3.1 was exact, the IC would be constant with time and the value derived at the beginning of the data taking could be exploited at present. Actually, the corrections for the time response do not completely remove all the time dependent effects and the intercalibration constants are derived separately for each year of data taking. Four independent methods are exploited:

- invariant mass of $\pi^0 \to \gamma\gamma$. Events with two photons in a mass range compatible with a π^0 decay are selected by a dedicated trigger stream. The invariant mass peak is reconstructed and its position is derived from a fit. The IC value is determined from the ratio between the nominal mass of the π^0 and the measured mass. The ICs are derived with an iterative procedure. For each crystal the invariant mass distribution of the π^0 candidates is obtained from all the selected events for which one of the photons is centred on the crystal. The mass shift determines the value which is used to update the ICs. The procedure is repeated until convergence;

- ratio between the energy measured by the ECAL and the momentum measured by the tracker for electrons following a W boson decay (E/p). This method uses the tracker as a reference to calibrate the ECAL. Events with isolated electrons are selected and the ICs are derived iteratively from the deviation of the E/p distribution from unity;

- invariant mass of $Z \to e^+e^-$ events. The $Z \to e^+e^-$ can be exploited to determine the ICs in addition to the absolute scale. The ICs are derived in an iterative procedure from fitting the Z invariant mass spectrum;

- the φ-symmetry method. Described in detail in Section 3.1.4, it exploits the approximate azimuthal symmetry of the detector and of the energy flow in collisions selected with a random trigger.

The four methods are strongly complementary, each one with peculiar strengths and weaknesses. The E/p method has the highest precision in the whole EB, while in the EE the large material budget in front of the ECAL limits the precision of the intercalibration. The precision of the E/p method ranges from better than 0.5% in EB to 1.5% in EE. Since the production of Z bosons has a lower cross section than the production of W bosons, the $Z \to e^+e^-$has a worse precision than the E/p one in EB (from 0.5 to 1%, depending on the η), where both the methods are limited by statistical uncertainty, but it dominates in the EE, where the precision is of about 1%. In addition, it is the only method that can be used beyond the tracker extension ($|\eta| = 2.5$) exploiting events with one electron in EB and one in the EE. The π^0 ICs have a subdominant precision, about 1% in EB, but are completely independent from the tracker. Finally, the φ-symmetry has an even lower precision (from 2% to 5%, depending of the η) but it has a very prompt processing, which allows the fast monitoring of the detector conditions and prompt ICs derivation.
The ICs derived from all the methods are then normalised so that in each η-ring, namely a ring of crystals running in φ at a constant η, the mean value of the ICs is one. For each crystal the final value of the IC is derived as the weighted average of the different methods. The weight is the precision of each method, measured as a function of the pseudorapidity from cross-comparing the different methods. The final precision of the ICs is better than 0.5% in EB and better 1% in EE.

After the derivation of the ICs, the absolute energy scale is measured from $Z \to e^+e^-$ events. The scale is derived for each η-ring equalising the Z mass fitted in data to the value expected for a Z boson.
During the data taking, a prompt reconstruction is performed contemporary to the acquisition. The conditions used for this reconstruction are defined at the beginning of the year, while the LM correction is evaluated during the collision data and it is applied to the data with 48 h delay. At the end of each year, a preliminary calibration is performed using the data collected during the whole year. This calibration is intended to be used for preliminary physics results or for analyses not affected by the ECAL energy resolution. At the end of the Run II, a recalibration campaign has started with the aim to provide the best possible resolution to the legacy results to be produced with the Run II data. It will impact all the measurements involving electrons and photons in the final state, as in the $H \to \gamma\gamma$ case. The left panel of 3.6 shows the comparison of the energy resolution of electrons as a function of the pseudorapidity between the preliminary calibration and the refined one for 2017 data, derived with $Z \to e^+e^-$ events. The electron resolution is improved of from 10% to 20% on the full pseudorapidity range, mainly for to the improved correction of the time dependent effects. The right panel of the same figure shows the expected impact on the energy resolution of photons following the decay of the Higgs boson. The per-photon resolution is estimated from the width if the invariant mass peak of simulated $H \to \gamma\gamma$ events, including all the production processes weighted for the respective cross section.

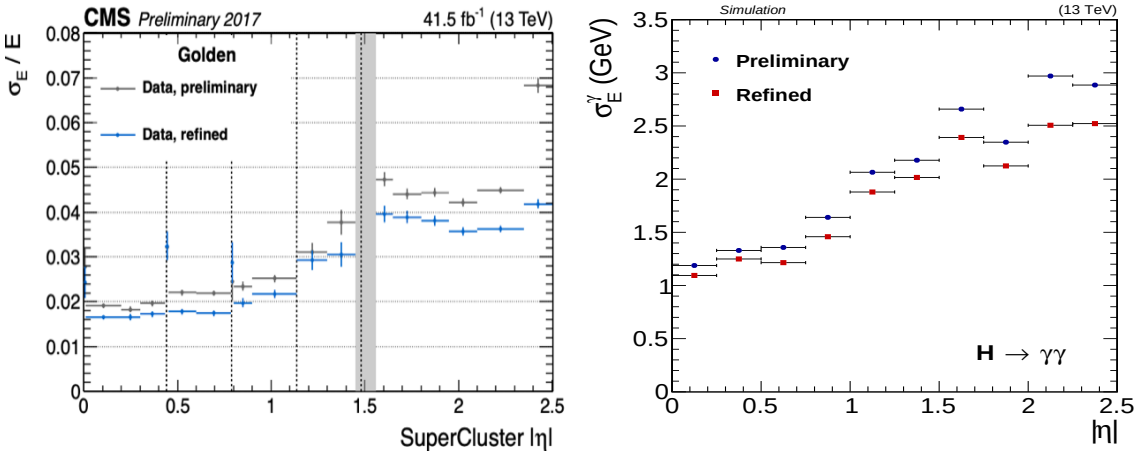


Figure 3.6: The relative electron energy resolution σ_E/E with the preliminary (black markers) and refined calibration (blue markers) is shown, measured on electrons following $Z \to e^+e^-$ decays, as a function of the pseudorapidity of the electron. The resolution is affected by the amount of material in front of the ECAL and is significantly worse near the gaps between ECAL modules, shown by vertical lines. Right: expected photon energy resolution with the preliminary (blue markers) and refined (red markers) calibration on simulated $H \to \gamma\gamma$ events as a function of the pseudorapidity of the photon.

The work devoted to the ECAL during this thesis had two main purposes. During data taking the main effort was the prompt monitor of the ECAL with the φ-symmetry method. After the end of Run II the effort was dedicated to the recalibration of the data. At the time of writing, the 2017 data have been recalibrated, while the processing with the new

calibration constants of 2016 and 2018 data is still ongoing.

3.1.4 The φ-symmetry calibration

The φ-symmetry method exploits the azimuthal symmetry of the detector and of the energy flow to correct inter-crystals response differences. At a fixed pseudorapidity the average energy deposition is independent from the azimuthal angle for symmetry reasons and this invariance is employed to derive the intercalibration constants.
The average energy deposition of each η-ring is used as a reference to correct for the response variation. For the ch$^{\text{th}}$ crystal in the t$^{\text{th}}$ time bin, the correction is derived as:

$$\text{IC}^{\text{ch, t}} = \left(\frac{\sum_{\text{t}} E_{\text{T}}^{\text{ch}}}{< E_{\text{T}}^{\text{ring, t}} >} \cdot \frac{1}{\kappa} \right)^{-1} . \tag{3.4}$$

The numerator is the sum of the transverse energy deposited in the ch crystal during the t time bin. The denominator is the average energy deposited in the same time bin in the η-ring, which the ch crystal belongs to. The κ-factor is a correction necessary to compensate the fixed energy threshold applied to select the events and will be described below in the text. For a detector with uniform response the value IC$^{\text{ch, t}}$ is expected to be one. Each deviation from the unity is interpreted as a variation in the crystal response and it is used as intercalibration constant. The time bins are chosen as a function of the number of events recorded by the φ-symmetry trigger. The binning is defined such that in each bin the statistical uncertainty on the IC value is smaller than the systematic one, while retaining the time granularity as high as possible. At the typical luminosity of LHC Run II, a time granularity of about one time bin per LHC fill ($\approx$ 10 h) is achieved.
The events used for this calibration are randomly selected by a dedicated trigger stream. The symmetry argument is valid for all the physics processes, as long as a large sample of events is collected. As described is Section 2.2.2, the proton-proton cross section is orders of magnitude higher than the one of every process relevant for the LHC physics program. Most of the interactions will produce just low energy deposits in the detector which can be exploited for the calibration. The randomness of the trigger prevents from inducing a bias in the choice of the events due to non-uniformity of the material along φ, which could introduce systematic effects in the derivation of the ICs. The energy spectrum of the events used for the calibration is shown in Fig. 3.7 for EB (left) and EE (right).
Events selected at the L1 trigger are sent to the HLT for processing. The HLT algorithm selects events with at least one crystal where the energy deposited is above a given threshold, set according to the electronic noise. The HLT trigger is 100% efficient with respect to the L1 selection, in other words there is always at least one crystal above the threshold. A dedicated format is employed to store φ-symmetry data. Instead of saving the information of the full CMS, only the energy of the ECAL crystals passing the HLT threshold is stored. The resulting event size is about a factor 1000 smaller than the usual one (about 2.5 kB per event instead of a few MB) and a much larger sample can be acquired and recorded, with a trigger rate of about 2 kHz during the data taking, to be compared with the less than 1 kHz of the usual trigger stream.
The thresholds are set to prevent the electronic noise from entering the calibration. Since the noise changes as the radiation damage increases, the thresholds are re-derived at the beginning of each year of data taking. An example of the evolution of the noise is provided in Fig. 3.8 for the 2017 data taking. The APD noise grows as the radiation damage

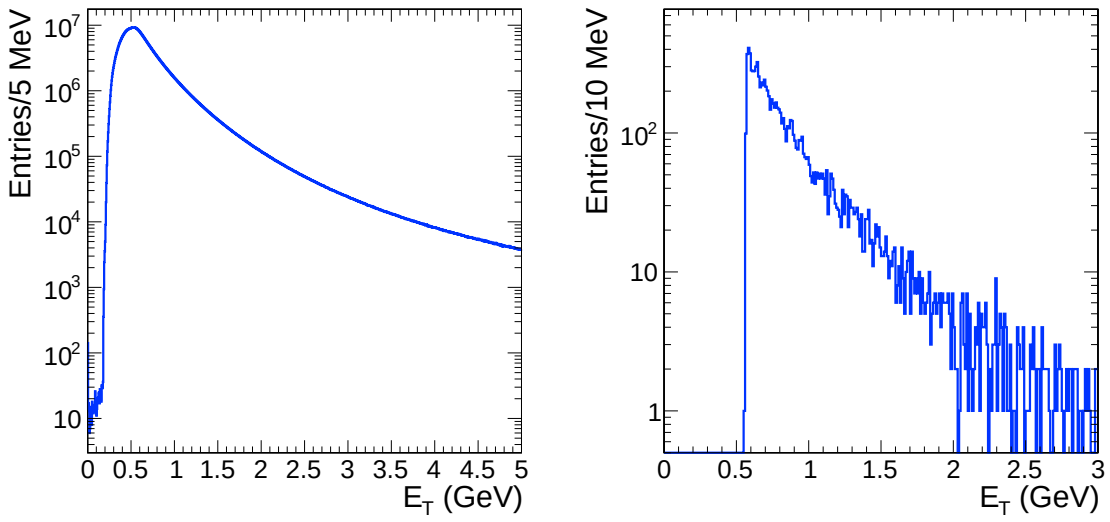


Figure 3.7: Distributrion of the per-crystal energy deposit of the events used for the φ-symmetry calibration in EB (left) and EE (right).

increases the dark current of the sensors. Instead, the VPT technology is more resistant to the radiation damage and no increase in the noise is observed. The effect of the electronic noise is further amplified by the transparency loss. The equivalent noise in energy units depends on the transparency of the crystals, as the energy necessary to produce a signal equivalent to the noise depends on the crystal transparency. As a consequence, the noise is dependent on the radiation damage and, in turn, on the pseudorapidity. In the very forward region, $2.5 < |\eta| < 3$, the noise forces the thresholds to values of a few GeV, preventing the application of the method as not enough events are collected.
The energy thresholds for the φ-symmetry trigger are set at the beginning of each year according to the foreseen noise at the end of the year. The thresholds applied at the HLT is set to seven times the RMS of the expected noise and it is computed independently for each η-ring. Events passing the trigger selection are further selected before calibrating the ECAL by applying an 'offline threshold' of ten times the RMS of the noise. The difference of the two selections is necessary for the computation of the κ-factors, as described below. An example of the threshold applied to select the events on 2018 data is shown in Fig. 3.9. The dependence on the pseudorapidity is clear, and it is very severe close to the beam pipe where the radiation damage is higher, with a crystal transparency reduced to less than 5% of the transparency of a new crystal.
An upper threshold is additionally applied to reject sporadic high energy events selected by the random trigger. As high energy events are rare compared to the low energy deposition used for the calibration, they would not represent a uniform sample in φ, inducing a bias in the measurement of the response variation. The upper threshold is set for each eta ring such that the energy window selected for the calibration is 1 GeV in E_{T}.
The application of a fixed energy window forces the usage of a correction called κ-factor. Events with an energy close to the window boundary are shifted by the presence of a mis-calibration and might fall outside the acceptance window. Figure 3.10 illustrates how the shift of the energy due to a mis-calibration influences the energy deposited in the acceptance window. As a consequence, the measured mis-calibration is bigger than the

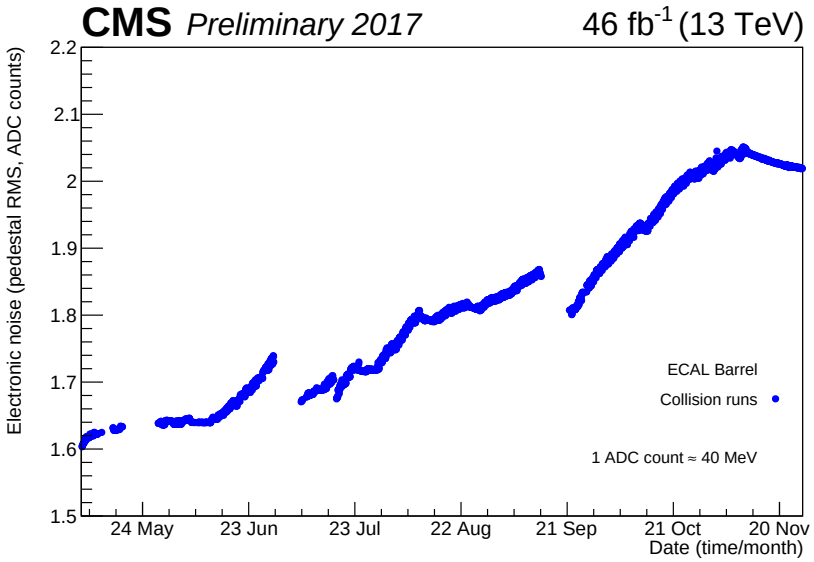


Figure 3.8: Evolution of the pedestal in EB during 2017 data taking. The y-axis shows the pedestal in ACD counts. The pedestal increases with the radiation damage.

one which is actually present in the detector and the κ-factor are necessary to compensate the effect. The κ-factors are computed from injecting a set of know mis-calibrations in the data and re-deriving them from Eq. 3.4. The derived miscalibration is fitted with a line as a function of the injected one, the slope being the κ-factor. An example of the fit for a crystal in EB is shown in Fig. 3.11. Typically κ-factors in EB have values around 2.
The φ-symmetry method is exploited for the monitoring of the ECAL conditions and to derive the values of the ICs, as described in the following sections. Finally the limitation of this methods and possible developments to exploit it in Run III and beyond are discussed.

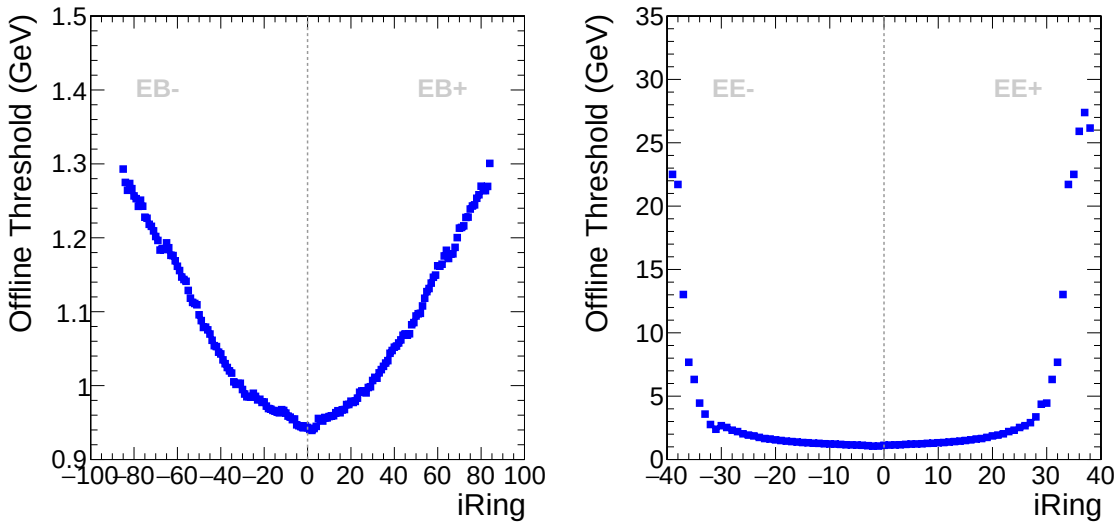


Figure 3.9: Energy thresholds used on 2018 data to select the events entering the φ-symmetry calibration as a function of the ring number for EB (left) and EE (right).

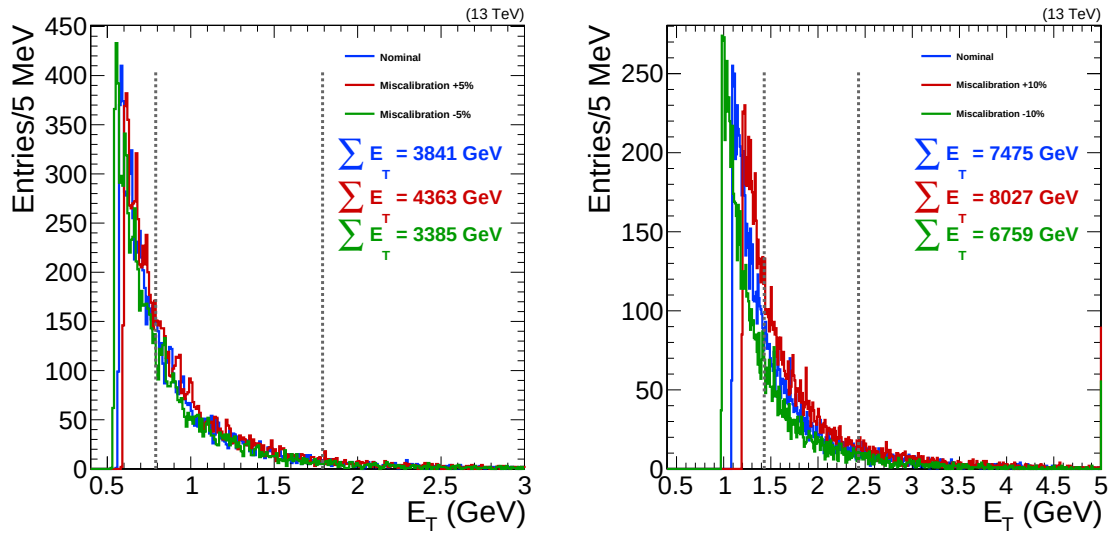


Figure 3.10: Energy spectrum of the events selected by the φ-symmetry trigger in an example crystal in the centre of EB (left) and EE (right). The blue histogram corresponds to the measured energy deposition, the red and green ones are obtained from the blue histogram by injecting a miscalibration of $\pm 5\%$ in EB and $\pm 10\%$ in EE. Vertical lines represent the energy window selected for the calibration in that crystal. For each histogram, the sum of the energy deposited between the lines is also reported.

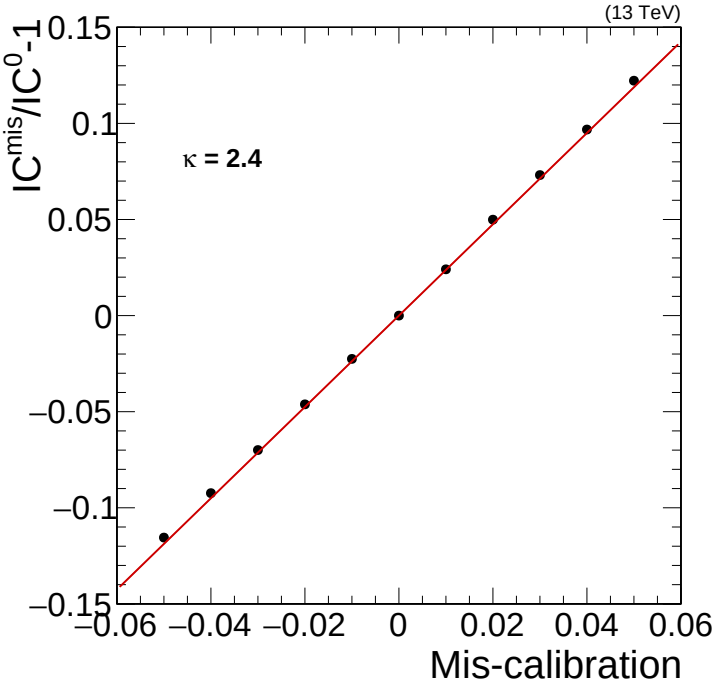


Figure 3.11: Example of the linear fit exploited to derive the κ-factor in a crystal in EB. The variation of the value of the IC relative the the one with 0 mis-calibration injected is shown as a function of the injected mis-calibration. The slope of the line is the κ-factor, 2.4 for this crystal.

The ECAL monitoring

The φ-symmetry method can be used throughout the data taking as a prompt tool to monitor the conditions of the ECAL. It provides both a per-crystal information on the time evolution of the response and a simple way to ascertain the quality of the whole ECAL calibration at a glance.
For each crystal, the IC computed according to Eq. 3.4 provides a direct access to the crystal response in a given time bin. The evolution of the IC for the same crystal during the data taking tracks the crystal response evolution. If the radiation damage is properly corrected, the IC value should be constant throughout the year. The ratio between the IC in a time bin to the IC of the first bin of each year, is used as a monitoring variable. Any deviation of this ratio from unity indicates a potential miscorrection to be addressed and investigated. The per-crystal tracking of the evolution is a unique feature of the φ-symmetry method which is used for anomaly detection, as shown in Fig. 3.12. The left panel shows the response of a crystal whose transparency loss is properly corrected by the LM system, while the right one shows a crystal with an under-correction of the transparency loss.

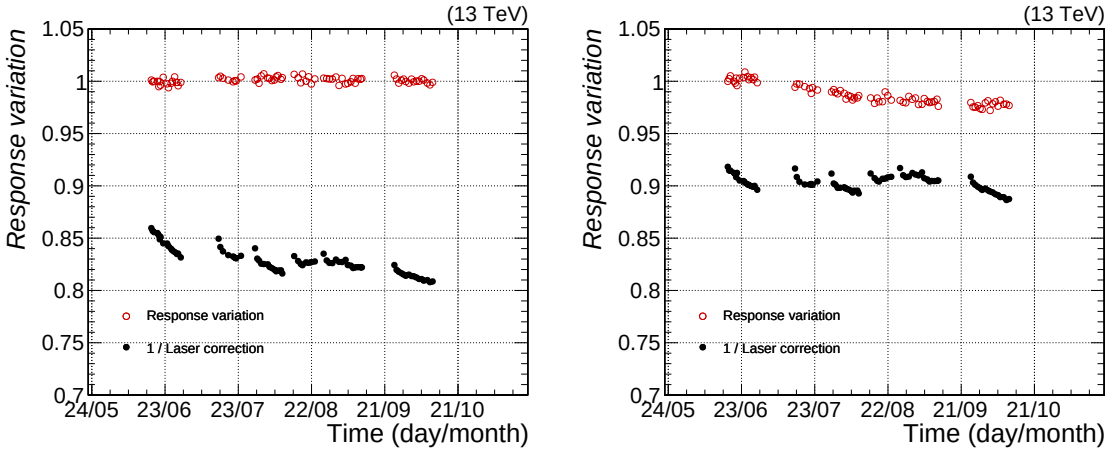


Figure 3.12: Crystal response (red markers), measured with the φ-symmetry method, as a function of the time for two crystals in EB. The left panel shows a crystal whose transparency loss is properly corrected for by the LM correction (black markers), the right one a crystal for which the LM system is under-correcting the response.

The ICs ratios evolution versus time allows the monitoring of the response of a single crystal in time, but gives no information on the general behaviour of the ECAL. The full detector (or a region of it) can be monitored from the RMS of the distribution of the ICs ratios for all the crystals versus time. With a perfectly uniform response of the whole detector the ratio would be 1 for each crystal, the RMS being zero. The wider is the spread, the worse is the uniformity of the response. Figure 3.13 shows on the left panel the ICs ratios distribution for two time bins, at the beginning and at the end of the 2017 data-taking. The right panel shows the evolution of the RMS of the ICs ratios of the ECAL barrel in time. If the radiation damage is completely corrected for, no time evolution of the detector should be visible and the spread should be constant as a function of time. For an ideal detector it should be at the level of the precision of the method. The precision is

estimated from randomly splitting the dataset in two and computing the RMS of each subset independently. The difference between the two computed RMSs, rescaled by a factor $\sqrt{2}$ to account for the halved number of events, is taken as lower limit of the method. The growth of the RMS in time is due to the miscorrection of the radiation-induced damage on the detector; the larger the variation, the worse the degradation. During the data taking, this variable is constantly monitored to check the stability of the ECAL conditions. A sudden change in the slope of the RMS growth is an indication of changing of the ECAL conditions which require immediate action to understand and solve the effect, restoring the data quality of the ECAL data. The recalibration of the detector partially corrected the time-dependent effects due to the miscorrection of the radiation damage, reducing in turn the growth of the ICs spread. Yet, the growth of the ICs spread is still a factor 3 bigger with respect to the beginning of the year. At present no other corrections are applied to correct for the time evolution. A preliminary study to exploit the φ-symmetry per-crystal monitoring to mitigate the time-dependent effects has been performed and it is discussed later in this document.

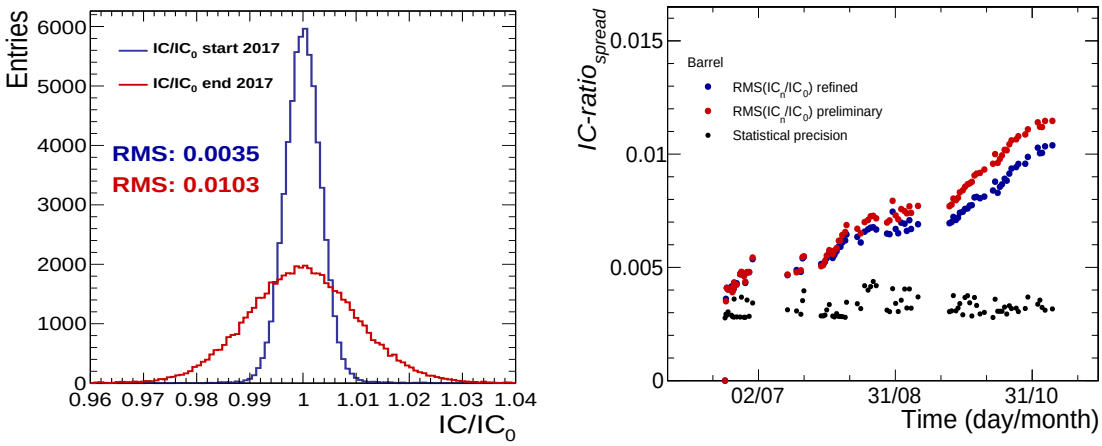


Figure 3.13: Left: distribution of the ICs of all the crystals of EB in 2017 data, normalised to the first time bin of the year, is shown for two time bins at the beginning (blue) and at the end (red) of the 2017 data taking. The RMS of the two distributions is also indicated. Right: RMS of the ICs in 2017, normalised to the first time bin of the year, as a function of the time. Red points are obtained with the preliminary ECAL calibration, blue ones with the refined one. Black points show the statistical precision of the method.

Derivation of the ICs

The intercalibration constants are derived independently for each year of operation in a single time bin. The values of the ICs are computed from integrating Eq . 3.4 over the whole year. Figure 3.14 shows a map of the ICs as a function of the crystals local coordinates (iη, iφ), defined as the number of the crystal in the η and φ direction. Several periodic structures appear due to the material upstream the ECAL and to the gaps between the crystals. The upper left figure illustrates the raw ICs values obtained from Eq. 3.4. Several lines running along the η direction are clearly visible. The services of the tracker (cooling, mechanical structures, electronics) runs along η from the centre of the

detector to the gaps between barrel and endcap, inducing a non-uniformity in φ in the amount of material in front of the ECAL. The particles travelling in regions with higher material budget deposits systematically more energy in the tracker and thus, at that particular azimuthal angles, the energy spectrum measured by the ECAL is slightly softer than in the other directions. The result is a bias in the value of the IC, which compensate for the different energy absorption in the tracker. In addition, the crystals surrounding the modules edges tends to collect more energy for the quasi-projective geometry of the ECAL.
In the barrel, the material of the tracker services is uniform along the η direction, therefore a correction can be derived exploiting the symmetry between the tracker and the ECAL mechanics. For each value of iφ, the average IC over the η direction $< IC_{\mathrm{i}\eta} >$ is computed. The correction is then applied as:

$$C_{\mathrm{i}\varphi} = \frac{1}{< IC_{\mathrm{i}\eta} >}. \tag{3.5}$$

The correction map is shown in the upper right panel of Fig. 3.14. Corrections are derived separately for the positive and negative η region of the EB and for the EB part facing the TB and the TE. The ICs corrected for the material budget are shown in the bottom panel of the same figure. The ICs are, as expected, generally close to unity with few structures arising. The red squares represent dead channels. The dark blue region corresponds to a miscorrection of the laser system, which is compensated by the ICs. The outer region of the EB has a worse precision compared to the inner one, due to the higher amount of material in front of that part of the ECAL, resulting in larger fluctuations of the ICs.
Figure 3.15 represents the maps of ICs values for the EE as a function of the crystals local coordinates (ix, iy). Even in this case some regions with suboptimal corrections of the LM system appear. Structures arising from the tracker and the ES are visible. The EE crystals are arranged in a $x - y$ grid, while the inner detectors features a $\eta - \varphi$ geometry, preventing from applying a simple correction as in the EB. Moreover, some support structures introduce variations with pattern difficult to identify. As a result, it has not been possible to find any reasonable material correction in the EE and thus the ICs derived with the φ-symmetry method can not be exploited in this region of the ECAL. The effect of the material does not impact the monitoring variable, as the material is constant as a function of time and its effect is cancelled out when normalising to the first bin of the year.
The ICs derived with the φ-symmetry method suffer from a systematic limitation due to the non-uniformity along φ of the material upstream the ECAL. The limited precision of the method, about 4 times less precise than the E/p in EB, makes the impact of the ICs subdominant with respect to the other methods. The ICs derived with the φ-symmetry are presently exploited as a cross-check of the more precise intercalibration methods and as a prompt tool to monitor the conditions of the ECAL.

Limitation and future of the φ-symmetry method

As the LHC keeps running, the evolution of the noise with the radiation damage will push the energy thresholds applied to select the events towards higher values. Since the energy spectrum is steeply falling, as shown in Fig. 3.7, the events available for the calibration are inevitably doomed to vanish in the near future, reducing the precision of the method

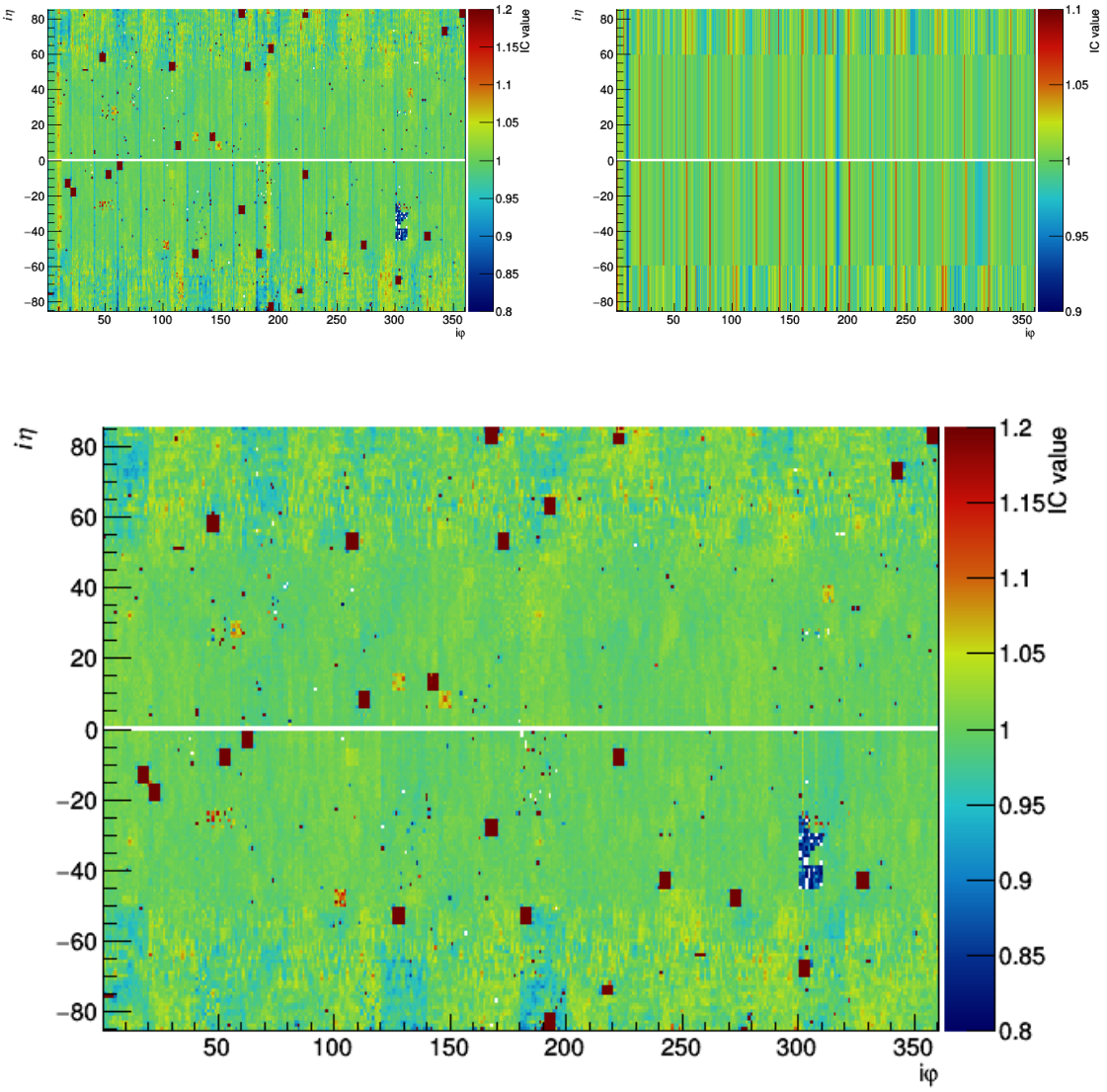


Figure 3.14: Maps of the ICs (z axis) in the EB as a function of the crystals local coordinates (iη, iφ). The upper left figure correspond to the ICs without correcting for the material non-uniformity upstream the ECAL, the upper right one represents the derived correction while the bottom one shows the corrected ICs values.

or even preventing its application.
Alternative possibilities for the usage of the φ-symmetry calibration are being investigated, to mitigate the loss of events and to provide a useful input to the ECAL calibration in view of Run III and beyond.
The first problem to be addressed is the drastic reduction of the number of events as the energy threshold is increased. At $\eta = 3$, as showed in Fig. 3.9, the energy threshold was already beyond 20 GeV during 2018, preventing from the application of the method. Since the noise can not be reduced nor the energy spectrum can be changed, the only viable way is to increase the trigger rate. The output rate of the HLT can be increased without any

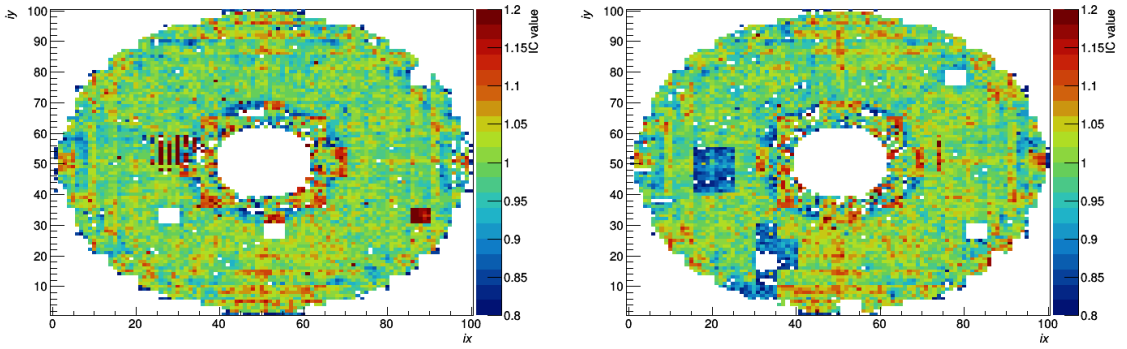


Figure 3.15: Maps of the ICs (z axis) in the EE as a function of the crystals local coordinates (ix, iy).

impact for the experiment, thanks to the small event size of the φ-symmetry output. The strongest limitation comes from the rate that the L1 can devote to random triggers and which can not be increased without reducing the one allocated for physics analyses. If no bias in φ is induced, potentially any L1 seed could be included in the φ-symmetry stream. Events triggered by electrons and photons can not be exploited, as the change in transparency of the crystals can induce a time-dependent bias in selection of the events used for the φ-symmetry intercalibration. On the other hand, events triggered at L1 by a muon does not involve triggering on ECAL objects. In such kind of events low energy deposits coming from pileup interactions useful for the calibration are not correlated with the direction of the triggered muon and could provide an unbiased samples of events for the φ-symmetry method. The feasibility study is started contemporary to the writing of this thesis and is going to assess the absence of any bias and the performance improvement in terms of ICs precision. Based on the L1 trigger rate of Run II, this approach would increase the φ-symmetry trigger rate by a factor up to 5.
As the growing levels of radiations keep damaging the diodes used as a reference by the LM system, the miscorrection of the crystal transparency is expected to worsen. At present the harness corrections (described in Section 3.1.2) are used to track the evolution of each diode. A similar approach can be applied with the φ-symmetry calibration, exploiting the capability of the method to track the response variation of each individual crystal. The response variation measured for each crystal, as in Fig. 3.12, can be used to mitigate the mis-correction of the LM system. The φ-symmetry method can provide a per-crystal correction as a function of time, while none of the other methods has a sufficient granularity to do the same. The per-crystal correction can be used instead of the harness correction or on top of it, to achieve a uniform response of the whole ECAL as a function of time. A first test has been performed on a part of the 2017 data. The resolution, measured from the invariant mass peak of $Z \to e^+e^-$events, showed a mild improvement of 0.3% in EB, suggesting a new interesting approach to the calibration. The impact of combining φ-symmetry and harness correction is being evaluated as the thesis is growing, to potentially benefit the Run III calibration.

3.2 The L1 electron and photon trigger

The luminosity delivered by the LHC increased by a factor 2 between 2016 and 2017; most of the trigger algorithms required a significant restyling to withstand the new conditions without affecting the physics program. Generally L1 seeds are built requiring a minimum number of reconstructed objects above given p_{T} thresholds. The full potential of the upgraded L1 system was necessary to avoid a large increase of the thresholds while retaining on acceptable trigger rate despite the increased PU.
The electron and photon L1 objects, generally referred to as EGamma (EG) candidates, are particularly sensitive to the PU energy deposition, as clusters from PU deposits can easily be misidentified as EG candidates. The separation of EGamma from clustered jet energy is particularly challenging as the L1 trigger has not access to tracking and the calorimeter information is delivered at a reduced granularity. Electrons and photons reconstructed offline (on recorded data where the full information is available) are identified exploiting several variables built on the energy deposition in the ECAL crystals and on tracks (see Section 4.3.1 and 4.3.3 for photons and electrons, respectively), which can not be exploited for the L1 algorithm. Additionally, as no track information is available, electrons and photons can not be distinguished.
For the 2016 algorithm to sustain a luminosity of $\mathcal{L} = 2 \times 10^{34}$ cm^{-2}s^{-1} within the allocated rate budget, the p_{T} threshold applied to select events with a single electron would have been of 42 GeV, largely reducing the acceptance of many physics channels as well as the acceptance for electrons following the decay of a W boson used for the ECAL calibration. The reduced size of the dataset for the ECAL calibration would have impacted the precision achieved in the intercalibration and, in turn, would have reduced the sensitivity of the $\mathrm{H} \rightarrow \gamma\gamma$ analysis.
In the following sections it is described the L1 trigger architecture (Section 3.2.1), the logic of the EG trigger (Section 3.2.2), the optimisation of the selections and the improvement achieved thanks to the work performed within this thesis (Section 3.2.3).

3.2.1 Architecture of the L1 trigger

The architecture of the L1 trigger is illustrated in Fig. 3.16. The information from the calorimeters is processed separately from the one of the muon detectors. The output of the two systems is combined in the global trigger which is responsible for accepting or rejecting the event.
The information is processed in Field-Programmable Gate Arrays (FPGAs), powerful electronic chips which can be remotely configured through a hardware description language. The programmability of the FPGAs grants a large flexibility in the system necessary to adapt it to the different LHC conditions, while their large processing power allows the implementation of sophisticated trigger algorithms. The communications between the boards are ensured by 10 Gb/s optical links.
The calorimeter trigger receives as input the energy deposited in the ECAL and HCAL with the granularity of a TT. For each event the TTs of the two detectors are read by 18 electronic boards of the Layer 1 system. At this stage, pre-processing operations are performed, such as the computation of the energy of each trigger tower, the energy calibration and the timing organisation of the data. The information are sent to one of the 9 processing nodes of the Layer 2 calorimeter trigger, where the reconstruction and identification algorithms are implemented. An additional node is present as backup in

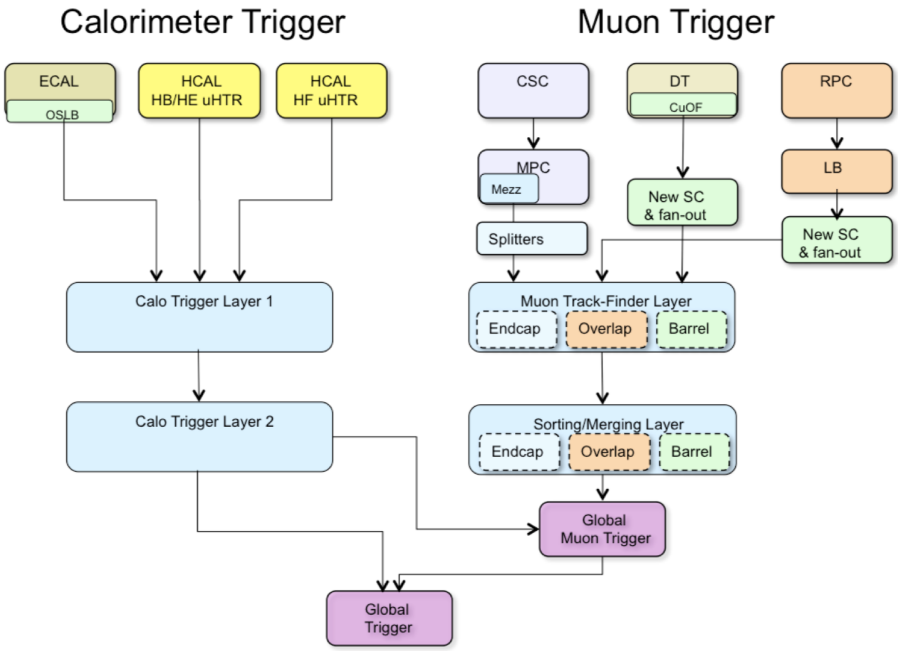


Figure 3.16: Scheme of the architecture of the L1 trigger system. The information from calorimeters and muon detectors is processed separately and combined in the global trigger, where the decision on the acceptance or rejection of the event is taken [80].

case of failure of any other node. The output of the Layer 2, after a reorganisation of the data, is sent to the global trigger for a decision on the event.
The muon trigger is based on the reconstruction of tracks primitives in the muon system. The information of the CSCs, DTs and RPCs, after some pre-processing, is sent to a track-finder layer. The layer is organised in three sections, each covering a different region in η and reflecting the different detectors that equip MB and ME. The barrel track finder system covers the region $|\eta| < 0.83$ and combines information from the DTs and RPCs. The redundancy of the two systems improves the determination of the muon hit. The number of hits associated with each track and the quality of the track extrapolation are used as a muon candidate quality criterion. The overlap and endcaps track finders cover the region $0.83 < |\eta| < 1.24$ and $|\eta| > 1.24$, respectively. Those regions are equipped with fast detectors with small latency time allowing the usage of fast pattern-recognitions algorithms. The number and the topology of hits associated to each track are used to assign a quality to each muon candidate. The tracks reconstructed are sent to an intermediate layer and to the global muon trigger to remove duplicate tracks reconstructed at the boundaries of the regions and to sort the tracks by p_{T} and quality before sending the information to the global trigger.
The global trigger is equipped with electronic boards of large computing power, capable of computing global event quantities and multi-object correlations. Decisions are taken based on the number of candidates passing specific p_{T} thresholds and on several other variables, such as invariant masses of pair of objects or angular separations. During the 2016 data taking the system was able to store more than 300 different seeds, increased to 500 in 2017.

3.2.2 The L1 EG trigger

The EG candidates reconstruction and identification, performed in the Layer 2 of the L1 calorimeter trigger, is based on a combination of ECAL and HCAL information. Due to bandwidth limitation, the TTs energy is sent from the Layer 1 to the Layer 2 calorimeter trigger in a compressed format. For each TT, composed by an ECAL and the corresponding HCAL TT, the information is sent in 16 bits divided as follows:

- nine bits for the transverse energy of the TT, computed as the sum of the HCAL (H) and ECAL (E) energy;
- three bits to store the ratio H/E. The information is saved as the $\log_2(H/E)$ if $H > E$ or $\log_2(E/H)$ if $E > H$;
- one bit to indicate if $E > H$ or the opposite;
- one bit to indicate if H or E are zero;
- one bit for the ECAL fine-grain veto described below in this section;
- one bit for an HCAL form factor which is currently unused.

The total TT energy ($H + E$) is stored with a precision of half a GeV and saturates for TT energies bigger than 255 GeV. The position of each TT is indexed by the position of the TT in the ($i\eta$, $i\varphi$) plane, with $-28 < |i\eta| < 28$ and $0 < |i\varphi| < 72$.
Clustering of EG candidates is based on dynamic clustering of adjacent TTs, as illustrated in the right panel of Fig. 3.17. At first a seed TT is identified as a local energy maxima in a window of 3×9 TTs in the ($i\eta$, $i\varphi$) plane. The seed is required to have a transverse energy of at least 2 GeV to mitigate the impact of the detector noise. Once the seed is identified, TTs with transverse energy of at least 1 GeV in the 3×3 window centred around the seed are added to the cluster. The window is extended in the φ direction to 5 TTs if a TT neighbouring with a tower already in the cluster has a transverse energy of at least 1 GeV. Finally, the size of the clustering is reduced to 2 TTs in the η direction by stripping the side with the lowest energy deposit. The energy of the EG candidate is derived as the sum of the TTs energies. The dynamic clustering allows the recovery of most of the *bremsstrahlung* energy loss, improving the energy resolution on the candidates. To further improve the energy resolution a calibration of the energy of each EG candidate is applied. The calibration factors are derived from the ratio between the energy of the EG candidates measured at the L1 trigger and the energy of the corresponding offline candidates. The calibration factors are derived in bins of uncalibrated energy of the EG candidate $E^{\mathrm{L1}}_{\mathrm{RAW}}$, $i\eta$ of the seed TT and according to the candidate shape:

$$E^{\mathrm{L1}} = E^{\mathrm{L1}}_{\mathrm{RAW}} \times C\left(E^{\mathrm{L1}}_{\mathrm{RAW}},\ |\eta|,\ \mathrm{Shape}\right). \tag{3.6}$$

As the clustering is performed with the sum of the hadronic and the electromagnetic energy, all the jets are also clustered as EG candidates. To retain the trigger rate within acceptable levels three selections based on the shape of the candidates and on the H/E ratio are applied.
The first selection exploits the fine-grain veto, computed by the ECAL front-end electronic and directly sent to the trigger as a single bit. It exploits the fact that electromagnetic

showers are generally much smaller than the hadronic ones. Within each ECAL TT, the sum of the energy in a strip, a row of 5 crystals along η, is computed. If at least 90% of the energy of the TT is deposited in two adjacent strips, the shower is judged compact enough to come from an EG candidate. The fine-grain can not be computed at the L1 trigger level since the granularity of the information is not enough. Candidates whose seed TT is vetoed are rejected.
The second selection is based on the H/E ratio. The threshold is set so that the efficiency on electrons is 99.5%.
Finally a topological veto is applied according to the shape of the candidate. Once more it exploits the different size of electromagnetic and hadronic showers. More than 100 different cluster shapes are coded, based on the average shape of electron and jets. Candidate with jet-like shapes are discarded, while candidate with small clusters compatible with originating from an electromagnetic shower are accepted. Figure 3.18 illustrate some examples of jet-like and electron-like shapes. All the selections are released for candidates with transverse energy higher than 128 GeV, to ensure the highest possible efficiency on high energy particles.
The rejection of energy deposits arising from pileup interactions is improved by the isolation criterion. Energy deposits from hard scattering are generally isolated while pileup energy deposits have large activity around them; the energy deposited around the EG cluster can be used for PU rejection. An isolation region, illustrated in the right panel of Fig. 3.17, is defined as the region of 6×9 TTs in the $i\eta$ and $i\varphi$ directions, both in ECAL and in HCAL, centred around the seed of the cluster. A window of 2×5 TTs in ECAL and of 2×1 in HCAL around the seed tower is removed form the region to prevent the energy of the EG candidate to be included in the computation of the isolation. The sum of energy of the TTs in the isolation region is used as a discriminating variable E^{Iso}. The threshold λ to be applied is defined according to the energy of the candidate, the position along η and the PU. As no track information is available, the PU can not be estimated from the number of reconstructed vertices, but the number of trigger towers with non-zero energy deposition n_{TT} is exploited as its proxy. The isolation criterion is satisfied if:

$$E^{\mathrm{Iso}} < \lambda \left(|\eta|,\ E^{\mathrm{L1}},\ n_{\mathrm{TT}} \right). \tag{3.7}$$

Once identified the EG candidates in the event, the decision whether to keep or to reject an event is taken accordingly to the number of EG candidates above a given p_{T} threshold. Trigger based on a single EG candidate (SingleEG) is composed by three different seeds. The SingleEG seed requires one EG candidate without any selection on the candidate isolation. The SingleEGIso seed applies a selection on the isolation of the candidate, slightly reducing the threshold in p_{T}. Finally the SingleEGIsoER, where ER means eta restricted, selects isolated EG candidates with $|\eta| < 2.1$, removing the region where the hadronic activity is higher to further reduce the p_{T} threshold. Generally events are selected exploiting the combination of the three seeds to grant the highest possible efficiency on genuine electron and photons. Other seeds are available, based on multiple EG candidates, such as the DoubleEG which requires two electrons or photons satisfying asymmetric p_{T} thresholds, or combination of EG candidates with other objets, generally referred to as cross-triggers.

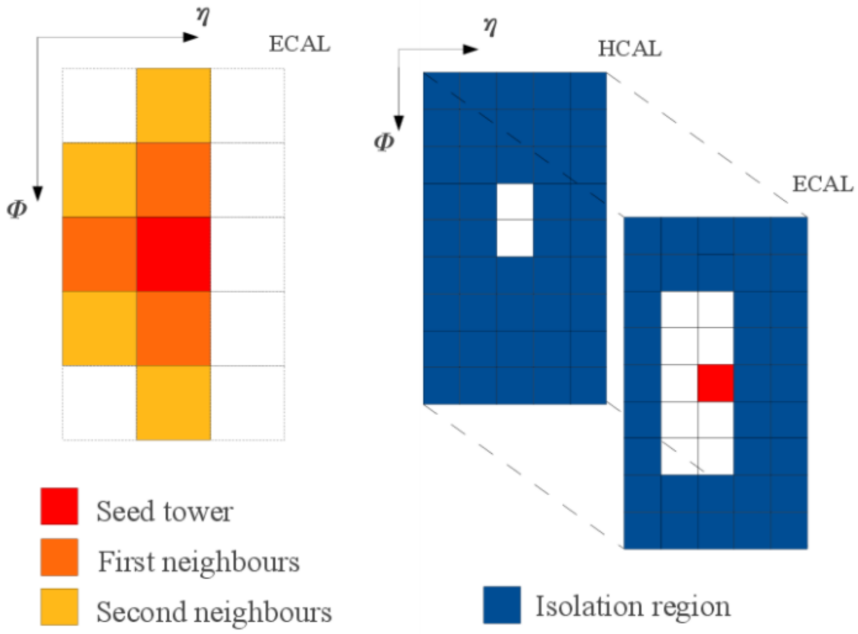


Figure 3.17: Left: illustration of the clustering for EG candidates in the L1 trigger. A seed tower, defined by a local energy maximum, can be associated to neighbouring trigger towers in a window of ±1 TT in the η direction and ±2 TTs in φ direction centred around the seed TT. Right: illustration of the isolation region. The isolation for the candidate whose seed is in the red TT is computed from summing the energy of all the TT in the blue region.

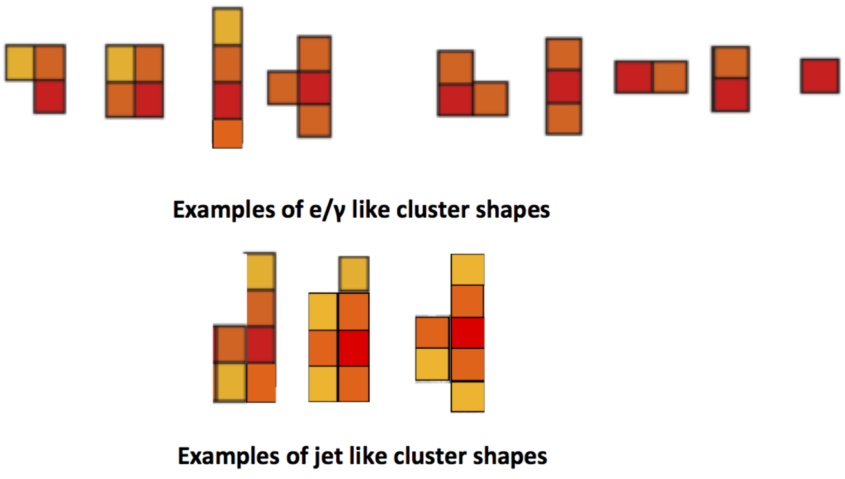


Figure 3.18: Example of possible shapes of the EG candidates. Candidates with EG-like shapes (top) are accepted by the L1 trigger while candidates with larger shapes, compatible with originating from a jet (bottom), are rejected.

3.2.3 Study of the EG trigger for 2017 data taking

The different LHC conditions in the 2016 and 2017 data forced a change in most of the L1 trigger seeds. The average number of reconstructed vertices per events jumped from 27 in the 2016 to 38 in 2017, reaching peaks up to 60 collisions per bunch crossing. To handle such a level of PU without a conspicuous increase of the p_{T} thresholds, considerable changes were introduced.
Each algorithm is a compromise between efficiency and rate; the highest possible efficiency should be granted within the allocated rate. Along with the efficiency, the slope of the 'turn-on' curve is also relevant. The turn-on curve represents the efficiency of the L1 trigger as a function of the transverse energy of the offline reconstructed particles (see Fig. 3.19 for an example of turn-on curve). The ideal turn-on curve would be a step function, with the step corresponding to the p_{T} threshold applied at the L1 trigger. Resolution and misidentification effects reduce the slope of the turn-on, introducing p_{T} dependency in the efficiency. Offline objects are generally selected for the data analyses setting a p_{T} threshold slightly higher than the trigger one, where the trigger efficiency is at its *plateau*. The introduction of p_{T} dependent efficiency in the analyses, which is difficult to model in the simulation, is avoided. A sharper turn-on curve is therefore mapped in an increased acceptance for offline analyses. Lowering the efficiency *plateau* is one of the targets of the trigger algorithms.
For electrons the efficiency is measured with respect to offline reconstructed electrons with the Tag & Probe method (T&P) from $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events. One electron is used as 'tag' electron by requiring it to satisfy trigger requirements and tight identification and isolation criteria. The second electron is the 'probe', which is required to pass specific criteria depending on the efficiency under study. The probe electron provides an unbiased samples to measure the trigger efficiency both on data and simulation. The rate is measured from ZeroBias events, which are events randomly chosen at the trigger level providing an unbiased sample of typical LHC collisions. The study presented in this section has been conducted on a special set of 4 pb^{-1} collected in 2016 with an average PU of 60; despite the small dimension the run is representative of the conditions expected on 2017.
The efficiency of the EG trigger algorithms has been measured on a part of the data collected in the RunF of 2016 data taking. The performance of the updated algorithm has been verified emulating the trigger decision with the improved algorithm on the 2016 data. Once the algorithm has been implemented on the hardware boards of the L1 trigger, the performance has been validated with early 2017 data. The two runs have been chosen to present the same average PU.
Several changes have been introduced in the EG trigger for the 2017 data taking. The recalibration of the L1 EG tigger kept the energy resolution at the same level of 2016. The retuning of the isolation threshold allowed a reduction in the energy threshold of the isolated candidates, while the definition of a new criterion for the H/E ratio induced a large reduction in the rate. Additionally, a second set of isolation thresholds, unexploited in 2016, has been studied targeting cross-triggers where the energy threshold on the EG candidate is lower. Finally a pilot study for a seed targeting W bosons decaying to an electron has been conducted, targeting the electrons used in the ECAL calibration.
The improvement in the performance of the SingleEG trigger is illustrated in Fig. 3.19, where the turn-on curve for the L1 trigger is shown. The efficiency is measured in the two selected runs of data with the T&P method on $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events for the combination

SingleEG, SingleEGIso and SingleEGIsoER seeds. Large improvements are observed: the energy threshold is 2 GeV lower, increasing the trigger efficiency by 7% on $Z \to e^+e^-$ events for the same trigger rate. The efficiency *plateau*, measured at 95% of efficiency, is reached 10 GeV earlier than in the 2016 case.

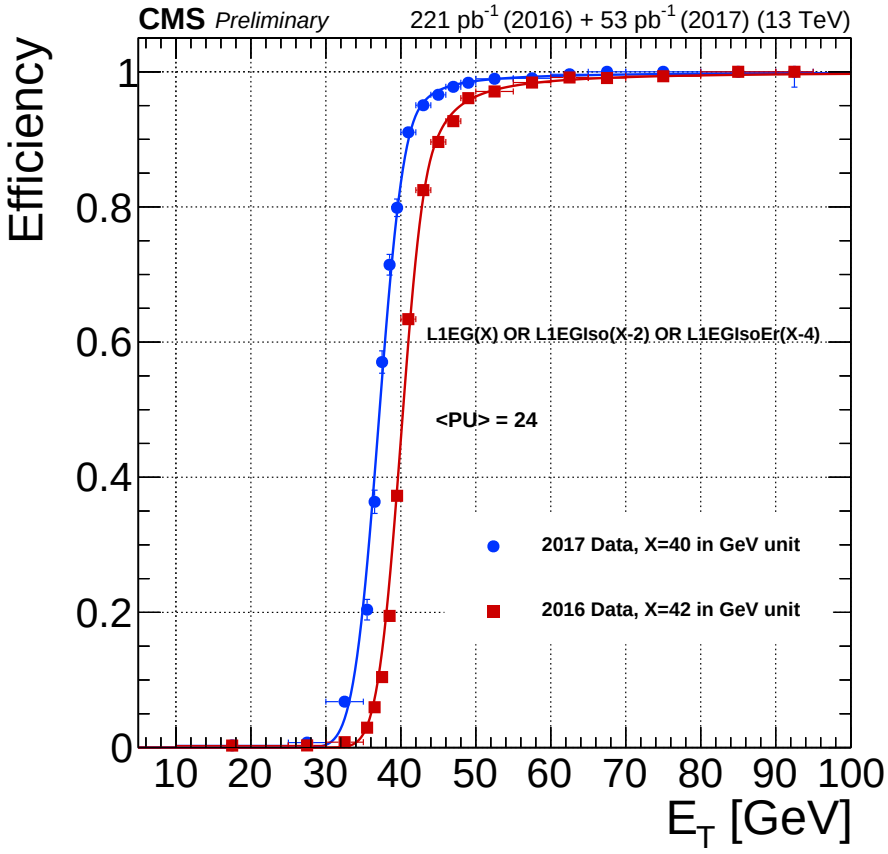


Figure 3.19: Efficiency of the L1 EG trigger as a function of the energy of the electrons reconstructed offline, measured with the T&P method on two selected runs of data with the same average pileup. The figure compare the efficiency achieved with the 2016 algorithm (red squares) and the 2017 one (blue circles). The two curves are obtained with the combination of the L1 seeds exploited to selected the EG events, combining SingleEG(X), SingleEGIso(X-2) and SingleEGIsoER(X-4), where X is the p_T threshold expressed in GeV. The two curves are drawn for the thresholds X, which grant the same trigger rate, of 42 GeV on 2016 data and of 40 GeV on the 2017 ones.

Calibration

The energy of L1 trigger EG candidates is calibrated with reference the offline electron candidates. Electrons reconstructed offline are matched to the L1 EG candidates exploiting the distance parameter ΔR. Events are divided in bins of pseudorapidity, energy of the L1 candidate and shape of the cluster. The number of bins is fixed by the hardware capability, while the choice of the boundaries of the bins is arbitrary and it can be tailored to ensure the optimal energy resolution. Figure 3.20 illustrates the energy resolution in 2016 and

2017 data with respect to the energy of L1 EG candidates computed offline. The energy resolution for candidates in EB is at the same level as in 2016 while it is slightly improved in EE.

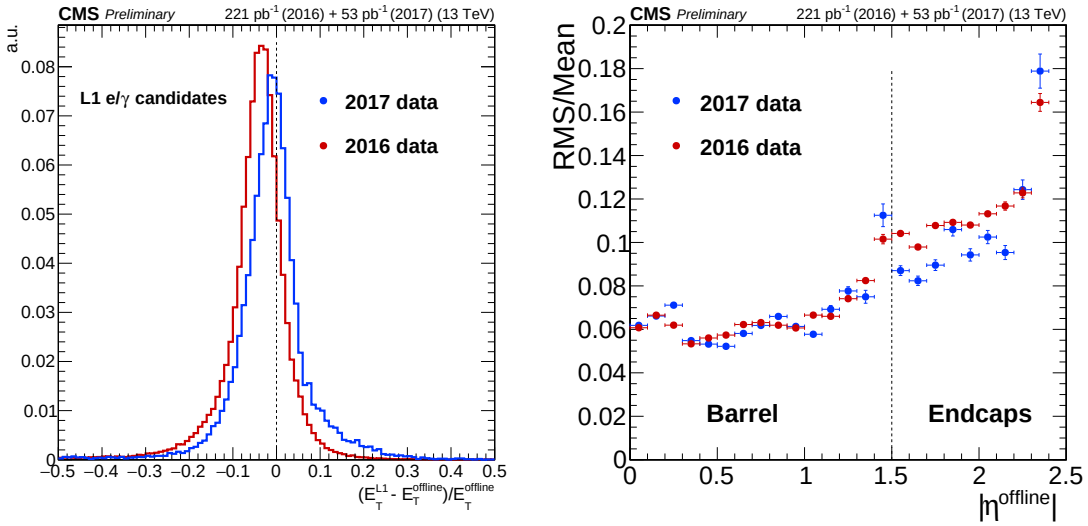


Figure 3.20: Energy resolution of the EG candidates reconstructed at the L1 trigger measured on electrons from the decay of a Z boson. Left: distribution of the difference between the energy of the electron reconstructed offline and the corresponding L1 EG candidate, normalised to the electron energy, integrated on the whole detector. Right: the RMS of the distribution, normalised to its mean, as a function of the electron pseudorapidity. The same energy resolution is achieved in EB, while in EE it is slightly improved.

Isolation

The isolation criterion is exploited at the L1 trigger to select candidates with a slightly lower p_{T} threshold compared to the SingleEG seed at the price of a small efficiency reduction. On offline electrons the selection based on the isolation is generally applied according to the ratio between the energy deposited in a cone around the candidate and its transverse momentum, as detailed in Eq. 4.2. The same approach can not be exploited at the L1 trigger, as the L1 hardware can not perform floating point operations. The offline selection, less severe as the p_{T} of the candidate increases, is emulated relaxing the threshold on the isolation as a function of the candidate transverse momentum. The tuning of the algorithm consists in defining how the selection is relaxed with increasing the p_{T} of the candidate so to achieve an efficiency close to unity for high energy candidates. As the L1 hardware can not apply a threshold varying continuously with the p_{T}, the selection is applied in bins of p_{T} as well as of η and n_{TT} to account for the different pileup conditions. The relaxation of the isolation with the p_{T} is modelled with a functional form and the effect of applying a particular relaxation scheme is evaluated emulating the trigger. Several functional forms and boundaries of the bins are tested and the one with the best combination of efficiency, low *plateau* and rate reduction is chosen. The improvement achieved with respect to the 2016 algorithm is illustrated in the left panel of Fig. 3.21. It shows the turn-on curve for the L1 trigger for isolated EG candidates with $p_{\mathrm{T}} > 38$ GeV

(the p_T threshold exploited for the data taking). The 2017 algorithm has a sharper curve and the efficiency is higher than the 2016 one on the whole p_T spectrum. The right panel demonstrates that no dependence on the PU is introduced.

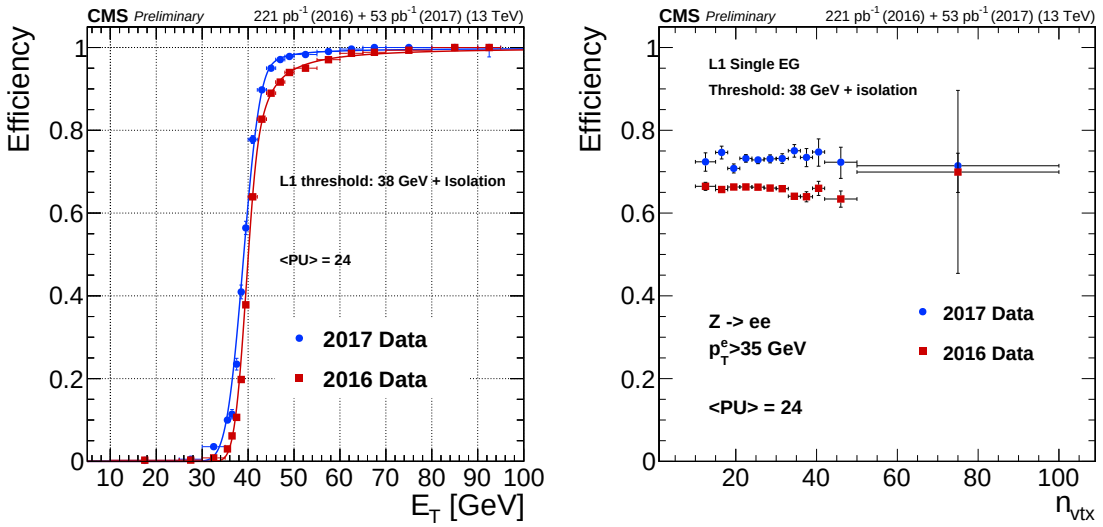


Figure 3.21: Left: efficiency of the L1 EG trigger for isolated EG candidates with $p_T > 38$ GeV as a function of the transverse energy of the offline reconstructed electron. The 2017 algorithm (blue) is more efficient on the whole p_T spectrum then the 2016 one (red). Right: efficiency of the L1 isolated EG seed as a function of the numebr of offline reconstructed vertices. The efficiency is computed on a sample of $Z \to e^+e^-$ events with the T&P method requiring the probe p_T to be grater than 35 GeV. No dependency on the PU is observed.

Selection on the H/E ratio

The selection on the H/E ratio is a powerful tool to discriminate jets from electrons (or photons). Since only three bits are available to store the H/E ratio, the information is saved as the logarithm of the ratio H/E. The 2016 H/E selection was based on the ratio computed on the seed TT; the candidate is accepted if $\log_2(E/H) > 5$ and it is rejected otherwise. As the low granularity of the stored information prevents from an effective tuning of the threshold as a function of the PU, the work focused on a possible extension of the selection to other TTs included in the cluster.
The average shape of electrons and jets suggests a possible extension of the H/E selection to the 3×3 matrix centred on the seed TT. Jets have sizeable hadronic activity around the seed TT, as opposite to electrons: the application of the selection on the ratio H/E on the 3×3 matrix centred on the seed TT provides a good discriminating power between the two. As most of the electromagnetic energy of the EG candidates is deposited in the seed tower, a looser selection has been adopted, requiring $\log_2(E/H) > 1$ for towers with an energy deposit of more than 5 GeV and no selection for TTs with lower energy deposition. The efficiency reduction on electrons is negligible, while a consistent reduction in the efficiency on jets has been observed, reflected in a reduction of the expected trigger rate. The usage of the same threshold for the 3×3 matrix as for the seed tower would

instead have given an efficiency loss of 30%. The left panel of Fig. 3.22 illustrated the L1 trigger rate reduction as a function of the p_{T} threshold applied on the EG candidate. For thresholds between 30 and 40 GeV, corresponding to the working point of the algorithm, a rate reduction of 20% has been found. As the rate budget is fixed, the reduction in the expected rate allows the decreasing of the p_{T} threshold applied to all the EG seeds.

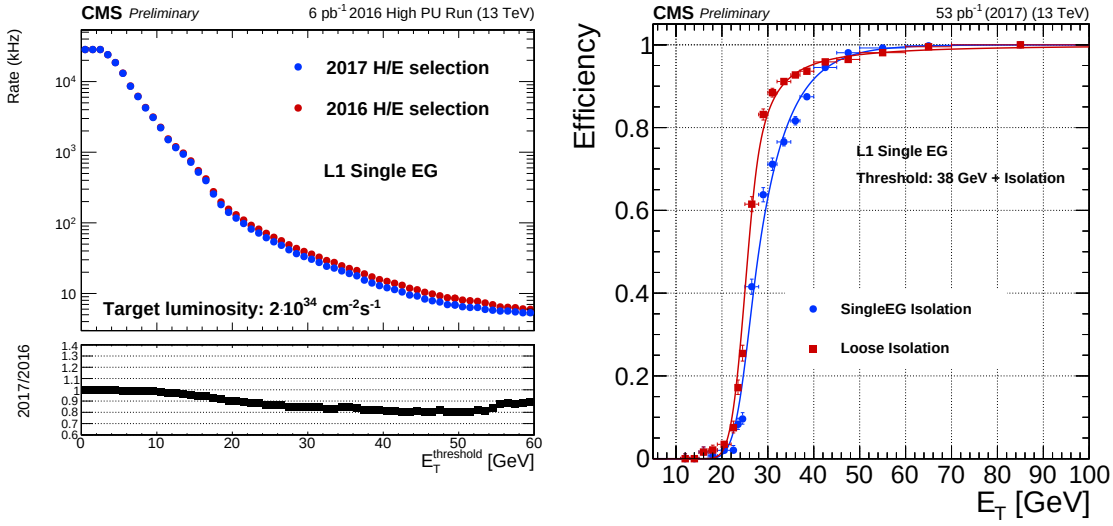


Figure 3.22: Left: L1 trigger rate for the EG seeds as a function of the p_{T} threshold applied. The red points are obtained applying the H/E selection exploited in 2016, while the blue ones with the refined H/E selection deployed for the 2017 data taking. The bottom panel is the ratio of the two curves. A rate reduction of 20% is observed for a threshold around 40 GeV. Right: efficiency of the L1 EG trigger for isolated EG candidates with $p_{\mathrm{T}} > 25$ GeV as a function of the transverse energy of the offline reconstructed electron. The blue curve is obtained with the isolation threshold designed for the SingleEG seed while the red one is realised with applying the loose isolation threshold specifically designed to increase the efficiency for the low p_{T} candidates exploited in cross-triggers.

Cross-triggers

Cross-triggers offer the possibility to select events based on multiple objects, lowering the p_{T} thresholds of the candidates and emulating the final states expected in the different analyses. As an example the DoubleEG seed used in 2017 applied two thresholds of 25 and 14 GeV to the EG candidates, to be compared with the 40 GeV of the SingleEG candidate. The requirements of isolated EG candidate would allow one to further reduce the threshold and benefit the analyses targeting low energy isolated objects. The isolation selection described above is tuned to be efficient combined with a p_{T} threshold of about 40 GeV, as the dependence of the isolation threshold as a function of the energy is designed to optimally work in that region. When lowering the p_{T} threshold at 25 GeV, the efficiency *plateau* is reached very slowly. To overcome the problem, a second set of thresholds has been implemented to be optimally efficient combined with a p_{T} threshold of about 25 GeV. The right panel of Fig. 3.22 compares the efficiency of the two isolation criteria, with the nominal one designed for the SingleEG seed and the loose one designed for the

cross-triggers. The efficiency is largely increased and the efficiency *plateau* is reached much faster.
Additionally, a pilot study has been conducted for a L1 seed specifically designed to increase the selection efficiency of $W \to e\nu_e$ events. The experimental signature of this channel is an electron produced with missing transverse momentum. Since no track information is available at the L1 trigger, the missing transverse energy E^{miss}, computed as the negative vector sum of all the objects reconstructed in the events, is exploited. To discriminate the production of the W boson from other processes it is generally convenient to define the transverse mass M_T as:

$$M_T = \sqrt{2E_T^{\text{miss}} E_T^e (1 - \cos\Delta\varphi)}, \tag{3.8}$$

where E_T^e is the transverse energy of the electron and $\Delta\varphi$ is the angle between the electron and the direction of the missing energy. For a W boson produced at rest $M_T \simeq 40$ GeV. Figure 3.23 shows the distribution of the transverse mass and of E_T^e reconstructed at the L1 trigger for events from the decay of a W boson and ZeroBias events. Despite the resolution on the transverse mass is rather poor, the variable is a powerful tool to discriminate the two processes.
The L1 seed has been designed to increase the selection efficiency on $W \to e\nu_e$ exploited for the calibration of the ECAL. The p_T spectrum of electrons following a W boson decay measured at L1 is shown in the right panel of Fig. 3.23. With a SingleEG p_T threshold above 40 GeV, the selection efficiency would have reduced, affecting the precision of the ECAL intercalibration. The seed requires the presence of an EG candidate with $p_T > 33$ GeV and $M_T > 40$ GeV. The two thresholds has been chosen performing a two-dimensional optimisation to achieve the highest possible signal efficiency within the allocated L1 trigger rate. The efficiency recovery on $W \to e\nu_e$ events is of about 10%.

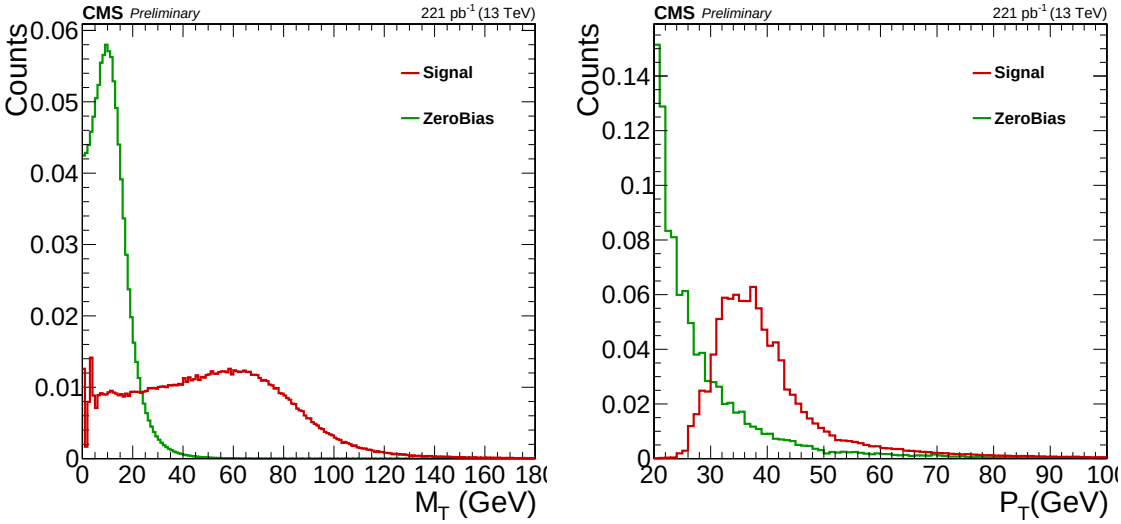


Figure 3.23: Distribution of the transverse mass M_T (left) and of the electron p_T (right) measured at the L1 trigger for $W \to e\nu_e$ events (red) and for ZeroBias events (green).

Chapter 4

Measurement of the $t\bar{t}H$ process

Stat Roma pristina nomine,
nomina nuda tenemus.

Bernardus Cluniacensis

This chapter describes in details the data analysis performed to measure the event rate of the $t\bar{t}H$ process in the channel with a diphoton decay of the Higgs boson. As already mentioned in the introduction, the search for a Higgs boson in the diphoton final state is the search for a peak in the invariant mass spectrum of the photon pairs ($m_{\gamma\gamma}$) arising over a continuous background due to spurious events. The data analysis hereby described concerns 35.9 fb^{-1} of data collected by the CMS experiment in 2016 and 41.5 fb^{-1} collected in 2017. The 2016 data analysis is summarised in Ref. [96] and, in combination with other finale states, contributed to the first observation of the $t\bar{t}H$ process [97]. The analysis of 2017 data, summarised in Ref. [98], goes in the direction of a precise measurement of the $t\bar{t}H$ cross section, with significant improvements compared to the 2016 one, resulting in a sensitivity increased by about 50%.
The two analyses present the structure of the $H \rightarrow \gamma\gamma$ searches. Events with two photon candidates are selected by the trigger. The analysis starts from photon identification, realised by applying a set of 'preselections' slightly tighter than the trigger. A Boosted Decision Tree (BDT) [99] is exploited to further suppress the contribution of jets mimicking photons while retaining good efficiency on genuine photons. The $Z \rightarrow e^+e^-$ process is used as a standard candle to validate the photon selections and to ensure good agreement between data and simulation. The $t\bar{t}H$ production is exclusively identified thanks to the decay products of the top-antitop pairs in the final state. Events are split in categories defined according to the expected levels of signal and backgrounds. Once the categorisation is defined, signal strength modifiers (see Section 1.3.4) are extracted from simultaneously fitting the diphoton invariant mass spectrum of all the categories. The fit function is composed by a signal model, derived from simulation, and a background model, whose shape and yield are defined from fitting the data. The data-driven background estimation relegates the role of background simulation to the definition of the categorisation, without directly influencing the final result and thus greatly reducing the systematic uncertainty associated with the background estimation. The data-to-simulation agreement is, instead, relevant for the signal model, which is derived from simulation. The full analysis is conducted 'blindly', without looking at the signal region ($115 < m_{\gamma\gamma} < 135$) during the

optimisation and the definition of the categories to prevent from inducing an artificial bias in the signal estimation. The analysis is 'unblinded' to estimate the signal yield from the data once the selections are fixed.
The two analyses performed on 2016 and 2017 data share the structure described above. The photon selections are similar, with minor differences due to the different trigger and ECAL conditions. The main difference between the two analyses is how the $t\bar{t}H$ production is exclusively identified. The 2016 analysis has been part of a comprehensive measurement of the Higgs boson properties in the diphoton final state. The event categorisation relative to the $t\bar{t}H$ process has been defined in order to maximise the sensitivity on the $t\bar{t}H$ process as well as on the other production processes. As opposite, the 2017 measurement has been an exclusive measurement of the $t\bar{t}H$ production, allowing a dedicated study of the categorisation capable to include the correlations between the final state objects, leading to a significant improvement in the sensitivity to the $t\bar{t}H$ process.
The detailed description of the two analyses follows. As a unique structure is adopted for the two years, the two are described together. Whenever a difference arises, it is highlighted in the text.
The work within this thesis focused on the categorisation of $t\bar{t}H$ events, necessary to achieve a precise determination of the $t\bar{t}H$ event rate. The largest effort was therefore devoted to study of the set of selections which provides the highest experimental sensitivity. In the analysis of the 2017 data, the study of the multivariate algorithms for event classification and the relevant optimisation has been a large part of the work. In addition, more advanced classification algorithms have been investigated to potentially benefit the analysis of the Run II data and future studies of the $t\bar{t}H$ channel based on Run III data.
The chapter introduces at first the topology of $t\bar{t}H$ events and of the main background processes affecting the measurement (Section 4.1). The data and the simulated samples used for the analyses are reported in Section 4.2, while Section 4.3 reviews in detail the definition of the final state objects exploited for the measurement, starting from photons and moving to the choice of the interaction vertex and to the definition of leptons, jets and b jets. Section 4.4 describes in detail how the $t\bar{t}H$ production has been identified in the two analyses based on 2016 and 2017 data. Section 4.5 describes the statistical procedure used to derive the result. The construction of the model to fit the data is described in Section 4.6, while Section 4.7 reports the systematic uncertainties affecting the measurement. The results of this study are presented in Sections 4.8 , followed by an outline of the future prospects for the current analysis and possible improvements to further increase the sensitivity, in Sections 4.9 and 4.10, respectively.

4.1 Final state topology and backgrounds

The $t\bar{t}H$ process presents a complex final state, with multiple topologies depending on the top quark pair decays. The possible decays for a pair of top quarks, reported in Section 1.5.2, are listed again here for sake of clarity:

- fully hadronic decays: $t\bar{t} \to b\bar{b}\ W^+W^- \to b\bar{b}\ q\bar{q}\ q'\bar{q}'$;
- semi-leptonic decays: $t\bar{t} \to b\bar{b}\ W^+W^- \to b\bar{b}\ q\bar{q}\ \ell\bar{\nu}_\ell$;
- fully leptonic decays: $t\bar{t} \to b\bar{b}\ W^+W^- \to b\bar{b}\ \ell\bar{\nu}_\ell\ \bar{\ell}'\nu_{\ell'}$.

The possible final states following a $t\bar{t}H$ event thus consist in a pair of photons with mass $m_{\gamma\gamma} \approx 125$ GeV from the diphoton Higgs boson decay, two b jets and additional jets or leptons. Missing transverse momentum can originates due to neutrinos in the leptonic channels or jets outside the acceptance of the detector.
Channels with leptons in the final state are generally characterised by very low levels of background and low signal yield, as opposite to the hadronic channel where both signal and background are enhanced. All the possible final state topologies are exploited for the $t\bar{t}H$ measurement, defining exclusive categories aiming at collecting events with different topologies.
Background events arise from multiple processes, involving photons, jets and leptons in the final state. The photons can be originated from non-resonant production (prompt photons) or from from jet fragments with high fraction of electromagnetic energy, capable of mimicking the signature of photons (fake photons). Three main processes are identified as backgrounds: the diphoton production, or prompt-prompt background, the γ + jet production, of prompt-fake, and the multijet production, or fake-fake background. The candidate photon pairs (either prompt or fake) should be accompanied by the presence of jets or leptons capable of reproducing the decay of a top quark pair. Another source of background is the production of $t\bar{t}$ events with additional photons in the final state. The $t\bar{t}\gamma\gamma$ constitutes the irreducible background of the analysis while $t\bar{t}\gamma$ + jet and $t\bar{t}$ + 2 jets are important sources of background in the leptonic channels.
A different source of background is due to the presence of events from other Higgs boson production processes in the $t\bar{t}H$ signal region. As all the events containing a Higgs boson decaying to photons resonate to the same value of $m_{\gamma\gamma}$, the number of estimated signal events from fitting the $m_{\gamma\gamma}$ spectrum is a mixture of different processes. This mixture can only be derived from simulation. The bigger the contamination from other processes, the larger is the uncertainty on the expected $t\bar{t}H$ event yield, therefore creating categories enriched in $t\bar{t}H$ with low contamination from the other Higgs boson production processes is one of the challenges of the analysis. For $t\bar{t}H$ events with fully hadronic decays, the main source of internal background is due to ggH events produced in association with jets, while, in the leptonic channel, it mostly arises from VH events. The tHq and tHW constitutes an additional source of internal background in the $t\bar{t}H$ channel (both in the hadronic and leptonic one), as their final state is very close to the $t\bar{t}H$ one. The low cross section of those processes makes their contribution subdominant with respect to the $t\bar{t}H$ one.

4.2 Data and simulation samples

Events included in the analysis are selected by the L1 trigger requiring the presence of two EG candidates with asymmetric p_{T} thresholds of 25 and 12 GeV, increased to 25 and 14 GeV in 2017 data. The HLT further filters the events by requiring two photon candidates with asymmetric p_{T} thresholds of 30 and 18 GeV in 2016 data and 30 and 22 GeV in 2017 ones. The increased trigger thresholds were necessary to retain the trigger rate at acceptable levels at the different operating conditions of the LHC. The two photon candidates must satisfy a set of requirements on shower shape and isolation variables in addition to a selection on the H/E ratio. Photons undergoing a conversion in the material upstream the ECAL are identified exploiting the R_9 variable, defined as the ratio between the energy deposited in the 3×3 crystals matrix centred on the cluster seed crystal divided

by the energy of the candidate. For photons not undergoing a conversion in the material upstream the ECAL, the R_9 variable is close to unity, while for converted photons the magnetic field spreads the two electrons widening the extension of the shower and lowering the value of R_9. The efficiency of the trigger algorithm is measured from $Z \to e^+e^-$ events with the T&P method both on data and simulation. The simulation is corrected from the ratio of the two values to match the efficiency observed in data, while the uncertainty on the correction factor is propagated to the expected event yield as a systematic uncertainty.
Simulated signal samples corresponding to the different Higgs boson production mechanisms are generated using MadGraph5_aMC@NLO [100] (version 2.2.2 for the 2016 analysis and 2.4.2 for the 2017 one) at NLO in perturbative QCD with FxFx merging [101]. The parton-level samples are interfaced with the Pythia 8.0 software [102] (version 8.205 and 8.230 for 2016 and 2017 samples, respectively) for parton showering and hadronisation. The PDFs are taken from the NNPDF 3.0 set [103]. For the $t\bar{t}H$ process a second sample generated with Powheg 2.0 [104–107] has been exploited for the training of the multivariate discriminators described in Section 4.4. The signal cross sections and branching ratios are the ones recommended by the LHC Higgs cross section working group [40].
Different generators are exploited to simulate background events. The prompt-prompt diphoton background is generated with the Sherpa generator [108] (version 2.2.1). It includes the Born-level process with up to three additional jets and the box process at LO accuracy in perturbative QCD. The multijet and γ + jet backgrounds are generated with Pythia 8.0 applying a filter at generator level to enrich the production of jets with high fraction of electromagnetic energy. Another sample of γ + jet background generated with MadGraph5_aMC@NLO is used for the 2017 analysis. It includes up to four inclusive jets and three exclusive jets and no filter is applied. Events with $t\bar{t}$ pairs are generated with MadGraph5_aMC@NLO.
A Drell-Yan sample of $pp \to Z/\gamma^* \to \ell\ell$ events (referred to as Drell-Yan for sake of simplicity), generated with MadGraph5_aMC@NLO is used for $Z \to e^+e^-$ validations.
The detailed response of the CMS detector is simulated with the Geant4 package [109]. The simulation include the addition of in-time and OOT pileup. The OOT contribution is simulated only in the window [-5, +4] bunch crossings around the nominal one, where the effect on the variables reconstructed in the detector is more relevant. Simulated events are weighted to match the pileup distribution observed in data (see Section 2.4).

4.3 Object identification

The PF algorithm, described in Section 2.3.1, merges the information coming from the different detectors to construct final state objects such as photons, electrons, muons and jets. The selections applied by the PF algorithm are generally loose, aiming to retain the reconstruction efficiency as high as reasonably achievable. When analysing the data, according to the expected level of background in the channel under study, more stringent requirements are applied to identify final state objets in order to adequately balance signal efficiency and background rejection. The following sections describe the algorithms exploited to identify the objects used in the $t\bar{t}H$ analysis, starting from the photons and the choice of the interaction vertex and moving to the different decay products of the top quark pair.

4.3.1 Photon identification

Photons are the central ingredient of the measurement. As the event rate is inferred from $m_{\gamma\gamma}$, special care should be devoted to their identification and to the measurement of their energy. Photons are clustered starting from isolated energy deposits in the ECAL not linked to any reconstructed track in the tracker, as described in Section 2.3.3. A multivariate procedure is exploited to recover the non-clustered energy, aiming at achieving an optimal energy resolution. Further corrections are applied to match the energy scale and the energy resolution in data and simulation. The procedure is validated with $Z \to e^+e^-$ decays, where electrons are used as proxies of photons. A set of preselections, slightly tighter than the trigger selection, is applied to the photons to discriminate genuine photons from fake ones. The discrimination is enhanced exploiting the photon identification BDT. The selected photons constitute a pure sample of photons with optimal energy resolution, capable of providing the basis for the $H \to \gamma\gamma$ analysis.

Energy correction

The photon energy is computed from the sum of the energy deposited in the ECAL crystals belonging to the photon supercluster, after proper detector calibration (see Section 3.1). The clustering starts from a seed crystal to which neighbouring crystals with energy deposits beyond a threshold are added. Crystals with energy below the threshold are not clustered, causing part of the photon energy to be lost. Additionally, mechanical structures supporting the ECAL induces dead regions where no active material is present, enhancing the non-reconstructed energy and the probability of *bremsstrahlung* emission. The energy loss systematically lowers the estimation of the photon energy, introducing a bias in its measurement and reducing the photon energy resolution.The effect of the PU can further bias the energy photon estimate, as pileup deposits can be clustered in the shower.
A multivariate tool has been implemented to estimate the true photon energy on a per-photon basis, ensuring the high energy resolution necessary for a successful $H \to \gamma\gamma$ analysis. A large simulated sample of photons, with p_T ranging from 0.25 to 100 GeV, is processed through a GEANT4 simulation of the CMS detector. Only photons which do not experience conversion to electron pairs in the material upstream the ECAL are exploited to derive the energy correction. In EB an analytical function $f(E, \eta) = g(E)h(\eta)$ is fitted to the two-dimensional distribution of the average ratio $< E/E^{\text{true}} >$ of the photons. The function f represents by construction the correction to be applied to the measured energy to obtain the true one. It is estimated from a multivariate regression technique, using as input position and shower shapes variables, sensitive to shower containment, and global variables sensitive to pileup. A similar procedure is exploited in the EE, with the addition of the energy deposited in the ES. The correction in the EB is close to unity for high energy photons, while it can be as high as 20% for low energy ones. In the EE the correction is 5% for high energy photons, as a photon deposit on average 5% of its energy in the ES absorbers, and up to 40% for low energy ones. The multivariate regression simultaneously estimate the true photon energy and a per-photon uncertainty on its energy. The effect of the energy regression is shown in figure 4.1, where the ratio between the reconstructed photon energy and the true generated energy is shown with and without applying the energy correction. The plots exploit photons from the decay of the Higgs boson not used in the training of the energy regression. The ratio between the raw energy of the photon and the true energy is the function f at the typical energy of the photons following the

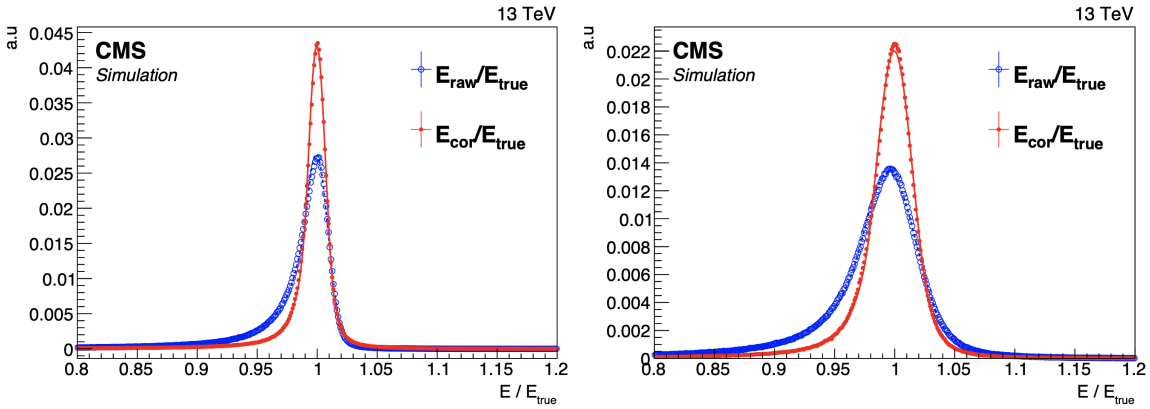


Figure 4.1: Ratio between the reconstructed photon energy and the true simulated photon energy for photons following the decay of an Higgs boson in EB (right) and in EE (left). The blue line represents the ration between the reconstructed photon energy and its true energy, while the red line represents the same ratio once the energy correction is applied. The blue line is by definition the function f at the typical energy of the photons following the decay of the Higgs boson, integrated over η.

decay of the Higgs boson, integrated over η.
After correcting the energy estimate, a multistep procedure is exploited to match energy scale and resolution in data and simulation. At first, the energy scale observed in data is matched to the one in simulation. The scale is derived from the position of the $Z \to e^+e^-$ invariant mass peak, where electrons are reconstructed as photons, ignoring the tracker information. A second correction is applied to simulated events in order to match the energy resolution measured in data. A gaussian smearing is applied to each event, representing all the detector non-idealities not included in the simulation. The correction applied to the energy scale ranges from 0.1 to 0.3% in EB and up to 2% in EE. The amount of smearing to be applied ranges between 0.1 to 3%, depending on the photon pseudorapidity and R_9.
The energy scale is derived differentially in time, to correct long term drifts, and in bins of η and in R_9. The amount of smearing to be applied is derived in the same categories of η and R_9. For the 2016 data the scale and smearing correction is derived in four bins in η (two in EB and two in EE) times two in R_9. The 2017 corrections are derived in 50 different categories, five in η times ten in R_9. The increased number of categories allows a more granular correction which is reflected in a a smaller uncertainty on the derived parameters. Figures 4.2 and 4.3 show the $Z \to e^+e^-$ invariant mass peak for non-showering electrons after scale and smearing corrections are applied in 2016 and 2017 data, respectively. An excellent agreement is found between data and simulation on the Z boson, ensuring a good control on the Higgs boson invariant mass peak.

Photon preselections

Events selected by the trigger are further required to pass a set of preselection criteria slightly stringent than the one used for triggering. The preselections are applied both on data and on simulated events, in order to get a uniform phase space for the analysis. Preselections target the rejection of fake photons while retaining the efficiency on prompt

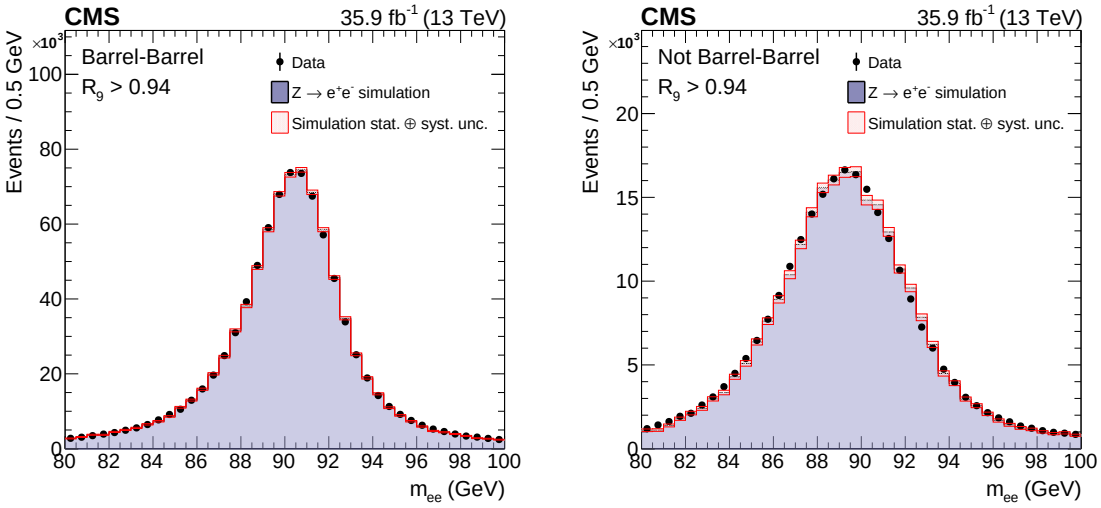


Figure 4.2: Comparison of the $Z \to e^+e^-$ invariant mass peak between data and simulation on 2016 data after the energy corrections are applied. Black markers represent the data while the solid histogram is the Drell-Yan simulation. The uncertainty on the simulation is shown by the red band. The two figures include non-showering electrons ($R_9 > 0.94$) reconstructed as photons (without the tracker information). The left figure is obtained with events where both the electrons are in EB while the right one represents all the other events. An excellent agreement is achieved between data and simulation.

ones as high as possible. The selections are based on shower shape variables, as jets are generally wider than photons, and isolation variables, as photons following a Higgs boson decay are isolated while electromagnetic deposits within jets present hadronic activity around them. Photons should satisfy the following requirements:

- minimum transverse momentum of the leading photon (the photon with the highest p_T) greater than 30 GeV in 2016 data and 35 GeV in 2017 ones;
- minimum transverse momentum of the subleading photon (the photon with the lowest p_T) greater than 18 GeV in 2016 data and 22 GeV in 2017 ones;
- the pseudorapidity of the photons must be $|\eta| < 2.5$ and not in the barrel-endcap transition region ($1.44 < |\eta| < 1.55$);
- a selection on the H/E ratio, the ratio between the energy deposited in the HCAL cell behind the supercluster and the energy of the supercluster;
- a selection on the R_9 variable and on $\sigma_{i\eta i\eta}$, the latter being the energy weighted extension of the shower in the η direction within the 5×5 crystal matrix centred on the seed crystal;
- an electron veto which rejects superclusters linked to a track with no missing hits in the innermost tracker layers;

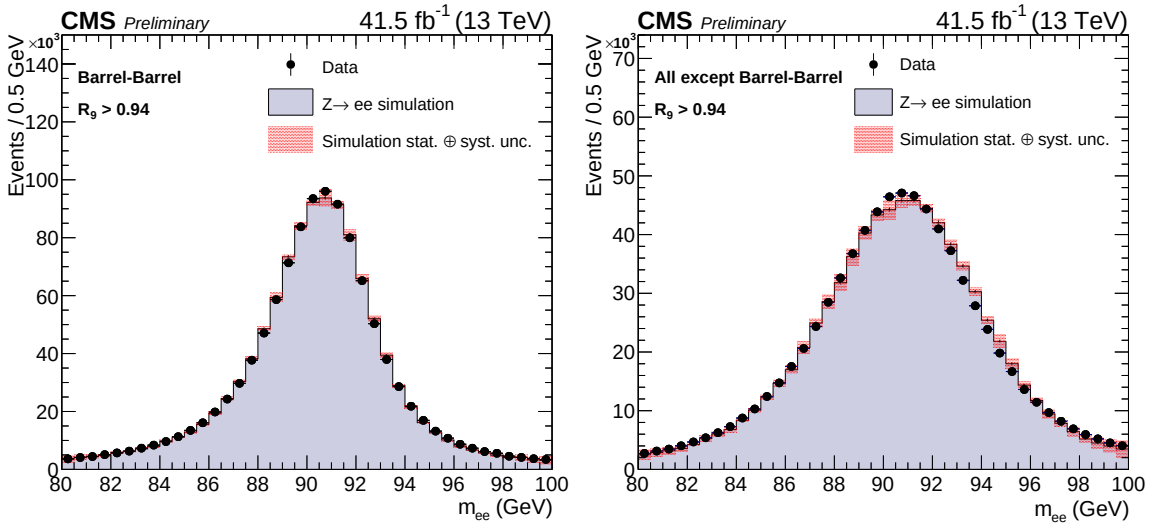


Figure 4.3: Comparison of the $Z \to e^+e^-$ invariant mass peak between data and simulation on 2017 data after the energy corrections are applied. Black markers represent the data while the solid histogram is the Drell-Yan simulation. The uncertainty on the simulation is shown by the red band. The two figures include non-showering electrons ($R_9 > 0.94$) reconstructed as photons (without the tracker information). The left figure is obtained with events where both the electrons are in EB while the right one represents all the other events. An excellent agreement is achieved between data and simulation.

- a requirement on the photon isolation ($\mathcal{I}_{\text{ph}}$), defined as the sum of the transverse momentum of all the PF candidates identified as photons in a cone with R=0.3 centred around the photon candidate;
- a requirement on the tracker isolation in a hollow cone ($\mathcal{I}_{\text{tk}}$), defined as the the sum of the transverse momentum of all the tracks in a cone with R=0.3 centred around the photon candidate. The inner cone with R=0.04 is excluded from the sum to use the same isolation criterium exploited for electron identification;
- a requirement on the charge-hadron isolation ($\mathcal{I}_{\text{ch}}$) defined as the the sum of the transverse momentum of all the charged particles in a cone with R=0.3 centred around the photon candidate. The requirement is redundant with respect to the previous one but it is added to the selection to match the one applied by the PF algorithm.

Additionally to the selections listed above, both the photons are required to satisfy either $R_9 > 0.8$ or $\mathcal{I}_{\text{ch}} < 20$ GeV or $\mathcal{I}_{\text{ch}}/p_{\text{T}} < 0.3$. Photon preselections are different for high R_9 and low R_9 photons, as the high R_9 photons are generally a pure sample of prompt photons, while the low R_9 ones are more subject to fake photons contamination. Table 4.1 summarises the preselection criteria, split for photons reconstructed in EB and EE.
The preselections efficiency is measured both in data and in simulation with the T&P method from $Z \to e^+e^-$ events, electrons being reconstructed as photons. Since the electron veto efficiency can not be measured on electrons, a sample of $Z \to \mu^+\mu^-\gamma$ is exploited. The process provides, once proper selections are applied, a sample of prompt

	R_9	H/E	$\sigma_{i\eta i\eta}$	$\mathcal{I}_{\text{ph}}$ (GeV)	$\mathcal{I}_{\text{tk}}$ (GeV)	$\mathcal{I}_{\text{ch}}$ (GeV)
Barrel	$R_9 > 0.85$	< 0.08	-	-	-	-
	$0.5 < R_9 < 0.85$	< 0.08	< 0.015	< 4.0	< 6.0	< 6.0
Endcap	$R_9 > 0.90$	< 0.08	-	-	-	-
	$0.80 < R_9 < 0.90$	< 0.08	< 0.035	< 4.0	< 6.0	< 6.0

Table 4.1: Summary of the preselection applied to photons.

photons with purity higher than 99%. The ratio between the measured efficiency in data and in simulation is exploited to correct the simulation, while its uncertainty is propagated to the expected signal yield as a systematic uncertainty. Tables 4.2 and 4.3 show the efficiency measured on data $\varepsilon_{\text{data}}$, the one on simulation ε_{sim} and their ratio for 2016 and 2017 data respectively. The scale factors to be applied to the simulation are generally close to unity. The efficiency measured with $\text{Z} \to \text{e}^+\text{e}^-$ events is relevant only for the agreement between data and simulation, while the efficiency on the Higgs boson is estimated directly from simulation, once proper corrections derived from Z bosons are applied.

Preselection category	$\varepsilon_{\text{data}}$ (%)	ε_{sim} (%)	$\varepsilon_{\text{data}}/\varepsilon_{\text{sim}}$
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 4.2: Efficiency of preselection criteria measured with $\text{Z} \to \text{e}^+\text{e}^-$ and $\text{Z} \to \mu^+\mu^-\gamma$ events on 2016 data with the tag and probe technique. The scale factors to be applied to simulation are generally close to the unity.

Preselection category	$\varepsilon_{\text{data}}$ (%)	ε_{sim} (%)	$\varepsilon_{\text{data}}/\varepsilon_{\text{sim}}$
Barrel; $R_9 > 0.85$	90.5 ± 0.9	91.3 ± 0.1	0.991 ± 0.010
Barrel; $R_9 < 0.85$	74.2 ± 1.8	76.5 ± 0.2	0.967 ± 0.024
Endcap; $R_9 > 0.90$	81.7 ± 0.3	83.9 ± 0.2	0.973 ± 0.004
Endcap; $R_9 < 0.90$	43.6 ± 0.6	46.6 ± 0.6	0.935 ± 0.018

Table 4.3: Efficiency of preselection criteria measured with $\text{Z} \to \text{e}^+\text{e}^-$ and $\text{Z} \to \mu^+\mu^-\gamma$ events on 2017 data with the tag and probe technique. The scale factors to be applied to simulation are generally close to the unity.

Photon identification BDT

The fake-photons contribution in preselected events is further suppressed by a multivariate discriminant called photon identification BDT. The BDT is trained on a γ + jet sample, where prompt photons are used as signal and fake ones as background. The variables used as inputs for the BDT are listed below:

- shower shape variables, after proper corrections are applied to mitigate the disagreement between data and simulation;
- the isolation variables $\mathcal{I}_{\text{ph}}$ and $\mathcal{I}_{\text{ch}}$. Two versions of the latter are exploited, the first one including only the hadrons originating from the chosen interaction vertex (see Section 4.3.2) and the second including the hadrons associated with the vertex providing the largest isolation sum. This version of $\mathcal{I}_{\text{ch}}$ is effective in rejecting jet fragments misidentified as photons originating from a vertex different from the photon one;
- the energy and pseudorapidity of the photon, as they are strongly correlated with the shower shape;
- the median energy density per unit area of the event ρ, sensitive to the pileup, helps in reducing the impact of the pileup of the other variables.

Variables concerning the shower shape of the photons are corrected exploiting $\text{Z} \rightarrow \text{e}^+\text{e}^-$ events, to mitigate the disagreement between data and simulation. The origin of the disagreement is due to the mis-modelling of the ECAL conditions, especially the pedestal. The correction is derived weighting the simulation in order to match the distributions observed in the data for the variables included in the training. The uncertainty on the correction is propagated through the analysis.
For each year of data included in the analysis a separate training is performed, as the different ECAL conditions can affect the input variables, particularly the one related to the shower shape. Similar performance are achieved by the two trainings. The output of the BDT for 2016 data is shown in the left panel of Fig. 4.4, comparing the output of the BDT for a $\text{H} \rightarrow \gamma\gamma$ sample, with all the production modes weighted according to their cross section, and the main background of the analysis for events in the mass range $100 < m_{\gamma\gamma} < 180$ GeV. The agreement between data and simulation of the output of the BDT is checked with $\text{Z} \rightarrow \text{e}^+\text{e}^-$ events, as shown in the right panel of Fig. 4.4 (for electron in EB in 2016 data). The systematic uncertainty on the output of the BDT is conservatively assigned to to cover the largest observed discrepancy on $\text{Z} \rightarrow \text{e}^+\text{e}^-$ events (in EE), indicated by the hatched area in the figure. Similar performance is achieved in 2017 data.
A loose selection on the output score of the BDT is added to the preselections by requiring the photon ID score to be greater than -0.9.

4.3.2 Identification of the interaction vertex

Interaction vertices are reconstructed from intersecting the collection of tracks, extrapolated to the beam line. The pileup level of Run II caused up to 60 vertices per events to be reconstructed, challenging the identification of the vertex of the primary interaction. The

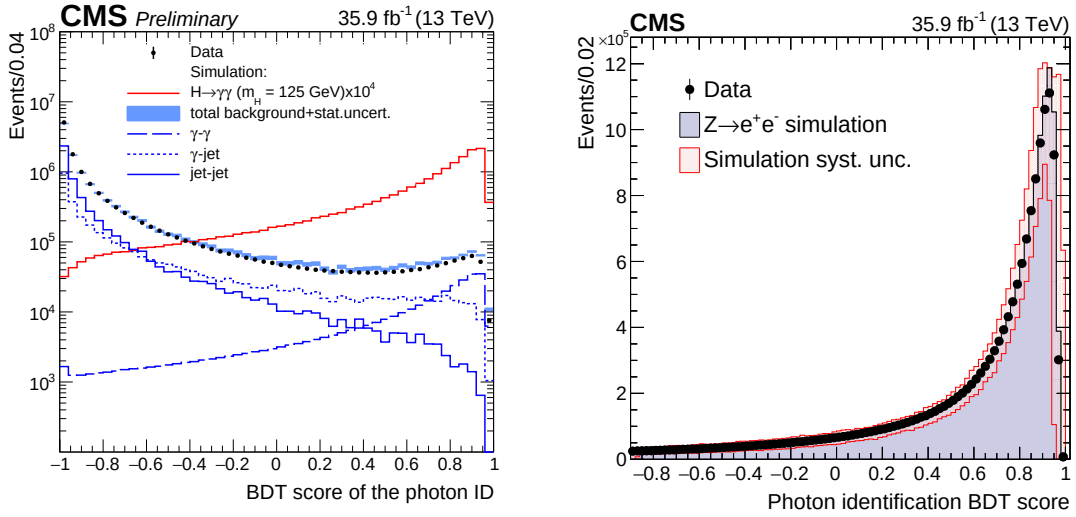


Figure 4.4: Left: output of the photon identification BDT for 2016 data (black markers), for simulated H $\rightarrow$ $\gamma\gamma$ signal (red histogram) and for the different backgrounds of the analysis (blue histograms). The background samples are normalised to match the number of events observed in data. Events are selected in the invariant mass range $100 < m_{\gamma\gamma} < 180$ GeV. Good separation is achieved for prompt photons and jets mimicking a photon signature. Right: output of the photon identification BDT for $Z \rightarrow e^+e^-$ in 2016 data (black markers) in comparison with simulation (full histogram) for electrons in EB reconstructed as photons. The red band shows the uncertainty, conservatively assigned to the BDT score to cover the largest observed discrepancy on $Z \rightarrow e^+e^-$ events (in EE).

correct assignment of the vertex in a H $\rightarrow$ $\gamma\gamma$ measurement impacts the invariant mass resolution of the photon, since the mass is reconstructed as:

$$m_{\gamma\gamma} = \sqrt{4E_1^{\gamma}E_2^{\gamma}\sin^2\frac{\theta}{2}}, \tag{4.1}$$

where $E_{1,2}^{\gamma}$ are the energies of the two photons and θ is the angle between them. The position of each photon is determined from the position of the supercluster, while the choice of the interaction vertex is necessary to derive θ, as pictorially shown in Fig. 4.5. Table 2.1 shows that, in the z direction, the LHC beam has an r.m.s. of 7.55 cm, thus the collision vertices are spread on approximately 20 cm. If the wrong vertex is assigned, the θ angle is wrongly measured and the invariant mass peak resolution is worsened. If the vertex position is found within 1 cm from the true interaction point, the vertex contribution to the mass resolution is negligible with respect to the one from the photon energy. Instead, when the vertex assignment is off by more than 1 cm from the true interaction point, the vertex contribution is the dominant term to the resolution. The determination of the vertex is therefore a central ingredient of the H $\rightarrow$ $\gamma\gamma$ measurement.
Two different strategies are exploited for the vertex assignment in 2016 and 2017 data. The default algorithm exploited in CMS to establish the correct interaction vertex is based on the variable $\sum p_{\mathrm{T}}^2$, defined as the square sum of the transverse momentum of all the tracks originating from a vertex. As the pileup vertices are characterised by low energy interactions, the vertex with the highest value of $\sum p_{\mathrm{T}}^2$, called vertex 0, is assumed to be

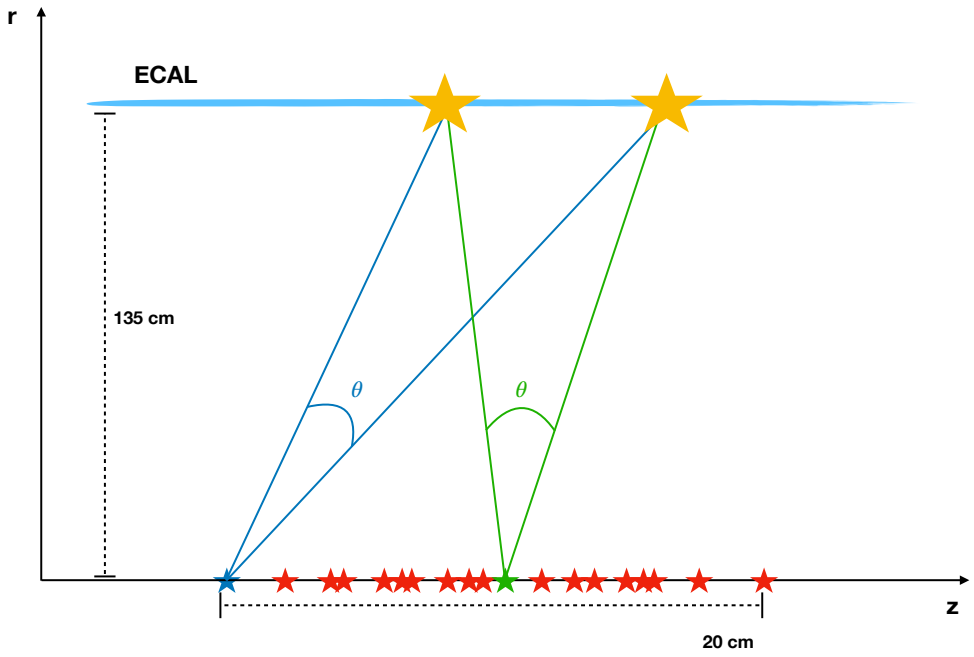


Figure 4.5: Pictorial representation of how the choice of the interaction vertex affects the measurement of the angle between the two photons. The z axis represents the beam line, with the stars being the reconstructed vertices. The light blue line is the ECAL surface, with the two yellow stars represents the position of the two superclusters. The green and blue lines illustrate how the angle θ changes according to the choice of the interaction vertex. Drawing is not to scale.

the hard interaction vertex.
In case of $H \rightarrow \gamma\gamma$ events produced via ggH, the choice of the vertex is quite challenging. The high momentum particles produced in the interaction are neutral and the vertex 0 is not granted to be the correct choice. Nevertheless, for momentum conservation, the production of two high energy photons force an asymmetry in the track distribution within the vertex. A multivariate algorithm has been implemented to identify the vertex in $H \rightarrow \gamma\gamma$ events, combining in a BDT the $\sum p_T^2$ and two other variables related to the vertex momentum imbalance. In case of converted photons, the tracks of the two electrons are also exploited.
In $t\bar{t}H$ events, the presence of top quarks induce the production of several charged particles, strongly enhancing the $\sum p_T^2$ of the vertex. Both the choice of the vertex 0 and of the vertex chosen with the $H \rightarrow \gamma\gamma$ algorithm provide an efficiency of finding the vertex within 1 cm from the true interaction vertex above 99%. In the 2016 analysis, where all the Higgs boson production processes have been measured, the $H \rightarrow \gamma\gamma$ vertex algorithm has been exploited, to ensure a uniform vertex selection with the ggH measurement. Instead, in the $t\bar{t}H$-only 2017 analysis the choice have been the usage of the vertex 0, which can be exploited without the necessity of training a BDT. No difference in the $m_{\gamma\gamma}$ resolution has been found with changing the vertex selection algorithm.

4.3.3 Electron identification

The electron reconstruction, described in Section 2.3.3, starts from clustering an ECAL energy deposit linked to a reconstructed track, fitted based on the GSF model. The electron identification [86] aims at distinguishing prompt electrons, originating directly in the interaction vertex, from electrons coming from secondary processes, such as decays of b or c quarks and photon conversions, and jets mimicking electron signatures. Three criteria are exploited:

- compatibility between the momentum measured in the tracker and the energy measured in the ECAL, including geometrical compatibility between the track extrapolation to the ECAL and the supercluster position;
- calorimetric observables, such as shower shape variables, as jet deposits are generally wider than electrons, and the H/E ratio;
- track quality variables, exploiting the difference between the GSF track and the Kalman filter track to improve the discrimination of charged hadrons.

Additionally, the relative isolation is particularly powerful in separating electrons within jets, with large activity around them, from prompt electrons. It is defined as the sum of the reconstructed PF candidates in a cone of R=0.3 around the electron candidate, normalised to the p_{T} of the electron:

$$\mathcal{I}^{\mathrm{rel}} = \frac{\sum p_{\mathrm{T}}^{\mathrm{charged}} + max\left(0, \sum p_{\mathrm{T}}^{\mathrm{neutral}} + \sum p_{\mathrm{T}}^{\gamma} - \frac{1}{2}\sum p_{\mathrm{T}}^{\mathrm{PU}}\right)}{p_{\mathrm{T}}}, \tag{4.2}$$

where $\sum p_{\mathrm{T}}^{\mathrm{charged}}$, $\sum p_{\mathrm{T}}^{\mathrm{neutral}}$ and $\sum p_{\mathrm{T}}^{\gamma}$ are the sum of the transverse momentum of the charged hadrons originating from the chosen interaction vertex, neutral hadrons and photons falling within the isolation cone. As for the neutral components no vertex can be assigned, the neutral pileup is estimated from $\sum p_{\mathrm{T}}^{\mathrm{PU}}$, the sum of the transverse momentum of all the charged hadrons originating in the pileup vertices. Since the energy deposition of the pileup is due to jets, it roughly consists of 65% of charged tracks and 35% of neutral deposits. Therefore half of $\sum p_{\mathrm{T}}^{\mathrm{PU}}$ gives an estimate of the neutral contribution of the pileup. The second term of the isolation is thus an estimate of the neutral activity coming from the chosen interaction vertex corrected for the pileup contribution.
Several variables, including the isolation, are exploited to train a BDT capable of identifying prompt electrons. The training is performed on a simulated sample of Drell-Yan events, with prompt electrons used as signal and jets and secondary electrons as background. The training is performed for electrons with $p_{\mathrm{T}} > 10$ GeV in three different regions of pseudorapidity, two in EB and one in EE.
The electron candidates are identified or rejected according to the output score of the BDT. Multiple 'working points' (WPs) are provided, corresponding to different signal efficiency and background rejection. The T&P method on $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events is used to match the identification efficiency measured in data to the one in simulation.
Electrons considered for the $\mathrm{t\bar{t}H}$ measurement are required to pass the 'medium' WP of the algorithm, providing 90% efficiency on prompt electrons, to have $p_{\mathrm{T}} > 20$ GeV and to fall within the tracker acceptance ($|\eta| < 2.5$). In the 2017 analysis, the p_{T} threshold has been lowered to 10 GeV to increase the signal efficiency. Electrons must not overlap with

the selected photons ($\Delta R\,(e,\gamma) > 0.2$) and the invariant mass of each selected photon and of the electron should not be compatible with the Z boson mass $|m_{e,\gamma} - m_Z| > 5$ GeV. The latter selection helps in reducing the background due to Z bosons decaying in electrons where one electron is misidentified as a photon. The electron selection criteria have been studied on $t\bar{t}H$ simulation to maximise the sensitivity of the analysis.

4.3.4 Muon identification

Muon reconstruction (see Section 2.3.2) is based on combining information from the tracker and from the muon system. A Kalman filter track reconstruction is performed and the resulting track is fitted with the shape expected for a muon. The curvature of the track allows the measurement of the muon momentum. Muon identification algorithms [84] are necessary to discriminate real muons from hadrons leaking the HCAL and depositing energy in the muon system. The discrimination is based on track quality variables, on the compatibility between the track and the selected vertex and on the number of hits associated with the track in the layers of the inner tracker and of the muon system. As for the electrons, several WPs of the algorithm are provided, balancing signal efficiency and background rejection. Muons exploited in the $t\bar{t}H$ analysis are required to satisfy the 'tight' selection criteria. The tight criterium targets the identification of prompt muons originating in the primary interaction vertex and the rejection of muons from in-flight decays of other particles, such as b and c quarks. A tight muon must satisfy the following criteria:

- the muon should be reconstructed by the PF algorithm as a global muon;
- the track should feature at least six hits in the tracker layers, at least one hit in the pixel detector and at least two hits in the muon station. The requirement on the pixel hit reduce the contribution of muon from in-flight decays, while the requirement on the muon station largely reduce the hadron contamination;
- the fit to the track should have a χ^2/degrees of freedom < 10. This requirement helps in the reduction of muons from in-flight decays;
- the track extrapolation to the beam line should be within 2 mm from the chosen vertex in the transverse plane and within 5 mm in the z direction, to suppress the contribution of cosmic muons.

The selection efficiency, measured on $Z \to \mu^+\mu^-$ events, is 96%. As for the electrons, the T&P method is exploited to correct the discrepancy in the simulation.
For a muon to be included in the $t\bar{t}H$ analysis, in addition to satisfy the tight identification criterium, it should present a minimum p_T of 20 GeV, lowered to 10 GeV in 2017, and be within the acceptance of the muon system ($|\eta| < 2.4$). To further suppress the contribution of muons within jets, a selection on the relative isolation, equivalent to the one used for electrons and defined in Eq. 4.2, is also applied.

4.3.5 Jet identification

Jet reconstruction, described in Section 2.3.4, is based on clustering PF candidates according to the anti-k_T algorithm with a distance parameter R=0.4. The jet identification [110] is necessary to distinguish jets due to the hadronisation of high energy quarks or gluons

from 'noise' jets and 'pileup' jets. The first are non-physical jets originated from clustering fluctuations in the electronic noise of the calorimeters, the second ones are jets obtained from clustering particles originating from the multitude of pileup vertices in the event.
Noise jets are suppressed by requiring the energetic composition of the jets to be compatible with what is expected for a physical jet. The variables involved in the identification are the fraction of energy of the jet carried by charged hadrons, neutral hadrons and electromagnetic deposits, as well as the multiplicity of charged tracks and of the neutral deposits within the jet. Minimum requirements on those variables are enough to reject about 100% of the noise jets with an efficiency of about 99% for physical jets, measured with the T&P technique on a sample of multijet and dijet events.
The discrimination of physical and pileup jets is realised exploiting the topological differences between the two. Physical jets are collimated jets originating from a single particle, as opposite to pileup jets, composed by several particles coming from different vertices and thus with a broad shape. The energy density profile of the jet is exploited for the discrimination. Tracking and jet shape variables are exploited to train a BDT with hard jets used as signal and pileup ones as background. The training is performed on a simulated sample of Z bosons produced in association with jets and it is validated on data exploiting the same process. For jets within the tracker volume, the algorithm provides 89% of pileup jets rejection for 96% of efficiency on physical jets. Without pileup jet suppression the number of reconstructed jets with $|\eta| < 1.4$ would be five times higher.
Jets satisfying the two identification criteria are included in the $t\bar{t}H$ analysis if they present $p_T > 25$ GeV and $|\eta| < 2.4$. The pseudorapidity range is restricted to the tracker region where the tagging of b jets is possible, as explained in the next section. It has been checked that excluding forward jets does not reduce the sensitivity of the analysis. Selected jets are required not to overlap with the two photons nor with leptons, if any is present, by applying $\Delta R\,(\text{jet}, \gamma/\ell) > 0.4$.

4.3.6 Identification of b jets

The exclusive identification of jets originating from bottom-type quarks is of particular interest for several measurement in the CMS experimental program. Within the $t\bar{t}H$ measurement, it is relevant for the identification of the top quarks decays. Jets from b-type quarks can be identified thanks to the long lifetime of the B mesons, which can travel for $\mathcal{O}(1\text{ mm})$ in the detector. The decay products of a B meson originate a secondary vertex within the jet, displaced from the beam line. Additionally, the sizeable mass of the b-type quark allows the presence of leptons within the jets with a sizeable transverse momentum with respect to the jet axis. The combination of those two elements can be exploited to identify b jets.
In the $t\bar{t}H$ measurement, b jets are identified with the Combined Secondary Vertex (CSV) algorithm [111]. The algorithm combines in a multivariate discriminant several variables related to the tracks distribution within the jet, the presence of secondary vertices, the distance between the primary and the secondary vertex and the presence of soft leptons. According to the output of the algorithm, three WPs are defined corresponding to a rate of jets originating from light quarks or gluons misidentified as b jets of 10% (loose WP), 1% (medium WP) and 0.1% (tight WP).
In the 2017 analysis, a second version of the CSV algorithm has been exploited, the

DeepCSV [112]. The same variables are combined in a Deep Neural Network (DNN) [113], resulting in a better discrimination against light jets. At the medium working point, the efficiency on b jets is of about 72% with the CVS algorithm and about 76% with the DeepCSV. The output score of the b jet identification algorithm is referred to as b-discriminant.
As for the other objets, the tagging efficiency is measured in data and simulation. The two samples used are $t\bar{t}$ events enriched with heavy flavour quarks (by requiring the presence of muons within jets) and multijet events. Scale factors, function of the jet p_T, η and flavour, are applied to simulation in order to correct the disagreement with the data. This disagreement was observed to be as high as 10% in some regions of the phase space.
All the three WPs have been exploited for the $t\bar{t}H$ analysis, using in each topology the one which provides the best sensitivity. In addition, the full shape of the b-discriminant distribution is adopted as input to train the BDTs used for $t\bar{t}H$ identification.

4.4 Events classification

Events with two photons satisfying the preselections described in Section 4.3.1 and in the invariant mass range $100 < m_{\gamma\gamma} < 180$ are included in the analysis. Photons are further required to satisfy $p_T^{\gamma_1} > m_{\gamma\gamma}/3$ and $p_T^{\gamma_2} > m_{\gamma\gamma}/4$. The requirement of a p_T selection depending on of the diphoton invariant mass prevents distortions of the low mass side of the $m_{\gamma\gamma}$ spectrum.
The $t\bar{t}H$ production is identified thanks to the final state objects arising from the decay of the top quarks. The most powerful variables in identifying the $t\bar{t}H$ production are shown in Figures 4.6 and 4.7. Events originating from the $t\bar{t}H$ process present high p_T photons, due to the recoil of the Higgs boson against the top quark pair, and a large number of high-p_T jets, b jets and leptons in addition to large p_T^{miss}. The background is taken from the 'data sidebands', the events in data in the invariant mass region $100 m_{\gamma\gamma} < 115$ or $135 or m_{\gamma\gamma} < 180$, excluding the signal region. The variables represented in Figures 4.6 and 4.7 are powerful not only in discriminating the $t\bar{t}H$ signal from the background but also in the discrimination of the $t\bar{t}H$ production from ggH, VBF and VH.
The topology of the final state changes drastically according to the decay of the top quarks, therefore events are split in leptonic categories, where at least one lepton is present, and hadronic ones, including events without leptons. Here leptons are intended as muons or electrons, since the exclusive reconstruction of τ leptons is not exploited in this work.
The following sections describe the categorisation of the events exploited in the analyses. In the 2016 analysis, several categories are defined to exclusively identify events produced in the ggH, VBF, VH and $t\bar{t}H$ channel. As the focus of this thesis is on the $t\bar{t}H$ production, the $t\bar{t}H$ categories are described in detail, while the other production modes are briefly summarised. The $t\bar{t}H$ selection is based on two categories; the leptonic category requires at least one lepton and one b jet in the final state, while the hadronic one requires the presence of jets and b jets. The signal-to-background ratio is enhanced thanks to a BDT common to all the categories aiming at the selection of pair of photons compatible with the decay of a Higgs boson.
The 2017 analysis, described in Section 4.4.2, largely improved the signal-to-background discrimination employing two BTDs, one for the leptonic category and one for the hadronic one. The BDTs are trained with a combination of photons, jets and leptons information to fully exploit the correlations among the variables.

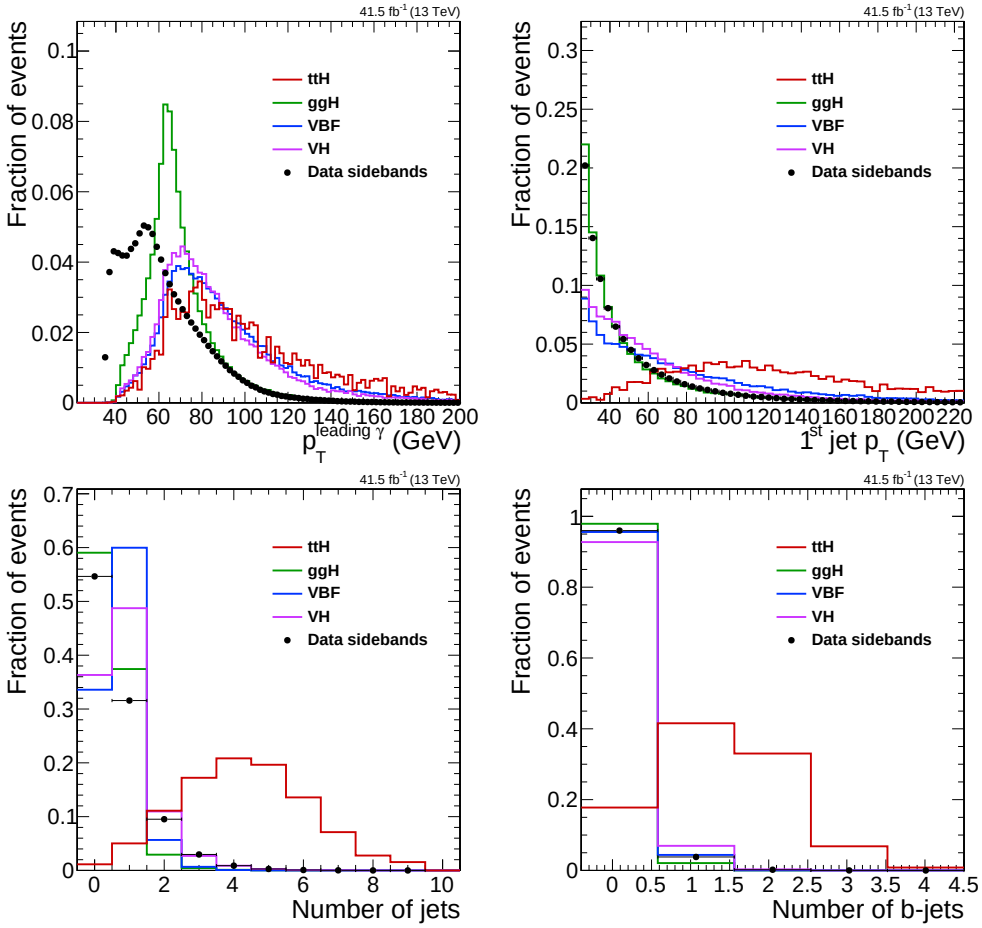


Figure 4.6: Most sensitive variables to separate $t\bar{t}H$ production from the other Higgs boson production processes. The data sidebands defined in the text (black markers) are compared with the $t\bar{t}H$ signal (red histogram) and with the ggH, VBF and VH production (coloured histograms). The p_T of the leading photon (top left) and leading jet (top right), the number of reconstructed jets (bottom left) and of b jets tagged with the medium WP of the DeepCSV algorithm (bottom right) are shown. All the histograms are normalised to have unitary area.

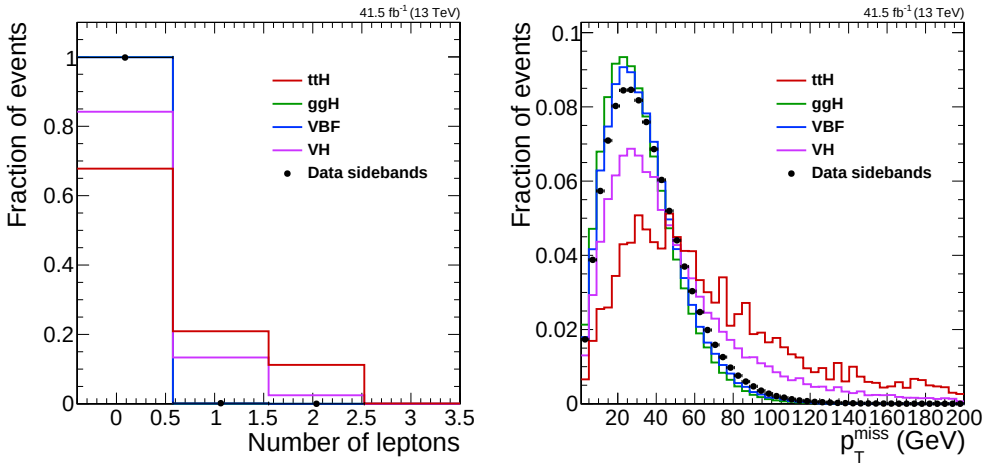


Figure 4.7: Most sensitive variables to separate $t\bar{t}H$ production from background and from the other Higgs boson production processes. The data sidebands defined in the text (black markers) are compared with the $t\bar{t}H$ signal (red histogram) and with the ggH, VBF and VH production (coloured histograms). The number of leptons (left) and the $p_{\mathrm{T}}^{\mathrm{miss}}$ (right) are shown. All the histograms are normalised to have unitary area.

4.4.1 Event categorisation for the 2016 analysis

The analysis of the data collected in 2016 aimed at a full and complete characterisation of the Higgs boson properties, including the measurement of the event yield of the ggH, VBF, VH and $t\bar{t}H$ processes. This is achieved by splitting the events in exclusive categories targeting the different production processes, according to the final state topology. At first events tagged as originating from $t\bar{t}H$, VH or VBF production are identified. The preselected events failing the exclusive categorisation are collected in the Untagged category, mainly populated by ggH production. In each category, specific selections are applied in order to increase the expected significance (see Section 4.5). Events within a given category are split in further subcategories to increase the signal-to-background ratio, when the number of expected signal events is high enough.
The analysis features a total of 14 categories, 2 for $t\bar{t}H$, 5 for VH, 3 for VBF and 4 Untagged. Five categories target VH production, three requiring leptons from the W and Z bosons, one requiring high $p_{\mathrm{T}}^{\mathrm{miss}}$ from Z boson decaying to neutrinos and one collecting events where the vector boson decays hadronically. Events produced through VBF are selected exploiting information related to the jets produced in association with the Higgs boson. The detail of the selections applied to the categories other than $t\bar{t}H$ can be found in Ref. [96], as their definition goes beyond the scope of this work. The two categories targeting the $t\bar{t}H$ production are described in detail below. An event that satisfies the selection of more than one category is assigned to the category with the highest expected signal to background ratio.
A BDT, common to all the categories, is trained to enhance the background rejection. The BDT is trained to distinguish events with two high-resolution, well-reconstructed and high-p_{T} photons compatible with originating from a Higgs boson decay from photons most likely originating from background processes. The training is performed using photons from ggH, VBF, VH and $t\bar{t}H$ events as signal, each process weighted for the respective cross section, and from QCD, γ + jet and diphoton processes as backgrounds. The input variables exploited in the training are:

- the p_{T}^{γ} of each photon;
- the pseudorapidity of each photon;
- the cosine of the angle in the transverse plane between the two photons;
- the output of the photon identification BDT for each photon;
- an estimate of the invariant mass resolution of the event;
- an estimate of the probability to correctly identify the interaction vertex.

Photons following the decay of a Higgs boson are generally high-p_{T} photons produced in the central region of the detector, as opposite to background photons that are uniform along η. The cosine between the two photons accounts for the relativistic boost of the photon pair, thus it is particularly useful in distinguishing the associated production mechanisms from the background, mainly produced at rest. The photon identification BDT is powerful in distinguish prompt photons from fake ones. Finally, the last two variables helps in identifying events with optimal invariant mass resolution.
The invariant mass resolution is estimated from the quadrature sum of the per-photon

estimates of the energy resolution of the two photons. The per-photon energy resolution is derived with the multivariate technique used to correct the photon energy as described in Section 4.3.1. In the hypothesis of wrong vertex assignment, a geometrical factor is added to the resolution, according to the position of the photons. The probability to correctly assign the vertex is extracted form a dedicated BDT, which takes as inputs information on the vertices, the number of converted photons and the transverse momentum of the diphoton system. The distribution of the output of the classifier is shown in Fig. 4.8. Events arising from $t\bar{t}H$ production peak at high scores of the BDT thanks to the recoil of the diphoton against the top quarks. The output of the classifier is validated on a sample of $Z \to e^+e^-$, with electrons reconstructed as photons, to check the level of agreement between data and simulation.
In the VBF and the Untagged categories, where the number of expected signal events is large, the diphoton BDT is exploited to define subcategories with different signal-to-background ratios. The number of categories and their boundaries are chosen to maximise the expected sensitivity. The boundaries of the Untagged categories are shown by the vertical lines in Fig. 4.8; events with a score lower than the last boundary are rejected as their inclusion add no sensitivity. In VH and $t\bar{t}H$ categories, where the events are split by topology, no further subdivision is performed. The diphoton BDT is exploited to improve the background rejection, discarding events lower than a boundary chosen independently in each category so to maximise the expected significance.

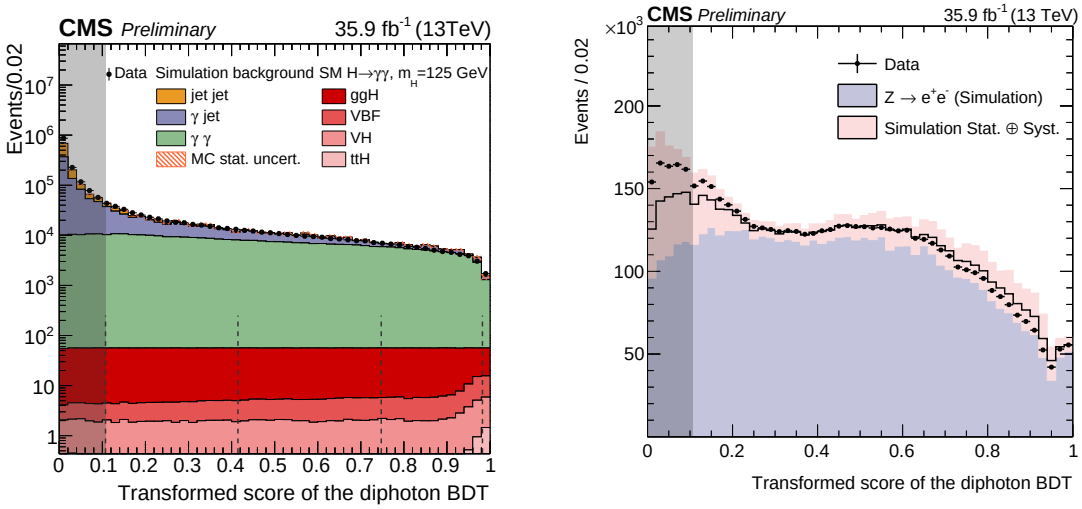


Figure 4.8: Distribution of the output of the diphoton BDT classifier. The score is transformed to have flat distribution on the signal. Left: data sidebands (black markers) are compared to the signal and to the backgrounds simulations. The vertical dotted lines represent the boundaries of the Untagged categories. Events shaded in grey are rejected. Right: distribution of the diphoton classifier output on $Z \to e^+e^-$ events with electrons reconstructed as photons. The pink band displays the simulation statistical and systematic uncertainty.

The $t\bar{t}H$ production is isolated through two categories, one collecting semi-leptonic and fully leptonic decays of the top quarks and one the fully hadronic ones. In each category, a set of selections is applied, defined to maximise the expected significance. Both the

categories exploit a control sample based on data to evaluate the background during the optimisation procedure. At first the definition of the control sample is explained, then the selections applied in each category are detailed.

Control sample definition

The figure of merit adapted to define the events in the categories is the significance. Therefore, an estimate of the signal and of the background expected after each selection under test is necessary. Once the categorisation is defined, the background model is derived directly from the data, as detailed in Section 4.6.2. If the estimate of the background comes from the same data even during the definition of the categories, the double usage of the data could induce a bias. The selections would be likely optimised on the statistical fluctuations of the sample used to extract the background estimation, and not on a independent sample with similar properties and a larger size.
To prevent this from happening, either a sample of simulated of background events is exploited or an independent data sample with properties similar to the data sidebands, referred to as control sample. At the time of performing the analysis not all the simulation background samples were available and the background has been modelled through a control sample.
The control sample is defined requiring the presence of a prompt photon and a fake one. The prompt photon is identified by applying the preselections described in Section 4.3.1. The fake photon has no preselection applied and it is required to have the photon identification BDT score lower than -0.9. As the kinematic properties of the prompt-fake photons are not the same of the preselected photons, a two-dimensional weighting in p_{T} and η of each photons is applied to match the control sample kinematic to the data sidebands one. Finally, the events in the control sample are normalised to the number of events observed in the data sidebands.
The resulting sample is a statistically independent proxy of the data sidebands with twice the number of events, providing a good model of the background for the definition of the $t\bar{t}H$ categories.

The $t\bar{t}H$ leptonic category

The $t\bar{t}H$ leptonic category targets semi-leptonic and fully leptonic decays of the top quarks, therefore events are required an electron or a muon, identified as explained in Section 4.3.3 and 4.3.4, respectively. To further suppress the contribution of the different background processes, events are required to have two jets and one b jet. The distributions of the number of jets and b jets in the signal and in the control sample are shown in Fig. 4.9.
All the selections have been defined to maximise the expected significance, estimated from fitting the diphoton invariant mass distribution. For each selection under test, the signal is modelled with the sum of two Gaussian functions whose parameters are derived from fitting the $t\bar{t}H$ simulation, while the background from an exponential fit to the $m_{\gamma\gamma}$ distribution in the control sample. The significance is thus estimated as explained in Section 4.5 and the combination of selection which provide the highest significance is exploited for the analysis. The $p_{\mathrm{T}}/m_{\gamma\gamma}$ selection of the two photons has been tuned, with the selection on the leading photon increase to $p_{\mathrm{T}}^{\gamma_1} > m_{\gamma\gamma}/2$ to enhance the rejection of the non-resonant diphoton production.

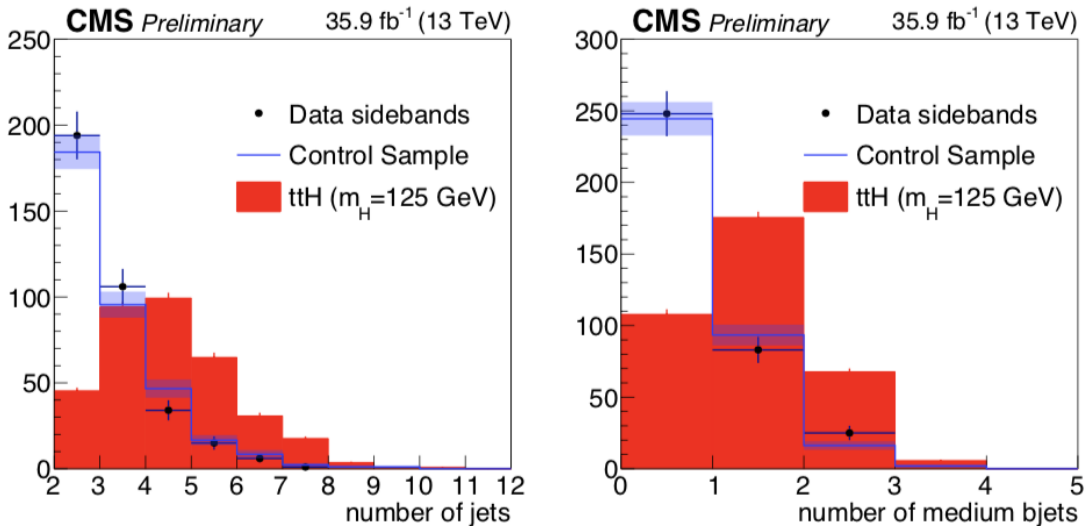


Figure 4.9: Distribution of the number of jets (left) and of b jets (right) in data sidebands (black markers), control sample (blue histogram) and in $t\bar{t}H$ signal (red histogram) for events with two photons and al least one lepton. The histograms are normalised to the number of events in the data sidebands.

An additional selection on the output of the diphoton classifier is applied to improve the background rejection. Events are required a BDT score higher than 0.11, corresponding to the lowest boundary of the Untagged categories, which ensure a signal efficiency of 95% for a background rejection of 40%. The summary of the requirements applied in this category is:

- two preselected photons with $p_{\mathrm{T}}^{\gamma_1} > m_{\gamma\gamma}/2$ and $p_{\mathrm{T}}^{\gamma_2} > m_{\gamma\gamma}/4$;
- at least one lepton (e or μ) with $p_{\mathrm{T}} > 20$ GeV;
- at least two jets with $p_{\mathrm{T}} > 25$ GeV;
- at least one b jet according to the medium working point of the CSV algorithm;
- diphoton BDT score greater than 0.11.

The $t\bar{t}H$ hadronic category

The $t\bar{t}H$ hadronic category targets fully hadronic decays of the top quarks, where the events are characterised by a large number of high-p_{T} jets and two b jets. Events with at least three jets, one b jet chosen according to the loose WP of the CSV algorithm and no leptons are preselected in this category. The background rejection is enhanced by a BDT, trained with the TMVA package [114], using variables related to the jet and b jet composition of the event as inputs.
The input variables of the BDT, shown in Fig. 4.10, are the number of reconstructed jets, the transverse momentum of the highest-p_{T} jet, and the output score of the CSV algorithm for the two jets with the highest CSV score. The training is performed on simulated $t\bar{t}H$ events as signal and non-resonant diphoton production as background.

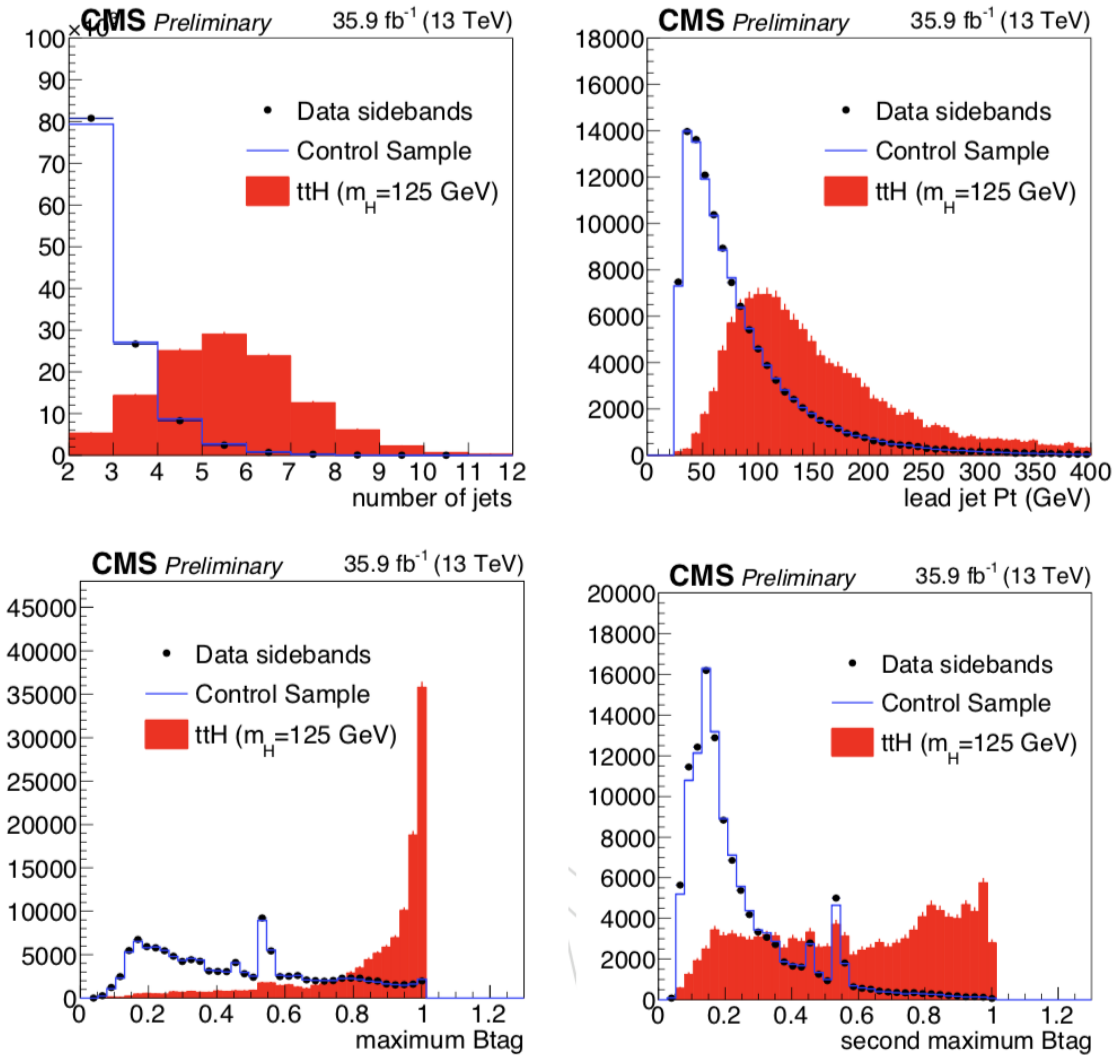


Figure 4.10: Distribution of the number of jets (top left), of the leading jet p_{T} (top right) and of the two highest scores of the CSV algorithm to identify b jets (bottom) in data sidebands (black markers), control sample (blue histogram) and in $\mathrm{t\bar{t}H}$ signal (red histogram) for events preselected in the $\mathrm{t\bar{t}H}$ hadronic category. The histograms are normalised to the number of events in the data sidebands.

The output of the BDT is shown in Fig. 4.11, comparing the control sample, the data sidebands, and the signal. Good separation arises among the $\mathrm{t\bar{t}H}$ signal and the background. Events in this category are ranked by two BDTs, one for the diphoton variables and one for the hadronic ones. To chose the boundaries to be applied on two BDT outputs, a two-dimensional scan of the thresholds is performed. The optimisation proceeded varying independently the boundaries on the two BDTs and measuring the expected significance for every selection. The significance is estimated as in the leptonic channel from fitting the $m_{\gamma\gamma}$ distribution of $\mathrm{t\bar{t}H}$ signal and of the control sample. The list of the selections applied in the $\mathrm{t\bar{t}H}$ hadronic category is reported below:

- two preselected photons with $p_{\mathrm{T}}^{\gamma_1} > m_{\gamma\gamma}/3$ and $p_{\mathrm{T}}^{\gamma_2} > m_{\gamma\gamma}/4$;
- no leptons, defined according to Sections 4.3.3 and 4.3.4;
- at least three jets with $p_{\mathrm{T}} > 25$ GeV;
- at least one b jet according to the loose working point of the CSV algorithm;
- score of the $t\bar{t}H$ hadronic BDT greater than 0.75;
- score of the diphoton classification BDT greater than 0.4.

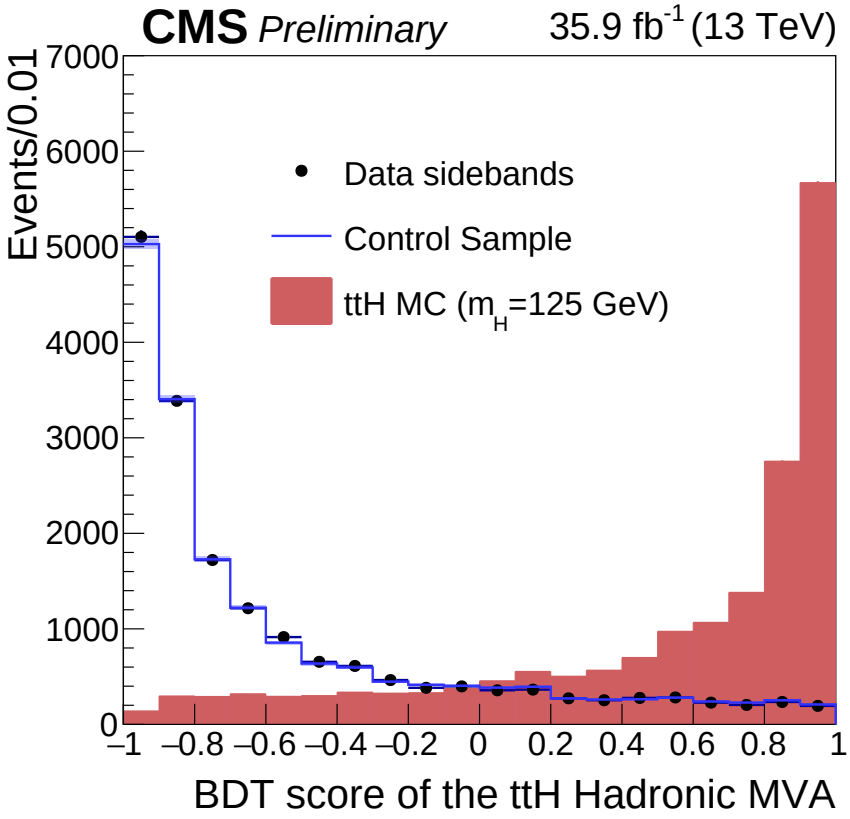


Figure 4.11: Distribution of the output score of the BDT trained in $t\bar{t}H$ hadronic category for events preselected in the that category. The data sidebands (black markers), control sample (blue histogram) and in $t\bar{t}H$ signal (red histogram) are compared. The histograms are normalised to the number of events in the data sidebands.

4.4.2 Event categorisation for the 2017 analysis

The analysis performed on 2017 data involved only the measurement of the $t\bar{t}H$ event rate, therefore only categories aiming at the $t\bar{t}H$ production have been defined. Events are split in leptonic and hadronic categories, according to the final state topology. Within each topology, events are further subdivided according to the expected signal-to-background ratio. Instead of the diphoton classification BDT, two BDTs are trained, one per topology, including variables related to photons, jets and leptons. A total of five categories is defined in the analysis, three for the hadronic events and two for the leptonic ones.
Events with two preselected photons in the invariant mass range $100 < m_{\gamma\gamma} < 180$ are selected. Both the photons are required to have the photon identification BDT score greater than -0.2. The increased threshold allows one to exclude the QCD background from the analysis phase space, as QCD events have two fake photons with low identification BDT score (see Fig. 4.4), while retaining more than 90% of the $t\bar{t}H$ signal.
The identification of the photons proceed as in the 2016 analysis with the addition of the Pixel Seed Veto (PSV) variable. The PSV improves the discrimination of electrons and photons exploiting the tracks reconstructed in the pixel detector. The position of each photon supercluster is extrapolated back to the Pixel detector. If in a cone around the extrapolated position of the photon hits in the Pixel detector not assigned to any track are present, the photon is interpreted as an electron. The variable is particularly useful in the leptonic channel, where electrons following the decay of a top quark can be misidentified as photons. The pixel seed veto efficiency on photons measured on $Z \to \mu\mu\gamma$ events is of about 70%. The radiative decay of the Z boson is used to correct the efficiency in the simulation to match the one observed in data. To avoid an efficiency reduction of 50% on signal, the PSV is not applied to the photon selection but it is exploited as input to the BDTs.
The modelling of the background for the category optimisation is handled differently than in the 2016 analysis. As the BDTs are trained using variables related to the photon pair, including the photon identification BDT score, the control sample defined in the 2016 analysis can not be exploited and the background estimation is derived directly from simulation.
The composition of the background is illustrated in Fig. 4.12, where the shapes of $m_{\gamma\gamma}$ in the data sidebands and in the background simulation are compared. The upper figure shows the $m_{\gamma\gamma}$ distribution for all the preselected photons, while the bottom figures shows $m_{\gamma\gamma}$ for events with two preselected photons with photon identification BDT score greater than -0.2 and additional selection targeting the hadronic and leptonic final state topologies. Events are tagged as hadronic, if they present at least two jets and no leptons. In leptonic events at least one lepton and one jet are required. The request of at least one lepton changes the background composition, enhancing the contribution of the $t\bar{t}$ plus photons (prompt or fake) process, where a lepton is produced, and reducing the contribution of the other processes, where the lepton candidate mainly comes from a misidentified jet.
The contribution of QCD is almost suppressed with increasing the photon identification BDT threshold, while the γ + jet component is strongly reduced. Both the processes vanish with the request of one lepton. The result is the presence of few simulated events, not suitable for training a BDT. The multijet sample is not exploited in the training of the two BDTs and the γ + jet one is not exploited in the leptonic category.
The lack of simulation for some background processes induces a mis-modelling of the

background and impacts the training of the BDTs and the choice of the selection. The chosen categorisation, based on background simulation, could be optimal on the simulated background but suboptimal on the background distribution in data. However, the $t\bar{t}H$ event rate is extracted directly from fitting the data, so the background simulation mis-modelling does not induce bias in the signal extraction.
To check the level of agreement between the data and the simulation, two procedures are adopted. The shape of the two BDTs is validated on an independent data sample of $Z \to e^+e^-$ events. The $t\bar{t}Z$ process is exploited as proxy of the $t\bar{t}H$ one to check the agreement in the shape of the BDTs. As an addition check, the categorisation based on background estimates derived from simulation is repeated with the background estimated from the data sidebands. If the boundaries of the categories are compatible between the two procedures, the mis-modelling in background simulation is judged small enough to be neglected.

The $t\bar{t}H$ leptonic categories

The leptonic categories target semi-leptonic and fully leptonic decays of the $t\bar{t}$. Events are included in this category if they feature at least one lepton (Section 4.3.3 and 4.3.4) and one jet in addition to a preselected photon pair.
A BDT is trained with the TMVA package to separate signal events from background, which in this category arises mainly from top quark pair production and non-resonant diphoton production. The BDT is trained on the $t\bar{t}H$ sample generated with POWHEG, on the diphoton and the $t\bar{t}$ samples described in Section 4.2.
The input variables, listed in Table 4.4, are related to photons, leptons, jets, b jets and to the $p_{\mathrm{T}}^{\mathrm{miss}}$ of the event. Photons variables target the identification of high-p_{T}, well-reconstructed photons; the usage of $p_{\mathrm{T}}/m_{\gamma\gamma}$ instead of p_{T} ensures that the BDT can not learn the mass of the diphoton. This assumption has been verified after the training checking the correlation between the BDT score and the invariant mass of the photon pairs. The variables related to jets target the identification of events with high number of high p_{T} jets and some jets with high probability of originating from b quarks. Finally the leptonic variables target high-p_{T} leptons and $p_{\mathrm{T}}^{\mathrm{miss}}$ originating from the decay of the top quarks. The p_{T} threshold applied to select the leptons has been lowered from the 20 GeV of the 2016 analysis to 10 GeV as the lepton p_{T} is given as input to the BDT.
The input variables have been chosen through several trainings of the BDT with changing the inputs. The performance of each training has been measured using the signal and background simulations to estimate the expected significance. The set of variables with the highest discriminating power have been chosen and a pruning of the inputs has been realised. Variables with less discriminating power have bees excluded from the list of inputs to get the smaller possible set of variables which do not compromise the sensitivity of the BDT.
Figures 4.13 to 4.16 show the distribution of all the input variables for the $t\bar{t}H$ events, the data sidebands, and the simulated backgrounds. The QCD and γ + jet samples are not shown as they are not exploited in the training. A fair agreement is found between data sidebands and background simulation for most of the variables. The residual mis-modelling mainly arises from the low number of events in the background simulation.
Some variables are not particularly useful in discriminating signal and background alone, but became powerful thanks to the correlations between each other. As an example, the φ

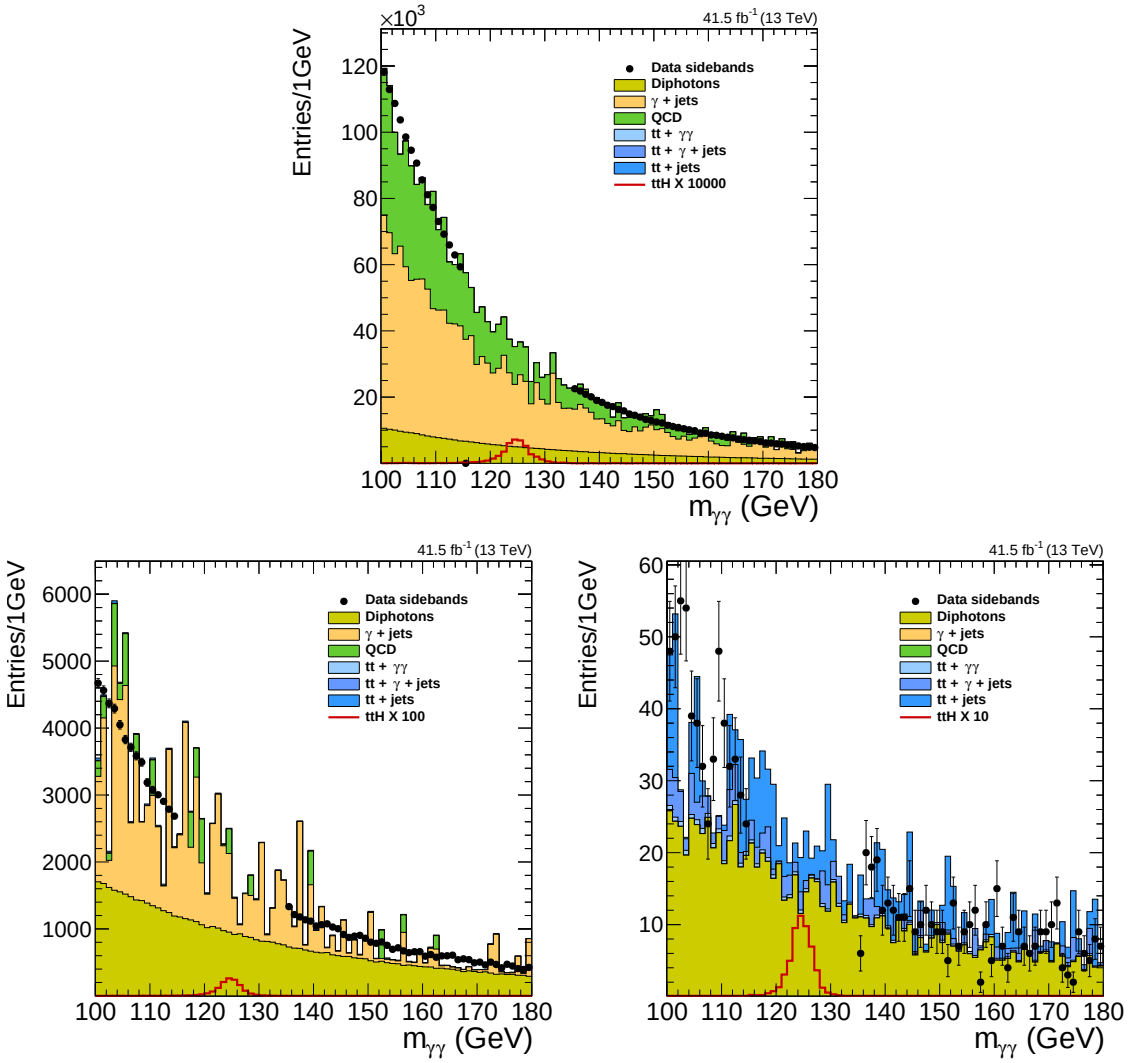


Figure 4.12: Distribution of $m_{\gamma\gamma}$ in preselected events (top), events with no leptons and two jets (bottom left) and events with at least one lepton and one jet (left). The 2017 data (black markers) are compared to the simulation of the background processes (stacked histograms). The $t\bar{t}H$ signal is also shown, rescaled by a proper factor.

distribution of the two photons is identical in data and simulation, but the difference in φ among the two photons and among each photon and the jets has high discriminating power thanks to the different kinematic of the events.

Photon variables	$p_{\mathrm{T}}^{\gamma}/m_{\gamma\gamma}$ of the two photons η^{γ} of the two photons Photon identification BDT score of the two photons $\Delta\varphi(\gamma\gamma)$ PSV of the two photons
Jet variables	Number of jets p_{T} of the three highest p_{T} jets η of the three highest p_{T} jets Number of b jets (DeepCVS at Medium WP) Score of DeepCSV algorithm for the two highest-scored jets
Lepton variables	p_{T} of the highest p_{T} lepton η of the highest p_{T} lepton
Missing momentum	$p_{\mathrm{T}}^{\mathrm{miss}}$ of the event

Table 4.4: List of the input variables of the $t\bar{t}H$ leptonic BDT.

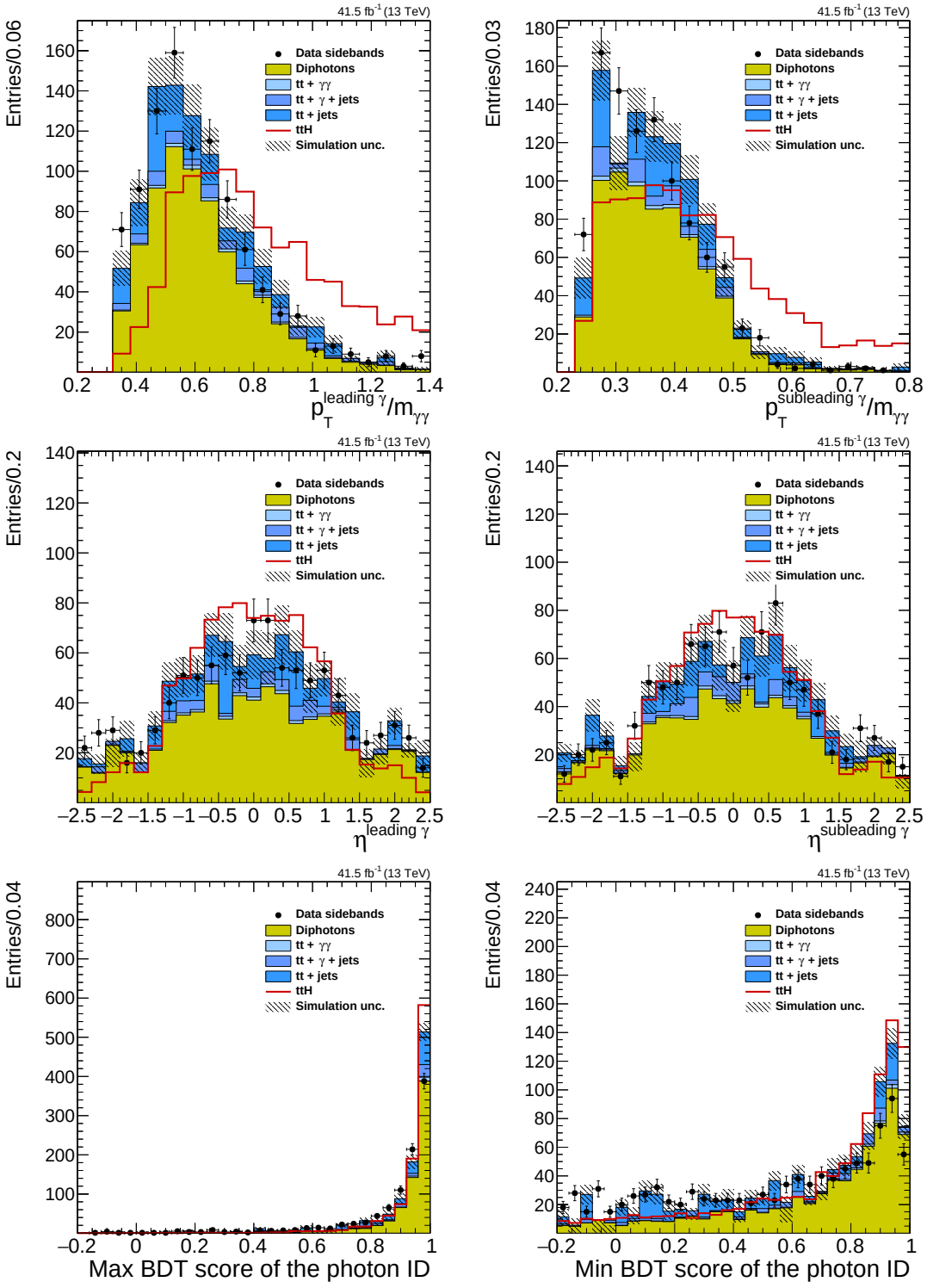


Figure 4.13: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the leptonic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The $p_T/m_{\gamma\gamma}$ of the two photons (top), their pseudorapidity (centre) and their photon identification BDT score (bottom) are shown.

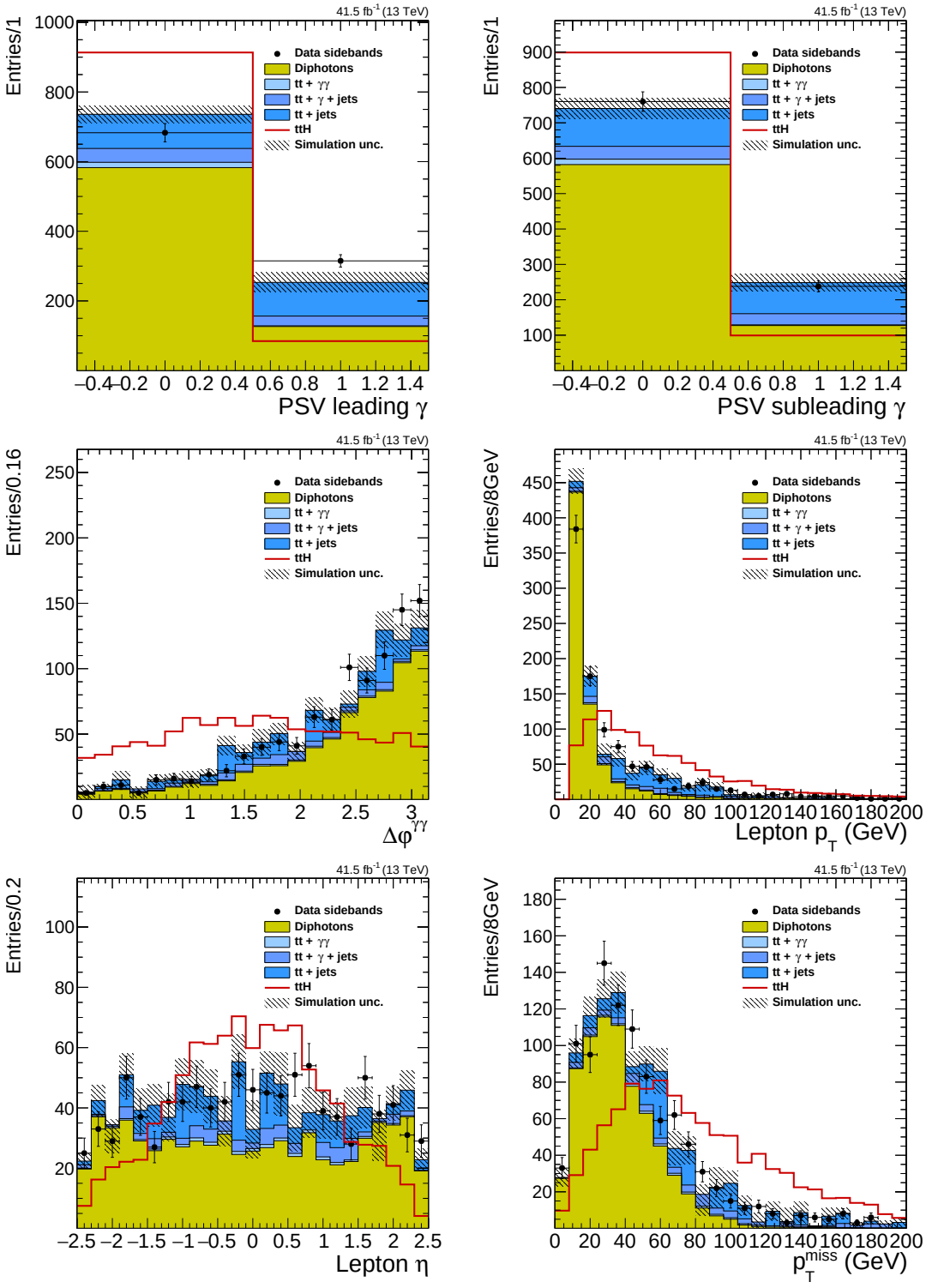


Figure 4.14: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the leptonic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The distribution of the PSV of the two photons (top), the $\Delta\varphi$ between the two photons (centre left), the $p_T$of the lepton (centre right), the η of the lepton (bottom left) and the p_T^{miss} (bottom right) are shown.

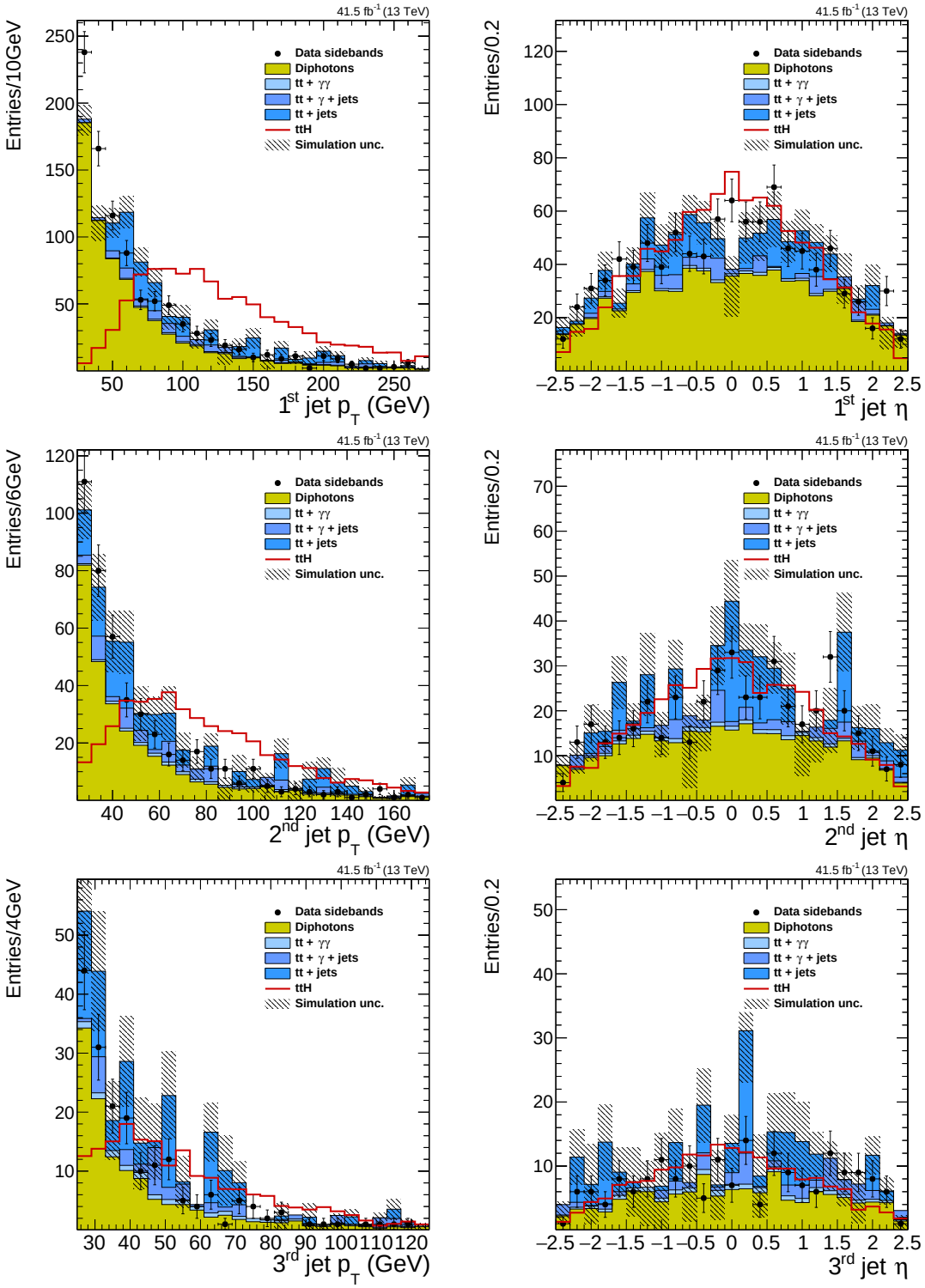


Figure 4.15: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the leptonic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The p_T (left) and η (right) of the first highest p_T jets are shown.

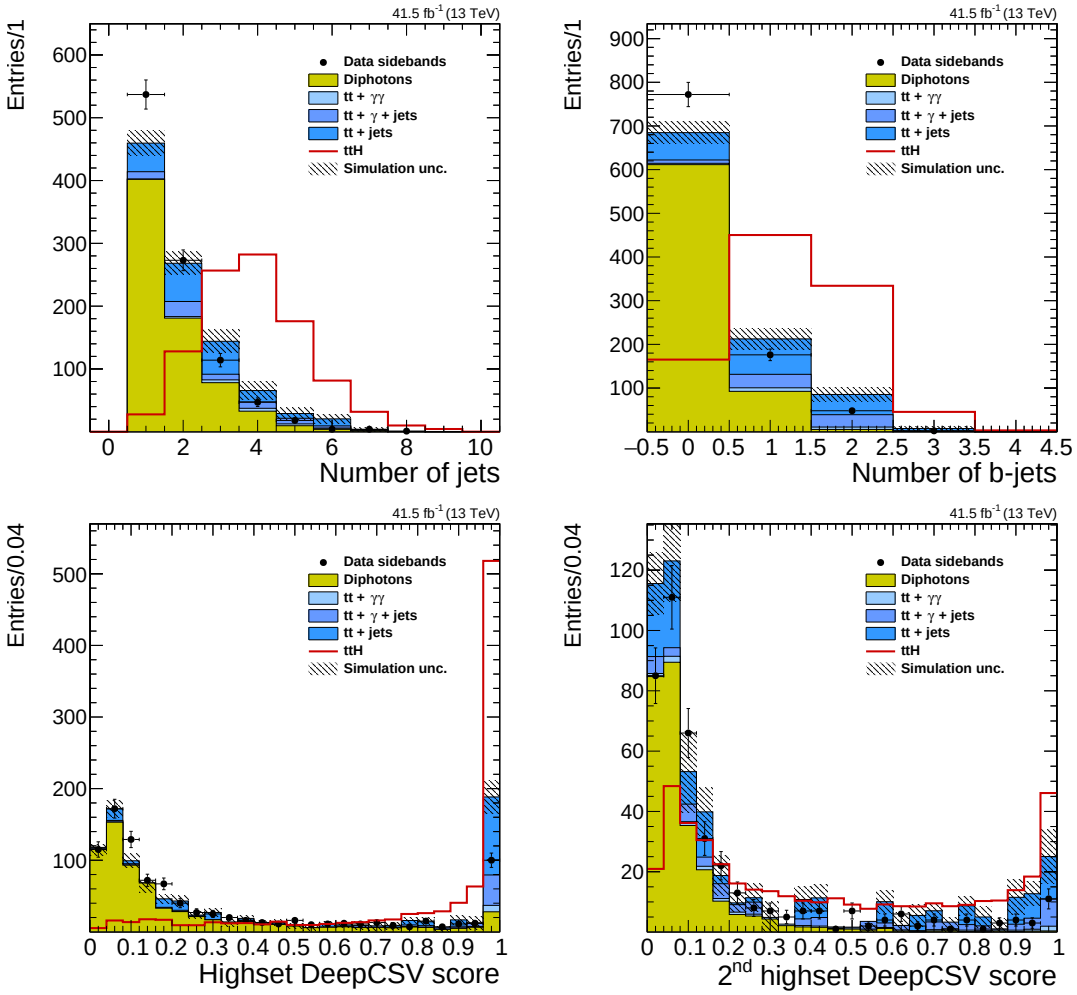


Figure 4.16: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the leptonic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The number of jets (top left), number of medium tagged b jets with the DeepCSV algorithm (top right), and the two highest DeepCSV scores are shown (bottom).

The output of the BDT is shown in Figure 4.17 while Table 4.5 shows the ranking of the input variables. The ranking is assigned during the training according to the frequency of appearance of a given variables in the BDT. The more a variable is called in the BDT the higher is its discriminating power. The training algorithm identify as most discriminating variables the PSV, the $p_T/m_{\gamma\gamma}$ of the two photons as well as the DeepCSV score of the highest scored jet. The PSV is helpful in distinguish events where one electron following the decay of the top quark is misidentified as photon and mainly suppress the $t\bar{t}$ background. The $p_T/m_{\gamma\gamma}$ of the photons is powerful in identifying photons from the $t\bar{t}H$ process recoiling on the top pair from low p_T ones coming from non-resonant production or radiated photons. Finally, the b-discriminant of the jets mainly discriminates events with the production of a $t\bar{t}$ from events with two photons and light-flavour jets.

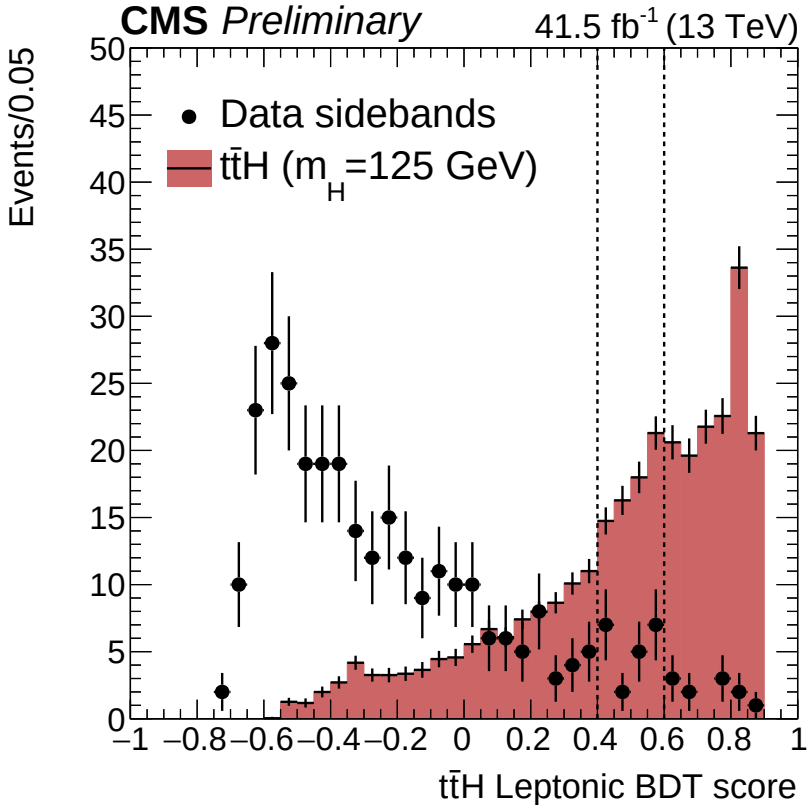


Figure 4.17: Distribution of the output score of the BDT trained to select $t\bar{t}H$ events in the leptonic categories. Events are selected requiring the presence of two preselected photons, one lepton and at least one jet. The data sidebands (black markers) and the $t\bar{t}H$ signal (red histogram) are compared. The histogram is normalised to the number of events in the data sidebands. Vertical lines displays the boundaries of the categories.

The signal region is identified by all the events preselected in this category (and used for the BDT training) with the additional requirement of the presence of one medium tagged b jet. The addition of this requirement improves the sensitivity of the category of about 10%. It has been tested that applying the requirement of the b jet before the training reduces the events in the simulation, worsening the performance of the BDT and, in turn,

Variable	Rank	Call frequency
PSV of the highest p_T photon	1	1×10^{-1}
$p_T^\gamma/m_{\gamma\gamma}$ of the highest p_T photon	2	9×10^{-2}
PSV of the lowest p_T photon	3	9×10^{-2}
DeepCSV score of the highest-scored jet	4	8×10^{-2}
$p_T^\gamma/m_{\gamma\gamma}$ of the lowest p_T photon	5	7×10^{-2}
p_T of the second highest p_T jet	6	7×10^{-2}
p_T of the lepton	7	6×10^{-2}
p_T of the third highest p_T jet	8	6×10^{-2}
Highest photon identification BDT score	9	5×10^{-2}
DeepCSV score of the second highest-scored jet	10	5×10^{-2}
Lowest photon identification BDT score	11	4×10^{-2}
η of the third highest p_T jet	12	4×10^{-2}
p_T^{miss}	13	4×10^{-2}
$\Delta\varphi(\gamma\gamma)$	14	4×10^{-2}
p_T of the highest p_T jet	15	2×10^{-2}
Number of jets	16	2×10^{-2}
η of the highest p_T photon	17	2×10^{-2}
η of the lowest p_T photon	18	1×10^{-2}
η of the second highest p_T jet	19	1×10^{-2}
η of the highest p_T jet	20	8×10^{-3}
η of the lepton	21	8×10^{-3}
Number of medium b-tagged jets (DeepCSV at the Medium WP)	22	2×10^{-3}

Table 4.5: Ranking and call frequency of the variables used as input to the $t\bar{t}H$ leptonic BDT. Highest ranked variables have the highest discriminating power.

Category	$t\bar{t}H$ events	Bkg/GeV	Significance
$t\bar{t}H$ Leptonic 0	2.4	0.3	1.3
$t\bar{t}H$ Leptonic 1	1.0	0.3	0.9

Table 4.6: The table shows the expected number of $t\bar{t}H$ events in the two leptonic categories, as well as the expected number of background events per GeV around 125 GeV of $m_{\gamma\gamma}$ and the expected significance of each category for the integrated luminosity of 41.5 fb^{-1}.

reducing the sensitivity of the analysis.
The events in the signal region are divided in two categories according to the score of the BDT. The number of categories and their boundaries are chosen to maximise the expected sensitivity, estimated from computing the significance by independently varying the boundaries of the categories. It has been verified that the addition of a third category does not bring any improvement to the significance. The significance is estimated using the $t\bar{t}H$ MadGraph5_aMC@NLO sample as a signal and the background simulation as described in Section 4.2. The procedure is repeated using the background estimate from the data sidebands as a cross check of the simulation. The optimal categorisation is the same when the significance is estimated on the background simulation or on the data sidebands. A summary of the selections applied for events included in the leptonic categories is:

- two preselected photons with $p_{\mathrm{T}}^{\gamma_1} > m_{\gamma\gamma}/3$ and $p_{\mathrm{T}}^{\gamma_2} > m_{\gamma\gamma}/4$;
- at least one lepton (e or μ) with$p_{\mathrm{T}} > 10$ GeV;
- at least one jet with $p_{\mathrm{T}} > 25$ GeV;
- at least one b jet according to the medium working point of the DeepCSV algorithm;
- output of the $t\bar{t}H$ leptonic BDT greater than 0.4.

The two subcategories are defined according to the output of the BDT as:

- $t\bar{t}H$ Leptonic 0: output of the $t\bar{t}H$ leptonic BDT greater than 0.6;
- $t\bar{t}H$ Leptonic 1: output of the $t\bar{t}H$ leptonic BDT between 0.4 and 0.6.

The number of expected $t\bar{t}H$ events and the estimated background in each category, as well as the expected significance of each category, are reported in Table 4.6. The purest category is the $t\bar{t}H$ Leptonic 0, with an expected significance of 1.3 standard deviations for 41.5 fb^{-1} of integrated luminosity.

The $t\bar{t}H$ hadronic categories

Events included in this category are required the presence of a preselected diphoton, at least two jets and no leptons. The background composition is different from the leptonic category, with a small contribution of the $t\bar{t}$ process over the dominant backgrounds due to γ + jet and to non-resonant diphoton events.
As for the leptonic case, a BDT is trained to separate the $t\bar{t}H$ signal from backgrounds. The training uses the $t\bar{t}H$ POWHEG sample as signal and the γ + jet, diphoton and $t\bar{t}$ samples as backgrounds. Since the ggH contamination is more relevant in the hadronic category than in the leptonic one, the ggH sample is also included in the training as a background.
The γ + jet sample generated in PYTHIA at LO in perturbative QCD is exploited for the training. This sample has a generator-level filter to enrich the production of jets with large electromagnetic energy, so to increase the selection efficiency. The data-to-simulation agreement is quite poor in several variables related to the jet modelling, as the sample is generated at LO. On the other hand, the MADGRAPH5_AMC@NLO γ + jet sample generated at the NLO can not be directly exploited for the training as no enrichment for the electromagnetic component of the fake photon is present at generator level. The number of events surviving the preselections in this sample is, thus, extremely limited and not suitable for the BDT training. A multistep procedure is applied to match the kinematic distributions of the γ + jet sample generated in PYTHIA at the LO with the ones generated al the NLO with MADGRAPH5_AMC@NLO. The result is a large sample generated in PYTHIA with the kinematic matched to the NLO sample to improve modelling of several variables. The increased data-to-simulation agreement is reflected in a training of the BDT with input closer to the data, and thus in a more realistic performance estimate. This procedure has not been applied in the leptonic tag, since the γ + jet component of the background has not been exploited for the BDT training.
The data-to-simulation agreement of several variables is checked with the LO γ + jet sample generated with PYTHIA. The variables with the highest discrepancy in the PYTHIA sample (number of jets, p_{T} and η of the fake photon, p_{T} of the two leading jets) are weighted to match the distribution of the MADGRAPH5_AMC@NLO one. A second step of the procedure is performed to adjust the relative contributions of the QCD, γ + jet and diphoton background in simulation to the proportion observed in data. The scale factors are derived from simultaneously fitting the data using the simulation as a template in three exclusive regions defined according to the photon identification BDT score of the two photons. Each region is enriched in one of the three background component: the QCD region is isolated by requiring both the photon candidates with photon identification score lower than -0.2, the γ + jet region is the one with one photon identification score above -0.2 and one below this value and the diphoton region is the remaining part of the data. The simulation is used as a template to fit the distribution of the photon identification BDT in data, with the normalisation of each background process left free to float in the fit. The outcome of the fit determines the relative contribution of each process in the data, and the simulation is scaled to match the measured value. As an example of the effect of the weighting procedure, Fig. 4.18 shows the distribution of lowest photon identification BDT score of the diphoton, the variable showing the largest discrepancy before the weighting procedure, before and after the procedure.
After this procedure is applied, the events are further selected by requiring the photon

identification BDT of both the photons to be greater than -0.2.
The training of the BDT proceeds in a similar way as the leptonic case, with several input variables related to photons and jets included in the training and a pruning of low rank variables. The variables included in the training are listed in Table 4.7. They involve photons variables, related to the p_{T} boost of the photons and to the quality of the photon reconstruction, and jet variables, accounting for the number and the transverse momentum of the jets. The score of the DeepCSV algorithm for the identification of b jets is also exploited for identification. The distributions of all the input variables are shown in figures from 4.19 to 4.23. The QCD sample, not exploited for the training, is not shown. Some data-to-simulation disagreement is present, especially in the modelling of the third and fourth jet distribution. The mis-modelling mainly arises from the lack of simulated γ + jet and QCD events, for which few events exists with four additional jets. As the event yield of the analysis is extracted from data, no bias in its estimation is induced by this disagreement. The level of performance on data is verified by checking the optimal categorisation on the data sidebands, as for the leptonic categories.

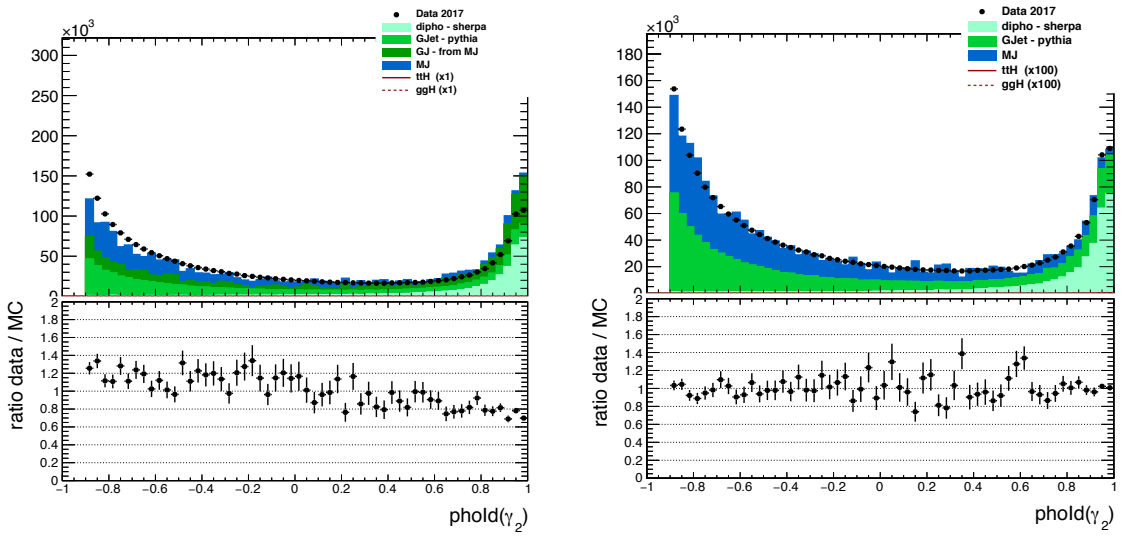


Figure 4.18: Comparison of the data to simulation agreement of the lowest photon identification BDT output score between the two photons before (left) and after (right) correcting the simulation discrepancy. The data sidebands (black markers) are compared to background simulation (stacked histograms). The bottom panels show the ratio between the data and the simulation.

Photon variables	$p_T^\gamma/m_{\gamma\gamma}$ of the two photons
	η^γ of the two photons
	φ^γ of the two photons
	$p_T^{\gamma\gamma}/m_{\gamma\gamma}$ of the diphoton
	$\eta^{\gamma\gamma}$ of the diphoton
	Photon identification BDT score of the two photons
	PSV of the two photons
Jet variables	Sum of the p_T of all the jets
	Score of DeepCSV algorithm for the three highest-scored jets
	p_T of the four highest p_T jets
	η of the four highest p_T jets
	Score of DeepCSV algorithm for the four highest p_T jets
	Number of jets
Missing momentum	p_T^{miss} of the event

Table 4.7: List of the input variables of the t$\bar{t}$H hadronic BDT.

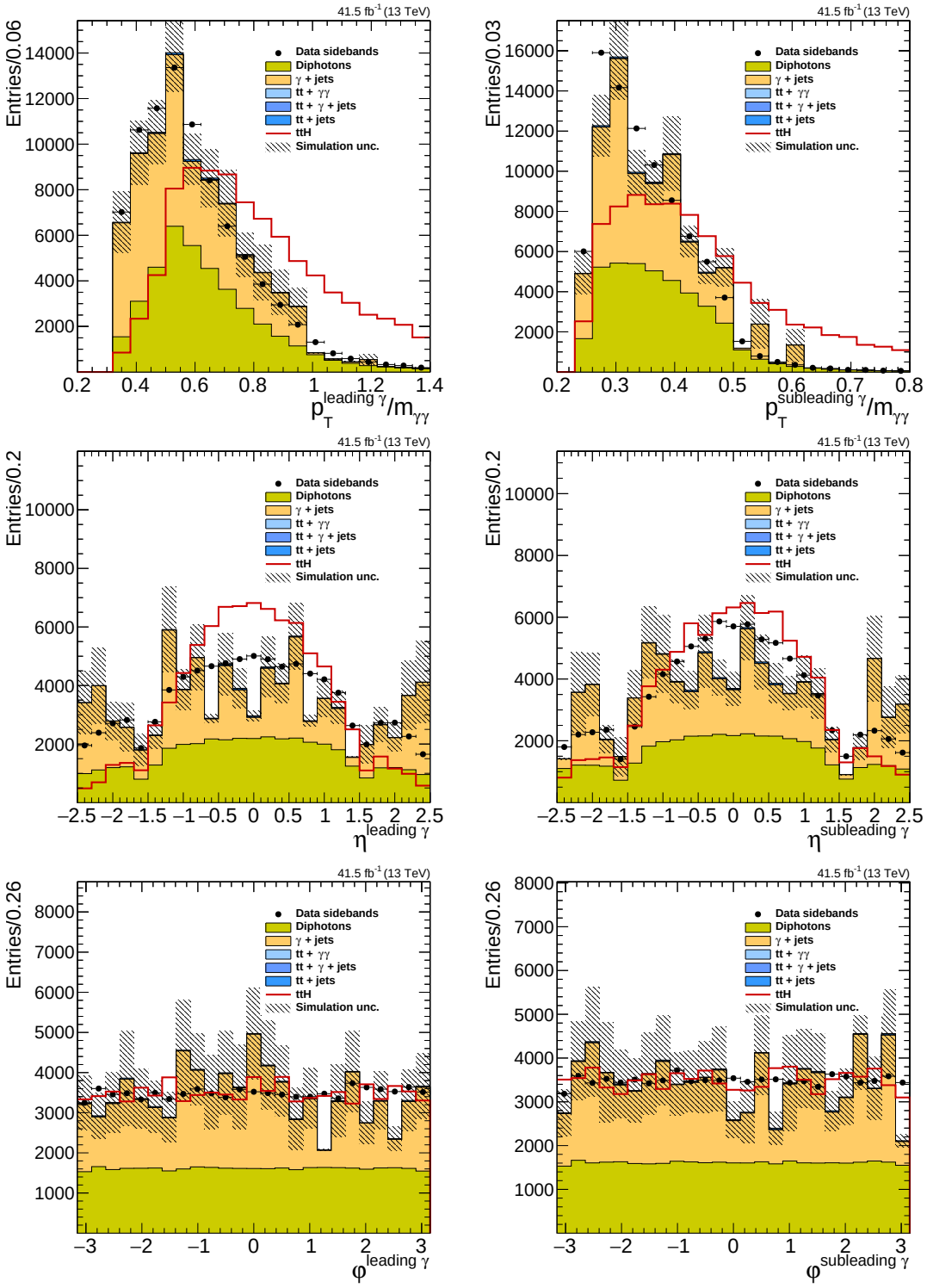


Figure 4.19: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the hadronic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The $p_T/m_{\gamma\gamma}$ of the two photons (top), their pseudorapidity (centre) and their azimuthal distribution (bottom) are shown.

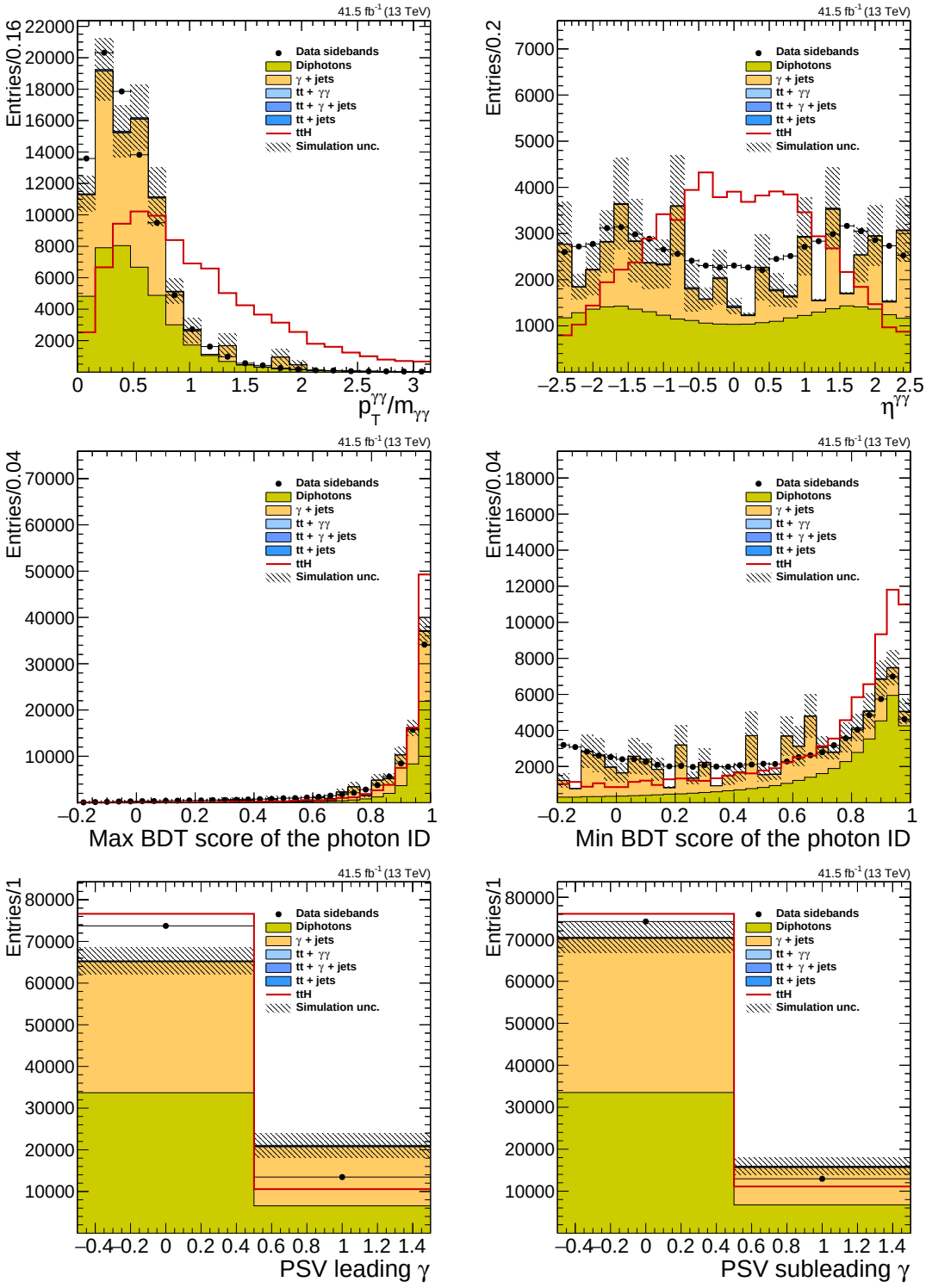


Figure 4.20: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the hadronic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The $p_T/m_{\gamma\gamma}$ of the diphoton (top left), the diphoton pseudorapidity (top right), the distribution of the photon identification BDT of the two photons (centre) and of the PSV (bottom) are shown.

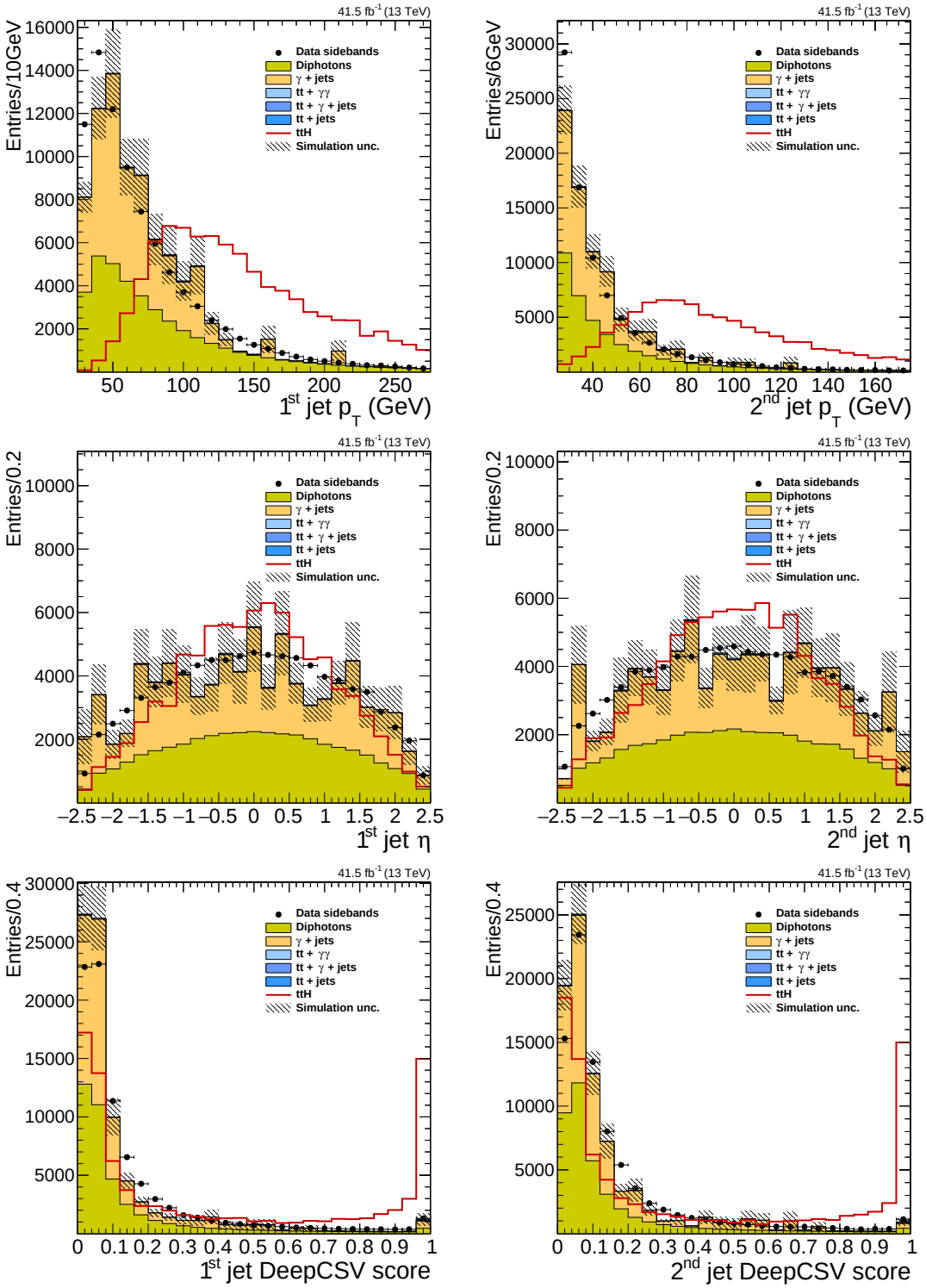


Figure 4.21: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the hadronic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The p_T (top), η (centre) and DeepCSV (bottom) score of the two highest p_T jets are shown.

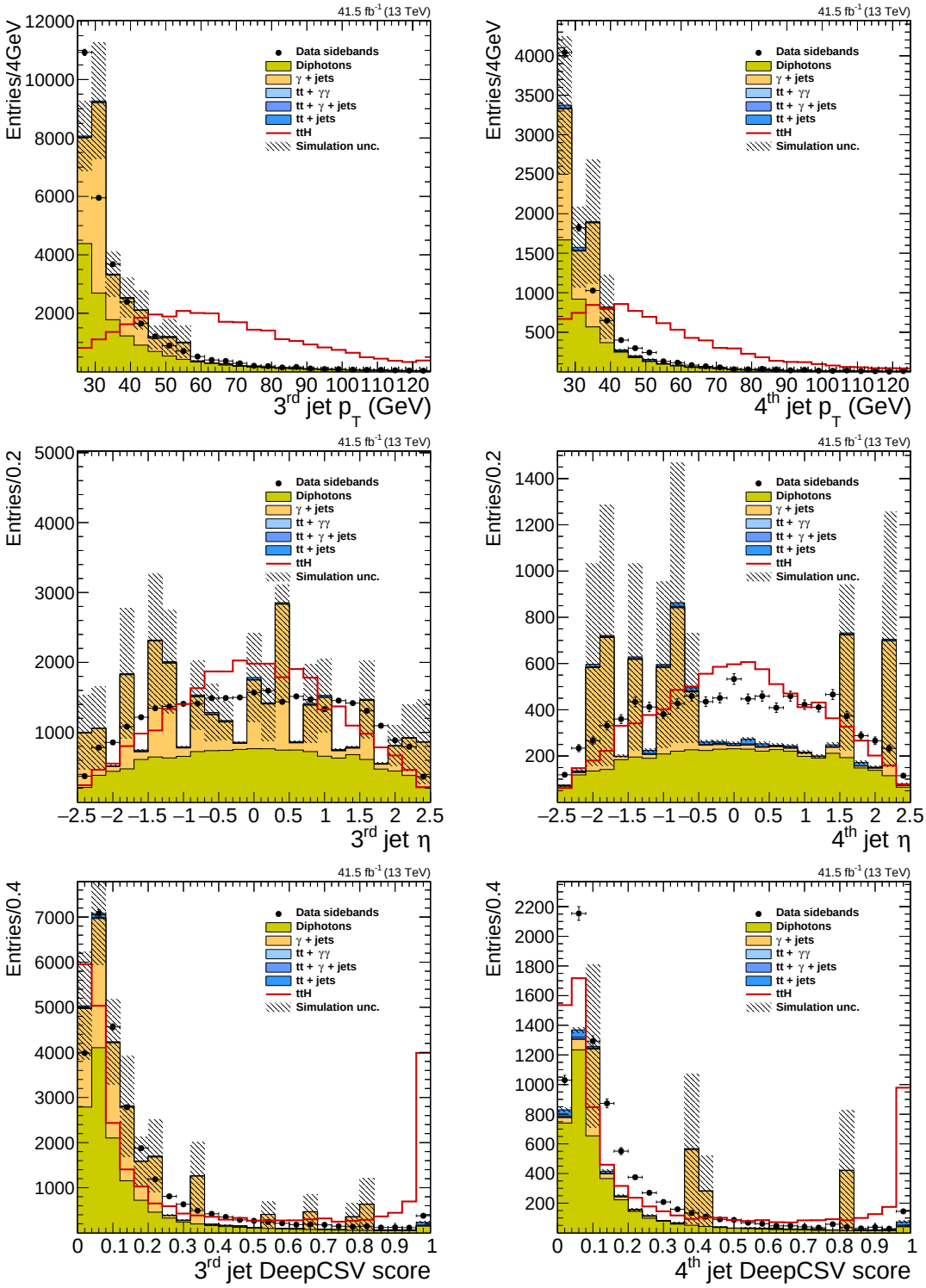


Figure 4.22: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the hadronic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The p_T (top), η (centre) and DeepCSV (bottom) score of the third and fourth highest p_T jets are shown.

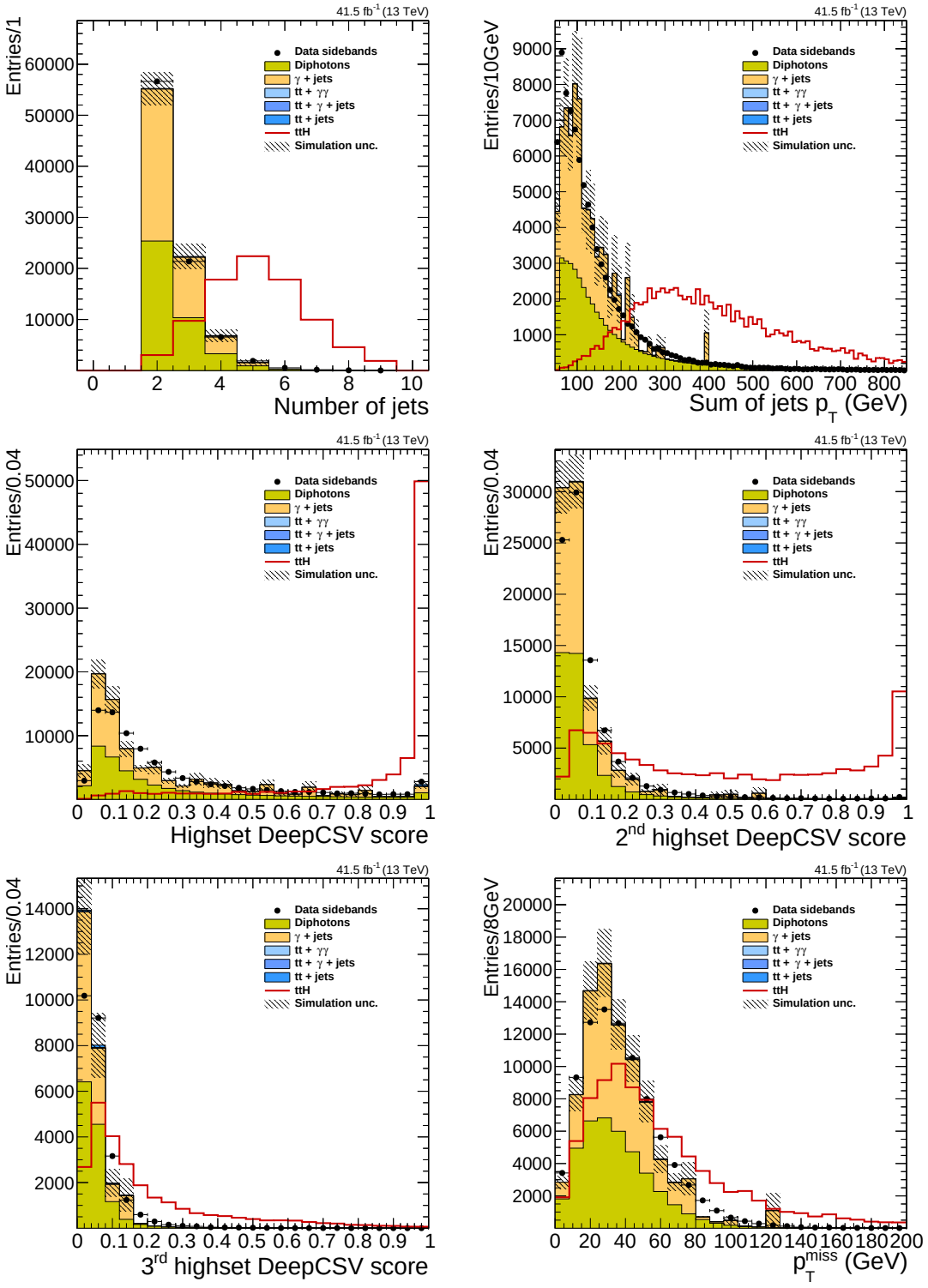


Figure 4.23: Distribution of the input variables of the BDT trained to select $t\bar{t}H$ events in the hadronic category. The data sidebands (black markers) are compared to the background simulation (stacked histograms) and to the shape expected for the $t\bar{t}H$ process (red histogram). The $t\bar{t}H$ histogram is scaled to the same area of the data. The number of jets (top left), the sum of the p_T of all the jets (top right), the highest (centre left), second highest (centre right) and third highest (bottom left) DeepCVS scores and the p_T^{miss} of the event (bottom right) are shown.

The output of the BDT is shown in Fig. 4.24, while the ranking of the input variables is listed in Table 4.8. The different ranking of the variables with respect to the leptonic channel reflects the different topology of the final state. The PSV has much less importance as the contamination of electrons in the final state is not relevant in the hadronic channel. The transverse momentum of the jets and their b-discriminant assume primary importance to distinguish the signal from the background, as they are helpful in identifying events where a $t\bar{t}$ pair is produced.

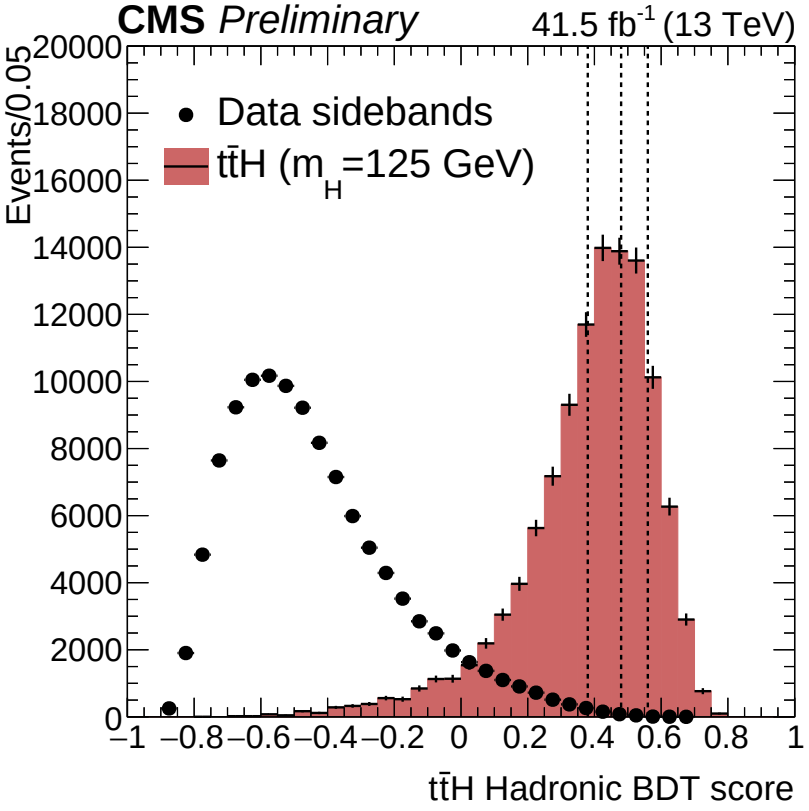


Figure 4.24: Distribution of the output score of the BDT trained to select $t\bar{t}H$ events in the hadronic categories. Events are selected requiring the presence of two preselected photons, no leptons and at least two jets. The data sidebands (black markers) and the $t\bar{t}H$ signal (red histogram) are compared. The histogram is normalised to the number of events in the data sidebands. Vertical lines displays the boundaries of the categories.

Events falling in this category are further split according to the output of the BDT. The number of categories and their boundaries are chosen as in the leptonic case in order to maximise the expected significance. Three categories are defined. The final categorisation is therefore:

- two preselected photons with $p_{\mathrm{T}}^{\gamma_1} > m_{\gamma\gamma}/3$ and $p_{\mathrm{T}}^{\gamma_2} > m_{\gamma\gamma}/4$;
- no leptons;
- at least two jets with $p_{\mathrm{T}} > 25$ GeV;

Variable	Rank	Call frequency
Sum of the p_T of all the jets	1	7×10^{-4}
DeepCSV score of the highest scored jet	2	7×10^{-4}
p_T of the third highest p_T jet	3	7×10^{-4}
p_T of the fourth highest p_T jet	4	7×10^{-4}
p_T of the highest p_T jet	5	7×10^{-4}
p_T^{miss}	6	7×10^{-4}
η of the highest p_T jet	7	7×10^{-4}
$p_T^{\gamma\gamma}/m_{\gamma\gamma}$ of the diphoton	8	6×10^{-4}
η of the second highest p_T jet	9	6×10^{-4}
Number of jets	10	5×10^{-4}
$p_T^{\gamma}/m_{\gamma\gamma}$ of the lowest p_T photon	11	5×10^{-4}
$\eta^{\gamma\gamma}$ of the diphoton	12	5×10^{-4}
DeepCSV score of the second highest scored jet	13	4×10^{-4}
p_T of the second highest p_T jet	14	4×10^{-4}
η of the third highest p_T jet	15	4×10^{-4}
η of the lowest p_T photon	16	4×10^{-4}
$p_T^{\gamma}/m_{\gamma\gamma}$ of the highest p_T photon	17	4×10^{-4}
η of the highest p_T photon	18	3×10^{-4}
Lowest photon identification BDT score	19	3×10^{-4}
DeepCSV score of the third highest scored jet	20	3×10^{-4}
DeepCSV score of the highest p_T jet	21	3×10^{-4}
φ of the highest p_T photon	22	2×10^{-4}
φ of the lowest p_T photon	23	2×10^{-4}
η of the fourth highest p_T jet	24	2×10^{-4}
PSV of the highest p_T photon	25	1×10^{-4}
Highest photon identification BDT score	26	1×10^{-4}
DeepCSV score of the second highest p_T jet	27	1×10^{-4}
PSV of the lowest p_T photon	28	1×10^{-4}
DeepCSV score of the third highest p_T jet	29	1×10^{-4}
DeepCSV score of the fourth highest p_T jet	30	1×10^{-4}

Table 4.8: Ranking and call frequency of the variables used as input to the t$\bar{t}$H hadronic BDT. Highest ranked variables have the highest discriminating power.

Category	$t\bar{t}H$ events	Bkg/GeV	Significance
$t\bar{t}H$ Hadronic 0	2.1	0.2	1.4
$t\bar{t}H$ Hadronic 1	2.6	1.1	0.8
$t\bar{t}H$ Hadronic 2	3.2	3.8	0.5

Table 4.9: The table shows the expected number of $t\bar{t}H$ events in the three hadronic categories as well as the expected number of background events per GeV around 125 GeV of $m_{\gamma\gamma}$ and the expected significance of each category for the integrated luminosity of 41.5 fb^{-1}.

- output of the $t\bar{t}H$ hadronic BDT greater than 0.38.

The three subcategories are defined according to the output of the BDT as:

- $t\bar{t}H$ Hadronic 0: output of the $t\bar{t}H$ hadronic BDT greater than 0.56;
- $t\bar{t}H$ Hadronic 1: output of the $t\bar{t}H$ hadronic BDT between 0.48 and 0.56;
- $t\bar{t}H$ Hadronic 2: output of the $t\bar{t}H$ hadronic BDT between 0.38 and 0.48.

The number of expected $t\bar{t}H$ events and the estimated background in each category, as well as the expected significance of each category, are reported in Table 4.9. The purest category is the $t\bar{t}H$ Hadronic 0, with an expected significance of 1.4 standard deviations for the integrated luminosity of 41.5 fb^{-1}.

Validation of the BDTs

The two BDTs trainings and the events categorisation is done with simulated events. While the estimate of the background from simulation is cross checked directly on data, the Higgs boson signal is taken from simulation. To ascertain that the level of performance estimated on the $t\bar{t}H$ simulation is achieved also on data, a validation of the BDTs on a sample of $Z \to e^+e^-$ events, with electrons reconstructed as photons, has been performed. The goal of the validation is to verify the data-to-simulation agreement in a sample independent from the one used for the training and where also the signal region is accessible. Events coming from the process $t\bar{t}Z$, with the Z boson decaying in a pair of electrons, are exploited as proxies of the $t\bar{t}H$. The shape of the BDT output evaluated on the $t\bar{t}Z$ sample is not expected to be the same as for the $t\bar{t}H$ process, as several variables are different (as an example the PSV has the opposite behaviour for electrons), but it offers the possibility to directly compare the agreement between the data and the simulation on a signal-like process.
Similarly to the $t\bar{t}H$, the dominant backgrounds for the $t\bar{t}Z$ process are the non-resonant production of electrons and the $t\bar{t}$ process. Minor contributions arises from production of two vector bosons (di-boson production). Data events are selected if they presents two electrons, reconstructed as photons, satisfying all the photon preselections with inverting the electron veto.
The events are split in the hadronic and in the leptonic categories according to the presence of an additional lepton, defined as in $t\bar{t}H$ events. The $t\bar{t}Z$ contribution is further enriched

Leptonic	Hadronic
Two preselected electrons with photon identification BDT > -0.2	
Invariant mass of the electrons between 70 and 110 GeV	
At least one b jet (Tight WP of the DeepCSV algorithm)	
At least one lepton	No leptons
At least one jet	At least two jets

Table 4.10: Selections applied to the events exploited for the $t\bar{t}Z$ validation in the leptonic and hadronic categories.

by applying additional selections in each category, as reported in Table 4.10. The shapes of the BDTs are compared between simulation and data, as shown in Fig. 4.25. Both the BDTs show a good discriminating power, with the $t\bar{t}Z$ signal assuming high values of the discriminant and the backgrounds presenting the opposite behaviour. The agreement between the data and the simulation is satisfactory. In the leptonic category, where the $t\bar{t}Z$ signal is easier to isolate, the data to simulation agreement is checked directly in the signal region and the shapes are found compatible within the statistical uncertainties. In the hadronic category the $t\bar{t}Z$ process is more difficult isolate because the control sample is dominated by the Drell-Yan process and a direct comparison of the $t\bar{t}Z$ with the data is not achieved. Nevertheless the BDT is proved capable to separate $t\bar{t}Z$ events from the background also in this category.

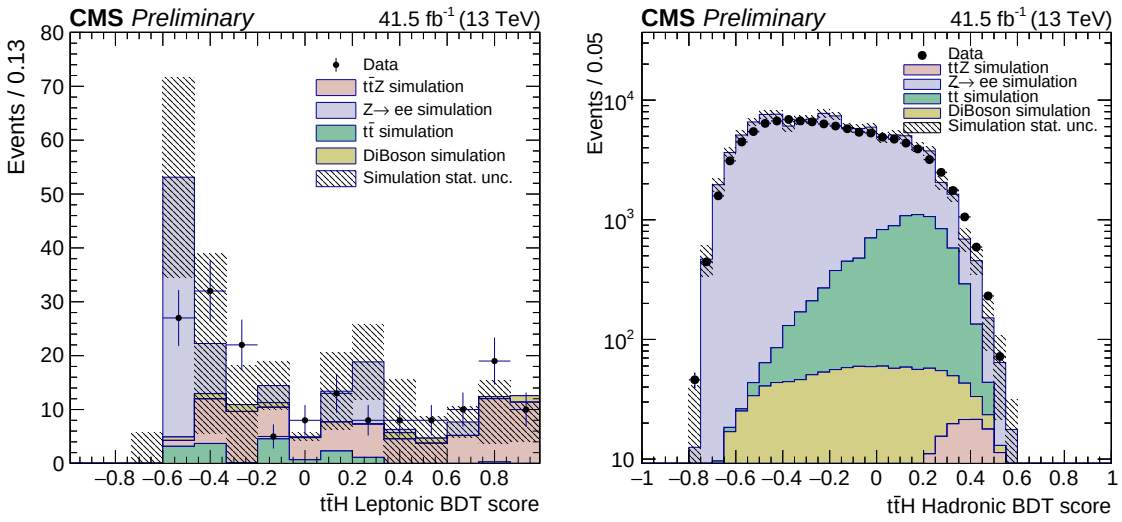


Figure 4.25: Distribution of the output score of the leptonic (left) and hadronic (right) BDTs evaluated on a sample of $Z \to e^+e^-$ events, where electrons are reconstructed as photons. The data (black markers) are compared to the $t\bar{t}Z$ and background simulation (stacked histograms).

4.5 Statistical interpretation

When performing a measurement, generally two information should be extracted from the data. The first one is how to determine the parameters of interest for the measurement from the data. The second one, relevant for the search of unobserved phenomena, is to establish if a new signal is present in the data. In the specific case of the $t\bar{t}H$ production, the two are necessary to establish which value of μ is preferred by the data and if the excess of events is enough to claim the observation of the process. The first task is performed by a maximum likelihood fit to the data, while the second one requires a hypothesis test performed according to a given test statistic. The ATLAS and CMS Collaborations adopted a common frequentist approach for all the measurement involving the Higgs boson [115]. The same approach is used in this thesis.
Let s and b be the total number of expected signal and background events. If the measurement is performed in different bins, s and b are two vectors, with s_{i} and b_{i} being the number of signal and background events in the i^{th} bin. The unbinned case is as the binned case with extremely small bin size. The number of signal events is fixed to the SM prediction and the signal strength modifier μ is used as a free parameter to scale the signal yield. The total number of expected events in the i^{th} bin is:

$$n_{\mathrm{i}}^{\mathrm{exp}} = \mu s_{\mathrm{i}} + b_{\mathrm{i}}. \tag{4.3}$$

If the SM holds, $\mu = 1$. The measurement consists in estimating μ, which means to infer which value of μ is preferred by the data. Different inference methods exist, the one adopted here consists in performing a maximum likelihood fit to the data. The likelihood function $\mathcal{L}$ is a function of the data and it is defined as the probability to observe $\mathrm{n}^{\mathrm{obs}}$ events when $\mathrm{n}^{\mathrm{exp}}$ are expected:

$$\mathcal{L}\left(n^{\mathrm{obs}}|\mu\right) = P(n^{\mathrm{obs}}|n^{\mathrm{exp}}) = P(n^{\mathrm{obs}}|\mu s + b). \tag{4.4}$$

In case of a binned distribution, the likelihood is the product of the probability computed on each bin. As the probability to observe a given number of events in a bin follows a Poisson distribution, the likelihood function can be written as:

$$\mathcal{L}\left(n^{\mathrm{obs}}|\mu\right) = \prod_i \frac{\left(\mu s_{\mathrm{i}} + b_{\mathrm{i}}\right)^{n_{\mathrm{i}}}}{n_{\mathrm{i}}!} e^{-(\mu s_{\mathrm{i}} + b_{\mathrm{i}})}, \tag{4.5}$$

where $\mathrm{n_i}$ is the number of events observed in the i^{th} bin and the product runs on all the bins included in the measurement. The best estimate of μ is assumed to be the one which maximise the likelihood function. It is common practice to minimise the negative logarithm of the likelihood (NLL) function instead of maximising the likelihood itself. As the logarithm is a monotonic function, the maximum of the likelihood coincides with the maximum of the log-likelihood and thus with the minimum of the NLL. The maximisation of the likelihood can easily incur into algorithmic problems, since the product of several factors can reach values too high to be represented on a computer. Moreover, most of the routines perform function minimisation and not maximisation, therefore the NLL is more suitable for computer application. The one and two standard deviations uncertainties for a function with a single parameter of interest can be computed as the intervals around the minimum of the NLL for which $2\Delta\mathrm{NLL} = 1$ and $2\Delta\mathrm{NLL} = 4$, respectively.

In the general case, the parameter μ is accompanied by a set of nuisances parameters, expressing other variables which affect the estimation of μ without being of direct interest for the measurement. Let θ be the vector containing all the nuisance parameters. Usually nuisance parameters are estimated from auxiliary measurements, performed using different data with respect to the one exploited to estimate μ. Auxiliary measurements have in general a Bayesian interpretation, with a function $\rho(\theta|\tilde{\theta})$ that represents the degree of belief of the real value of θ given the constraint $\tilde{\theta}$, the constraint being the auxiliary measurement. The Bayes' theorem allows a frequentist reinterpretation of the auxiliary measurements, resulting a frequentist probability density function $p(\tilde{\theta}|\theta)$:

$$\rho(\theta|\tilde{\theta}) = p(\tilde{\theta}|\theta) \cdot \pi_\theta(\theta). \tag{4.6}$$

where $\pi_\theta(\theta)$ is the prior over the auxiliary measurement which is assumed to be flat. The likelihood function in presence of nuisance parameters θ (uncorrelated between each other) is therefore modified as:

$$\mathcal{L}\left(n^{\text{obs}}, \tilde{\theta}|\mu, \theta\right) = P(n^{\text{obs}}|\mu s + b) \cdot p\left(\tilde{\theta}|\theta\right). \tag{4.7}$$

The function ρ for all the nuisance parameters involved in the t$\bar{\text{t}}$H measurement is a log-normal distribution:

$$\rho(\theta|\tilde{\theta}, \kappa) = \frac{1}{\sqrt{2\pi}\log\kappa} \exp\left(-\frac{\log^2 \theta/\tilde{\theta}}{2\log^2\kappa}\right)\frac{1}{\theta}. \tag{4.8}$$

The log-normal constraint is similar to a Gaussian constraint with the advantage of having a support in $[0, +\infty)$. This support avoids the complication arising from truncated Gaussian functions. For small uncertainties, the Gaussian function with variance ϵ and a log-normal distribution with $\kappa = 1 + \epsilon$ are identical.
When the likelihood has been defined, the fit on the data determines the best values for the parameters of interest as well as for all the nuisances parameters. The effect of nuisances parameters is to widen the likelihood phase space, resulting in larger uncertainties on the parameters of interest.
Once the value of μ has been established, the hypothesis test is performed to determine whether the measurement is compatible with a new signal. Two hypothesis are tested, the background-only or null hypothesis H_b and signal-plus-background or alternate hypothesis $H_{\mu s+b}$. The outcome of the hypothesis test is summarised in a single value representing the probability to observe an equal or greater incompatibility with the hypothesis under test. If the H_b is rejected (the data show high incompatibility with the null hypothesis), the signal is observed and a discovery is claimed. In the opposite case, if the $H_{\mu s+b}$ hypothesis can be rejected, no signal is observed and a limit on μ is set. The level of disagreement is quantified by a p-value, computed under some test statistics. It is common to express the p-value in terms of significance Z, defined such that the p-value is equal to the probability to measure a Gaussian distributed variable Z standard deviations from its mean.
The test statistic used in particle physics is built from the profile likelihood ratio $\lambda(\mu)$:

$$\lambda(\mu) = \frac{\mathcal{L}\left(n^{\text{obs}}, \tilde{\theta}|\mu, \hat{\hat{\theta}}\right)}{\mathcal{L}\left(n^{\text{obs}}, \tilde{\theta}|\hat{\mu}, \hat{\theta}\right)}. \tag{4.9}$$

The denominator is the absolute maximum of the likelihood function and $\hat{\mu}$ and $\hat{\theta}$ are the values of μ and θ which maximise the likelihood. The numerator is the maximum of the likelihood function when setting μ to a given value under test, while $\hat{\hat{\theta}}$ is the value of θ which maximise the likelihood when μ is fixed to the predefined value. In case of discovery of a positive signal, the value of μ under test is zero, since the goal is to reject H_b. The test statistic $\lambda(\mu)$ is close to unity if the chosen value of μ is close to $\hat{\mu}$. A more convenient way to express the same information is to use the test statistic q_μ:

$$q_\mu = -2 \log \lambda(\mu). \tag{4.10}$$

It is definite positive and higher values of q_μ means higher incompatibility between the data and the hypothesis under test. The p-value to quantify the level of disagreement between the hypothesis and the data is given by the probability of finding a value of q_μ greater than the observed one:

$$p_\mu = \int_{q_\mu^{\text{obs}}}^{\infty} f\left(q_\mu|\mu\right) dq_\mu, \tag{4.11}$$

where $f\left(q_\mu|\mu\right)$ denotes the probability density function of q_μ under the assumption of the value of μ. For the specific case of a discovery of a positive signal, the test statistics takes the form of:

$$q_0 = \begin{cases} -2 \log \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}, \tag{4.12}$$

since the aim is to reject the null hypothesis for which $\mu = 0$. Setting the test statistics to zero in case of negative measured signal strengths prevents from using downward fluctuations of the background against the background-only hypothesis.
To express the level of disagreement between the data and the null hypothesis, the probability density function for the test statistics q_0 under the null hypothesis should be computed. Either an analytical expression is found or it can be sampled generating pseudo-data distributed according to the null hypothesis and, for each pseudo-experiment, sampling the value of q_0. It is also necessary to compute the distribution $f\left(q_0|\mu'\right)$ with $\mu' \neq 0$, to know how the test statistics is distributed under the signal-plus-background hypothesis. In the estimation of the median expected sensitivity, it is sufficient to set $\mu' = 1$. Once the two distributions are known, the value of p_0 is computed on data and if it is value is less than $p_0 < 2 \times 10^{-7}$ (huge incompatibility between the data and the null hypothesis), a discovery of a signal is claimed. The threshold value for the p-value corresponds to a significance of Z=5.
The analyses described in this thesis makes use of the asymptotic approach [116] to express analytically the distribution of $f\left(q_0|\mu\right)$. The analytical knowledge of the distribution of q_0 avoids the necessity to generate a huge amount of pseudo-experiments, saving conspicuous CPU time. The derivation of the distribution of $f\left(q_0|\mu\right)$ starts from the Wald identity [117]. From that it can be derived that, under the hypothesis of a Gaussian distributed $\hat{\mu}$ with mean μ' and standard deviation σ_μ, the test statistics q_μ is expressed as:

$$q_\mu = \frac{\mu - \hat{\mu}}{\sigma_\mu} + \mathcal{O}\left(\frac{1}{\sqrt{\text{N}}}\right), \tag{4.13}$$

where N is the size of the data sample. Using this assumption and neglecting the asymptotic term which tends to zero for large samples, it can be proven that the test statistics q_μ follows a non-central χ^2 distribution with one degree of freedom. In the specific case of q_0 under the null hypothesis, the distribution is given by:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\chi^2(q_0), \tag{4.14}$$

where δ is the Dirac distribution to keep into account that for all the negative values of $\hat{\mu}$, q_0 is set to zero. Starting from this, the significance is expressed as:

$$Z = \sqrt{q_0}. \tag{4.15}$$

The asymptotic approach allows a simple and elegant way to test the presence of new phenomena in the data. When the median expected sensitivity of the analysis is estimated, the test statistics q_0 is built with $\hat{\mu} = 1$, while for the measurement on the data it is computed with the value inferred from the maximum likelihood estimation.

4.6 Signal and background modelling

This section describes how the signal-plus-background model is derived. The model consists of a function suitable to fit the $m_{\gamma\gamma}$ distribution. The function is composed by a background model, to describe the smoothly falling continuous background, and by a signal one, to describe the resonant Higgs boson peak. The signal model is derived from fitting the shape of the simulation and it is normalised to the SM expectation. The background is modelled directly from the data, therefore its shape and normalisation do not rely on simulation.

4.6.1 Signal model

The shape of the $m_{\gamma\gamma}$ peak is parametrised separately for each category included in the analysis. As the width of the Higgs boson with $m_{\mathrm{H}} \approx 125$ GeV is as small as 4 MeV, the shape is completely dominated by the experimental resolution and the Breit-Wigner contribution due to the resonance can be omitted.
The signal model is derived from fitting the signal simulation with the sum of at most five Gaussian functions. The model accounts for all the Higgs boson production mechanisms, weighted for the respective cross sections, and it includes the tuning of the simulation described in the previous sections to mitigate the observed discrepancies with the data in photon energy scale and resolution, trigger efficiency and object identification efficiencies. The choice of the vertex impacts considerably the shape of $m_{\gamma\gamma}$, as explained in Section 4.3.2. Consequently events where the vertex is assigned within 1 cm from the true interaction vertex (correct vertex scenario) are fitted separately from the others (wrong vertex scenario). The two contributions are summed in the signal model with the vertex assignment efficiency measured from simulation. The uncertainty of the vertex finding efficiency is an additional source of systematic uncertainty.
For each category, process, and vertex scenario, a simultaneous fit to the simulation with the Higgs boson mass m_{H} in the range from 120 to 130 GeV is performed, in order to obtain a parametric variation of the parameters of the Gaussian functions with m_{H}. The variation as a function of the mass is described with polynomials. Figure 4.26 shows an

example of the fit to the simulation in one of the categories included in the 2017 analysis as well as the variation of the signal model as a function of m_{H}.

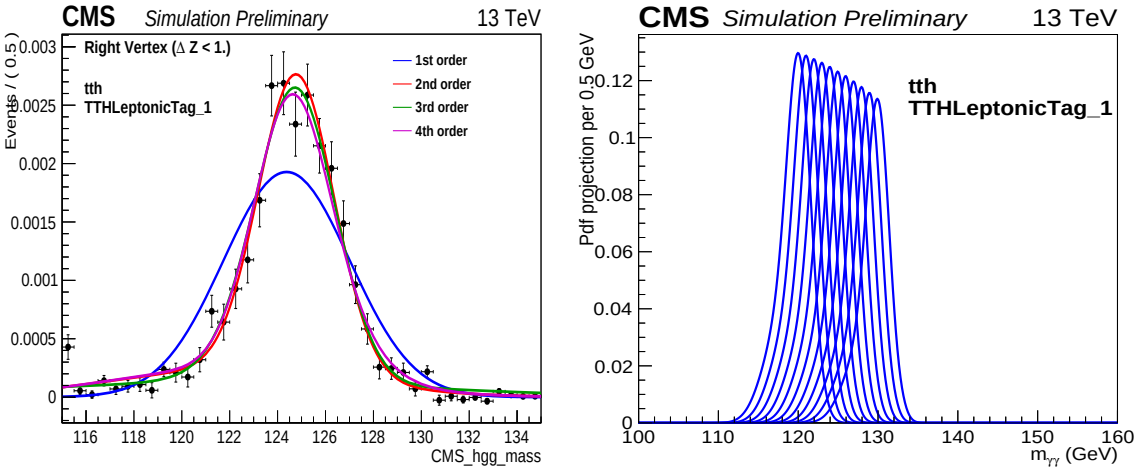


Figure 4.26: Left: signal fit to the $t\bar{t}H$ Leptonic 1 category of the 2017 analysis. The fit is shown for the $t\bar{t}H$ events included in this category in the right vertex scenario. The fit is performed with one, two, three or four Gaussian functions (solid lines). The agreement with the simulation (black markers) improves when adding more functions. Left: variation of the signal model as a function of m_{H} for the same category of the left panel.

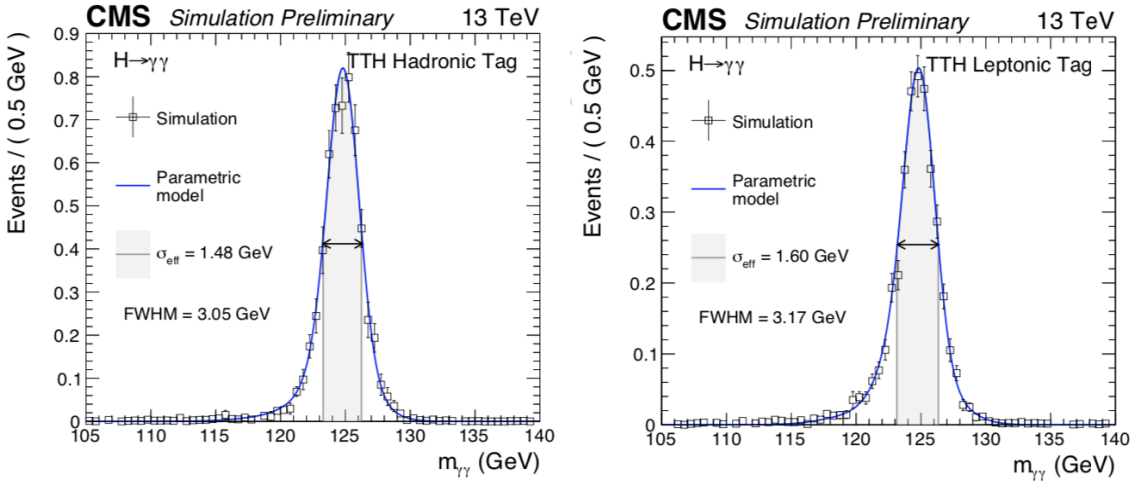


Figure 4.27: Signal model for $m_{\mathrm{H}} = 125$ GeV derived for the different categories included in the 2016 analysis. The signal model (blue line) is shown superimposed to the simulation (white markers). The resolution of each category is quantified by σ_{eff}, half of the narrowest interval containing 68.3% of the events, and by FWHM, the width of the peak at half of its maximum.

For each category, the final fit function is obtained from summing the functions for each production process and vertex scenario, weighted for the relative contribution expected from the simulation. The signal models are shown Fig. 4.27 for the 2016 categories and in

Fig. 4.28 for the 2017 ones. The signal model is shown superimposed to the simulation used to derive the fit function. The resolution of the signal is quantified by σ_{eff}, half of the narrowest interval containing 68.3% of the events, and by the Full Width of the peak at Half of its Maximum (FWHM). The two values are also shown in the figures. As expected, the resolution in 2016 is better than in 2017, due to the different ECAL conditions. The two resolutions are expected to become comparable after the final recalibration of the ECAL, described in Section 3.1. The efficiency times acceptance of the analysis on $t\bar{t}H$ events is of 21% and 24% in the 2016 and 2017 analyses, respectively.

4.6.2 Background model

The background model is built from fitting directly the data. The data-driven technique avoid the usage of simulation for the background estimation, eliminating all the systematic uncertainties related to the background simulation. As no theoretical model is capable to predict the exact shape for the background, some arbitrariness is related with the choice of the fit function. When a small signal arises over the background, even a tiny variation in the background function can significantly change the estimation of μ. For this reason, when choosing a particular fit function, an uncertainty rises due to the arbitrariness of the choice. The 'envelope' method [118] has been specifically designed to estimate the uncertainty due to the choice of a particular functional form.
The basic idea is to treat the choice of the fit function as a discrete nuisance parameter in the likelihood function. The effect of the nuisance parameters is to widen the likelihood function, causing a larger uncertainty on the estimation of the parameter of interest, as illustrated by Fig. 4.29. A NLL minimisation is performed with a model including a parameter of interest x as well as several nuisance parameters free to float in the fit, within their constraints. The value of the NLL as a function of x is depicted by the black curve. The red dotted curves represent the same minimisation when fixing the nuisance parameters to some values (fixing their uncertainties to zero). The result is a narrower NLL function, with the minimum varying as the nuisance parameters are changed. The blue solid line represents the NLL when the nuisance parameters are fixed at their best-fit value. Its minimum is in the same position of the black curve but the NLL is narrower. The idea of the method is to exploit the envelope of the likelihood functions computed at fixed values of the nuisance parameters (green dotted line) as an approximation of the black solid curve. As can be seen from Fig. 4.29, even with few points in the scan of the nuisance parameters, the envelope is already a good approximation of the full likelihood scan. If one of the red curves happens to be always above the others, it does not contribute to the creation of the envelope and thus it is automatically ignored.
In principle, the choice of the function could be treated as the other nuisance parameters, so to derive directly the black curve of Fig. 4.29. However, minimisation algorithms capable to work with discrete nuisance parameters are generally not reliable. The envelope method provides a good compromise between algorithmic performance and statistical reliability. Several intensive tests of the method have been performed in Ref. [118]. A large set of pseudo-experiment has been generated, checking the coverage of the method. The coverage is computed as how many times the generated value of the parameter of interest falls within the uncertainty of the fitted one. Excellent coverage was found, compatible with the one expected by generating a background according to a known function and using the same function to refit the generated data.

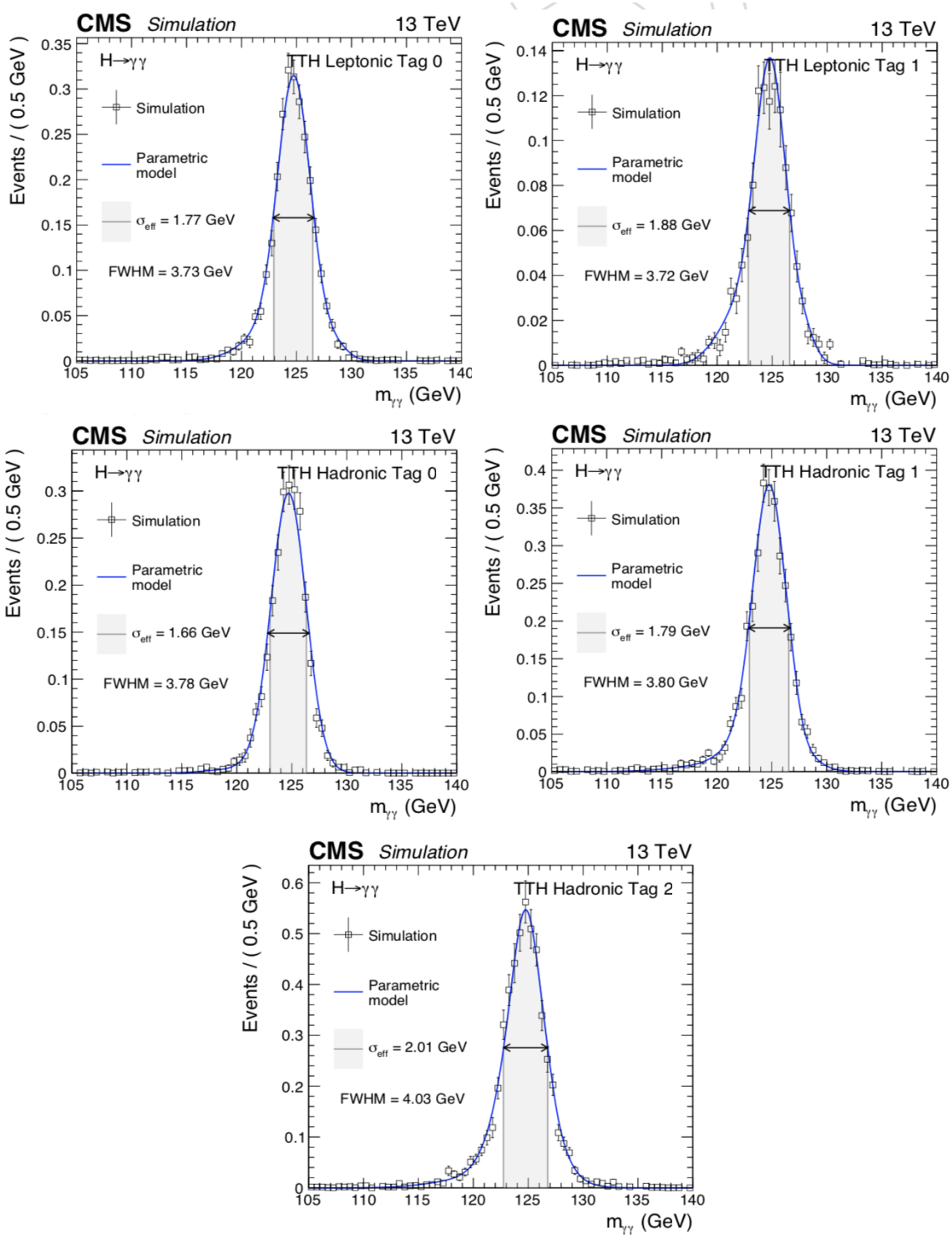


Figure 4.28: Signal model for $m_{\mathrm{H}} = 125$ GeV derived for the leptonic (top) and hadronic (centre and bottom) categories included in the 2017 analysis. The signal model (blue line) is shown superimposed to the simulation (white markers). The resolution of each category is quantified by σ_{eff}, half of the narrowest interval containing 68.3% of the events, and by the FWHM, the width of the peak at half of its maximum.

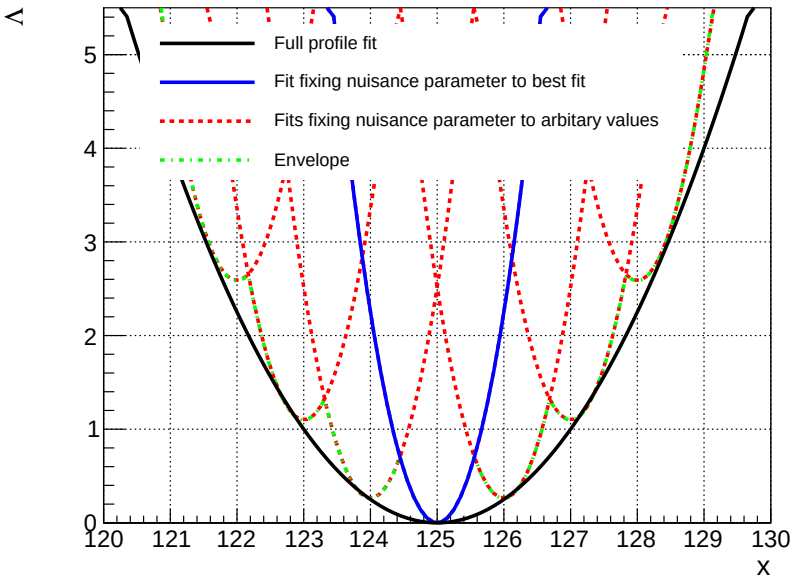


Figure 4.29: Illlustration of the envelope method. The red dotted lines represent twice the negative log-likelihood (Λ) as a function of the parameter of interest x when fixing the nuisance parameters to different values while the blue solid represents the same curve when the nuisance parameters are fixed to their best-fit value. The envelope (green dotted line) of the NLLs is built by taking for each value of x the minimum log-likehood function among the red lines. Even with a coarse scan of the parameters, the envelope is a good approximation of the full likelihood, with all the nuisance parameters free to float (black solid line) [118].

This method is applied to the choice of the background function for the H $\rightarrow$ $\gamma\gamma$ analysis. Each possible function capable of describing the background is numbered with an integer. The choice of the function is treated as a discrete nuisance parameter in the NLL minimisation. The best-fit value of μ coincides with the background function that provides the best fit (as the blue curve in Fig .4.29), but its uncertainty is widened by the envelope due to the other functions. The method keeps into account the uncertainty due to the arbitrariness of the choice of the functional form.
For the H $\rightarrow$ $\gamma\gamma$ analysis, four families of functions are considered to describe the smoothly falling $m_{\gamma\gamma}$ spectrum:

- Sum of power law functions: $f(x) = \sum_{i=0}^{N} p_{2i} x^{p_{2i+1}} = p_0 x^{p_1} + p_2 x^{p_3} + \ldots$;
- Sum of exponential functions: $f(x) = \sum_{i=0}^{N} p_{2i} e^{p_{2i+1} x} = p_0 e^{p_1 x} + p_2 e^{p_3 x} + \ldots$;
- Laurent series: $f(x) = \sum_{i=0}^{N} p_i / x^{i+4} = p_0/x^4 + p_1/x^5 + \ldots$;
- Sum of polynomials (in Bernstein basis): $f(x) = \sum_{i=0}^{N} p_i \binom{N}{i} x^i (1-x)^{N-i}$.

All the p_i and i are parameters free to float in the fit. In principle, a full orthonormal basis in the functional space should be included, up to infinite degree. In practice those four families of functions provide already a good coverage with few degrees included in

the method.
The likelihood is just a measure of agreement between the data and a function, and naturally tends to chose the function which has a better description of the data. As more flexible function better follows the shape of the data, the likelihood estimate tends to privilege functions with high number of free parameters (N_{par}), at least within the same functional family. To prevent this from happening and to obtain a fair comparison between functions with different number of parameters, the NLL is corrected with a penalty depending on N_{par}:

$$NLL_{corr} = NLL + \kappa N_{par}. \tag{4.16}$$

There is no general rule to determine the value of κ. The χ^2 approximation of the likelihood suggests $\kappa = 1$ as is shown in Ref. [118]. Other interpretations are possible, and in general any value of κ could be used. Lower value of κ means low penalty to higher order function, resulting in a bias estimate of μ as the likelihood is mainly driven by N_{par}. Instead, large values of κ penalise higher order functions, forcing the choice of low order function and thus narrowing the NLL. For the $H \rightarrow \gamma\gamma$ analysis the choice of $\kappa = 1$ has been adopted and validated through the usage of pseudo-experiments.
In principle, when fitting the background, for each of the four functional families, function up to an arbitrary large order, can be exploited. Functions with poor fit quality are automatically ignored by the method, while the penalty to the likelihood ensures the same for extremely flexible functions. In practice, to retain the computing time within acceptable levels, a preselections of the functions to be exploited is performed. For each category included in the analysis, the lowest order function of each family is tested. A p-value according to the χ^2 test statistic is computed and functions incompatible with the data are rejected. When a function of order N is accepted, the function of order N+1 is tested. The quantity

$$2\Delta NLL_{N+1} = 2\,(NLL_{N+1} - NLL_N) \tag{4.17}$$

is exploited to establish whether to accept or not the higher order function. The value of $2\Delta NLL_{N+1}$ is approximately distributed as a χ^2 distribution with M degrees of freedom, where M is the difference between the degrees of freedom of the two functions. For exponential and power law functions M=2, while for the other two families M=1. A p-value is computed as the probability to find a value of $2\Delta NLL$ greater than the one observed. If the p-value is below 0.05 the function is judged too flexible and it is discarded. Once for each category the set of functions is defined, the background is modelled from the signal-plus-background fit to the data. An example of fit in one of the categories included in the 2017 analysis is shown in Fig. 4.30. The left panel shows the different functions considered for the fit, while the right panel shows the result of the fit superimposed with the data sidebands.

4.7 Systematic uncertainties

As the goal of the measurement is to compare the experimental result with the SM prediction, an excellent theoretical and experimental control of the process under investigation is needed. The SM prediction for the $t\bar{t}H$ signal is derived from simulation, therefore theory mis-modelling of the process or imperfect detector simulation can lead to a biased

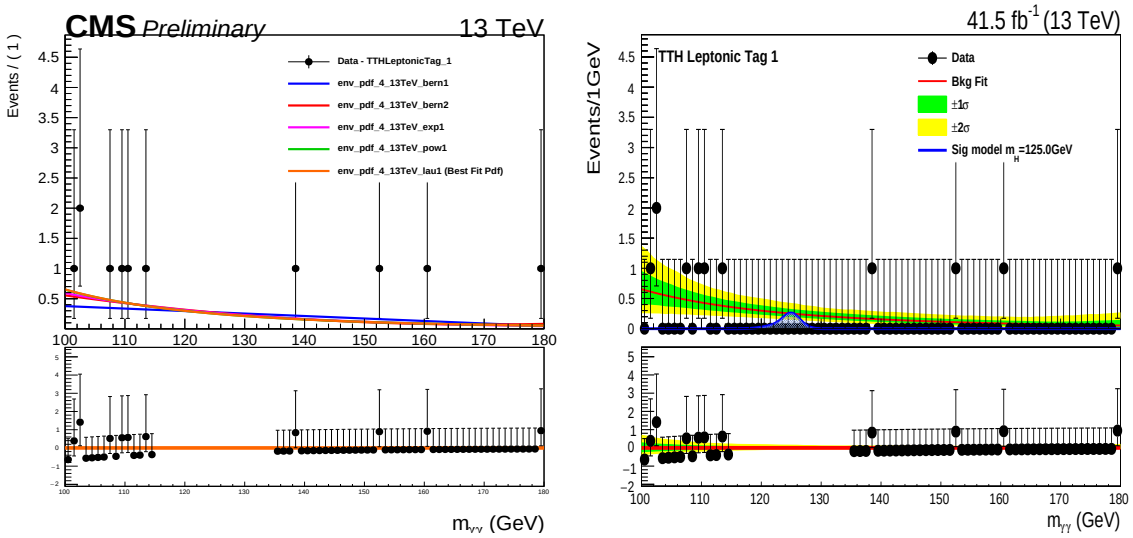


Figure 4.30: Left: fit of the different background models to the $t\bar{t}H$ Leptonic 1 category of the 2017 analysis. The data sidebands (black markers) are shown together with the functions considered for the background fit, chosen according to the procedure described in the text. Right: result of the best-fit in the same category of the left panel. The data sidebands (black markers) are shown with the result of the fit. The one and two standard deviations uncertainties (green and yellow bands) include the uncertainty on the choice of the fit function and on the fitted parameters. They do not include the Poisson uncertainty due to the number of events in the category. The signal model for $m_{\mathrm{H}} = 125$ GeV is also shown. The bottom panel shows the residuals after background subtraction.

estimate of the expected number of signal events after analysis selections are applied. Any potential source of bias has to be investigated and, whenever possible, corrected. All the uncorrected biases should be estimated and their effect has to be covered with systematic uncertainties affecting the SM prediction for the event yield.
In the general case, uncertainties affect both the signal and the background predictions. Since the background contribution is estimated directly from data, the only uncertainty to be considered is the one associated with the choice of the fit function. As this uncertainty is already included in the envelope method described in Section 4.6.2, systematic uncertainties affects only to the modelling of the signal.
A source of uncertainty can affect the measurement in two possible ways. Either it can induce a modification in the shape of the variable adopted to perform the measurement ($m_{\gamma\gamma}$), or it can affect the prediction on the event yield. Uncertainties affecting the shape of $m_{\gamma\gamma}$, as the energy corrections described in Section 4.3.1, are included as nuisance parameters in the signal model and are constrained directly by fitting the data. Instead, uncertainties that affect just the event yield, as the uncertainty on the integrated luminosity, are incorporated in the likelihood as log-normal uncertainties. Uncertainties that cause a migration of the events between categories are handled so that the event yield in each category can vary but the overall event yield (including events outside the acceptance of the analysis) is unchanged.
The description of the different sources of systematic uncertainties is reported in the following sections, separating the theory uncertainties from the experiments ones. Section 4.7.3

describes the impact of each uncertainty on the expected signal strength of the $t\bar{t}H$ process ($\mu_{t\bar{t}H}$).

4.7.1 Theoretical uncertainties

Theory uncertainties are related to QCD predictions (see Section 1.1.1) for the Higgs boson cross section. They generally cause a variation in the total signal yield as well as a migration between categories. The two effects are factorised in the likelihood. The uncertainties considered in the analyses are:

- **scale uncertainty of QCD**: it is related to the variation of the QCD renormalisation and factorisation scales μ_R and μ_F. They are handled as two nuisance parameters, binned according to the number of jets in the event, affecting the overall event yield. It is the dominant systematic uncertainty, with an impact on $\mu_{t\bar{t}H}$ of 9%;
- **uncertainty on PDF modelling**: it is computed following the prescription from PDF4LHC [40, 119]. It causes both a variation on the global event yield and a migration of the events between the categories, as it affects the prediction on the number of jets of the event. The category migration due to the presence of additional jets is computed with the NNPDF 3.0 set with the MC2HESSIAN method [120]. The overall variation of $\mu_{t\bar{t}H}$ is of 5%;
- **uncertainties on the strong coupling constant**: the uncertainty on α_s is reflected in an uncertainty on the expected number of events. It is computed following the PDF4LHC prescription [40, 119] and its impact on $\mu_{t\bar{t}H}$ is 3%;
- **uncertainty on the ggH contamination**: theory predictions in the regime where a Higgs boson is produced in association with many jets are not reliable. The ggH event yield entering the $t\bar{t}H$ categories is affected by a considerable uncertainty and, consequently, the $t\bar{t}H$ prediction is affected by a 2% uncertainty. Three different sources are considered:
  - **uncertainty of the parton shower modelling**: it is estimated from the difference between data and MADGRAPH5_AMC@NLO prediction in $t\bar{t}$ events [121], mainly produced by gluon fusion $gg \to t\bar{t}$. The largest observed discrepancy is up to 35% in events with more than 5 jets;
  - **uncertainty on the gluon splitting**: it is related to the uncertainty on the probability of the process $g \to b\bar{b}$. It is estimated by scaling the number of events in the ggH simulation with real b jets by the observed difference between data and simulation in the gluon splitting probability to b quarks. The simulation correction factor is derived from the ratio between data and simulation of $\sigma(t\bar{t}b\bar{b})/\sigma(t\bar{t}jj)$ [122]. This uncertainty implies a variation in the number of ggH events produced with b jets of about 50%;
  - **uncertainty due to the limited size of the simulated ggH sample**: it is accounted as 10% variation in the expected ggH event yield;
- **uncertainty on the $H \to \gamma\gamma$ branching fraction**: it is estimated from Ref. [40] and its impact is of 2%.

- **uncertainty on the modelling of the underlying event**: it is obtained from modifying the generator parameters related to the modelling of the underlying event. It affects mainly jet production and its impact on the $t\bar{t}H$ event yield is less than 1%.

4.7.2 Experimental uncertainties

Experimental uncertainties account for imperfections in the simulation of the detector response that cause a disagreement between data and simulation on some observables relevant to the analysis. The level of disagreement is generally measured on control regions, exploiting processes different from the Higgs boson production, as the $Z \to e^+e^-$ process. Correction factors are derived and the signal simulation is corrected to match the data. The uncertainty on the correction factor is then propagated through the analysis and the expected variation of the event yield is used as a constraint on the systematic uncertainty. The sources of experimental uncertainties with larger impact on the $t\bar{t}H$ measurement are:

- **uncertainty on the photon identification BDT**: as described in Section 4.3.1, the data-to-simulation agreement of the output score of the photon identification BDT is checked on $Z \to e^+e^-$ events. An uncertainty is conservatively assigned to cover the largest observed discrepancy in the validation region. It impacts the prediction of $\mu_{t\bar{t}H}$ by 3% in the 2016 data and 6% in the 2017 ones. The difference between the two values is due to the different level of mis-modelling of the ECAL conditions;

- **uncertainty on the jet energy scale and resolution**: it is related to the mis-modelling of the detector response to jets. Scale factors are derived to match the energy scale and energy resolution of simulated jets to the one observed in data, using $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$ events, as well as in γ + jet, dijet and multijet events [110]. The impact on the signal yield, of 2% and 4% in 2016 and 2017 data respectively, is evaluated from propagating the uncertainty to the expected event yield;

- **uncertainty on the shape of the b-discriminant**: the observed discrepancy between data and simulation is corrected through the application of scale factors (see Section 4.3.6) derived from control regions. The associated uncertainty on $t\bar{t}H$ prediction is derived from propagating the scale factor uncertainty through the analysis. Its value is of 2% on the 2016 data and of 3% on the 2017 ones;

- **Uncertainty on the integrated luminosity**: it is estimated from auxiliary measurements (Ref. [123] for 2016 and Ref. [124] for 2017) and its impact is 2.5% on the 2016 analysis and 2.3% on the 2017 one;

- **uncertainty on the photon energy scale and resolution**: it accounts for the discrepancy of the photon energy scale and resolution after applying the corrections described in Section 4.3.1. Several factors are considered, including differences between showers generated by electrons used to derive the corrections and photons, non-linearities of the light collection, different trainings in the energy regression and variation of the binning used to derive the correction. Its value is of 3% on the 2016 data and less than 1% on the 2017 ones, thanks to the refinement of the photon energy corrections.

Several other sources of experimental uncertainty are included in the analyses, whose impact on the result is less than 1% and thus it is small compared to the dominant uncertainties. Additional sources of experimental uncertainty are:

- **uncertainty on trigger efficiency**: it is evaluated with the T&P method from $Z \to e^+e^-$ events;
- **uncertainty on the lepton identification**: it is evaluated from propagating the uncertainty on the relevant simulation scale factors to the expected result;
- **uncertainty on photon preselection**: it is evaluated from the uncertainty on the scale factors derived with $Z \to e^+e^-$ and $Z \to \mu^+\mu^-\gamma$ events;
- **uncertainty on PSV efficiency**: it is evaluated from the uncertainty on the scale factors derived with $Z \to \mu^+\mu^-\gamma$ events;
- **uncertainty on $p_{\mathrm{T}}^{\mathbf{miss}}$** : it is evaluated from scaling the p_{T} of the particles entering in the computation of $p_{\mathrm{T}}^{\mathrm{miss}}$ within the respective uncertainties [125];
- **uncertainty on pileup jet identification**: it is estimated from propagating the uncertainty in the simulation scale factors derived from Z+jets events to the expected event yield;
- **uncertainty on vertex assignment**: it is handled as an additional nuisance parameter in the signal model which allows events in the correct vertex and wrong vertex scenario to vary;
- **uncertainty on Higgs boson mass**: in the 2017 analysis the mass of the Higgs boson is constrained to the Run I measurement $m_{\mathrm{H}} = 125.09 \pm 0.24$ GeV, while it is free to float in the 2016 one. In both the cases the uncertainty on the mass causes an uncertainty on the predicted event yield as the cross section and branching ratio are function of m_{H}.

4.7.3 Impact of the systematic uncertainties

The impact on $\mu_{\mathrm{t\bar{t}H}}$ for every source of uncertainty described in the previous sections is illustrated in Table 4.11, separated for the 2016 and 2017 data. The theoretical uncertainties are the one with the larger impact on the expected signal strength. The QCD scale uncertainty is the dominant one, causing a variation of $\mu_{\mathrm{t\bar{t}H}}$ of 9%.
The experimental uncertainties present some differences between the 2017 and the 2016 analyses. The 2017 data presents in general a worse data-to-simulation agreement of photon related variables, as the evolution of the ECAL pedestals is not correctly modelled in the simulation. The uncertainty on the photon identification BDT is correspondingly enhanced moving from 2016 to 2017 data, as the discrepancy on the input variables is larger. The impact of the scale and resolution corrections is largely reduced in the 2017 analysis, thanks to the refined granularity of the corrections. The impact of the jet and b jet uncertainties in the 2017 analysis is enhanced, as more variables related to jets are exploited to select the events. As the PSV and the $p_{\mathrm{T}}^{\mathrm{miss}}$uncertainty are not exploited in 2016, no uncertainty associated to those quantities is considered.
At the present integrated luminosity the impact of the systematic uncertainties is extremely

reduced compared to the large statistical uncertainty, of about 50%, expected on the measurement.

	Uncertainty	$\Delta\ \mu_{t\bar{t}H}$(2016 data)	$\Delta\ \mu_{t\bar{t}H}$(2017 data)
Theoretical	QCD scale	9%	9%
	PDF modelling	5%	5%
	Strong coupling constant	3%	3%
	Contamination from ggH	2%	2%
	Branching fraction of $H \rightarrow \gamma\gamma$	2%	2%
	Underlying event modelling	<1%	<1%
Experimental	Photon identification BDT	3%	6%
	Jet energy scale and resolution	2%	4%
	Shape on the b-discriminant	2%	3%
	Integrated luminosity	2.5%	2.3%
	Photon energy scale and resolution	3%	< 1%
	Trigger efficiency	< 1%	< 1%
	Lepton identification	< 1%	< 1%
	Photon preselection	< 1%	< 1%
	Efficiency of the PSV	-	< 1%
	Measurement of $p_{\mathrm{T}}^{\mathrm{miss}}$	-	< 1%
	Identification of pileup jets	< 1%	< 1%
	Assignment of the vertex	< 1%	< 1%
	Mass of the Higgs boson	< 1%	< 1%

Table 4.11: Impact on the measurement of the different sources of systematic uncertainties described in Section 4.7.1 and 4.7.2. The impact is quantified as the expected variation on the signal strength modifier $\mu_{t\bar{t}H}$, separately on 2016 and 2017 data.

4.8 Results

The work described in this thesis led to two results. The 2016 analysis, in combination with other Higgs boson decay channels, allows the first observation of the $t\bar{t}H$ process, establishing the coupling of the Higgs boson with the top quark and, for the first time, with an up-type quark. The 2017 analysis produced a sizeable improvement in the expected significance, reducing in turn the uncertainty on y_t, moving towards a precise determination of the Higgs boson couplings.
The results are extracted performing a binned maximum likelihood fit (see Section 4.5) to the invariant mass spectrum of the photon pairs. The fit is performed simultaneously to all the categories included in each analysis, with a common signal strength modifier free to float in the fit. The binning for the fit is chosen to be 250 MeV, much smaller than the invariant mass resolution, to avoid any loss of information. It has been verified that the results are identical to the one obtained from an unbinned fit, with the additional advantage of a processing time reduced by about a factor ten.
The results are derived both as signal strength modifiers and in the Stage 0 STXS framework (see Section 1.3.4). The main differences between the two approaches comes

from the selection applied on the acceptance of the Higgs boson. The reduced acceptance prevents large theory uncertainties arising from the extrapolation of the cross section to the full phase space.
The results of the 2016 and 2017 analyses are presented in Section 4.8.1 and 4.8.2, respectively. The combination of the 2016 analysis with the other channels is presented in Chapter 5, alongside with the future prospects for the $t\bar{t}H$ measurement.

4.8.1 Results of 2016 data analysis

In the 2016 analysis, several categories targeting the four main Higgs boson production processes are included, as explained in Section 4.4.1. An overall signal strength modifier μ scaling all the processes together is defined, as well as a signal strength modifier for each of the production processes. Along with $\mu_{t\bar{t}H}$, signal strength modifiers for the ggH (μ_{ggH}), VBF (μ_{VBF}) and VH processes (μ_{VH}) are left free to float in the fit, while the bbH, tHq and tHW processes are constrained to the SM expectations (within the respective theoretical uncertainties) [40]. The Higgs boson branching fraction to photons is also constrained to the SM expectation, within its uncertainty [40], while the Higgs boson mass is free to float (profiled) in the fit.
The signal strength modifiers are extracted from the simultaneous fit to all the categories of the analysis. The $\mu_{t\bar{t}H}$ is mainly constrained by the exclusive categories targeting $t\bar{t}H$ production. The fit to all the categories allows us to simultaneously measure the contribution of each of the Higgs boson production processes and, thus, to determine from data the contamination of ggH, VBF and VH in the $t\bar{t}H$ categories.
The signal-plus-background fit to the two $t\bar{t}H$ categories is shown in the upper panels of Fig. 4.31, while the fit to all the other categories can be found in Ref. [96]. The combination of all the categories included in the analysis is shown in the bottom panel of Fig. 4.31. Here, each category is weighted for the expected sensitivity, estimated as S/(S+B), where S and B are the number of signal and background events is a window of $\pm 1\sigma_{eff}$ centred around m_H. Figure 4.32 shows, for each category included in the analysis, the relative composition in terms of different Higgs boson production processes. The two $t\bar{t}H$ categories are enriched in $t\bar{t}H$ events, with a purity higher than 80%, while the $t\bar{t}H$ contamination in the other categories is rather small. Therefore, the capability of the latter categories to constrain $\mu_{t\bar{t}H}$ is rather limited. The width of the signal peak, in term of σ_{eff} and σ_{HM}, defined as the FWHM divided by 2.35, is also reported, as well as an estimate of the signal to background ratio. The same information are also reported in Table 4.12.

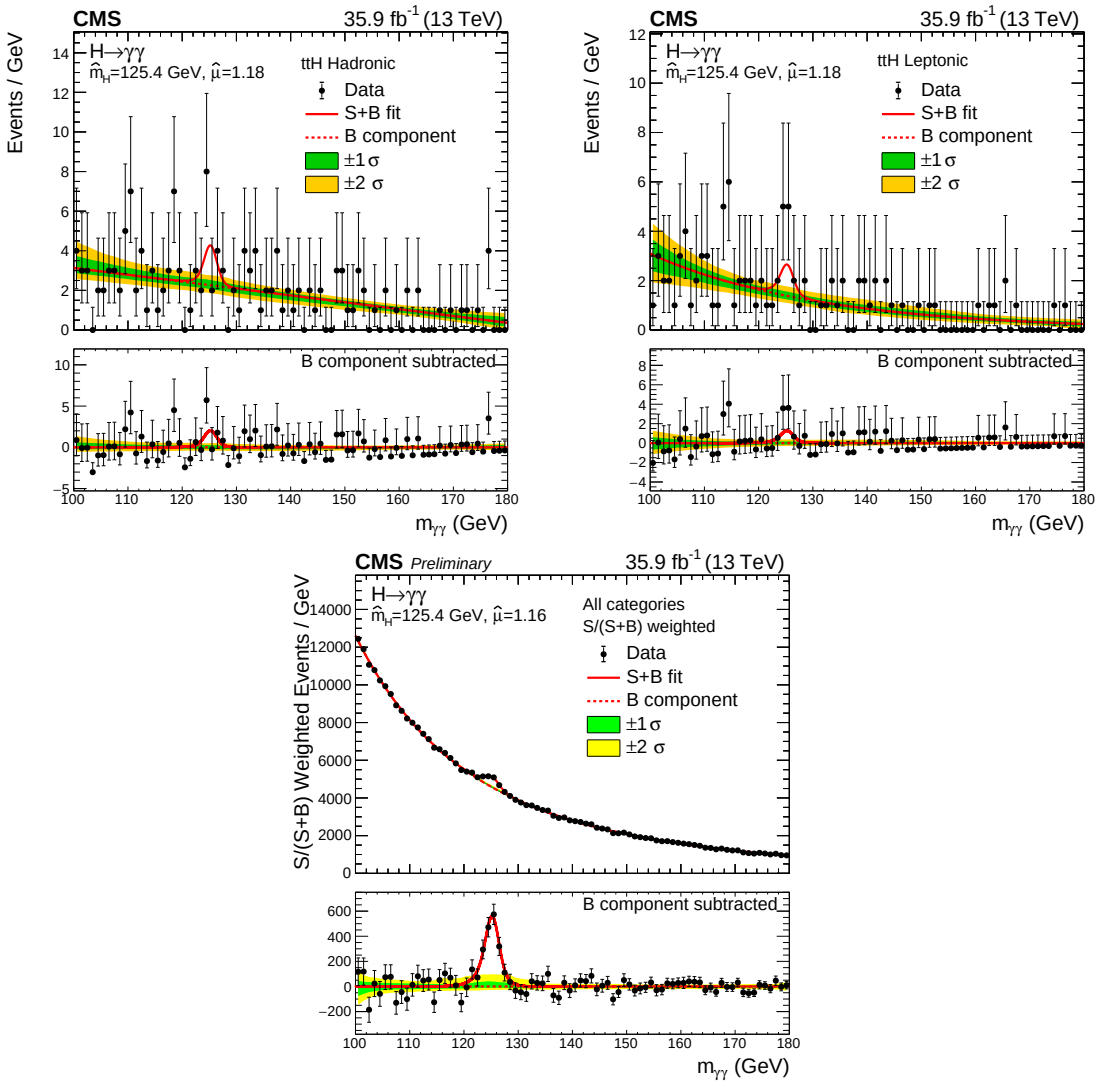


Figure 4.31: Diphoton invariant mass spectrum of the $t\bar{t}H$ categories of the 2016 analysis (top) and of the combination of all the categories (bottom) included in the analysis. In the combination, each category is weighted for the expected expected sensitivity, as explained in the text. The data (black markers) are shown together with the signal-plus-background fit (red line). The one and two standard deviations uncertainties (green and yellow bands) include the uncertainty on the choice of the fit function and on the fitted parameters. They do not include the Poisson uncertainty due to the number of events in the category. The bottom panel of each figure shows the residuals after background subtraction.

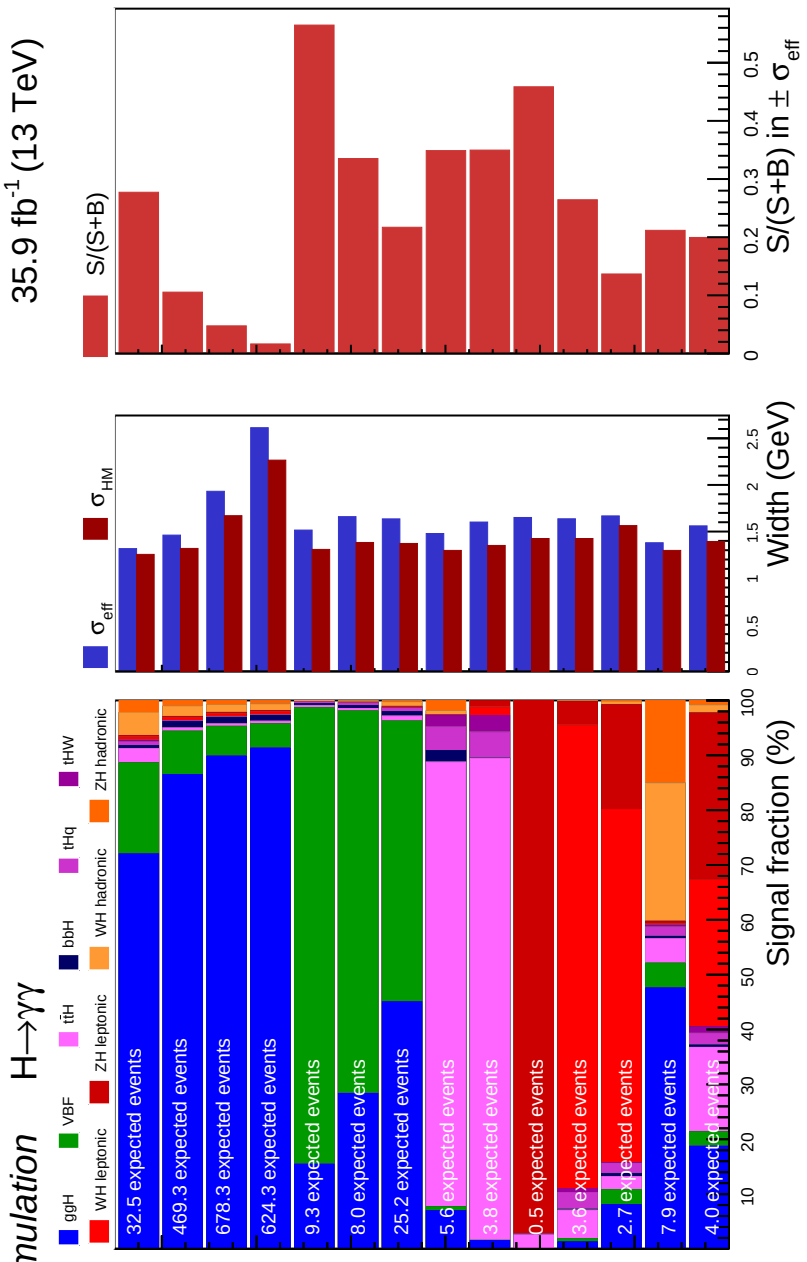


Figure 4.32: Expected fraction of events per production process in each of the categories included in the analysis. The $t\bar{t}H$ contribution is shown in pink. The σ_{eff} and the FWHM, divided by 2.35, are reported in the second panel, while the third panel gives an estimate of the sensitivity of each category by the ratio between the expected number of signal events (S) and the signal plus the background events (S+B) in a window of $\pm 1\sigma_{\text{eff}}$ around m_H.

Categories	Events	$t\bar{t}H$	ggH	VBF	VH	bbH	tHq	tHW	σ_{eff}	σ_{HM}	Bkg/GeV
$t\bar{t}H$ Hadronic	5.6	81.1 %	7.0 %	0.7 %	2.8%	2.1 %	4.3 %	2.1 %.	1.48	1.30	2.4
$t\bar{t}H$ Leptonic	3.8	87.8 %	1.5 %	<0.05 %	2.7%	0.1 %	4.7 %	3.1 %	1.60	1.35	1.5
Untagged 0	32.5	2.6 %	72.0 %	16.6 %	7.3%	0.6 %	0.7 %	0.3 %	1.32	1.26	21.8
Untagged 1	469.3	0.6 %	86.5 %.	7.9%	3.8%	1.2 %	0.1 %	<0.05 %	1.46	1.32	925.1
Untagged 2	678.3	0.4 %	89.9 %	5.4 %	3.0%	1.2 %	0.1 %	<0.05 %	1.93	1.67	2391.7
Untagged 3	624.3	0.5 %	91.3 %	4.4 %	2.7%	1.0 %	0.1 %	<0.05 %	2.61	2.27	4855.1
VBF 0	9.3	0.4 %	15.5 %	83.2 %	0.2%	0.4 %	0.3 %	<0.05 %	1.52	1.31	1.6
VBF 1	8.0	0.4 %	28.4 %	69.7 %	0.5%	0.6 %	0.4 %	<0.05 %	1.66	1.38	3.3
VBF 2	25.2	0.9 %	45.1 %	51.2 %	1.4%	0.8 %	0.6 %	0.1 %	1.64	1.37	18.9
ZH Leptonic	0.5	2.6 %	<0.05%	<0.05 %	97.3%	<0.05 %	<0.05 %	0.1 %	1.65	1.43	0.1
WH Leptonic	3.6	5.2 %	1.3 %.	0.6%	89.0%	0.2 %	3.0 %	0.7 %	1.64	1.43	2.1
WH Leptonic Loose	2.7	2.4 %	8.1 %	2.7 %	84.3%	0.6 %	1.8 %	0.1 %	1.67	1.56	3.5
VH Hadronic	7.9	4.4 %	47.6 %	4.5 %	41.0%	0.4 %	1.7 %	0.3 %.	1.38	1.30	7.2
VH MET	4.0	15.4 %	18.7 %	2.6 %	59.5%	0.4 %	2.1 %	1.2 %	1.56	1.39	3.5
Total	1875.0	1.0 %	86.9 %	7.1 %	3.7%	1.1 %	0.2 %	<0.05 %	1.96	1.62	8237.8

Table 4.12: Expected number of signal events per category split by production mode. An estimate of the resolution on $m_{\gamma\gamma}$ of each category is reported as σ_{eff} and σ_{HM}, defined as the FWHM divided by 2.35. The expected number of background events per GeV around 125 GeV is also listed.

The likelihood function, when scaling all the processes together as a function of μ, is shown in the left panel of Fig. 4.33. The best-fit value for μ is $\hat{\mu} = 1.18^{+0.17}_{-0.14} = 1.18^{+0.12}_{-0.11}(\text{stat.})^{+0.09}_{-0.07}(\text{syst.})^{+0.07}_{-0.06}(\text{theo.})$. The contribution of the statistical and systematic uncertainty is computed performing a likelihood scan with all the systematic uncertainties fixed to zero to derive the statistical contribution. The systematic component is then derived from subtracting in quadrature the statistical uncertainty from the total uncertainty. The results of the fit with a signal strength modifier for each Higgs boson production process is summarised in Fig. 4.34. The likelihood scan for $\mu_{t\bar{t}H}$ is depicted in the right panel of Fig. 4.33. The best-fit value is $\hat{\mu}_{t\bar{t}H} = 2.2^{+0.9}_{-0.8} = 2.2^{+0.9}_{-0.8}(\text{stat.})^{+0.2}_{-0.1}(\text{syst.})^{+0.2}_{-0.1}(\text{theo.})$. The corresponding significance is 3.2 standard deviations, while 1.5 is expected assuming the SM. The measurement is strongly limited by the statistical uncertainty.

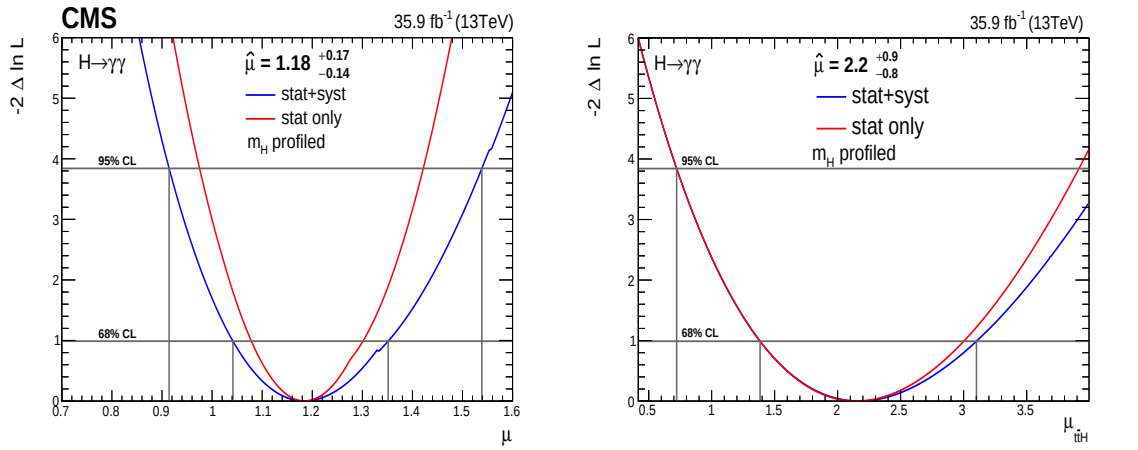


Figure 4.33: Value of the likelihood function as a function of μ (left) and for $\mu_{t\bar{t}H}$ (right) on the 2016 data. In the first case all the Higgs boson production processes are scaled together, while in the second one each process is scale with a dedicated signal strength modifier. The Higgs boson mass is profiled in the fit. The red line represents the likelihood function when only the statistical uncertainty is accounted for, while the blue one includes all the relevant systematic uncertainties.

Figure 4.35 shows the ratio of the observed to the expected cross sections extracted within the Stage 0 STXS framework. The VH process is further split according to the decay of the vector boson. The result for the $t\bar{t}H$ bin is $\sigma_{\text{obs}}/\sigma_{\text{exp}} = 2.0^{+0.8}_{-0.7}$.
All the results are found in agreement with the SM expectation, despite a small tension in the $t\bar{t}H$ and VH rate is found. The tension of about two standard deviations in the $t\bar{t}H$ rate is consistent with what measured on Run I data.

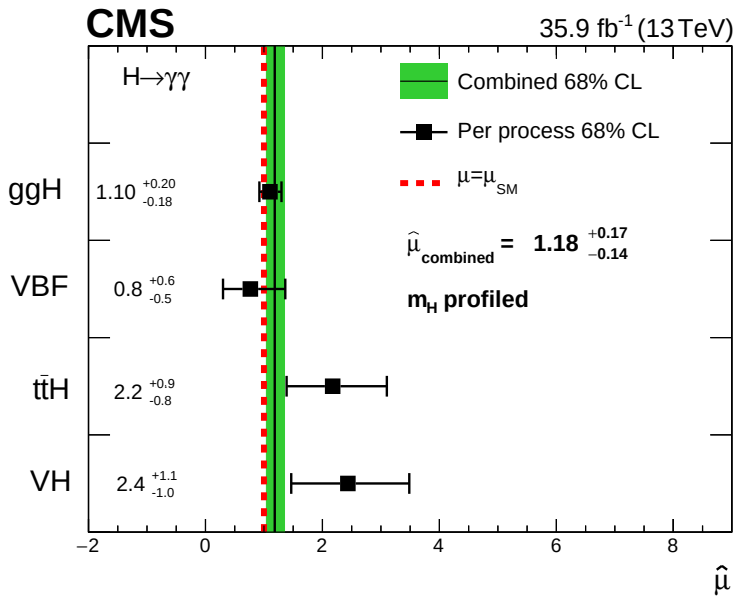


Figure 4.34: Measured value of the signal strength modifier for each of the Higgs boson production mode (black markers) with the Higgs boson mass profiled in the fit. The overall signal strength μ with its 68% uncertainty is depicted by the green band, while the SM expectation is the red dotted line.

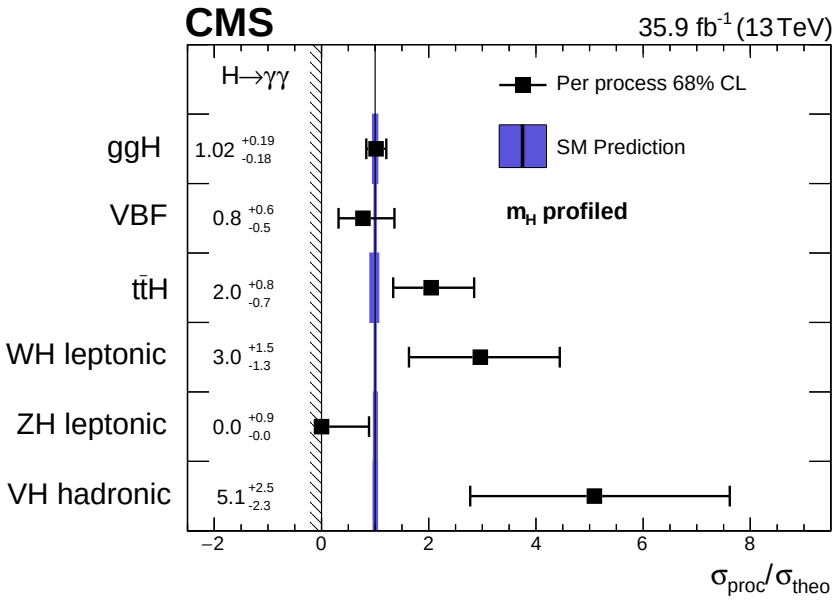


Figure 4.35: Measured cross sections, normalised to the theory expectations σ_{theo}, in the different STXS bins with the Higgs boson mass profiled in the fit. The SM expectation and its uncertainty is shown by the blue band. The measured values are constrained to be non-negative, as indicated by the vertical pattern at zero.

4.8.2 Results of 2017 data analysis

As the main target of this analysis is the measurement of $\mu_{t\bar{t}H}$, no categories are defined targeting the other Higgs boson production processes, forcing a slightly different approach in extracting $\mu_{t\bar{t}H}$. The contribution of the ggH, VBF and VH process can not be fitted directly from data, as the constraint from the $t\bar{t}H$ categories would be extremely poor. The contribution of all the processes other than $t\bar{t}H$ is constrained to the SM expectation and it is free to float within its uncertainty [40]. The Higgs boson mass is also constrained to the ATLAS and CMS Run I combined measurement $m_H = 125.09 \pm 0.24$ GeV, as the constrain from the $t\bar{t}H$ categories only would be poor.
A simultaneous fit to the five categories included in the analysis is performed. The composition of each category is shown in Fig. 4.36 and in Table 4.13, while the signal-plus-background fit to the five categories is shown in Fig. 4.37 and 4.38. As for the 2016 data analysis, the categories are rather pure in $t\bar{t}H$. The degraded invariant mass resolution compared to 2016 in mainly due to the preliminary ECAL calibration.

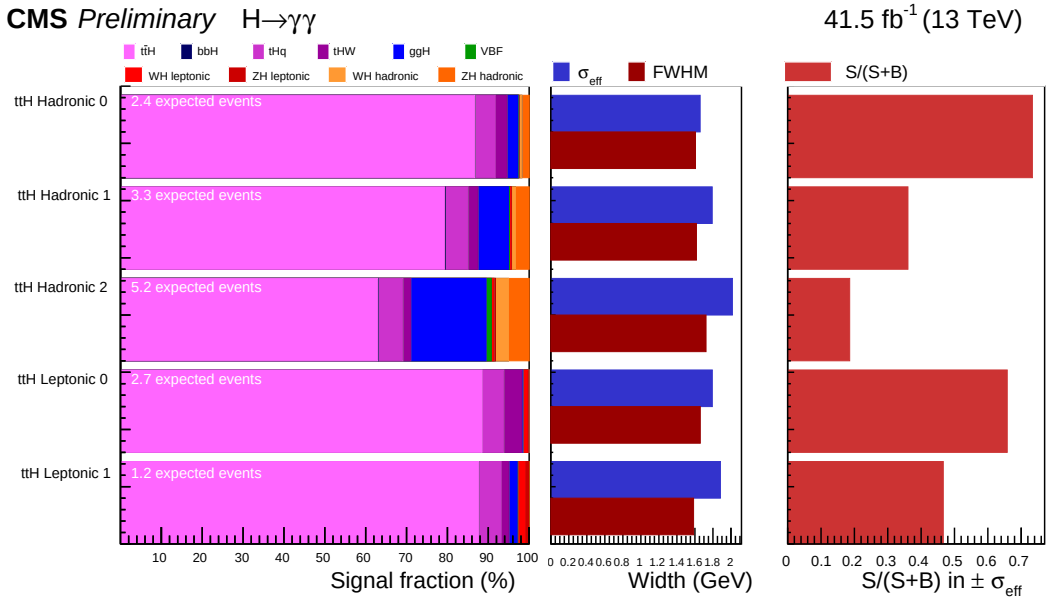


Figure 4.36: Expected fraction of events per production process in each of the categories included in the analysis. The $t\bar{t}H$ contribution is shown in pink. The σ_{eff} and the FWHM, divided by 2.35, are reported in the second panel, while the third panel gives an estimate of the sensitivity of each category by the ratio between the expected number of signal events (S) and the signal plus the background events (S+B) in a window of $\pm 1\sigma_{\text{eff}}$.

Categories	Events	$t\bar{t}H$	ggH	VBF	VH	bbH	tHq	tHW	σ_{eff}	σ_{HM}	Bkg/GeV
$t\bar{t}H$ Hadronic 0	2.4	86.7 %	2.6 %	0.1 %	5.7%	<0.05 %	5.0 %	2.8 %	1.66	1.61	0.2
$t\bar{t}H$ Hadronic 1	3.3	79.2 %	7.5 %	0.2 %	10.5%	0.2 %	5.6 %	2.4 %	1.79	1.62	1.1
$t\bar{t}H$ Hadronic 2	5.2	62.9 %	18.4 %	1.3 %	15.2%	0.2 %.	5.9 %	1.9 %	2.02	1.72	3.8
$t\bar{t}H$ Leptonic 0	2.7	88.5 %	0.2 %	<0.05 %	1.8%	<0.05 %	5.2 %	4.4 %	1.77	1.59	0.3
$t\bar{t}H$ Leptonic 1	1.2	87.6 %	2.0 %	<0.2 %	6.1%	<0.05 %	5.5 %	1.8 %	1.88	1.59	0.3
Total	14.8	77.2 %	8.7 %	0.5 %	9.4%	1.5 %	5.5 %	2.6 %	1.84	1.65	5.6

Table 4.13: Expected number of signal events per category in the 2017 analysis split by production mode. An estimate of the resolution on $m_{\gamma\gamma}$ of each category is reported as σ_{eff} and FWHM, divided by 2.35. The expected number of background events per GeV around 125 GeV is also listed.

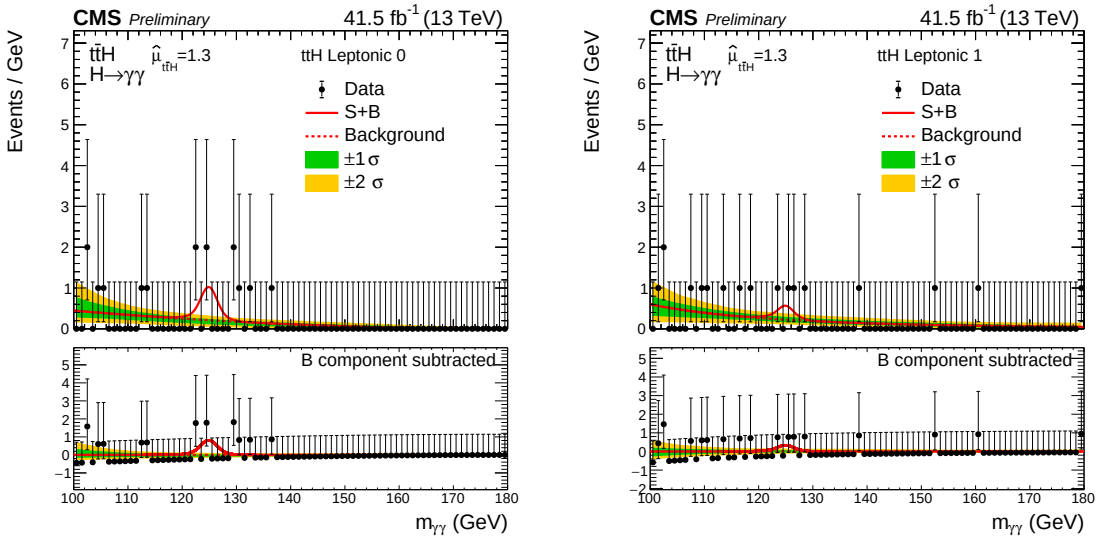


Figure 4.37: Diphoton invariant mass spectrum of the leptonic categories of the 2017 analysis. The data (black markers) are shown together with the signal-plus-background model (red line). The one and two standard deviations uncertainties (green and yellow bands) include the uncertainty on the choice of the fit function and on the fitted parameters. They do not include the Poisson uncertainty due to the number of events in the category. The bottom panel shows the residuals after background subtraction.

The likelihood scan of $\mu_{t\bar{t}H}$ is depicted in Fig. 4.39. As for the 2016 result, the contribution of the statistical and systematic uncertainty is computed performing a likelihood scan removing all the systematic uncertainties to derive the statistical contribution. The systematic component is then derived from subtracting in quadrature the statistical uncertainty from the total uncertainty. The best-fit value is $\hat{\mu}_{t\bar{t}H} = 1.3^{+0.7}_{-0.5} = 1.3^{+0.6}_{-0.5}(\text{stat.})^{+0.3}_{-0.1}(\text{syst.})$, in agreement with the SM expectation. The result is still strongly dominated by the statistical uncertainty. A signal strength modifier per category included in the analysis is shown in Fig. 4.40.

The measurement corresponds to a rejection of the background-only hypothesis of 3.1 standard deviations, while 2.2 is expected assuming the SM. The cross section normalised to the theory expectation, measured within the STXS framework, is measured in $1.3^{+0.6}_{-0.5}$, in agreement with the SM expectation. The STXS measurement is performed scaling the tHq and tHW processes with the $t\bar{t}H$ one, in order to derive a signal strength associated with tha top-quark,

The result is combined with the 2016 one. Only the two categories targeting the $t\bar{t}H$ production are used for the combination. All the uncertainties are assumed to be correlated between the two years. The assumption has been verified to have negligible impact on the result. The likelihood scan for $\mu_{t\bar{t}H}$ when combining the two analyses is shown in Fig. 4.41. The best-fit value is $\hat{\mu}_{t\bar{t}H} = 1.7^{+0.6}_{-0.5}$, corresponding to a significance of 4.1 standard deviations, while 2.7 were expected. Figure 4.42 shows the sum of all the categories included in the analyses (2 from 2016 and 5 from 2017 analysis). Each category is weighted for the expected sensitivity, estimated as S/(S+B), where S and B are the number of signal and background events is a window of $\pm 1\sigma_{\text{eff}}$ centred around m_H.

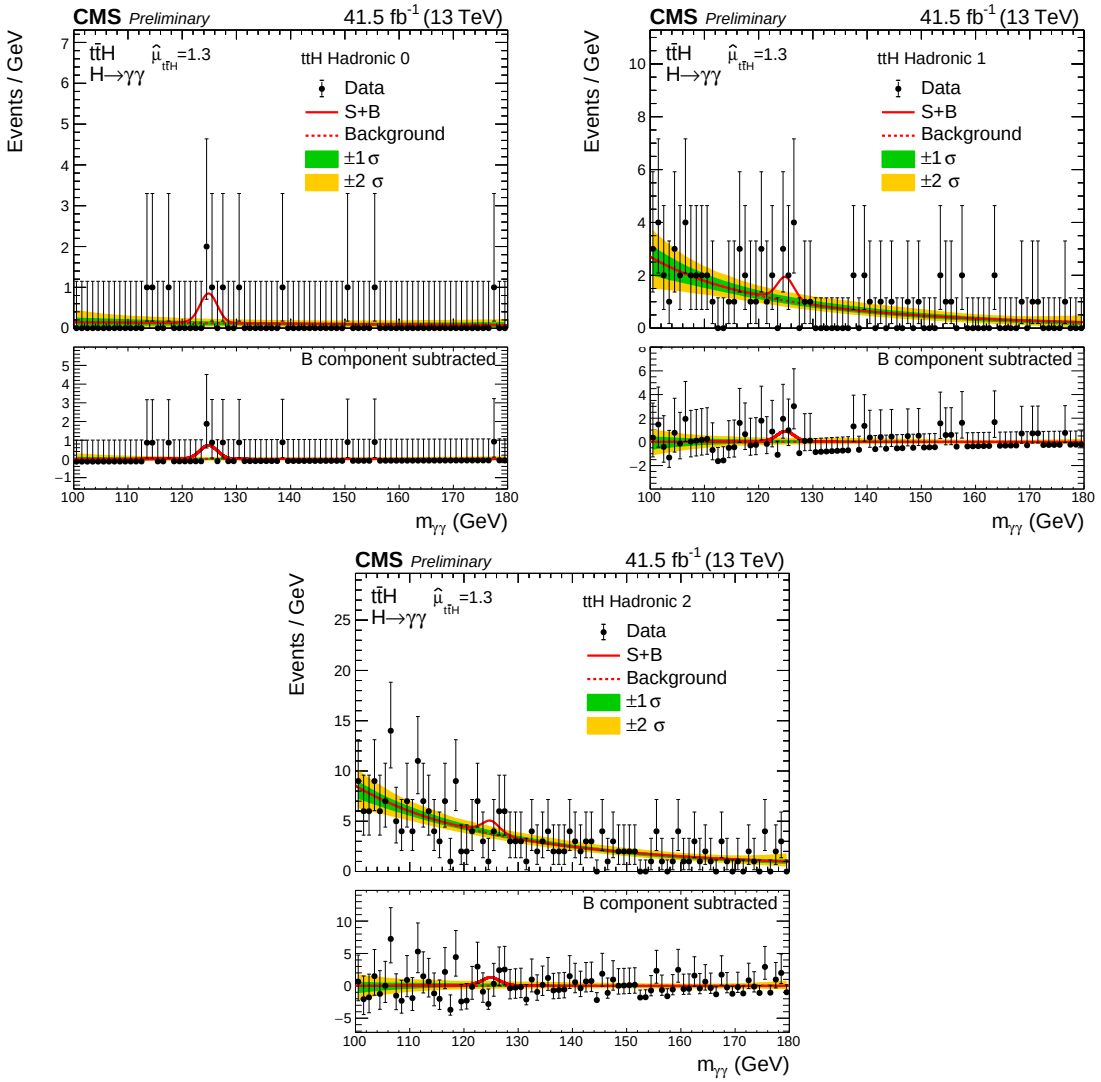


Figure 4.38: Diphoton invariant mass spectrum of the hadronic categories of the 2017 analysis. The data (black markers) are shown together with the signal-plus-background model (red line). The one and two standard deviations uncertainties (green and yellow bands) includes the uncertainty on the choice of the fit function and on the fitted parameters. They do not include the Poisson uncertainty due to the number of events in the category. The bottom panel shows the residuals after background subtraction.

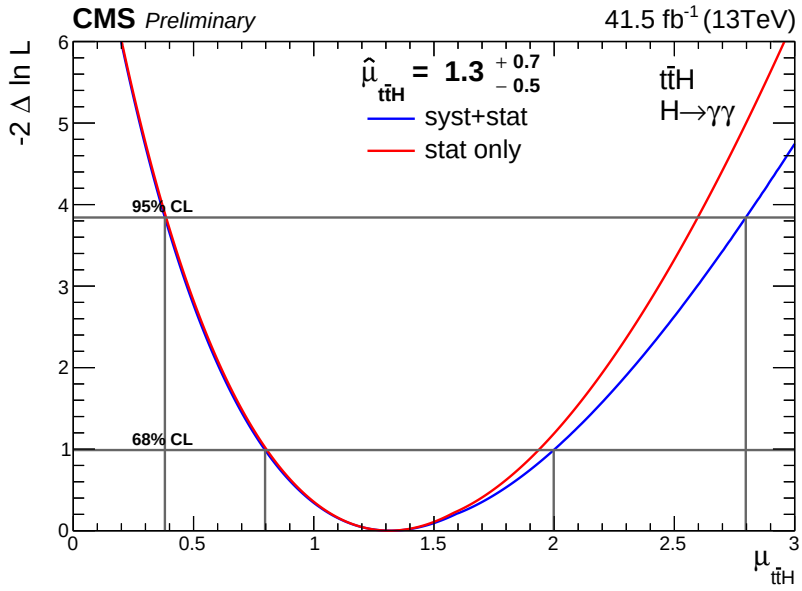


Figure 4.39: Value of the likelihood function in proximity of its minimum for $\mu_{t\bar{t}H}$ on the 2017 data. The Higgs boson mass is constrained to the Run I best-fit value $m_H = 125.09 \pm 0.24$ GeV. The red line represents the likelihood function when only the statistical uncertainty is accounted for, while the red one includes all the systematic uncertainties related to the measurement.

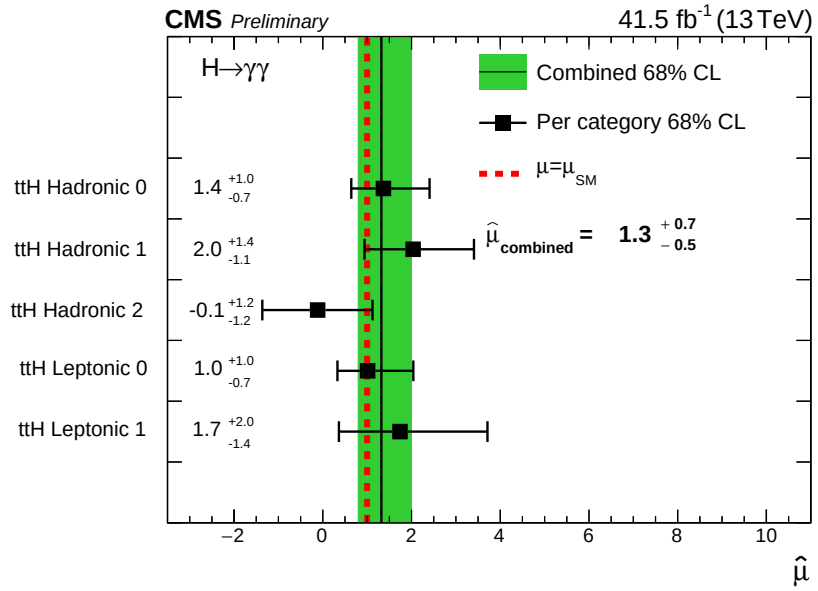


Figure 4.40: Measured value of the signal strength modifier for each of the categories included in the analysis. The $t\bar{t}H$ signal strength $\mu_{t\bar{t}H}$ with its 68% uncertainty is depicted by the green band, while the SM expectation is the red dotted line.

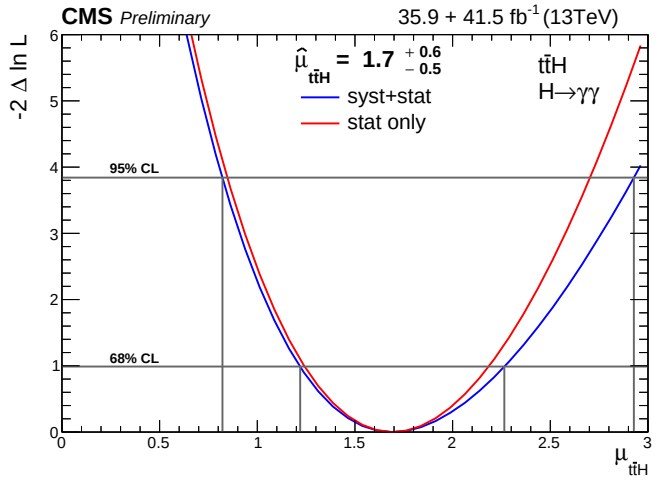


Figure 4.41: Value of the likelihood function in proximity of its minimum for $\mu_{t\bar{t}H}$ when combining the 2016 and 2017 analyses. The Higgs boson mass is constrained to the Run I best-fit value $m_H = 125.09 \pm 0.24$ GeV. The red line represents the likelihood function when only the statistical uncertainty is accounted for, while the red one includes all the systematic uncertainties related to the measurement.

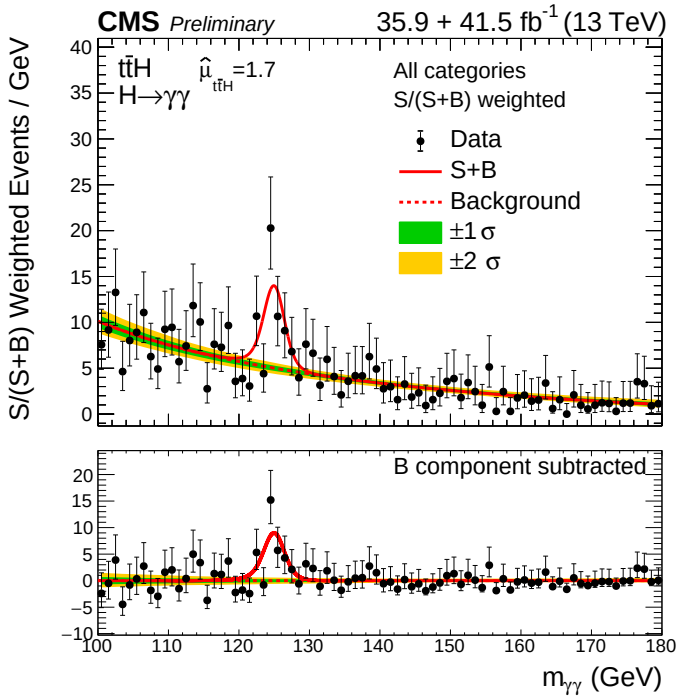


Figure 4.42: Diphoton invariant mass spectrum of the combination of all the categories of the 2016 and 2017 analyses. Each category is weighted for the expected sensitivity, estimated as explained in the text. The data (black markers) are shown together with the signal-plus-background model (red line). The one and two standard deviations uncertainties (green and yellow bands) include the uncertainty on the choice of the fit function and on the fitted parameters. They do not include the Poisson uncertainty due to the number of events in the category. The bottom panel shows the residuals after background subtraction.

4.9 Prospects for the $t\bar{t}H$ measurement in the diphoton channel

This section highlights the foreseen precision for the measurement of the $t\bar{t}H$ process in the $H \rightarrow \gamma\gamma$ channel in coming years. As the $H \rightarrow \gamma\gamma$ channel is still dominated by the statistical uncertainty, as opposite to the other production modes, it will become the leading channel in the near future.
Figure 4.43 reports the likelihood scan for the 2017 analysis extrapolated to the integrated luminosity accumulated in Run II and to the expected integrated luminosity accumulated by the end of Run III. The extrapolations are based on the 2017 data used for the work presented in this thesis. The systematic uncertainties are assumed to be unchanged with respect to the current analysis. When extrapolating the result to the 140 fb^{-1} accumulated in Run II, the expected result is $\mu_{t\bar{t}H} = 1.0^{+0.4}_{-0.3} = 1.0^{+0.2}_{-0.3}(\text{stat.})^{+0.3}_{-0.1}(\text{syst.})$. The statistical uncertainty is still the dominant component and it is reduced to about 40% the current value. The extrapolation to 300 fb^{-1} shows an expected value of $\mu_{t\bar{t}H} = 1.0^{+0.3}_{-0.2} = 1.0^{+0.2}_{-0.2}(\text{stat.})^{+0.3}_{-0.1}(\text{syst.})$ with the statistical component close to the systematic one. Further improvement in the precision are expected thanks to the refined ECAL calibration, which will restore the invariant mass resolution of the 2017 data to the level of 2016, improving the signal to background ratio. The expected improvement thanks to the ECAL calibration on the 2017 dataset is a sensitivity increase of 8 to 10% and between 3 and 5% for the 2016 and 2018 data.
As the dominant source of systematic uncertainty is the theoretical prediction on the rate of the process, there is no way to improve the precision of the measurement beyond the 10% uncertainty. The dominant experimental systematic uncertainty is due to the data-to-simulation disagreement in the distribution of the photon identification BDT. This uncertainty can be reduced by improving the noise modelling of the ECAL, which affects the data-to-simulation agreement of the shower shapes variables. Preliminary studies showed that the uncertainty associated with the photon identification BDT can be reduced by a factor 2 with respect to the 2017 one with refining the noise model of the ECAL simulation. Furthermore, new ideas are being tested to improve the sensitivity of the analysis (Section 4.10), therefore the result of Fig. 4.43 can be seen as a lower limit for the precision of $\mu_{t\bar{t}H}$ expected for the Run II result.

4.10 Possible improvements of the current analysis

At the time of writing, several studies are being finalised aiming at further improving the sensitivity of the $t\bar{t}H$ analysis. The jump in the sensitivity from the 2016 to the 2017 analysis was mainly due to the introduction of BDTs, capable to exploit the correlations among the final state variables. The largest limitation, especially in the leptonic channel, has been the lack of events in the simulation to train the BDT. At present the impact of introducing deep learning techniques is being tested. Deep learning could improve the separation of the signal from the background thanks to a better usage of the correlations among the variables and to the possibility to explicitly reconstruct the top quarks starting from the final state objects.
The hadronic channel, where the lack of events in the background simulation is less severe, presents a more favourable situation for the application of deep learning techniques. The

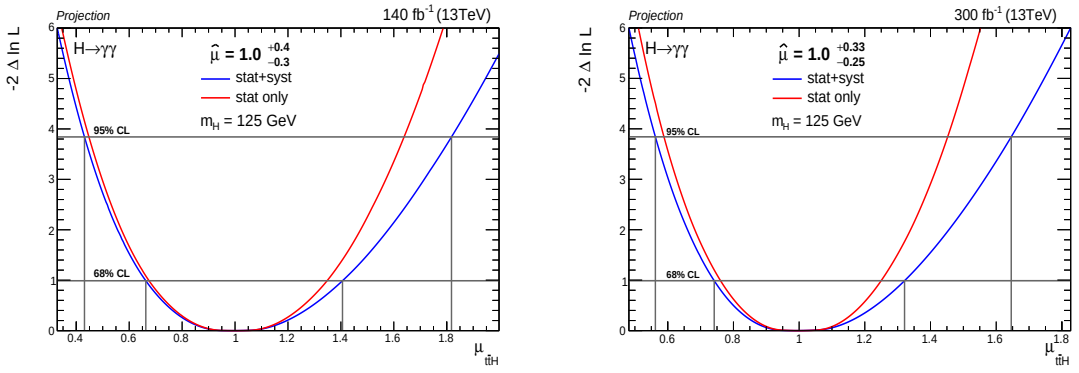


Figure 4.43: Likelihood as a function of $\mu_{t\bar{t}H}$ for the 2017 analysis extrapolated to the integrated luminosity of the Run II (left) and to the foreseen integrated luminosity at the end of the LHC Run III (right). The Higgs boson mass is fixed to $m_H = 125$ GeV. The red line represents the likelihood function when only the statistical uncertainty is accounted for, while the red one includes all the systematic uncertainties related to the measurement.

application of DNNs instead of BDTs could lead to a better separation between signal and background thanks to the capability of DNN to apply non-linear discrimination among the variables. The drawbacks of DNN are the large number parameters of the model, which requires a large training sample, and the presence of several hyper-parameters to be optimised. A preliminary test with DNN and optimised parameters, using as input variables the same ones used for the BDT of the 2017 analysis, showed a mild improvement of about 5% in the sensitivity of the analysis. The sensitivity could be further boosted, using as input of the DNN the four-vectors with the moments of all the final state objects together with some global variables in order to achieve optimal kinematic separation between signal and background.

To further improve the sensitivity of the analysis, additional input variables can be exploited, to help the direct identification of the top quarks. The introduction of the top-tagger, described in Ref. [58], contributes with another 5% of improvement in the sensitivity of the analysis. The top-tagger is a DNN trained to identify triplets of jets coming from the decay of a top quark. For each triplet, a score is assigned by the top-tagger, close to unity for jets very likely to originate from the decay of a top quark and close to zero for randomly chosen triplets. The output of the top-tagger can be used as an additional input variable to the DNN or BDT exploited for the identification of $t\bar{t}H$ events. The training of a classifier is expected to benefit also from additional simulation samples which became available after the conclusion of this work. The simulation of diphoton events with one or two b jets in MadGraph5_aMC@NLO largely increases the number of events which could be used for the training. In addition, a sample of γ + jet events generated in MadGraph5_aMC@NLO with an electromagnetic filter applied at generator level has been created. This sample provides a large number of events passing the preselection generated at the NLO, as opposite to the Pythia sample at LO, improving the data to simulation agreement. The additional samples are expected to further improve the performance of the multivariate tools exploited in the analysis, and hence to slightly increase the sensitivity.

The leptonic categories are less suitable for DNN application, as training of DNNs with

low number of events generally produce poor results. No benefit has been observed with respect to the BDT in this channel. The addition of the above mentioned simulation samples is expected to benefit the channel, despite no preliminary results are available now.
From preliminary studies, a significance of ≈ 5 standard deviations is expected from the analysis of the $H \rightarrow \gamma\gamma$ channel alone. Figure 4.44 shows the expected likelihood as function of $\mu_{t\bar{t}H}$ for the analysis of the full Run II dataset. The expected value of $\mu_{t\bar{t}H}$, including the analysis improvements discussed above, is $\hat{\mu}_{t\bar{t}H} = 1.0^{+0.32}_{-0.27}$, with an uncertainty reduced by about 10% with respect the 2017 analysis extrapolated to the full Run II dataset (see Fig. 4.43).

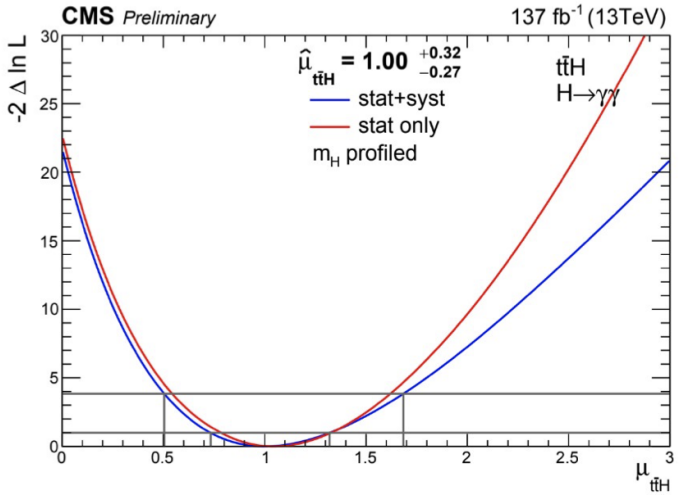


Figure 4.44: Value of the likelihood function in proximity of its minimum for $\mu_{t\bar{t}H}$. The Higgs boson mass is profiled in the fit.

Chapter 5

Observation of the $t\bar{t}H$ process and future prospects

Il y a un spectacle plus grand que la mer, c'est le ciel;
il y a un spectacle plus grand que le ciel,
c'est l'intérieur de l'âme.

Victor Hugo

The work on the 2016 data presented in Chapter 4, in combination with other channels, directly contributed to the first observation of the $t\bar{t}H$ process. The $t\bar{t}H$ event rate is measured exploiting multiple final states, following the different decays of the Higgs boson. Five independent searches performed by the CMS Collaboration are combined to extract a single $t\bar{t}H$ rate measurement. In addition to the diphoton final state, two searches are conducted in the $H \to bb$ channel, one targeting fully hadronic decays of the top quarks [56] and one with at least one of the top quarks decaying to leptons [57]. The channel $H \to ZZ^* \to 4\ell$ is exploited as well, selecting events with four leptons following the Higgs boson decay and additional jets, b jets and leptons [44]. Final states with electrons, muons or hadronically decaying τ leptons coming from decays of the Higgs boson in vector bosons or tau leptons are combined in a single analysis targeting events with leptons in the final state [58]. Similar searches have been performed both with the 2016 data collected at centre-of-mass energy of 13 TeV and with the Run I data collected at 7 and 8 TeV.
The combination of the different analyses is performed by fitting all the categories of the different channels with a single signal strength modifier $\mu_{t\bar{t}H}$, after properly correlating the systematic uncertainties. The tHq and tHW processes are treated as background and normalised to the SM expectation in the fit. Figure 5.1 shows the outcome of the fit. The upper part of the figure shows the result when a signal strength modifier per decay channel is fitted to the data. The best-fit value is depicted with the one and two standard deviations uncertainty bands. Good compatibility among the different processes is observed. The central panel of Fig. 5.1 shows the outcome of the fit when a single signal strength modifier $\mu_{t\bar{t}H}$ is fitted to all the channels, with the decay branching ratios fixed to the SM expectations, separately for the 7 and 8 TeV data and for the 13 TeV ones. Finally, the last entry of Fig 5.1 shows the result of the fit combining the data collected at different centre-of-mass energies. The best-fit value is $\hat{\mu}_{t\bar{t}H} = 1.26^{+0.31}_{-0.26} = 1.26^{+0.16}_{-0.16}(\text{stat.})^{+0.17}_{-0.15}(\text{syst.})^{+0.21}_{-0.13}(\text{theo.})$. The value is in agreement with the SM expectation

and its uncertainty is dominated by the uncertainty on the theoretical modelling of the background. The background modelling is the current limitation of all the analyses involved but $H \to ZZ^* \to 4\ell$, limited by statistical uncertainty, and $H \to \gamma\gamma$, where the background estimation is derived directly from data. The dominant sources of experimental uncertainties are due to the identification efficiency of leptons and b jets.

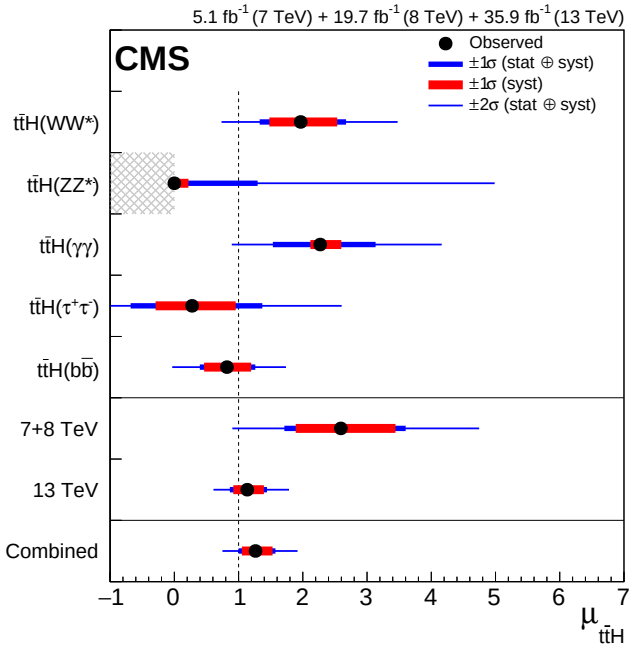


Figure 5.1: Best-fit value of $\mu_{t\bar{t}H}$ with the corresponding one and two standard deviations uncertainty bands. The upper part is the result of the fit when the different decay channels are fitted independently. The mid section is obtained from scaling all the events with a single signal signal strength modifier, separately for the 13 TeV data and for the 7 and 8 TeV ones. The bottom panel shows the overall combination. The signal strength for the $H \to ZZ^*$ process is constrain positive. The SM expectation is shown by the vertical dotted line.

The value of the test statistics q, defined in Eq. 4.10, as a function of $\mu_{t\bar{t}H}$ is depicted in Fig. 5.2. The full combination, as well as the results split according to the centre-of-mass energy, is shown. The null hypothesis is rejected with a significance of 5.2 standard deviations, where 4.2 are expected for a SM Higgs boson. This result constitutes the first direct observation of the $t\bar{t}H$ process. It establishes not only an unobserved Higgs boson production process, but also it proved the tree-level coupling of the Higgs boson with the top quark, and hence to an up-type quark.

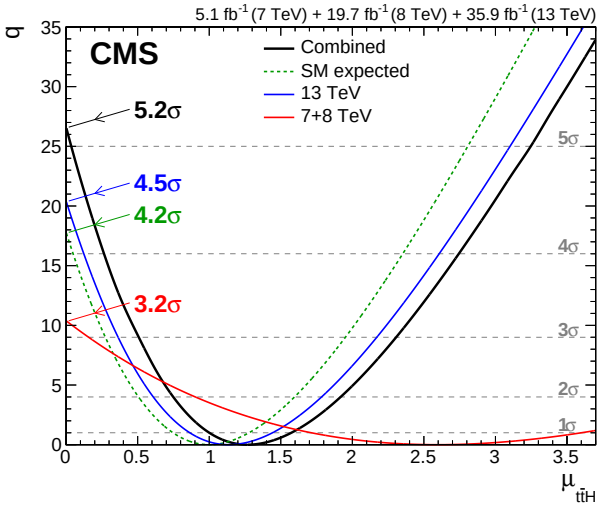


Figure 5.2: Value of test statistic q, defined in Eq. 4.10, as a function of $\mu_{t\bar{t}H}$. The value of q is shown for the combination of all the production modes at 13 TeV (blue line), at 7 and 8 TeV (red line) and for the overall combination (black line). The SM expectation is shown by the green dotted line. The horizontal dotted lines are the p-values for the background-only hypothesis expressed in units of significance computed in the asymptotic limit.

5.1 Status of the Higgs boson couplings

The analysis of the 2016 data resulted in precise measurement of the four Higgs boson production processes in several final states involving multiple decays of the Higgs boson. The results obtained in the $H \rightarrow \gamma\gamma$ channel are described in Section 4.8.1. Analogue results are available by the CMS experiment in final states involving decays of the Higgs boson to bottom quarks [126, 127], τ leptons [128] and vector bosons [44, 55]. The combination of all the information gives an update to the knowledge of the Higgs boson couplings [35], largely improving the precision achieved in the LHC Run I combination that was available at the beginning of this work (see Section 1.3.4).
Figure 5.3 summarises the knowledge of the Higgs boson couplings achieved with 2016 data. The left figure illustrates the signal strength modifiers for the different production processes, with the VH split in WH and ZH. Large improvements are obtained compared to the Run I result, with an uncertainty reduced by 50% in the ggH and $t\bar{t}H$ processes and of 20% on the VBF and VH ones. The mild excess observed on the $t\bar{t}H$ production is no longer present, as the $t\bar{t}H$ production rate is compatible with the SM expectation within one standard deviation. The right figure illustrates the best-fit values for the coupling modifiers derived within the κ-framework. In this parameterisation, the loop involved in the ggH production and the one involved in the $H \rightarrow \gamma\gamma$ decay are scaled with the respective coupling modifiers κ_g and κ_γ. The decay branching fraction of the Higgs boson to unpredicted particles is assumed to be zero ($\mathcal{B}_{BSM} = 0$) and the coupling to the top quark is assumed positive, while the couplings to the vector bosons are free to assume either positive or negative values. The improvement on the Run I result in the precision of κ_t is of about 40%, thanks to the improved precision in the measurement of the $t\bar{t}H$ cross section. All the measurements are compatible with the SM expectation.

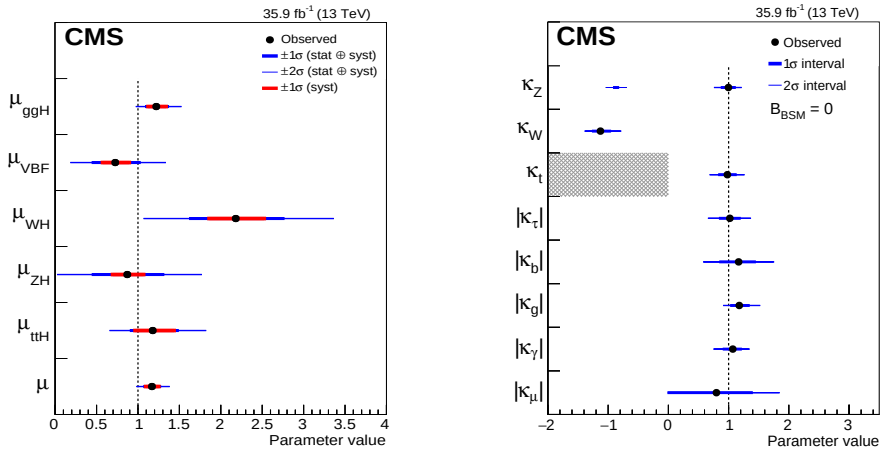


Figure 5.3: Best-fit values of the Higgs boson coupling measured by the CMS experiment with 35.9 fb^{-1} of data collected in 2016. Left: signal strength modifiers for the Higgs boson main production processes. The values are shown with the one and two standard deviation uncertainty bands. The vertical dotted line indicate the SM prediction. Right: best-fit value for the coupling modifiers extracted from the κ-framework (see Section 1.3.4). The details about the two parameterisations are given in the text. The coupling to the top quark is assumed positive, while the coupling to the vector bosons are free to assume either positive or negative values. The one and two standard deviations uncertainty bands are shown. The vertical dotted line indicates the SM prediction [35].

Figure 5.4 shows the dependence of the couplings from the particle mass, updating the result presented in Fig. 1.4. The top panel illustrates the CMS measurement performed with 35.9 fb^{-1} of data collected in 2016. A similar result is also available from the ATLAS Collaboration [129], based on about 79.8 fb^{-1}of data collected in 2016 and 2017. The ATLAS result is illustrated in the bottom figure. All the result are in agreement, within their uncertainties, with the expectation of the SM.
An update of the measurement of the Higgs boson couplings, and thus of the $t\bar{t}H$ event rate, is expected by the end of 2020, based on the analysis of the data collected during the LHC Run II in all the accessible final states.

5.2 The $t\bar{t}H$ measurement at the HL-LHC

After the end of the Run III, the LHC will enter its high luminosity phase. The upgraded triplet magnets will increase the luminosity delivered by the LHC by a factor 5 to 7 the present luminosity. The HL-LHC is expected to deliver 3000 fb^{-1} of data in ten years of operations (see Section 2.1.2). The impressive amount of data that will be delivered by the HL-LHC will determine a large reduction of the uncertainty on the measurement of the Higgs boson couplings. The CMS experiment quantified the expected improvements on the Higgs boson properties in Ref. [130]. The study is performed extrapolating the analyses performed on 2016 data to 300 and 3000 fb^{-1}, corresponding to the expected amount of data collected by the end of the LHC Run III and of the HL-LHC. The extrapolations make different assumptions. In the first scenario (S1) the systematic uncertainties are assumed to be unchanged with respect to the present. Instead, the second scenario (S2)

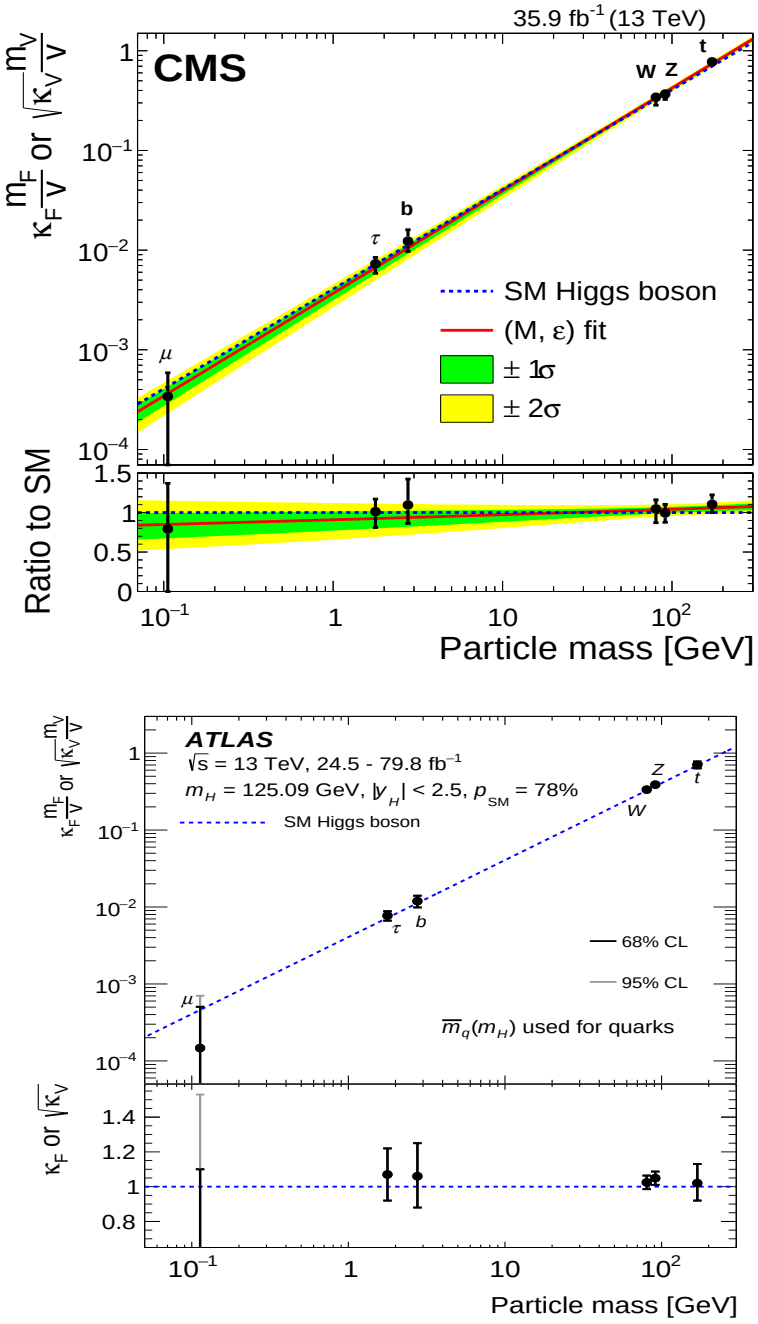


Figure 5.4: Best-fit values of the Higgs boson couplings as a function of the particle mass. The y axis reports $\kappa_{\mathrm{F}} \cdot m_{\mathrm{F}}/v$ for fermions and $\sqrt{\kappa_{\mathrm{V}}} \cdot m_{\mathrm{V}}/v$ for bosons, where m is the mass of the particle, v is the Higgs field vacuum expectation value and κ is the coupling modifier of the κ-framework defined in the text. The dotted line shows the SM prediction. The bottom panel of each figure is the ratio to the SM expectation. The top figure shows the CMS measurement based on 35.9 fb^{-1} of data collected in 2016, the bottom one shows the ATLAS measurement based on 79.8 fb^{-1} of data collected in 2016 and 2017. All results are compatible with the SM within the experimental uncertainties.

assume all the systematic uncertainties are reduced to half of the present one, considering improvements in the experimental control of the uncertainties thanks to the large amount of data, as well as improvements in the precision of the theory computations.
The projection for the expected results is shown in Fig. 5.5 and 5.6. The first figure shows the expected uncertainty on the signal strength modifiers for the Higgs boson production processes with 300 and 3000 fb^{-1}. The second one shows the expected uncertainty on the coupling modifiers extracted in the κ-framework for the same integrated luminosities and assuming no contributions from unpredicted particles ($\mathcal{B}_{\text{BSM}} = 0$). Already at the end of the LHC Run III, the uncertainty on $\mu_{\text{t}\bar{\text{t}}\text{H}}$ is expected to be dominated by the systematic uncertainty, with major contributions from the theory uncertainties. The uncertainty on the signal strength modifier is expected to be 15% in the S1 scenario and 10% in the S2, further reduced to 5 and 10% at the end of the HL-LHC. The uncertainty on the top coupling modifier is expected to be of about 8% in S1 and 6% in S2 at the end of the Run III and 6% or 4% at the end of HL-LHC. The main limitation will be due to systematic uncertainties, therefore a careful study of the experimental sources of uncertainties will be mandatory, in addition to the development of improved theoretical predictions for the processes.
The uncertainties on the other production processes is expected to be less than 10% at the end of the HL-LHC, while the uncertainty on the couplings will be less than about 5%, including the muon coupling.

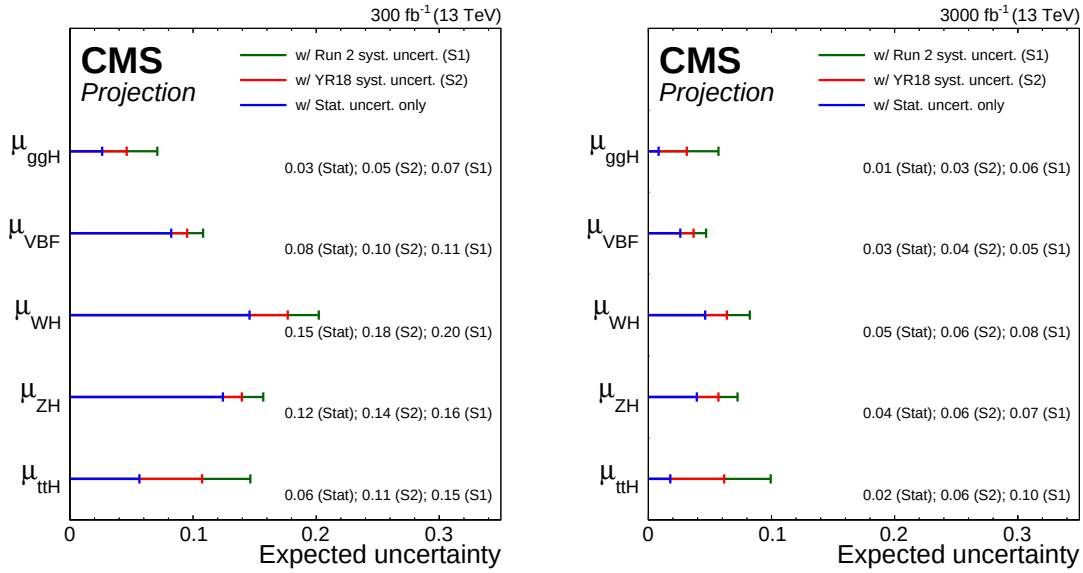


Figure 5.5: Expected one standard deviation uncertainty on the signal strength modifiers for the Higgs boson production processes after 300 (left) and 3000 fb^{-1} (right) of data are collected. The uncertainty is shown in the S1 and S2 scenarios (defined in the text). The statistical component of the uncertainty is indicated by the blue band [130].

5.3 The $t\bar{t}H$ measurement beyond the HL-LHC

After the HL-LHC, the landscape of the experimental high energy physics is still undefined. Several hypothesis are under investigation, both linear and circular colliders of protons or electrons. The electron-positron (ee) colliders mainly target to become 'Higgs boson factories', capable of providing an extremely clean experimental environment and large number produced of Higgs bosons, in order to achieve ultimate precision in the determination of the Higgs boson properties. Instead, the proton colliders aim at breaking the energy frontier of the LHC.
The Future Circular Collider (FCC) [131] is one of the possibilities under study. It would be a 80 to 100 km accelerator with the LHC as injector, featuring two separate experimental phases. At the beginning, an electron-proton collider (FCC-ee) would work as an Higgs boson factory, providing a precise determination of the Higgs boson properties and improving the precision of the measurements in the electroweak sector. After about 10 years of operations, the accelerator would be dismantled and replaced by a 100 TeV proton-proton collider (FCC-hh), to complement the measurement of the FCC-ee thanks to the unprecedented luminosity and centre-of-mass energy. In one of the experimental caverns of the FCC-hh, it would be possible to install an electron accelerator so to realise parasitic electro-proton collisions, the so called FCC-eh. The latter would provide complementary measurement of the Higgs boson properties, as well as a measurement of the PDF at extremely high energies. The full experimental program of the FCC would consist of about 50 years of operations, from the end of the HL-LHC until the end of the XXI century.

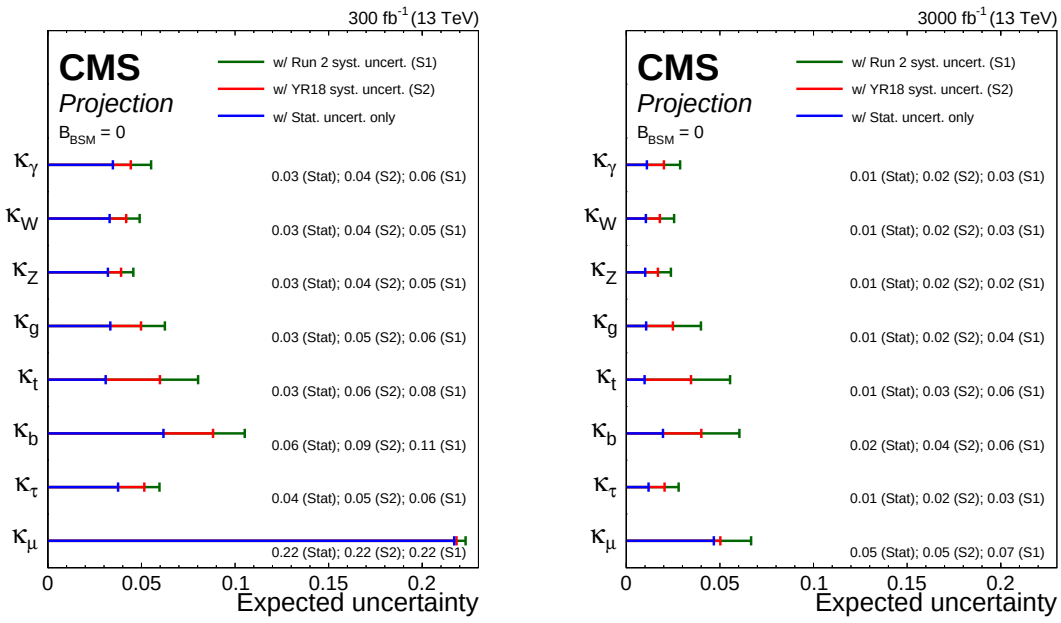


Figure 5.6: Expected one standard deviation uncertainty on the coupling modifiers extracted in the κ-framework after 300 (left) and 3000 fb^{-1} (right) of data are collected. The uncertainty is shown in the S1 and S2 scenarios (defined in the text). The statistical component of the uncertainty is indicated by the blue band. The results are derived assuming no decay branching fraction in unpredicted particles ($\mathcal{B}_{\text{BSM}} = 0$) [130].

Despite the project is still in its conceptual design phase, a first projection on the precision of the Higgs boson couplings after FCC is shown in Fig. 5.7. The combination of the three experimental programs could reduce the uncertainty on all the measurable Higgs boson couplings to less than 1% and the uncertainty on the Higgs boson self coupling to less than 10%. Such a level of precision would be enough to finally assess the SM nature of the Higgs boson and a possible special role of y_t in the model.

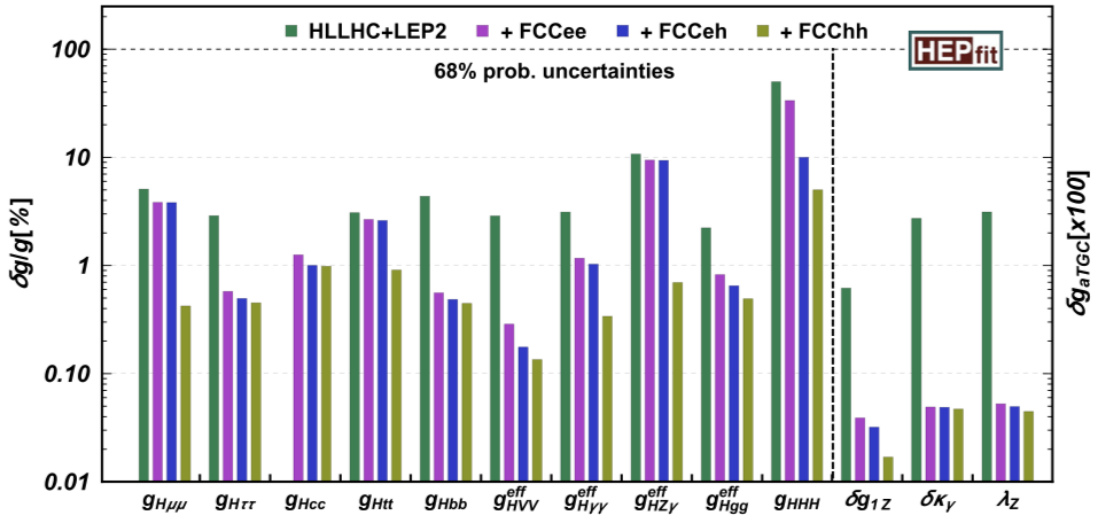


Figure 5.7: Expected one standard deviation uncertainty on the coupling modifiers ($\delta g/g$) extracted in the κ-framework for the Higgs boson production processes after the FCC program. The green bar is the precision expected after the HL-LHC, the purple one adding the FCC-ee program, the blue one with the addition of FCC-eh and finally the yellow bar is the expected precision at the end of the FCC experimental program [131].

Chapter 6

Conclusions

A l'alta fantasia qui mancò possa;
ma già volgeva il mio disio e 'l velle,
sì come rota ch'igualmente è mossa,
l'amor che move il sole e l'altre stelle.

Dante Alighieri

This thesis presents the analysis of the data collected by the CMS detector at the Large Hadron Collider to perform a precise measurement of the cross section of the Higgs boson production in association with a pair of top-antitop quarks ($t\bar{t}H$). The diphoton decay channel of the Higgs boson is exploited to perform the measurement because it is one of the most sensitive channels. The measurement is based on 35.9 fb^{-1} and 41.5 fb^{-1} of data collected in 2016 and 2017 at a centre-of-mass energy of 13 TeV.
The diphoton decay channel offers a rather clean experimental environment to study the Higgs boson, where the final state is fully reconstructed. As the photon energy resolution directly affects the precision of the measurement, a large effort has been dedicated to the calibration of the ECAL. In particular, the φ-symmetry method has been exploited and refined during this thesis. The refined calibration of the ECAL for the Run II data is expected to improve the sensitivity of the $t\bar{t}H$ analysis, with the $H \rightarrow \gamma\gamma$, of about 10% on the 2017 data and around 5% on the 2016 and 2018 ones. In addition, the effort in the development of refined L1 trigger algorithms ensured highly-efficient triggering of photons and electrons throughout the Run II data taking.
The work involved in this thesis contributed to the measurement of the $t\bar{t}H$ production cross section, with two separate analyses, one based on the 2016 data and one on the 2017 ones.
The analysis of the 2016 data measured a value of $\hat{\mu}_{t\bar{t}H} = 2.2^{+0.9}_{-0.8}$, corresponding to significance of 3.2 standard deviations, where 1.5 is expected for a SM Higgs boson. The 2017 analysis measured a signal strength of $\hat{\mu}_{t\bar{t}H} = 1.3^{+0.7}_{-0.5}$, rejecting the background-only hypothesis at the level of 3.1 standard deviations, where 2.2 are expected for a SM Higgs boson. The combination of the two analyses resulted in an observed signal strength of $\hat{\mu}_{t\bar{t}H} = 1.7^{+0.6}_{-0.5}$, corresponding to a significance of 4.1 standard deviations.
The analysis of the 2016 data, together with the analyses targeting final states with the Higgs boson decaying to b quarks, vector bosons, and τ leptons, allowed the first experimental observation of the $t\bar{t}H$ process. The combined result exploited the analyses of the data collected in 2016 at centre-of-mass energy of 13 TeV, as well as the data

collected in 2011 and 2012 at a centre-of-mass energy of 7 and 8 TeV, respectively. The combination of the different channels measured $\hat{\mu}_{\mathrm{t\bar{t}H}} = 1.26^{+0.31}_{-0.26}$, in agreement with the SM expectation. The background-only hypothesis is rejected at the level of 5.1 standard deviations. This result proves for the first time the tree-level coupling of the Higgs boson with the top quark and, hence, with an up-type quark.
The work presented in this document constitutes the basis for the analysis of the full Run II data, whose result is expected within a few months from the end of this thesis. The precision on the $\mathrm{t\bar{t}H}$ cross section will benefit from the increased integrated luminosity, as the statistical uncertainty is still the dominant contribution in the diphoton channel.
The LHC is just at the first stages of its history, with the present delivered integrated luminosity that is less than 10% of the amount expected at the end of its operations in 2035. The precision of the measurement of y_{t} is expected to improve to better than 5% at the end of the LHC era. Future colliders could further reduce the uncertainty below the 1% level, probing the Higgs boson sector of the SM at an unprecedented precision. More efforts will be necessary in the future to constantly improve the performance of the detectors, to find innovative hardware solutions and to refine the techniques exploited for the analysis of the data. The work presented in this document is just a small piece in the characterisation of the Higgs boson properties but, as the wise latin writer Seneca wrote, '*Minus ex crastino pendeas, si hodierno manum inieceris*'.

List of abbreviations

ADC	Analogue-to-Digital Converter
ALICE	A Large Ion Collider Experiment
APD	Avalanche PhotoDiode
ATLAS	A Toroidal Apparatus
bbH	Associated production of Higgs bosons and bottom quark pair
BDT	Boosted Decision Tree
BSM	Beyond the Standard Model
CERN	European Organisation for Nuclear Research
CKM	Cabibbo-Kobayashi-Maskawa
CMS	Compact Muon Solenoid
CSC	Cathode Strip Chamber
CSV	Combined Secondary Vertex
DAQ	Data Acquisition system
DNN	Deep Neural Netweork
DT	Drift Tubes
EB	ECAL Barrel
ECAL	Electromagnetic Calorimeter
EE	ECAL Endcaps
EG	EGamma (electrons and photons L1 objects)
ES	ECAL preshower
EW	ElectroWeak
FPGA	Field-Programmable Gate Arrays
FWHM	Full Width at Half of the Maximum
ggH	Gluon fusion
GSF	Gaussian Sum Filter
HB	HCAL Barrel
HCAL	Hadronic Calorimeter
HE	HCAL Endcaps
HF	HCAL Forward
HL-LHC	High Luminosity - Large Hadron Collider
HLT	High Level Trigger
HO	HCAL Outer
IC	Intercalibration Constant
JEC	Jet Energy Correction
LHC	Large Hadron Collider
LHCb	Large Hadron Collider bottom
LINAC	Linear Accelerator

LM	Laser Monitoring
LO	Leading Order (in perturbation theory)
LS1 (2,3)	Long Shutdown 1 (2,3)
L1	Level 1 Trigger
MB	Muon Barrel
ME	Muon Endcaps
NLL	Negative Log-Likelihood
NLO	Next to Leading Order (in perturbation theory)
NNLO	Next to Next to Leading Order (in perturbation theory)
N^{3}LO	Next to Next to Next to Leading Order (in perturbation theory)
OOT	Out Of Time
PDF	Parton Density Function
PF	Particle Flow
PS	Proton Synchrotron
PSV	Pixel Seed Veto
PU	Pile Up
QCD	Quantum ChromoDynamics
RPC	Resistive Plate Chamber
SM	Standard Model of particle interaction
SPS	Super Proton Synchrotron
STXS	Simplified Template Cross Sections
TEC	Tracker EndCaps
T&P	Tag & Probe
tHq	Associated production of Higgs boson and single top quark
tHW	Associated production of Higgs boson, single top quark and W boson
TIB	Tracker Inner Barrel
TID	Tracker Inner Disks
TOB	Tracker Outer Barrel
TT	Trigger Tower
t$\bar{\text{t}}$H	Associated production of Higgs boson and top-antitop quark pair
VBF	Vector Boson Fusion
VH	Associated production of Higgs boson and vector bosons
VPT	Vacuum PhotoTriode
WH	Associated production of Higgs boson and W bosons
WP	Working Point
ZH	Associated production of Higgs boson and Z bosons

Bibliography

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

[2] S. Weinberg, *A Model of Leptons*, *Phys. Rev. Lett.* **19** (1967) 1264–1266.

[3] A. Salam, *Weak and Electromagnetic Interactions*, *Conf. Proc.* **C680519** (1968) 367–377.

[4] S. L. Glashow, J. Iliopoulos and L. Maiani, *Weak Interactions with Lepton-Hadron Symmetry*, *Phys. Rev.* **D2** (1970) 1285–1292.

[5] ATLAS Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the detector at the LHC*, *Phys. Lett.* **B716** (2012) 1–29, [1207.7214].

[6] CMS Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett.* **B716** (2012) 30–61, [1207.7235].

[7] Particle Data Group Collaboration, M. Tanabashi, K. Hagiwara, K. Hikasa, K. Nakamura, Y. Sumino, F. Takahashi et al., *Review of particle physics*, *Phys. Rev. D* **98** (Aug, 2018) 030001.

[8] C. A. Baker et al., *An Improved experimental limit on the electric dipole moment of the neutron*, *Phys. Rev. Lett.* **97** (2006) 131801, [hep-ex/0602020].

[9] V. N. Gribov and L. N. Lipatov, *Deep inelastic e p scattering in perturbation theory*, *Sov. J. Nucl. Phys.* **15** (1972) 438–450.

[10] L. N. Lipatov, *The parton model and perturbation theory*, *Sov. J. Nucl. Phys.* **20** (1975) 94–102.

[11] G. Altarelli and G. Parisi, *Asymptotic Freedom in Parton Language*, *Nucl. Phys.* **B126** (1977) 298–318.

[12] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, *Sov. Phys. JETP* **46** (1977) 641–653.

[13] ATLAS Collaboration, M. Aaboud et al., *Measurement of charged-particle distributions sensitive to the underlying event in $\sqrt{s} = 13$ TeV proton-proton collisions with the ATLAS detector at the LHC*, *JHEP* **03** (2017) 157, [1701.05390].

[14] CMS Collaboration, A. M. Sirunyan et al., *Measurement of the underlying event activity in inclusive Z boson production in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *JHEP* **07** (2018) 032, [1711.04299].

[15] N. Cabibbo, *Unitary symmetry and leptonic decays*, *Phys. Rev. Lett.* **10** (Jun, 1963) 531–533.

[16] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, *Prog. Theor. Phys.* **49** (1973) 652–657.

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

[18] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, *Phys. Lett.* **12** (1964) 132–133.

[19] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, *Phys. Rev. Lett.* **13** (1964) 508–509.

[20] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, *Phys. Rev. Lett.* **13** (1964) 585–587.

[21] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, *Phys. Rev.* **145** (1966) 1156–1163.

[22] T. W. B. Kibble, *Symmetry breaking in nonAbelian gauge theories*, *Phys. Rev.* **155** (1967) 1554–1561.

[23] J. Goldstone, *Field Theories with Superconductor Solutions*, *Nuovo Cim.* **19** (1961) 154–164.

[24] H. G. et al., *Naturalness, Chiral Symmetry, and Spontaneous Chiral Symmetry Breaking*, *NATO Advanced Study Institutes Series (Series B. Physics)* **59** (1980) .

[25] G. F. Giudice, *Naturalness after LHC8*, *PoS* **EPS-HEP2013** (2013) 163, [1307.7879].

[26] E. Witten, *Dynamical breaking of supersymmetry*, *Nuclear Physics B* **188** (1981) 513 – 554.

[27] S. Dimopoulos and S. Raby, *Supercolor*, *Nuclear Physics B* **192** (1981) 353 – 368.

[28] S. Dimopoulos and H. Georgi, *Softly broken supersymmetry and su(5)*, *Nuclear Physics B* **193** (1981) 150 – 162.

[29] S. Dimopoulos, H. Georgi, R. K. Kaul and P. Majumdar, *Cancellation of quadratically divergent mass corrections in globally supersymmetric spontaneously broken gauge theories*, *Nuclear Physics B* **199** (1982) 36 – 58.

[30] M. Dine, W. Fischler and M. Srednicki, *Supersymmetric technicolor*, *Nuclear Physics B* **189** (1981) 575 – 593.

[31] M. J. Dugan, H. Georgi and D. B. Kaplan, *Anatomy of a Composite Higgs Model*, *Nucl. Phys.* **B254** (1985) 299–326.

[32] R. Contino, *The Higgs as a Composite Nambu-Goldstone Boson*, in *Physics of the large and the small, TASI 09, proceedings of the Theoretical Advanced Study Institute in Elementary Particle Physics, Boulder, Colorado, USA, 1-26 June 2009*, pp. 235–306, 2011. 1005.4269. DOI.

[33] K. Agashe, R. Contino and A. Pomarol, *The Minimal composite Higgs model*, *Nucl. Phys.* **B719** (2005) 165–187, [hep-ph/0412089].

[34] ATLAS and CMS Collaborations, Aad et al., *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, *JHEP* **08** (2016) 045, [1606.02266].

[35] CMS Collaboration, A. M. Sirunyan et al., *Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J.* **C79** (2019) 421, [1809.10733].

[36] ATLAS Collaboration, *Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment*, Tech. Rep. ATLAS-CONF-2019-005, CERN, Geneva, Mar, 2019.

[37] MuLan Collaboration, D. M. Webber et al., *Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision*, *Phys. Rev. Lett.* **106** (2011) 041803, [1010.0991].

[38] ATLAS and CMS Collaborations, *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** (2015) 191803, [1503.07589].

[39] F. Demartin, F. Maltoni, K. Mawatari and M. Zaro, *Higgs production in association with a single top quark at the LHC*, *Eur. Phys. J.* **C75** (2015) 267, [1504.00611].

[40] LHC Higgs Cross Section Working Group, D.de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, 1610.07922.

[41] ATLAS Collaboration, M. Aaboud et al., *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*, *Phys. Lett.* **B786** (2018) 59–86, [1808.08238].

[42] CMS Collaboration, A. M. Sirunyan et al., *Observation of Higgs boson decay to bottom quarks*, *Phys. Rev. Lett.* **121** (2018) 121801, [1808.08242].

[43] LHC Higgs Cross Section Working Group, David, A. and Denner, A. and Duehrssen, M. and Grazzini, M. and Grojean, C. and Passarino, G. and Schumacher, M. and Spira, M. and Weiglein, G. and Zanetti, M., *LHC HXSWG interim recommendations to explore the coupling structure of a Higgs-like particle*, 1209.0040.

[44] CMS Collaboration, A. M. Sirunyan et al., *Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **11** (2017) 047, [1706.09936].

[45] ATLAS Collaboration, M. Aaboud et al., *Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector*, *Phys. Lett.* **B784** (2018) 345–366, [1806.00242].

[46] CMS Collaboration, A. M. Sirunyan et al., *Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state*, 1901.00174.

[47] ATLAS Collaboration, *Combination of searches for Higgs boson pairs in pp collisions at 13 TeV with the ATLAS experiment.*, Tech. Rep. ATLAS-CONF-2018-043, CERN, Geneva, Sep, 2018.

[48] CMS Collaboration, A. M. Sirunyan et al., *Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. Lett.* **122** (2019) 121803, [1811.09689].

[49] S. Dimopoulos, M. Dine, S. Raby and S. D. Thomas, *Experimental signatures of low-energy gauge mediated supersymmetry breaking*, *Phys. Rev. Lett.* **76** (1996) 3494–3497, [hep-ph/9601367].

[50] S. Ambrosanio, G. L. Kane, G. D. Kribs, S. P. Martin and S. Mrenna, *Search for supersymmetry with a light gravitino at the Fermilab Tevatron and CERN LEP colliders*, *Phys. Rev.* **D54** (1996) 5395–5411, [hep-ph/9605398].

[51] K. T. Matchev and S. Thomas, *Higgs and z-boson signatures of supersymmetry*, *Phys. Rev. D* **62** (Sep, 2000) 077702.

[52] K. Howe and P. Saraswat, *Excess Higgs Production in Neutralino Decays*, *JHEP* **10** (2012) 065, [1208.1542].

[53] J. Ellis, D. S. Hwang, K. Sakurai and M. Takeuchi, *Disentangling Higgs-Top Couplings in Associated Production*, *JHEP* **04** (2014) 004, [1312.5736].

[54] ATLAS Collaboration, G. Aad et al., *Measurement of the production cross section for a Higgs boson in association with a vector boson in the $H \to WW^* \to \ell\nu\ell\nu$ channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett.* **B798** (2019) 134949, [1903.10052].

[55] CMS Collaboration, A. M. Sirunyan et al., *Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV*, *Phys. Lett.* **B791** (2019) 96, [1806.05246].

[56] CMS Collaboration, A. M. Sirunyan et al., *Search for $\mathrm{t\bar{t}}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **06** (2018) 101, [1803.06986].

[57] CMS Collaboration, A. M. Sirunyan et al., *Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with leptonic $\mathrm{t\bar{t}}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **03** (2019) 026, [1804.03682].

[58] CMS Collaboration, A. M. Sirunyan et al., *Evidence for associated production of a Higgs boson with a top quark pair in final states with electrons, muons, and hadronically decaying τ leptons at $\sqrt{s} = 13$ TeV*, *JHEP* **08** (2018) 066, [1803.05485].

[59] ATLAS Collaboration, M. Aaboud et al., *Measurement of the top quark mass in the $t\bar{t} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results*, *Eur. Phys. J.* **C79** (2019) 290, [1810.01772].

[60] CMS Collaboration, V. Khachatryan et al., *Measurement of the top quark mass using proton-proton data at $\sqrt{(s)} = 7$ and 8 TeV*, *Phys. Rev.* **D93** (2016) 072004, [1509.04044].

[61] CMS Collaboration, V. Khachatryan et al., *Measurement of the ratio $\mathcal{B}(t \to Wb)/\mathcal{B}(t \to Wq)$ in pp collisions at $\sqrt{s} = 8$ TeV*, *Phys. Lett.* **B736** (2014) 33–57, [1404.2292].

[62] LHC Study Group, *The Large Hadron Collider: conceptual design*, Tech. Rep. CERN-AC-95-05-LHC, No, Oct, 1995.

[63] E. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, Jul, 2016.

[64] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole et al., *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004, 10.5170/CERN-2004-003-V-1.

[65] L. Evans and P. Bryant, *LHC machine*, *Journal of Instrumentation* **3** (aug, 2008) S08001–S08001.

[66] The HL-LHC project. http://hilumilhc.web.cern.ch/about/hl-lhc-project, 2018.

[67] Apollinari et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1*. CERN Yellow Reports: Monographs. CERN, Geneva, 2017, 10.23731/CYRM-2017-004.

[68] CMS Collaboration, C. Bayatian et al., *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. CERN, Geneva, 2006.

[69] CMS Collaboration, S. Chatrchyan et al., *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (aug, 2008) S08004–S08004.

[70] CMS Collaboration. https://cms.cern/news/cms-detector-design, 2016.

[71] CMS Collaboration, S. Chatrchyan et al., *Description and performance of track and primary-vertex reconstruction with the CMS tracker*, *JINST* **9** (2014) P10009, [1405.6569].

[72] *CMS Technical Design Report for the Pixel Detector Upgrade*, Tech. Rep. CERN-LHCC-2012-016. CMS-TDR-11, NO, Sep, 2012.

[73] CMS Collaboration, G. Bayatian et al., *The CMS electromagnetic calorimeter project: Technical Design Report*. CERN, Geneva, 1997.

[74] CMS Collaboration, S. Chatrchyan et al., *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[75] Q. Ingram et al., *Energy resolution of the barrel of the CMS electromagnetic calorimeter*, *Journal of Instrumentation* **2** (apr, 2007) P04004–P04004.

[76] *CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter*, Tech. Rep. CERN-LHCC-2012-015. CMS-TDR-10, NO, Sep, 2012.

[77] USCMS, ECAL/HCAL Collaborations, S. Abdullin et al., *The CMS barrel calorimeter response to particle beams from 2-GeV/c to 350-GeV/c*, *Eur. Phys. J.* **C60** (2009) 359–373.

[78] CMS Collaboration, S. Chatrchyan et al., *The Performance of the CMS Muon Detector in Proton-Proton Collisions at $\sqrt{s}$ = 7 TeV at the LHC*, *JINST* **8** (2013) P11002, [1306.6905].

[79] CMS Collaboration, S. Chatrchyan et al., *Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV*, *JINST* **7** (2012) P10002, [1206.4071].

[80] CMS Collaboration, *CMS Technical Design Report for the Level-1 Trigger Upgrade*, Tech. Rep. CERN-LHCC-2013-011. CMS-TDR-12, NO, Jun, 2013.

[81] CMS Collaboration, A. M. Sirunyan et al., *Particle-flow reconstruction and global event description with the CMS detector*, *JINST* **12** (2017) P10003, [1706.04965].

[82] R. E. Kalman, *A New Approach to Linear Filtering and Prediction Problems*, *Journal of Fluids Engineering* **82** (03, 1960) 35–45.

[83] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[84] CMS Collaboration, A. M. Sirunyan et al., *Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$ = 13 TeV*, *JINST* **13** (2018) P06015, [1804.04528].

[85] K. Ito and K. Xiong, *Gaussian filters for nonlinear filtering problems*, *IEEE Transactions on Automatic Control* **45** (May, 2000) 910–927.

[86] CMS Collaboration, V. Khachatryan et al., *Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV*, *JINST* **10** (2015) P06005, [1502.02701].

[87] CMS Collaboration, V. Khachatryan et al., *Performance of Photon Reconstruction and Identification with the CMS Detector in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV*, *JINST* **10** (2015) P08010, [1502.02702].

[88] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, [0802.1189].

[89] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J. C* **72** (2012) 1896, [1111.6097].

[90] CMS Collaboration, V. Khachatryan et al., *Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV*, *JINST* **12** (2017) P02014, [1607.03663].

[91] CMS Collaboration, A. M. Sirunyan et al., *Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector*, *JINST* **14** (2019) P07004, [1903.06078].

[92] P. Adzic et al., *Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter*, *Eur. Phys. J.* **C46S1** (2006) 23–35.

[93] A. Annenkov, E. Auffray, M. Korzhik, P. Lecoq and J.-P. Peigneux, *On the Origin of the Transmission Damage in Lead Tungstate Crystals Under Irradiation*, .

[94] M. Anfreville et al., *Laser monitoring system for the CMS lead tungstate crystal calorimeter*, *Nucl. Instrum. Meth.* **A594** (2008) 292–320.

[95] CMS Collaboration, S. Chatrchyan et al., *Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV*, *JINST* **8** (2013) P09009, [1306.2016].

[96] CMS Collaboration, A. M. Sirunyan et al., *Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **11** (2018) 185, [1804.02716].

[97] CMS Collaboration, A. M. Sirunyan et al., *Observation of $\mathrm{t\bar{t}}H$ production*, *Phys. Rev. Lett.* **120** (2018) 231801, [1804.02610].

[98] CMS Collaboration, A. M. Sirunyan et al., *Measurement of the associated production of a Higgs boson and a pair of top-antitop quarks with the Higgs boson decaying to two photons in proton-proton collisions at $\sqrt{s} = 13$ TeV*, 2018.

[99] Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, (San Francisco, CA, USA), pp. 148–156, Morgan Kaufmann Publishers Inc., 1996.

[100] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079, [1405.0301].

[101] R. Frederix and S. Frixione, *Merging meets matching in MC@NLO*, *JHEP* **12** (2012) 061, [1209.6215].

[102] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten et al., *An Introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159–177, [1410.3012].

[103] NNPDF Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040, [1410.8849].

[104] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, *JHEP* **11** (2007) 070, [0709.2092].

[105] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, *JHEP* **11** (2004) 040, [hep-ph/0409146].

[106] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, *JHEP* **06** (2010) 043, [1002.2581].

[107] H. B. Hartanto, B. Jager, L. Reina and D. Wackeroth, *Higgs boson production in association with top quarks in the POWHEG BOX*, *Phys. Rev.* **D91** (2015) 094003, [1501.04498].

[108] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert et al., *Event generation with SHERPA 1.1*, *JHEP* **02** (2009) 007, [0811.4622].

[109] GEANT4 Collaboration, S. Agostinelli et al., GEANT4—*a simulation toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.

[110] CMS Collaboration, *Jet algorithms performance in 13 TeV data*, Tech. Rep. CMS-PAS-JME-16-003, CERN, Geneva, 2017.

[111] CMS Collaboration, A. M. Sirunyan et al., *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, *JINST* **13** (2018) P05011, [1712.07158].

[112] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban and D. Whiteson, *Jet Flavor Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev.* **D94** (2016) 112002, [1607.08633].

[113] Y. Lecun, Y. Bengio and G. Hinton, *Deep learning*, *Nature* **521** (May, 2015) 436–444.

[114] A. Hocker et al., *TMVA - Toolkit for Multivariate Data Analysis*, physics/0703039.

[115] *Procedure for the LHC Higgs boson search combination in Summer 2011*, Tech. Rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug, 2011.

[116] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J.* **C71** (2011) 1554, [1007.1727].

[117] A. Wald, *Tests of statistical hypotheses concerning several parameters when the number of observations is large*, *Transactions of the American Mathematical Society* **54** (1943) 426–482.

[118] P. D. Dauncey, M. Kenzie, N. Wardle and G. J. Davies, *Handling uncertainties in background shapes*, *JINST* **10** (2015) P04015, [1408.6865].

[119] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, *J. Phys.* **G43** (2016) 023001, [1510.03865].

[120] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre and J. Rojo, *An Unbiased Hessian Representation for Monte Carlo PDFs*, *Eur. Phys. J.* **C75** (2015) 369, [1505.06736].

[121] CMS Collaboration, *Measurement of the differential cross section for* $\mathrm{t\bar{t}}$ *production in the dilepton final state at* $\sqrt{s} = 13$ TeV, Tech. Rep. CMS-PAS-TOP-16-011, CERN, Geneva, 2016.

[122] CMS Collaboration, *Measurement of the cross section ratio ttbar+bbbar/ttbar+jj using dilepton final states in pp collisions at 13 TeV*, Tech. Rep. CMS-PAS-TOP-16-010, CERN, Geneva, 2016.

[123] CMS Collaboration, *CMS Luminosity Measurements for the 2016 Data Taking Period*, Tech. Rep. CMS-PAS-LUM-17-001, CERN, Geneva, 2017.

[124] CMS Collaboration, *CMS luminosity measurement for the 2017 data-taking period at* $\sqrt{s} = 13$ TeV, Tech. Rep. CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[125] CMS Collaboration, *Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector*, Tech. Rep. CMS-PAS-JME-16-004, CERN, Geneva, 2016.

[126] CMS Collaboration, A. M. Sirunyan et al., *Evidence for the Higgs boson decay to a bottom quark–antiquark pair*, *Phys. Lett.* **B780** (2018) 501–532, [1709.07497].

[127] CMS Collaboration, A. M. Sirunyan et al., *Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair*, *Phys. Rev. Lett.* **120** (2018) 071802, [1709.05543].

[128] CMS Collaboration, A. M. Sirunyan et al., *Observation of the Higgs boson decay to a pair of* τ *leptons with the CMS detector*, *Phys. Lett.* **B779** (2018) 283–316, [1708.00373].

[129] ATLAS Collaboration, G. Aad et al., *Combined measurements of Higgs boson production and decay using up to* 80 fb^{-1} *of proton-proton collision data at* $\sqrt{s} = 13$ *TeV collected with the ATLAS experiment*, 1909.02845.

[130] CMS Collaboration, *Sensitivity projections for Higgs boson properties measurements at the HL-LHC*, tech. rep., 2018.

[131] FCC Collaboration, A. Abada et al., *FCC Physics Opportunities*, *Eur. Phys. J.* **C79** (2019) 474.