

NNT : 2017SACLS181

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE AU COMMISSARIAT À L'ÉNERGIE ATOMIQUE ET
AUX ÉNERGIES ALTERNATIVES

École doctorale n°576

Particules, Hadrons, Énergie, Noyau, Instrumentation, Imagerie, Cosmos et Simulation (PHENIICS)
Spécialité de doctorat: Physique des particules

par

INNA KUCHER

**Search for the Higgs boson decaying to two photons and produced in
association with a pair of top quarks in the CMS experiment**

Thèse présentée et soutenue à l'Orme de Merisiers le 17 juillet 2017
devant le jury composé de :

José Ocariz	LPNHE, Paris	Président du jury
Arnaud Lucotte	LPSC, Grenoble	Rapporteur
Fabrice Hubaut	CPPM, Marseille	Rapporteur
Chiara Mariotti	INFN, Torino	Examineur
Lydia Roos	LPNHE, Paris	Examineur
Adam Falkowski	LPT, Orsay	Examineur
Gautier Hamel de Monchenault	CEA-Saclay	Directeur de thèse
Julie Malclès	CEA-Saclay	Co-encadrante de thèse

Contents

Résumé	ix
Introduction	xii
1 The Standard Model of particle physics	1
1.1 Particles and their interactions	3
1.1.1 Fermions	3
1.1.2 Gauge bosons	3
1.1.3 Composite particles	4
1.2 Mathematical formalism	5
1.2.1 The Yang-Mills sector	5
1.2.2 The Dirac sector	5
1.2.3 Higgs mechanism	7
1.2.4 The Yukawa sector	9
1.3 Shortcomings of the Standard Model	10
2 Higgs boson at Large Hadron Collider	12
2.1 Large Hadron Collider	12
2.1.1 LHC design and detectors	12
2.1.2 LHC operations for proton-proton collisions	16
2.2 The Higgs boson production and decay modes at LHC	18
2.2.1 Production modes	18
2.2.2 Decay modes	22
2.3 The Higgs boson discovery	24
2.4 The $t\bar{t}H, H \rightarrow \gamma\gamma$ channel	26
3 Compact Muon Solenoid experiment at Large Hadron Collider	32
3.1 Compact Muon Solenoid experiment overview	32
3.2 The tracker	34
3.3 The electromagnetic calorimeter	37
3.4 The hadronic calorimeter	40

3.5	The muon system	41
3.6	The trigger system	43
3.7	Event reconstruction	46
4	Electromagnetic calorimeter laser monitoring system in CMS	48
4.1	ECAL signal pulse reconstruction and calibration	49
4.2	ECAL laser monitoring system	53
4.2.1	General description	53
4.2.2	Laser pulse reconstruction	57
4.2.3	Laser monitoring performance	60
4.3	Possible ECAL laser monitoring system upgrade for LHC Phase II	63
4.3.1	Description	64
4.3.2	Installation into the existing laser monitoring system and first tests .	67
4.3.3	"SpyBox" stability and precision study	73
4.4	Conclusion	75
5	$H \rightarrow \gamma\gamma$ analysis at 13 TeV	85
5.1	Samples	86
5.1.1	Data samples	86
5.1.2	Simulation samples	86
5.2	Photon reconstruction	89
5.2.1	Reconstruction of photon conversions	90
5.3	Trigger	93
5.3.1	Level 1 trigger	94
5.3.2	High Level Trigger	94
5.3.3	Trigger performance	96
5.4	Photon energy resolution	97
5.5	Event preselection	100
5.6	Diphoton vertex identification	103
5.7	Photon identification	103
5.8	Event classification	103
5.9	Diphoton multivariate classifier	104
5.9.1	Setup and performance	104
5.9.2	Systematic uncertainties	107
5.9.3	Event categorization with diphoton MVA output	109
5.10	VBF selection	110
5.10.1	Analysis strategy	110
5.10.2	Jet definition	111
5.10.3	Kinematic dijet MVA	112
5.10.4	Combined dijet MVA and categorization	114

5.10.5	Systematic uncertainties	115
5.11	VH selection	116
5.11.1	Objects definition	116
5.11.2	Leptonic channel	119
5.11.2.1	Selections	119
5.11.2.2	Systematic uncertainties	120
5.11.3	Missing transverse energy channel	121
5.11.3.1	Selection	121
5.11.3.2	Systematic uncertainties	121
5.11.4	Hadronic channel	121
5.11.4.1	Selection	122
5.11.4.2	Systematic uncertainties	123
5.12	$t\bar{t}H$ selection	123
5.13	Statistical analysis	123
5.13.1	Signal model	123
5.13.2	Background model	125
5.13.3	Results of event selection	127
5.13.3.1	Event yields	127
5.13.3.2	Event distributions	128
5.13.4	Systematic uncertainties	134
6	Diphoton vertex identification in the $H \rightarrow \gamma\gamma$ analysis	140
6.1	Vertex identification algorithm	142
6.1.1	Principle	142
6.1.2	Investigation for possible improvements	149
6.1.3	Performance in $H \rightarrow \gamma\gamma$ MC simulated events	158
6.1.4	Validation with $Z \rightarrow \mu\mu$ events for unconverted photons	160
6.1.5	Validation with $\gamma + \text{jet}$ events for converted photons	167
6.2	Per-event probability of correct diphoton vertex choice	171
6.2.1	Principle	171
6.2.2	Performance in $H \rightarrow \gamma\gamma$ MC simulated events	171
6.2.3	Validation with $Z \rightarrow \mu\mu$ events for unconverted photons	171
6.2.4	Validation with $\gamma + \text{jet}$ events for converted photons	175
6.3	Corrections and systematic uncertainties	179
6.3.1	Corrections to simulation	179
6.3.2	Systematic uncertainties	181
6.4	Conclusion	183

7	Photon identification in CMS	184
7.1	Photon identification in CMS	185
7.1.1	Discriminating variables	185
7.1.2	Method	188
7.1.3	Results	189
7.1.4	Corrections for simulation	197
7.2	Photon identification in the $H \rightarrow \gamma\gamma$ analysis	201
7.2.1	Method	202
7.2.2	Results	202
7.2.3	Data-simulation comparison	210
7.3	Conclusion	213
8	The $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis	216
8.1	Strategy for $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis	217
8.2	$t\bar{t}H, H \rightarrow \gamma\gamma$ analysis for ICHEP'16	217
8.2.1	Object definition	217
8.2.2	Leptonic channel	219
8.2.3	Hadronic channel	220
8.2.4	Diphoton MVA cut optimization	221
8.2.5	Results	223
8.3	The $t\bar{t}H, H \rightarrow \gamma\gamma$ search new strategy in 2017	223
8.3.1	Object definition	227
8.3.2	Leptonic channel	228
8.3.2.1	The $ m_{e,\gamma} - m_Z $ cut optimization	228
8.3.2.2	Muon isolation	228
8.3.2.3	Final cuts	230
8.3.2.4	Expected improvement	231
8.3.3	Hadronic channel	232
8.3.3.1	Multivariate discriminator $t\bar{t}H$ MVA	232
8.3.3.2	Optimization	238
8.3.3.3	Expected improvement	246
8.4	Systematic uncertainties	246
8.5	Conclusion	248
9	The $H \rightarrow \gamma\gamma$ and $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis final results	249
9.1	Final results with the full 2016 dataset of 35.9 fb^{-1}	249
9.2	Conclusion	258
	Conclusions	258
	Appendix A Statistical methodology	261

Bibliography	277
Acknowledgements	279

Résumé

Le mécanisme de la brisure spontanée de symétrie électrofaible est un élément essentiel du Modèle Standard. Il explique l'origine de la masse de particules dans le Modèle Standard. Le champ de Higgs provient de ce mécanisme, et son excitation quantique, le boson de Higgs, est une particule scalaire. Le boson de Higgs se couple à d'autres particules avec une force proportionnelle à leurs masses. Sa masse, quant à elle, est un paramètre libre de la théorie. Un candidat au boson de Higgs a été découvert avec une masse de 125 GeV par les expériences ATLAS et CMS dans le Run I du Grand Collisionneur de Hadron au CERN (LHC) à une énergie de centre de masse de 7-8 TeV. Après la découverte, une mesure précise des propriétés du boson de Higgs est devenue l'un des principaux objectifs du LHC.

Bien que la découverte repose sur une combinaison d'études dans cinq canaux de désintégration différents, deux d'entre eux ont fourni la plus grande sensibilité et une mesure de la masse: la désintégration en une paire de photons ($H \rightarrow \gamma\gamma$) et la désintégration en une paire de bosons Z qui se désintègrent à leurs tours en paires d'électrons ou de muons ($H \rightarrow ZZ^* \rightarrow 4l, l = e, \mu$). La nouvelle particule a rapidement montré une parité de spin de 0^+ . Jusqu'à présent, toutes les mesures de propriété subséquentes des deux collaborations sont conformes aux attentes du Modèle Standard, avec un seul boson de Higgs.

Dans cette thèse, la mesure des propriétés du boson de Higgs dans le canal de désintégration en deux photons est présentée. Ce canal de désintégration bénéficie de la reconstruction complète de l'état final avec une excellente résolution avec le détecteur CMS. Dans ce travail, l'accent est mis sur le mode de production $t\bar{t}H$ afin de mesurer le couplage de Yukawa du quark top auquel on peut accéder directement en utilisant le taux de production $t\bar{t}H$. Plusieurs nouveaux scénarios de physique prédisent l'existence de partenaires lourds du quark top, qui se décomposent en un quark top et un boson de Higgs. L'observation d'une déviation significative dans le taux de production $t\bar{t}H$ par rapport à la prédiction du Modèle Standard serait une indication indirecte de la nouvelle physique. La mesure du mode de production $t\bar{t}H$ tel que prévu par le Modèle Standard est difficile car il s'agit d'un processus extrêmement rare. Seuls 1% de tous les bosons de Higgs produits au LHC sont produits en association avec une paire de quarks tops. En conséquence, pendant le LHC Run I, il n'y avait aucune observation du processus $t\bar{t}H$.

En 2015, le LHC a redémarré avec une énergie dans le centre de masse de 13 TeV. En raison de l'augmentation de l'énergie, les sections efficaces de production du boson de Higgs augmentent, ce qui donne la possibilité de mesurer les propriétés du boson de Higgs avec une meilleure précision. Les données recueillies en 2015 et en 2016 à 13 TeV par l'expérience CMS sont utilisées dans l'analyse $H \rightarrow \gamma\gamma$.

Ce travail de thèse a débuté en septembre 2014, dans le contexte du *Long Shutdown 1*. Cela m'a permis de participer à la base même de la future analyse $H \rightarrow \gamma\gamma$, c'est-à-dire à l'algorithme d'identification des vertex. Il est essentiel d'identifier le vertex primaire correct pour avoir une bonne résolution de masse de diphoton. En même temps, ceci n'est pas trivial puisque les photons non convertis ne laissent pas de traces dans le trajectographe, et le calorimètre électromagnétique de CMS (ECAL) n'est pas segmenté longitudinalement. Dans les données de 2016, l'efficacité d'identification de vertex est de 81%.

Une autre contribution aux efforts de création d'un nouveau cadre d'analyse $H \rightarrow \gamma\gamma$ était de reproduire les sélections LHC Run I du mode de production $t\bar{t}H$. Le calorimètre électromagnétique joue le rôle central de l'analyse $H \rightarrow \gamma\gamma$. Au cours de la Phase II du LHC, l'ECAL continuera à jouer un rôle crucial dans de nombreuses analyses physiques, y compris le $H \rightarrow \gamma\gamma$. La principale préoccupation du système de surveillance actuel est sa tolérance au rayonnement, qui est le plus important pour le système de distribution de lumière et les diodes PN. Une mise à niveau possible pour le système de surveillance laser est présentée avec une étude de sa précision de mesure de transparence en utilisant les données 2015 et 2016. Il semble que l'effet de vieillissement des fibres se produise en raison des rayonnements. Il doit être modélisé et corrigé. Si après la correction, la précision de la mesure de transparence est supérieure à 1 %, le système de surveillance laser actuel doit être mis à niveau à l'aide de diodes PN avec une résistance au rayonnement plus élevée.

En 2016, mon travail a porté sur l'algorithme d'identification des photons pour l'analyse $H \rightarrow \gamma\gamma$, qui est l'une des étapes essentielles, et aussi pour le groupe d'objets physiques CMS EGamma. L'efficacité de l'algorithme général d'identification des photons CMS, réalisée avec le groupe EGamma, a été améliorée de 30% - 10% dans la partie centrale et de 15% - 10% dans les extrémités par rapport à sa version précédente. En même temps, l'analyse $t\bar{t}H$ a encore été améliorée en optimisant les coupures de sélection et en introduisant une approche multivariée au lieu d'un ensemble de coupure afin d'augmenter la sensibilité. Dans le canal hadronique la contamination de ggH a été réduite de 7% et la sensibilité augmentée de 20%, alors que dans le canal leptonique, la sensibilité a augmenté de 7%. L'intensité du signal dans le canal $H \rightarrow \gamma\gamma$ a été mesurée à $\mu = 1.16_{-0.14}^{+0.15} = 1.16_{-0.10}^{+0.11}(\text{stat.})_{-0.08}^{+0.09}(\text{syst.})_{-0.05}^{+0.06}(\text{theo.})$. Les meilleures valeurs d'ajustement pour les modificateurs de l'intensité du signal associés aux mécanismes de production ggH et $t\bar{t}H$ et aux processus de production VBF et VH ont également été mesurées, $\mu_{ggH}, \mu_{t\bar{t}H} = 1.19_{-0.18}^{+0.20}$ et $\mu_{VH}, \mu_{VBF} = 1.01_{-0.51}^{+0.57}$. Lorsque le processus $t\bar{t}H$ est considéré séparément, le meilleur ajustement est $\mu_{t\bar{t}H} = 2.22_{-0.8}^{+0.9}$, correspondant à un excès de 3.3σ par rapport à l'hypothèse de bruit de fond et à un excès de 1.6σ par rapport à la prédiction du Modèle Standard.

$t\bar{t}H$.

Les perspectives comprennent des recherches supplémentaires sur les propriétés du boson de Higgs en utilisant des données de 13 TeV, en particulier sur sa mesure de masse en utilisant une reconstruction de données améliorée. Cette mesure aidera à éclairer la (méta) stabilité du vide électrofaible [63, 64]. D'autres mesures de $\mu_{t\bar{t}H}$, accumulant davantage de données, révéleront sa compatibilité avec le Modèle Standard. La production $t\bar{t}H$ ouvre également un nouveau champ de tests du Modèle Standard. Par exemple, la violation de CP peut être sondée via le couplage supérieur de Yukawa. La séparation des composants scalaires et pseudoscalaires peut se faire en utilisant une combinaison de mesures de $t\bar{t}H$, tH , et $\bar{t}H$ modes de production [198].

Introduction

The mechanism of the electroweak symmetry breaking is an essential element of the Standard Model (SM). It explains the origin of SM particles mass. The Higgs field originates from this mechanism, and its quantum excitation, Higgs boson, is a scalar particle. The Higgs boson couples to other particles with a strength proportional to their masses. Its mass is instead a free parameter of the theory. A Higgs boson candidate was discovered with a mass of 125 GeV by the ATLAS and CMS experiments in the Run I of the CERN Large Hadron Collider (LHC) at a centre-of-mass energy of 7-8 TeV. After the discovery, accurate measurement of the Higgs boson's properties has become one of the main goals of the LHC.

While the discovery relied on a combination of studies in five different decay channels, two of them provided the highest sensitivity and a measurement of the mass: the decay to a pair of photons ($H \rightarrow \gamma\gamma$) and the decay to a pair of Z bosons that both decay into pairs of electrons or muons ($H \rightarrow ZZ^* \rightarrow 4l, l = e, \mu$). The new particle was soon shown to have spin-parity 0^+ . So far, all subsequent property measurements from both collaborations are consistent with expectations from the SM, with a single Higgs boson.

The measurements of the properties of the Higgs boson in the $H \rightarrow \gamma\gamma$ decay channel are presented in this thesis. This decay channel benefits from the complete reconstruction of the final state with excellent resolution of the CMS detector. In this work, the emphasis is put on the $t\bar{t}H$ production mode in order to measure the top-Yukawa coupling, which can be accessed directly using the rate of the $t\bar{t}H$ production. Several new physics scenarios predict the existence of heavy top-quark partners, that would decay into a top quark and a Higgs boson. The observation of a significant deviation in the $t\bar{t}H$ production rate with respect to the SM prediction would be an indirect indication of new physics. Measuring the $t\bar{t}H$ production mode as predicted from the SM is challenging as it is an extremely rare process. Only 1% of all Higgs bosons produced at LHC, is produced in association with a pair of top quarks. As a result, in the LHC Run I, there was no observation of the $t\bar{t}H$ process.

In 2015 the LHC restarted with a centre-of-mass energy of 13 TeV. Due to the energy increase, the Higgs boson production cross sections increase giving the possibility to measure Higgs boson's properties with a better precision. The data collected in 2015 and in

2016 at 13 TeV by the CMS experiment are used in the $H \rightarrow \gamma\gamma$ analysis.

The present thesis work started in September 2014, in the context of the Long Shutdown 1. This allowed me to participate into the very base of the future $H \rightarrow \gamma\gamma$ analysis, i.e. the vertex identification algorithm. It is crucial to identify the correct primary vertex to have a good diphoton mass resolution. At the same time, it is not trivial since non-converted photons do not leave tracks in the tracker and the CMS electromagnetic calorimeter (ECAL) is not segmented longitudinally. Another contribution to the common efforts of creating new $H \rightarrow \gamma\gamma$ analysis framework was to reproduce the LHC Run I selections of the $t\bar{t}H$ production mode. As a service task for the CMS collaboration, the stability and precision studies of the possible ECAL laser monitoring system upgrade for the LHC Phase II, were performed using the 2015 and 2016 laser data. In 2016, my work was focused on the photon identification algorithm for the $H \rightarrow \gamma\gamma$ analysis, which is one of the essential steps, and also for the CMS EGamma physics object group. At the same time, the $t\bar{t}H$ analysis was further improved by optimizing the selection cuts and introducing a multivariate approach instead of the cut-based one in order to increase the sensitivity.

This thesis is structured as follows. A theory-oriented introduction to $H \rightarrow \gamma\gamma$ studies is given in Chapter 1, while an experiment-oriented introduction to the LHC and phenomenology of the Higgs boson at LHC are shown in Chapter 2. The CMS detector is introduced in Chapter 3. Chapter 4 focuses on the ECAL laser monitoring system, and its possible upgrade for the LHC Phase II. In Chapter 5, I present $H \rightarrow \gamma\gamma$ analysis, omitting personal contributions. My work on the vertex identification is reported in Chapter 6, and on photon identification in Chapter 7. Chapter 8 shows my contribution to the $t\bar{t}H$ analysis as a part of inclusive $H \rightarrow \gamma\gamma$ one. The most recent results, based on the full 2016 dataset of 35.9 fb^{-1} , are shown in Chapter 9 where all benefits of this thesis work are collected.

At the moment of writing, an improved reconstruction of the 2016 CMS data is ongoing. This new reconstruction will be used for the $H \rightarrow \gamma\gamma$ publication, thus the results presented here after will be further improved.

Chapter 1

The Standard Model of particle physics

"Excuse me... how can you discover a particle so small that nobody has ever seen one?"

J.J. Thomson, 1897

Introduction

Current understanding of particle physics relies on discoveries made in the last century [1]. It started with Wilhelm Röntgen, who made an experiment with cathode-ray tube, that revealed the X-radiation [2]. In 1897 J. J. Thomson discovered the electron [3, 4], by constructing glass tube and applying a high electrical voltage at both ends. Exploiting uranium, two types of radioactivity were detected by Ernest Rutherford in 1899: alpha and beta rays [5]. They differed from X-rays by their penetrating power. In 1900 Paul Villard discovered γ -rays, while studying radiation emitted by radium [6]. Ernest Rutherford made the discovery of an atom nucleus in 1911 [7], demonstrating the nuclear nature of atoms by deflecting alpha particles passing through a thin gold foil. Eight years after he became the first person to deliberately transmute one element into another. Using pure nitrogen, Ernest Rutherford exploited alpha radiation to convert nitrogen into oxygen, doing so he discovered the proton [8]. Several experiments with "beryllium radiation" [9] triggered neutron discovery by James Chadwick in 1932 [10, 11].

Quantum mechanics was born in 1900 from Max Planck's solution to the black-body radiation problem [12]. Albert Einstein's work [13] in 1905 offered a quantum-based theory to explain the photoelectric effect. Also he was the inventor of the theories of special [14, 15] and general [16] relativity. In 1925 Werner Heisenberg established a basis for quantum mechanics [17]. In 1926 Max Born, Pascual Jordan and Werner Heisenberg developed Heisenberg's pioneering theory into the first complete formulation of quantum

mechanics [18]. In 1926 Erwin Schrödinger established wave mechanics and shown its first applications [19, 20]. He also proved that quantum mechanics of Heisenberg, Born, and Jordan and wave mechanics he introduced are equivalent [21]. In 1927 Dirac's equation took into account the newly discovered spin of the electron [22] and using a new scheme of four-dimensional matrices succeeded in expressing relativity into suitable form. In 1931 Dirac predicted the "mirror image" of the electron, positron [23]. At the same time at the California institute Carl Anderson studied cloud chamber photographs of cosmic rays. He noticed curved tracks appearing like electrons, but curved in an opposite way. In 1932 he suggested that those tracks were the new kind of electron [24]. Muon and π -meson were also discovered in exploiting cosmic rays [25–27]. Wolfgang Pauli proposed the existence of neutrinos as neutral weakly interacting particles, in order to explain the continuous spectrum of electrons in nuclear beta decay. It was incorporated in the theory of beta decay applied by Enrico Fermi in 1934 [28]. However, it was more than two decades later that a more complete theory of weak interactions was formulated with the inclusion of the V-A theory, that explained the parity violating nature of the weak interaction, by George Sudarshan and Robert Marshak along with Richard Feynman and Murray Gell-Mann [29, 30]. Chen-Ning Yang and Robert Mills introduced local gauge invariance in quantum field theory in 1954 [31].

The myriad of discoveries led to an organisation of particles into octets and subsequently decuplets by Murray Gell-Mann and Yuval Ne'eman and subsequently the proposal of the quark model by Murray Gell-Mann and George Zweig [32, 33] in 1964. According to this model the newly discovered particles were classified as hadrons and quarks were proposed to be the elementary particles that were the building blocks of the hadrons. Experimental evidence for the presence of quarks was found at the Stanford Linear Accelerator Center (SLAC) in 1969 [34, 35]. The unification of the theories of weak and electromagnetic interactions by Sheldon Glashow, Adam Salam and Steven Weinberg into the theory of electroweak interaction [36–38] was a major breakthrough in the understanding of these fundamental interactions. Their electroweak theory postulated the existence of W (necessary to explain beta decay) and Z bosons. To introduce the mass of W and Z bosons electroweak symmetry breaking mechanism was developed by Anderson, Brout, Englert, Guralnik, Hagen, Higgs and Kibble [39–41] in 1964. The Higgs boson was the outcome of this mechanism. Quantum Chromodynamics (QCD) was established in the 1970s by David Gross, Frank Wilczek, and David Politzer [42, 43]. Gerard 't Hooft and Martinus Veltman introduced universal regularization and renormalization method for gauge field theories in 1972 [44, 45].

Discoveries continued with tau lepton detection [46, 47] in 1974-1976. Gluon was experimentally found by TASSO [48] and Pluto [49] collaborations in 1979. W and Z bosons were discovered at CERN by UA1 and UA2 collaborations [50–53] in 1983. The evidence of a top quark was reported in 1995 by CDF and $D\bar{0}$ collaborations at the Fermi National Accelerator Laboratory [54, 55].

Most of particles described by the theory were detected, the only missing piece was the Higgs boson. The discovery of the Higgs boson at CERN is briefly described in section 2.3.

1.1 Particles and their interactions

The standard model (SM) of particle physics is a paradigm of quantum field theory with special relativity that describes the elementary particles discovered in nature and the interactions between them, that is, the electromagnetic, weak and strong interactions. The gravitational interaction is not included in the SM. Its strength is much smaller compared to the other fundamental interactions at particles scale.

1.1.1 Fermions

Fermions are matter particles with spin $1/2$. They are classified into two categories: leptons and quarks, according to the way they interact. All fermions couple via weak forces, all charged fermions also couple via electromagnetic force, and only quarks also interact via strong force. They are arranged in three generations as shown in Table 1.1. Each generation is made of two leptons of electric charge -1 (top left column) and 0 (top right column), and two quarks of electrical charge $+2/3$ (left bottom column) and $-1/3$ (right bottom column). The electric charges mentioned are in units of the charge of an electron. Each fermion particle has an associated antiparticle with opposite quantum numbers. Each fermion has particular weak hypercharge (Y), weak isospin (I_3), and only quarks carry colour charge (c) that is associated with strong interactions.

Name	Mass	Name	Mass
Leptons			
electron (e)	0.511 MeV	electron neutrino (ν_e)	< 2 eV
muon (μ)	105.7 MeV	muon neutrino (ν_μ)	< 2 eV
tau (τ)	1777 MeV	tau neutrino (ν_τ)	< 2 eV
Quarks			
up (u)	$2.2^{+0.6}_{-0.4}$ MeV	down (d)	$4.7^{+0.5}_{-0.4}$ MeV
charm (c)	1.27 ± 0.03 GeV	strange (s)	96^{+8}_{-4} GeV
top (t)	$173.21 \pm 0.51 \pm 0.71$ GeV	bottom (b)	$4.18^{+0.04}_{-0.03}$ GeV

Table 1.1: Leptons and quarks generations. Masses are taken from [56].

1.1.2 Gauge bosons

Three types of interactions are accounted for in the SM and each type is carried by a particular mediator:

- Electromagnetic interaction is responsible for cohesion of atoms, and it is carried by the photon (γ). The photon is massless and neutral. The electromagnetic interaction has an infinite action range.
- Weak interaction is responsible for nuclear decays. Its mediators are massive¹ W, Z bosons with an action range of $\sim 10^{-18}$ m.
- Strong interaction is responsible for the cohesion of atoms' nucleus and hadrons, and it is carried by 8 massless gluons. They carry a color charge and have an action range of $\sim 10^{-15}$ m.

1.1.3 Composite particles

Hadrons are particles assembled with quarks. In nature the physical state particles are colorless, so that hadrons are composed in a way to have no color. Four types of hadrons can be distinguished :

- mesons, which valence quarks are quark-antiquark pair (or pairs linear superposition). For example π^0 meson is a combination of $(\frac{u\bar{u}-d\bar{d}}{\sqrt{2}})$.
- baryons, which valence quarks are 3 quarks or antiquarks. For example proton is composed from uud quarks, and neutron from udd quarks.
- tetraquarks, which valence quarks are two quarks and two antiquarks. LHCb collaboration at CERN confirmed the existence of the Z(4430) state [57].
- pentaquarks, which valence quarks are four quarks and one antiquark bound together. LHCb collaboration reported results consistent with pentaquark states in the decay of bottom Lambda baryons (Λ_b^0) [58].

Most of hadrons have a small life time, e.g. $\sim 10^{-17}$ s for π^0 and $\sim 10^{-8}$ s for $\pi^\pm$, and can be assumed as decaying immediately from the experimental point of view. The exception is a proton, which lifetime is more than 2.1×10^{29} years. It is not possible to observe single, isolated and coloured quark. When two quarks move away and reach a separation distance of 10^{-15} m (diameter of a hadron) their strong interaction is so great that quark-antiquark pairs are produced. These pairs join together in different combinations to create mesons and baryons (hadronization), that are recognized in the detectors as hadronic showers.

¹ $m_W = 80.363 \pm 0.020$ GeV and $m_Z = 91.1876 \pm 0.0021$ GeV [56]

1.2 Mathematical formalism

Particle physics is described by quantum field theory. Each particle corresponds to a field $\psi(x, t)$ which depends on position and time. The fields of interaction are described by gauge bosons. The three fundamental interactions correspond to invariances under the symmetry groups $SU(2)_L \times U(1)_Y$ for both the weak and the electromagnetic forces, and $SU(3)_c$ for the strong force. The notation L means left-handed fermion, Y stands for hypercharge and c for color charge. In order to describe matter particles and their interactions the Lagrangian formalism is used, which includes fields and their partial derivatives. It permits to know at the same time the dynamics of the system and the different interactions. The SM Lagrangian consists of four terms:

$$\mathcal{L} = \mathcal{L}_{Yang-Mills} + \mathcal{L}_{Dirac} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa} \quad (1.1)$$

where $\mathcal{L}_{Yang-Mills}$ represents kinematics of the gauge fields, $\mathcal{L}_{Dirac}$ describes fermions and their interaction with gauge bosons, $\mathcal{L}_{Higgs}$ introduces the Higgs field that allows to describe the electroweak spontaneous symmetry breaking, $\mathcal{L}_{Yukawa}$ represents the interaction between the Higgs and fermion fields.

1.2.1 The Yang-Mills sector

$$\mathcal{L}_{Yang-Mills} = -\frac{1}{4g_1^2} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4g_2^2} W_{\mu\nu}^\alpha W^{\alpha\mu\nu} - \frac{1}{4g_3^2} G_{\mu\nu}^A G^{A\mu\nu} \quad (1.2)$$

where

- g_1, g_2, g_3 are the couplings associated to hypercharge, weak isospin and colour respectively.
- Hypercharge tensor $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ when B_μ is the vector boson field of $U(1)_Y$
- Isospin tensor $W_{\mu\nu}^\alpha = \partial_\mu W_\nu^\alpha - \partial_\nu W_\mu^\alpha - \epsilon^{abc} W_\mu^b W_\nu^c$ when $W_\mu^\alpha, (\alpha = 1, 2, 3)$ are the vector bosons of $SU(2)$ and ϵ^{abc} represents asymmetric structure of $SU(2)_L$.
- Color tensor $G_{\mu\nu}^A = \partial_\mu A_\nu^A - \partial_\nu A_\mu^A - f^{ABC} A_\mu^B A_\nu^C$, where A_μ^A are the gluon fields ($A = 1, 2, \dots, 8$ and f^{ABC} are the asymmetric structure constants of $SU(3)_c$)

1.2.2 The Dirac sector

Relativistic fermions satisfying the Pauli principle are described by spinors in quantum field theory. In particular, the relativistic spin 1/2 fermion is described by a four component spinor ψ :

$$\psi = \begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix} \quad (1.3)$$

where ψ_R and ψ_L are the left and right chirality components. In the SM, the left and right chirality fermions have different interactions. This is related to the parity violation in the weak interactions and it is at the heart of the construction of the SM. As a result, among the fermions, only left-handed fields are doublets and the group is therefore noted $SU(2)_L$. In the classical version of the SM, neutrinos are considered as being massless and left-handed. The $U(1)$ symmetry group acts only on fields with non-zero hypercharge Y , and is therefore noted $U(1)_Y$.

In the gauge group representations leptons and quarks can be written as $(SU(3)_c, SU(2)_L)_{U(1)_Y}$ components, index i means generation and $y_1, \dots, y_5$ is hypercharge:

$$l_i = \begin{pmatrix} \nu_i \\ e_i \end{pmatrix}_L : (\mathbf{1}, \mathbf{2})_{y_1}, \quad e_{iR} = (\mathbf{1}, \mathbf{1})_{y_2}$$

$$q_i = \begin{pmatrix} u_i \\ d_i \end{pmatrix}_L : (\mathbf{3}, \mathbf{2})_{y_3}, \quad u_{iR} = (\mathbf{3}, \mathbf{1})_{y_4}, \quad d_{iR} = (\mathbf{3}, \mathbf{1})_{y_5}$$

The term in Dirac free Lagrangian:

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

is not invariant under local gauge transformations $\psi(x) \rightarrow e^{i\alpha(x)}\psi(x)$, where α is a scalar. The term $\partial^\mu \psi$ is not invariant under $U(1)$ transformations, but invariance is achieved by adding a compensating field B_μ . The covariant derivative is written as:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu - ig_1 B_\mu$$

the B_μ field couples with the parameter g_1 , the electrical charge, and it can be identified with the photon field.

The W_μ and A_μ gauge fields are expressed in terms of matrices:

$$\tilde{W}_\mu = \frac{1}{2} W_\mu^\alpha \tau_\alpha$$

$$\tilde{A}_\mu = \frac{1}{2} A_\mu^A \lambda_A$$

where τ_α are Pauli SU(2) matrices and λ_A are Gell-Mann SU(3) matrices. The fermion couplings to gauge fields are expressed as covariant derivatives:

$$\begin{aligned} D_\mu l_i &= (\partial_\mu + ig_2 \tilde{W}_\mu + ig_1 \frac{y_1}{2} B_\mu) l_i \\ D_\mu e_{iR} &= (\partial_\mu + ig_1 \frac{y_2}{2} B_\mu) e_{iR} \\ D_\mu q_i &= (\partial_\mu + ig_3 \tilde{A}_\mu + ig_2 \tilde{W}_\mu + ig_1 \frac{y_3}{2} B_\mu) q_i \\ D_\mu u_{iR} &= (\partial_\mu - ig_3 \tilde{A}_\mu^* + ig_1 \frac{y_4}{2} B_\mu) u_{iR} \\ D_\mu d_{iR} &= (\partial_\mu - ig_3 \tilde{A}_\mu^* + ig_1 \frac{y_5}{2} B_\mu) d_{iR} \end{aligned}$$

The Lagrangian of the Dirac sector can be written as:

$$\mathcal{L}_{Dirac} = \sum_{i=1}^3 l_i^\dagger i \gamma^\mu D_\mu l_i + e_{iR}^\dagger i \gamma^\mu D_\mu e_{iR} + q_i^\dagger i \gamma^\mu D_\mu q_i + u_{iR}^\dagger i \gamma^\mu D_\mu u_{iR} + d_{iR}^\dagger i \gamma^\mu D_\mu d_{iR}$$

So far, the W and Z bosons are massless to preserve the gauge invariance of SU(2)_L × U(1)_Y.

1.2.3 Higgs mechanism

Within the mathematical framework of the SM, all particles have to be massless, as introducing a mass term to the Lagrangians would violate local gauge symmetry. However, the observed masses of the W and Z bosons, seems to be in conflict with this requirement. Based on the work of Anderson, Brout, Englert, Guralnik, Hagen, Higgs and Kibble, the Higgs mechanism [39–41] was developed to explain the masses of W and Z bosons via the spontaneous breaking of electroweak symmetry.

The mechanism introduces a complex scalar field (Higgs field) ϕ . It is chosen such that it only affects the SU(2)_L group symmetry from the electroweak theory, as the photons should remain massless. The effective potential of the Higgs field has a local extremum at $\phi = 0$, but has an infinite number of global minima at $|\phi| > 0$ that represent the vacuum. At high energies the gauge bosons are located at $\phi = 0$ and the local gauge symmetry of the SM is conserved. At lower energies the symmetry is spontaneously broken with a distinct ground state, as illustrated in Fig. 1.1.

For the introduced field, four degrees of freedom are postulated. According to the Goldstone theorem [60, 61] for every spontaneously broken continuous symmetry there is

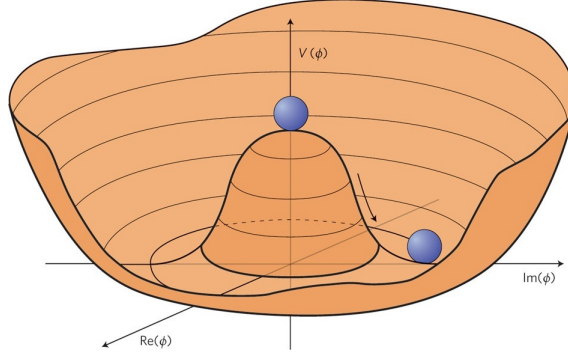


Figure 1.1: Sketch of the effective potential of the Higgs field taken from [59]. At high energies particles are located at $\phi = 0$ and do not interact with the Higgs field. The cylindrical symmetry of the system is conserved. At lower energies this symmetry is spontaneously broken, as the state of particle chooses one distinct minimum of the potential.

a massless Goldstone boson. The Higgs mechanism explains how three Goldstone bosons are absorbed by the W and Z bosons, giving them masses. The fourth degree of freedom predicted the existence of a massive spin-zero particle. The Higgs field also generates masses for fermions through their Yukawa couplings (see next section). The Lagrangian which generates masses for gauge bosons looks like:

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

where

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2$$

ϕ transforms as $\phi \rightarrow e^{i\alpha_i \tau_i} \phi$ under $SU(2)_L$ transformations, where τ_i are the Pauli matrices, and as $\phi \rightarrow e^{iY\beta} \phi$ for $U(1)_Y$ transformations. The simplest field that can be introduced is a $SU(2)_L$ doublet of complex scalar fields:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \quad \phi^\dagger \phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2)$$

With $-\mu^2 > 0$, spontaneous symmetry breaking appears with a minimum (vev - vacuum expectation value) given by:

$$\phi^\dagger \phi = \frac{v^2}{2}, \quad v^2 = \frac{\mu^2}{\lambda}$$

A more general way to express the ϕ field is:

$$\phi = e^{i\theta_i(x)\tau_i} \begin{pmatrix} 0 \\ \frac{v+H(x)}{\sqrt{2}} \end{pmatrix}$$

The $\theta_1, \theta_2, \theta_3$ and $H(x)$ are functions defining $\phi(x)$ in a given point of space-time x . To define the W and Z bosons masses it is enough to use the ϕ term for the vacuum expectation value:

$$\phi = \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix}$$

One can put this expression into the Lagrangian $\mathcal{L}_{Higgs}$ and obtain :

$$\begin{aligned} m_W &= g_2 \frac{v}{2} \\ m_Z &= \sqrt{g_2^2 + g_1^2} \frac{v}{2} \\ m_H &= \sqrt{2\lambda} v \\ m_\gamma &= 0 \end{aligned}$$

The Higgs field does not interact with electromagnetic field, and consequently the Higgs boson is neutral. The theory does not tell how many scalar doublets there should be. It is possible to obtain the same mechanism with several scalar doublets. For example, with two doublets one gets a pair of neutral bosons (scalar and pseudo-scalar) and charged H^+ , H^- bosons in addition to the Higgs boson H . In this thesis I focus on the case of only one doublet in SM.

1.2.4 The Yukawa sector

The electromagnetic gauge invariance allows the coupling between fermions and scalars. One can write the Lagrangian for the Yukawa interaction with the Higgs field:

$$\mathcal{L}_{Yukawa} = h_{ij}^u \bar{q}_L^i u_R^j \tilde{\phi} + h_{ij}^d \bar{q}_L^i d_R^j \phi + h_{ij}^e \bar{l}_L^i e_R^j \phi$$

where $\tilde{\phi}$ is charge-conjugate Higgs field, $h_{ij}^{u,d,e}$ (or λ_f) are Yukawa couplings and $i, j = 1, 2, 3$ are flavour indices. Developing the new Lagrangian around its minimum results in the appearance of mass:

$$\mathcal{L}_{mass} = m_{ij}^u \bar{u}_L^i u_R^j + m_{ij}^d \bar{d}_L^i d_R^j + m_{ij}^l \bar{e}_L^i e_R^j + c.c$$

where fermions interact with the Higgs field with a coupling λ_f proportional to their mass and to the average value of the vacuum expectation value. However, the coupling terms must be added "by hand":

$$m_f = \frac{\lambda_f v}{\sqrt{2}}, \quad v = (\sqrt{2}G_F)^{-\frac{1}{2}}, \quad v = 246 \text{ GeV}$$

The Yukawa couplings are not independent, as redefinitions of fields are possible using global symmetries of $\mathcal{L}_{Yang-Mills} + \mathcal{L}_{Dirac}$.

Thus, besides providing the needed fermion mass terms, this procedure implies that the Higgs boson field interacts with fermions, with couplings proportional to their masses. It should however be mentioned that these masses are not predicted by the theory. It should also be mentioned that the overall symmetry of the theory gets reduced since the three generations of matter no longer appear as identical.

Finally, the introduction of the Higgs boson also has the virtue of ensuring the calculability of the SM. In particular, the perturbative unitarity of the scattering matrix is preserved at high energies, as the longitudinal W/Z boson scattering amplitude no longer grows as the centre-of-mass energy increases.

1.3 Shortcomings of the Standard Model

The Standard Model is presently considered as an effective theory, working only at electro-weak energy scale. It has been very successful in describing many experimental results with high precision. However it has several numbers of issues. Physics "Beyond Standard Model" (BSM) seems to be necessary to solve them. BSM refers to any possible extension of the SM, typically expressed either in terms of coupling constants for new interactions, new charges or other quantum numbers, and parameters describing possible new degrees of freedom or new symmetries. The following list provides some types of shortcomings of the SM:

- The hierarchy problem, which is essentially the huge discrepancy between the electroweak ($\sim 10^2$ GeV) and Planck scales ($\sim 10^{19}$ GeV) [62];

-
- The measured values of the top quark mass and the Higgs boson mass indicate that the electroweak vacuum is probably unstable [63, 64];
 - The SM provides a mechanism of symmetry breaking of the SM interactions under charge and parity transformations (CP symmetry) within the CKM matrix. However, the CP violating phase (in the CKM matrix) is not large enough to explain baryon-antibaryon asymmetry [65];
 - Neutrinos are massless in the SM. The experiments with solar, atmospheric, reactor and accelerator neutrinos have provided compelling evidence for the existence of neutrino oscillations [66, 67], transitions in flight between the different flavour neutrinos ν_e, ν_μ, ν_τ (antineutrinos $\bar{\nu}_e, \bar{\nu}_\mu, \bar{\nu}_\tau$), caused by non-zero neutrino masses and neutrino mixing.
 - Cosmological inflation requires physics BSM, as the effective Higgs potential would become negative at high scales;
 - Gravitational effects cannot be included in the SM [68];
 - There is no explanation for the Dark Matter [69–72] and Dark Energy [73, 74] in the SM.

The Higgs boson is a possible portal to new physics. The Higgs boson could be a composite particle [75]. A phenomenological framework to characterize the experimental constraints of this possibility could be provided by some effective Lagrangian, as for example shown in [76, 77]. In general, the effective field theory approach to BSM is discussed in [78]. In this thesis work, possible deviations from the SM expectations are studied by measuring the production rate of the Higgs boson produced with a pair of top-antitop quarks and decaying into a pair of photons. A large deviation would indicate a presence of a new physics, e.g. some theories predict a heavy top-quark partner decaying into top quarks and Higgs boson [79–81].

Chapter 2

Higgs boson at Large Hadron Collider

2.1 Large Hadron Collider

The Large Hadron Collider (LHC) [82, 83] is a two-rings superconducting hadron accelerator and collider. It was designed to collide proton beams with a nominal centre-of-mass energy of $\sqrt{s} = 14$ TeV (i.e. 7 TeV per beam) and an instantaneous luminosity of 10^{34} cm⁻²s⁻¹, and heavy ion beams (lead nuclei) with a nominal energy of 2.76 TeV per nucleon and a peak luminosity of 10^{27} cm⁻²s⁻¹, making it the most powerful machine ever built for particle physics research. Fig. 2.1 gives an overview of the event rates of known SM processes in proton-proton collisions at the LHC, as compared to the Tevatron¹ [84, 85]. The magnitude of the nominal 14 TeV energy was not only chosen to probe the scalar sector over the whole Higgs boson mass range allowed by the unitarity constraint, but also to probe new physics in general, even in the hypothesis of the absence of any Higgs boson, a case where theoretical constraints would have required new physics to appear at a scale $\Lambda < 1.2$ TeV.

2.1.1 LHC design and detectors

The LHC is installed near Geneva in a 26.7 km-long circular tunnel passing through Switzerland and France, which was constructed between 1984 and 1989 for an earlier CERN collider, the Large Electron Positron (LEP). The LHC project was given a first approval by the CERN Council in 1994, and updated two years later to its definitive form of a 14 TeV machine. The LEP was closed in 2000 and the tunnel was cleared for the LHC construction. As a particle-particle collider, the LHC uses two separate parallel rings with clockwise and

¹Tevatron was circular proton-antiproton accelerator and collider with maximum energy of 980 GeV per-beam. It was located at the Fermi National Accelerator Laboratory functioning from 1987 to 2011.

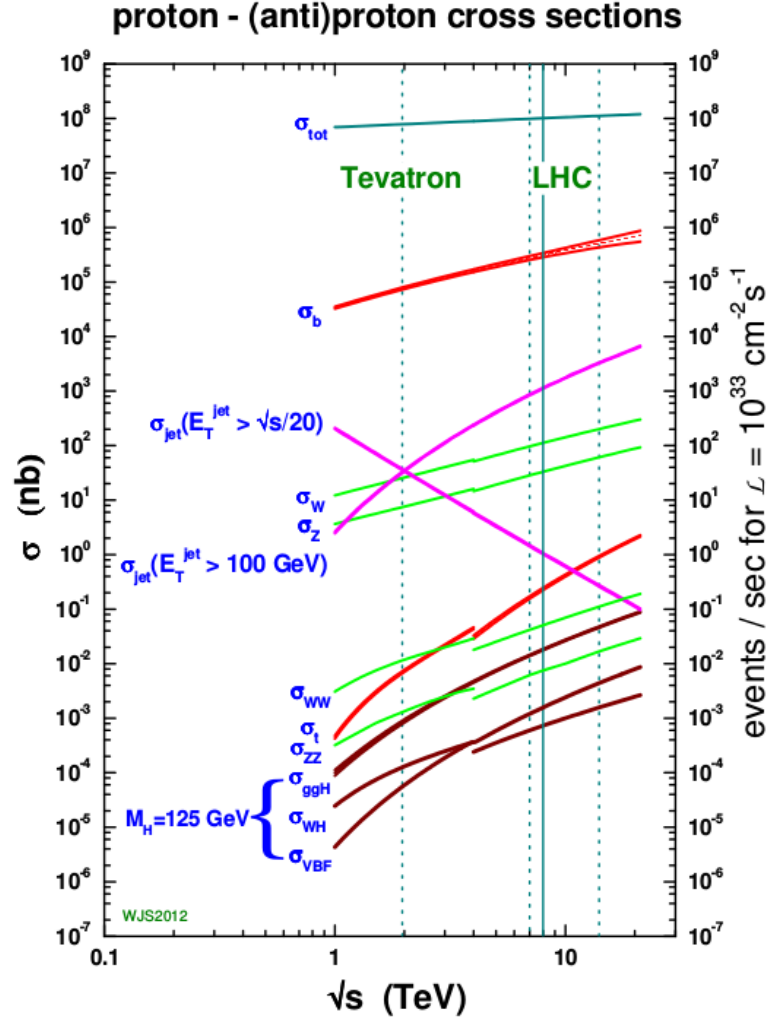


Figure 2.1: Expected production cross sections and event rates for signal and background processes at hadron colliders, as a function of the centre-of-mass energy [86]. Discontinuities are due to the Tevatron being a proton-antiproton collider while the LHC is a proton-proton collider.

anticlockwise circulating beams. Owing to limited space in the 3.7-metre-large arc section of the LEP tunnel, these two rings are included in one single twin-bore magnet system. They intersect at four interaction points. The beams are bent by 1232 15-metre-long dipole magnets made of copper-clad niobium-titanium. They are kept focused by 392 quadrupole magnets, each 5-7 metres long, while some stronger quadrupole electromagnets squeeze them further close to the intersection points to maximize the probability of interaction.

Superfluid helium-4 is used to cool the magnets and maintain them at their operating temperature of 1.9 K (-271.25°C). Prior to being injected into the LHC, proton beams are prepared by a chain of pre-accelerators that increase the energy in steps. These systems are visible in the diagram of the CERN accelerator complex shown in Fig. 2.2.

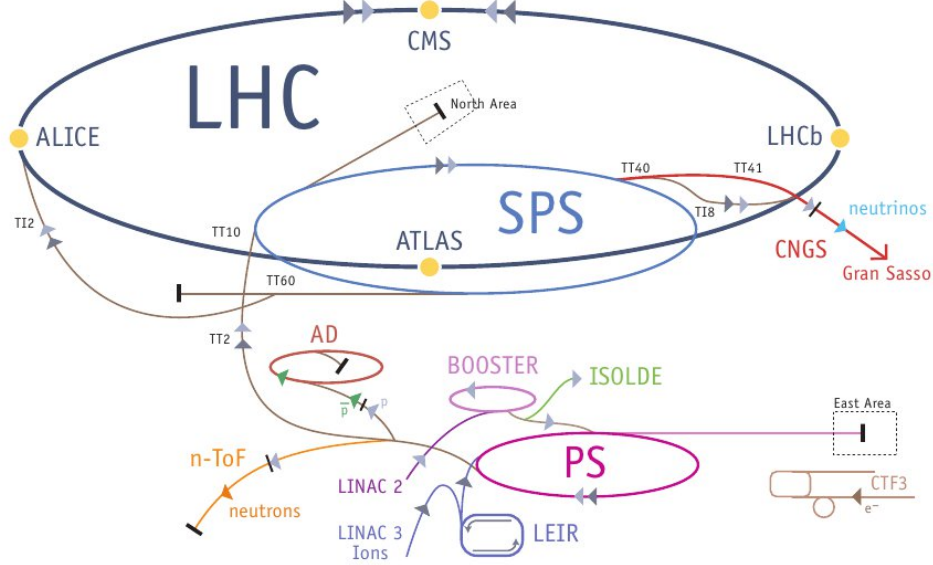


Figure 2.2: The CERN accelerator systems [87]. The proton injection chain for the LHC starts from the LINAC2 and proceeds through the Booster, PS, and SPS.

Protons are first accelerated to an energy of 50 MeV in the Linear Accelerator (LINAC2), which feeds the Proton Synchrotron Booster (PSB), where they are accelerated to 1.4 GeV. They then reach 26 GeV in the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS) further increases their energy to 450 GeV before they are being injected into the LHC ring at Point 2 and Point 8. In the LHC, the acceleration process and shaping of proton bunches is done by 16 radio-frequency cavities, where the electromagnetic field oscillates at 400 MHz. The machine instantaneous luminosity only depends on the beam parameters, and can be written as:

$$L = \frac{f_{rev} N_b^2 n_b \gamma_r}{4\pi \epsilon_n \beta^*} F \quad (2.1)$$

where $f_{rev} = 11$ kHz is the bunch crossing frequency, N_b is the number of particles per bunch, n_b is the number of bunches per beam, ϵ_n the normalized transverse beam emittance, β^* is the beta function at the collision point, which measures the beam focalization and is corrected by the relativistic gamma factor γ_r , and F is a geometric luminosity reduction

factor that accounts for the crossing angle at the interaction point [82]. The nominal value of the beta function is $\beta^* = 0.55$ m, and the nominal luminosity is reached with $n_b = 2808$ bunches per beam, and $N_b = 1.15 \times 10^{11}$ protons per bunch. This choice corresponds to a spacing of 25 ns (or 7.5 m) between bunches, i.e. a bunch crossing rate of 40 MHz.

The record-breaking collision parameters of the LHC machine have important consequences on the design of detectors. Under nominal conditions, the number of inelastic collision events is of the order of 10^9 per second, with approximately 20 collisions per bunch crossing. The occurrence of such overlapping proton-proton interactions is called pileup, and calls for a high granularity to distinguish particles from different interactions. Moreover, fast response and good time resolution are needed in order to distinguish events from consecutive crossings, i.e. avoiding the phenomenon of overlap of consecutive signals called out-of-time pileup. Four main particle detectors are installed in underground caverns at the four beam intersection points, they are shown in Fig. 2.3.

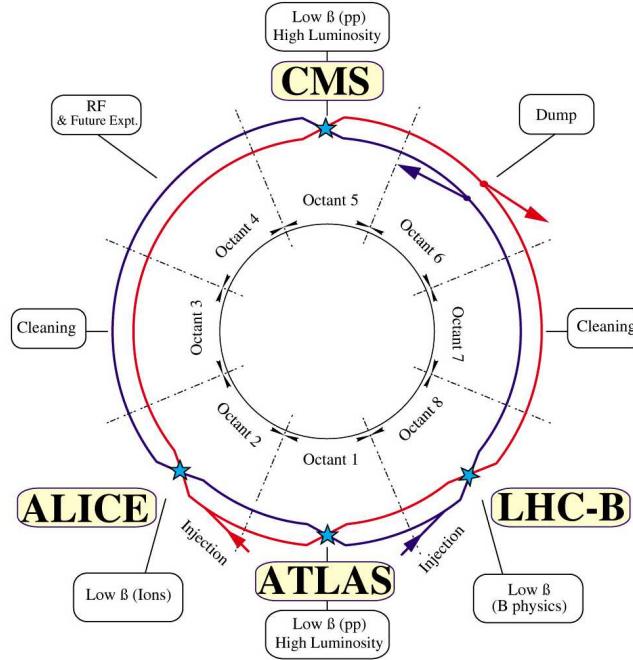


Figure 2.3: Particles detectors at LHC [88].

The two largest ones, A Toroidal LHC ApparatuS (ATLAS) and Compact Muon Solenoid (CMS) are located at the symmetrically opposite Point 1 and Point 5 of the LHC ring, respectively, where the provided instantaneous luminosities are expected to be the highest. ATLAS and CMS are multi-purpose experiments, designed to cover a wide physics program in the scalar, electroweak, and strong sectors, with optimized sensitivity for Higgs boson searches and possible physics beyond the standard model (BSM) at the

TeV scale. The two other detectors have more specific goals: Large Hadron Collider beauty (LHCb, located at Point 8) is aimed at studying CP violation in B-hadron interactions using lower peak luminosities ($\sim 10^{32} \text{ cm}^{-2}\text{s}^{-1}$), while A Large Ion Collider Experiment (ALICE, Point 2) is dedicated to heavy-ion collisions, studying quark-gluon plasma.

2.1.2 LHC operations for proton-proton collisions

The first injections of proton beams into the LHC were carried out on 10th September 2008, but the initial testing was interrupted on the 19th of that month by a major magnet quench incident due to a faulty electrical connection between two magnets, which caused extensive damage to over 50 superconducting magnets and delayed operations for more than one year. The injections started over in late November 2009, and were shortly followed by the first high-energy collisions. Beam energies were progressively increased, reaching 3.5 TeV in March 2010, when the first physics runs started. It was decided to not immediately aim at the design LHC beam parameters yet, and to only operate the collider with a 50 ns bunch spacing (1404 bunches) and intermediate centre-of-mass energies until 2012, hoping to discover the long-sought Higgs boson in this first data taking era referred to as Run I. Design energies would have to wait for Run II in 2015, after a two-year upgrade period called Long Shutdown 1 (LS1). Following a first data sample of 47 pb^{-1} in 2010 at $\sqrt{s} = 7 \text{ TeV}$, the LHC delivered a high-luminosity data set of about 6 fb^{-1} during the 2011 runs, a large fraction of which was collected by ATLAS and CMS, allowing the exclusion of most of the allowed range for the Higgs boson mass (shown in section 2.3). Fig. 2.4 shows the cumulative integrated luminosities delivered to the CMS experiment in every year of data taking.

In early 2012, it was decided to increase the centre-of-mass energy to 4 TeV per beam for the year 2012. By the end of that year, the LHC had delivered a data sample corresponding to an integrated luminosity of 23 fb^{-1} at $\sqrt{s} = 8 \text{ TeV}$, whereby the Higgs boson discovery was already accomplished using the first $\sim 5 \text{ fb}^{-1}$. The exploitation of the large collected data samples went on during LS1, providing a quantity of precision measurements in the electroweak and strong sectors. LS1 brought considerable upgrade and consolidation to the LHC, with the goal of starting Run II in 2015 at a centre-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$, and reaching beyond-design instantaneous luminosities of $1.4 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$. After several months of training of the dipole magnets, the first Run II beam was injected on 5th April 2015, and the first 6.5 TeV beam was obtained on 10th April. The year 2015 was intended as a commissioning year and started with 50 ns collisions, before moving to the nominal bunch spacing of 25 ns over the summer. In total, a data sample corresponding to more than 4 fb^{-1} was delivered in 2015 to ATLAS and CMS, as illustrated on the left in Fig 2.5 for the case of CMS. 2016 started as a production year, still at $\sqrt{s} = 13 \text{ TeV}$ and with a bunch spacing of 25 ns, but the experience gained in 2015 led to a choice of relatively bold set of operational parameters. Even though some problems with the SPS beam dump

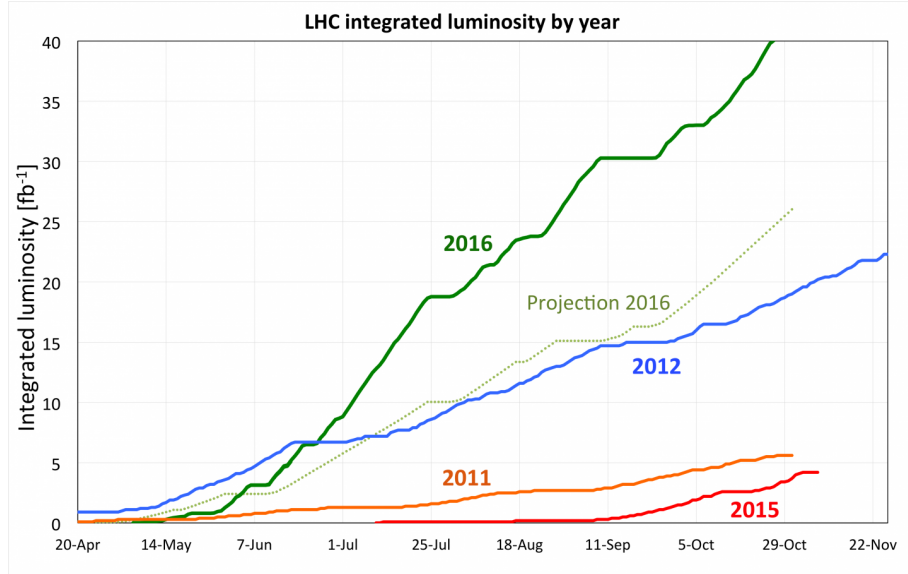


Figure 2.4: Cumulative integrated luminosity versus day delivered to CMS during stable beams for proton-proton collisions, for the 2011 (orange), 2012 (blue), 2015 (red), and 2016 (green) data taking [89].

made it necessary to restrain the number of bunches to about 2100, the nominal LHC luminosity of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$ was reached in June 2016, notably thanks to a reduction of the β^* parameter from 80 cm to 40 cm. Instantaneous luminosities further grew over the summer, reaching more than $1.4 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$. The 2016 proton-proton data taking ended in late October, with a total delivered integrated luminosity of about 41 fb^{-1} , as shown on the right in Fig. 2.5.

The LHC Run II is scheduled till 2018. It will be followed by the Long Shutdown 2 (LS2), during which the machines will be upgraded to deliver designed centre-of-mass energy ($\sqrt{s} = 14 \text{ TeV}$) and instantaneous luminosity of $2 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$. The LHC Run III (called also LHC Phase I) will last from 2018 till 2021.

The LHC Phase II is scheduled to start in 2022 with the Long Shutdown 3 (LS3). In this second phase of the LHC physics program, the accelerator will provide to CMS and ATLAS an additional integrated luminosity of about 3000 fb^{-1} over 10 years of operation, starting from 2025. This will substantially enlarge the mass reach in the search for new particles and will also greatly extend the potential to study the properties of the Higgs boson discovered at the LHC in 2012. In order to meet the experimental challenges of unprecedented proton-proton luminosity, the CMS collaboration will need to address the ageing of the present detector and to improve the ability of the apparatus to isolate and precisely measure the products of the most interesting collisions.

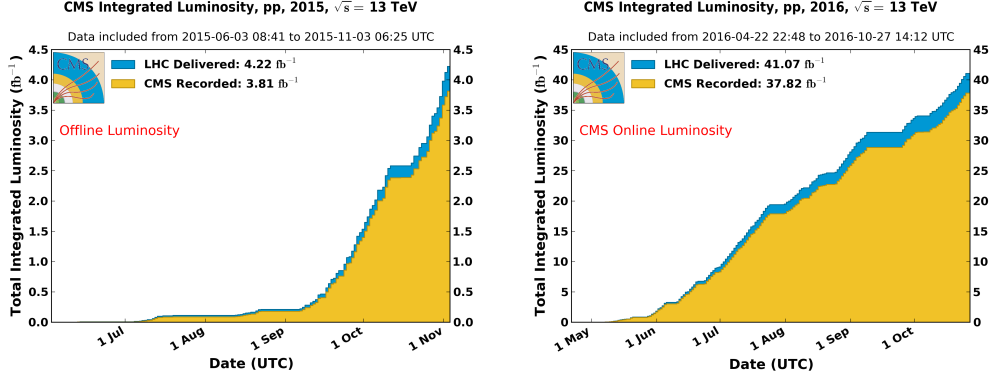


Figure 2.5: Cumulative offline luminosity versus day delivered to CMS (blue), and recorded by CMS (yellow) during stable beams for pp collisions at a centre-of-mass energy of $\sqrt{s} = 13$ TeV, in 2015 (left) and 2016 (right). The delivered luminosity accounts for the luminosity delivered from the start of stable beams until the LHC requests CMS to turn off the sensitive detectors to allow a beam dump or beam studies. The luminosity is given as determined from counting rates measured by the luminosity detectors after offline validation.

2.2 The Higgs boson production and decay modes at LHC

At the LHC, the SM Higgs boson is produced in a variety of processes [90], four of them are studied in this thesis. As shown in the Feynman diagrams of Fig. 2.6, their final states can involve other particles than just the decay products of the Higgs boson, hence leading to different experimental signatures that help extracting these processes.

2.2.1 Production modes

Gluon fusion. The gluon fusion process (noted as ggH) is induced by a pair of gluons that fuse into a Higgs boson via an intermediate quark loop, as illustrated in Fig. 2.6 (on the top left). As the most massive quark, the top quark gives the largest contribution to the loop. Since the gluon luminosity is very large in proton-proton collisions at the high centre-of-mass energies provided by the LHC, ggH is the leading Higgs boson production mechanism and dominates all the other ones by more than one order of magnitude. Higher-order QCD corrections have been found to increase the LO ggH cross section by as much as a factor 2-2.5, which makes it crucial to use state-of-the-art computations.

Vector boson fusion. Vector boson fusion (VBF) is the second most important production mode at the LHC, its cross section being an order of magnitude smaller than that of ggH. At leading order, it is a process where two vector bosons are radiated off quarks and merge into a Higgs boson (Fig. 2.6, on the top right). This results in a very clean experimental signature featuring two forward and backward energetic jets with a large di-

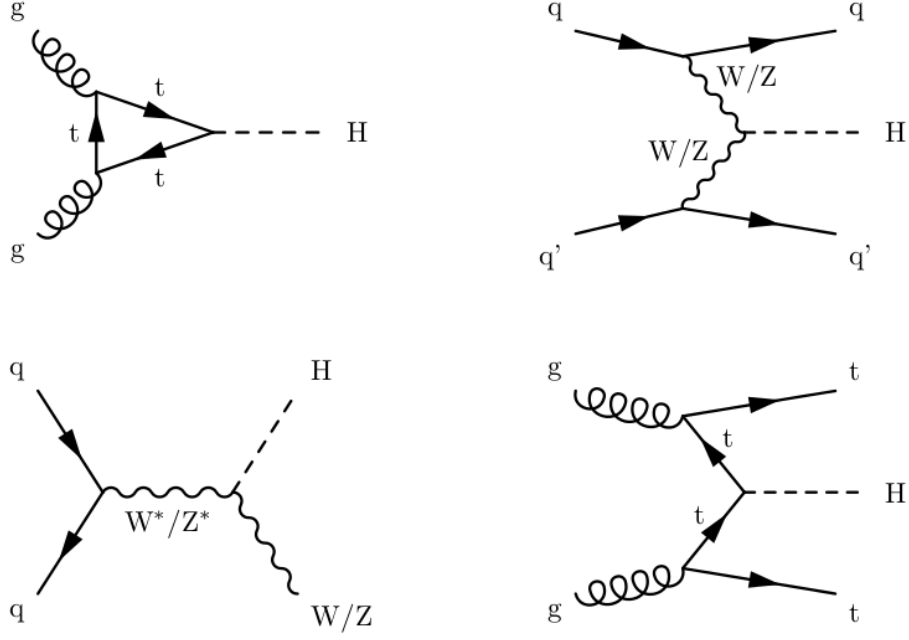


Figure 2.6: Leading order Feynman diagrams for the main production modes of the SM Higgs boson at the LHC: gluon fusion (top left), vector boson fusion (top right), VH associated production (bottom left) and $t\bar{t}H$ associated production (bottom right).

jet invariant mass, while the Higgs boson decay system is boosted and ends up in a more central region of the detector. Moreover, since no colour is exchanged, central hadronic activity is also suppressed. This characteristic topology helps rejecting backgrounds from SM processes and ggH production in association with two jets. W boson fusion and Z boson fusion cannot be distinguished experimentally. The VBF cross section is known to a good accuracy, with higher-order QCD corrections being smaller than for ggH .

Associated production with a vector boson. Associated production with a vector boson, which is often referred to as VH associated production or Higgsstrahlung, and further split into WH and ZH production, is the third most prominent Higgs boson production mechanism at the LHC (about twice less frequent than VBF). In this qq -induced process, the Higgs boson is radiated off a W or Z boson (Fig. 2.6, on the bottom left), leading to an experimental signature where the Higgs boson decay products are boosted, and accompanied by the decay products of the associated W or Z. When this vector boson decays hadronically, a pair of nearby boosted jets with invariant mass close to the nominal m_W or m_Z can be sought for, whereas leptonic decays either provide one lepton and missing transverse energy (for WH), or either a pair of leptons or missing transverse energy (for

ZH). The high-order QCD corrections are quite large for this process. As for VBF, the cross sections are currently computed to NNLO QCD and NLO EW accuracy [83].

Associated production with a top quark pair. Associated production with a top quark pair ($t\bar{t}H$) is about 100 times rarer than ggH , but it is the first production mode that allows direct probing of the Yukawa coupling between the top quark and the Higgs boson, which is the key to SM precision tests. Indeed, although the Higgs-top vertex is already involved in gluon fusion through the top quark loop, one cannot exclude that an unknown heavy fundamental particle contributes to the loop. In this mainly gluon-induced $t\bar{t}H$ process, the Higgs boson is accompanied by a $t\bar{t}$ pair in the final state, as illustrated in Fig. 2.6, on the bottom right. Each associated top quark then decays into a bottom quark and a W boson which can in turn decay leptonically or hadronically, leading to several possible experimental signatures and extraction strategies. This production mode is the focus of this thesis and it is discussed in chapter 8.

Other Higgs boson production mechanisms. The Higgs production in association with a single top quark, though subdominant, can bring valuable information, in particular regarding the sign of the top Yukawa coupling. This process has been computed at NLO in a five-flavor scheme [91] and amounts to about 90 fb at $\sqrt{s} = 14$ TeV (with the opposite sign of the top Yukawa coupling, the cross section increases by one order of magnitude).

The Higgs boson production in association with bottom quarks is known at NNLO in the case of five quark flavors [92–94]. The coupling of the Higgs boson to a b quark is suppressed in the SM by the bottom-quark mass over the Higgs vev, m_b/v , implying that associated production of a SM Higgs boson with b quarks is small at the LHC. Still at high energy, large logarithms are present and need to be resummed, leading to an enhancement of the inclusive cross section. At $\sqrt{s} = 14$ TeV the $b\bar{b}H$ cross section can be as large as 600 fb, still two orders of magnitude below the ggH production cross section.

Main Higgs boson production modes cross sections. Fig. 2.7 presents the centre-of-mass-energy dependency of the cross sections of the main Higgs boson production modes at the LHC [95]. Table 2.1 numerically shows by how much the cross sections of the four processes studied in this thesis increase from 8 TeV to 13 TeV, i.e. from Run I to Run II of the LHC. Most of them gain a factor of about 2, except the $t\bar{t}H$ which benefits from a factor 4 because of the large mass of the involved objects.

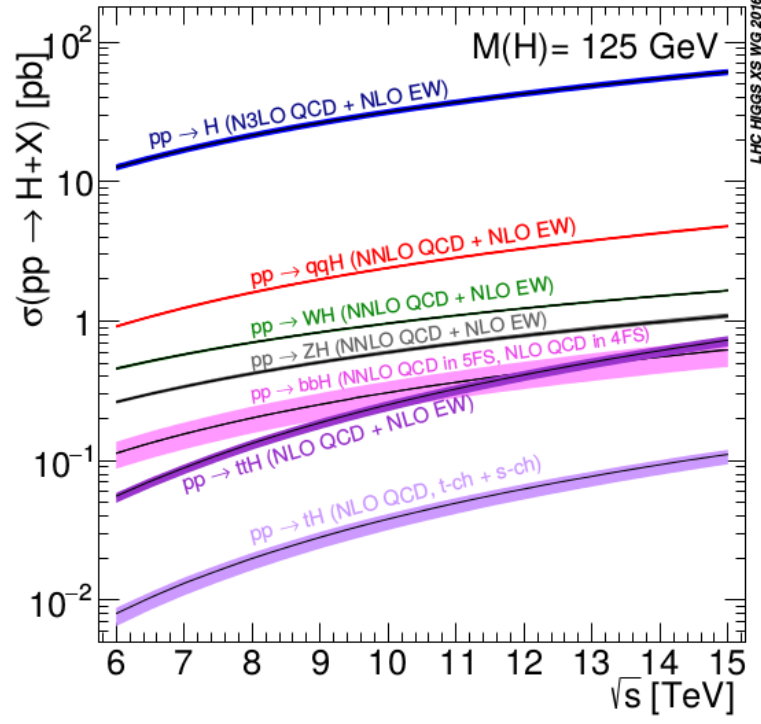


Figure 2.7: Total production cross section for a 125 GeV SM Higgs boson at the LHC as a function of the centre-of-mass energy [95].

Process	Computation order	$\sigma_{8 \text{ TeV, pb}}$	$\sigma_{13 \text{ TeV, pb}}$	$\sigma_{8 \text{ TeV}}/\sigma_{13 \text{ TeV}}$
ggH	N3LO QCD, NLO EW	$19.50^{+10\%}_{-11\%}$	$44.10^{+11\%}_{-11\%}$	2.26
VBF	NNLO QCD, NLO EW	$1.60^{+2\%}_{-2\%}$	$3.78^{+2\%}_{-2\%}$	2.36
WH	NNLO QCD, NLO EW	$0.70^{+3\%}_{-3\%}$	$1.37^{+2\%}_{-2\%}$	1.96
ZH	NNLO QCD, NLO EW	$0.42^{+5\%}_{-5\%}$	$0.88^{+5\%}_{-5\%}$	2.09
$t\bar{t}H$	NLO QCD, NLO EW	$0.13^{+8\%}_{-13\%}$	$0.51^{+9\%}_{-13\%}$	3.92

Table 2.1: Variations of the cross sections of the main Higgs boson production modes at the LHC for $m_H = 125$ GeV, moving from late Run I (8 TeV) to Run II (13 TeV), taken from [56]

2.2.2 Decay modes

Since the SM Higgs boson directly couples to all massive particles of the SM and can also couple to massless particles via intermediate loops, it can decay in a variety of channels. The total decay width of the Higgs boson and the relative branching fractions of its decay channels are fully determined by the value of its mass (m_H). Fig. 2.8 presents the values of the branching fractions as a function of the hypothesized m_H [96], thus illustrating the strategic issues that existed before the discovery. Five main decay channels were studied at that time, and are still being exploited these days. Their respective relevance to the discovery and to the property measurements does not only depend on their branching fraction, but also on the experimental capability of extracting the corresponding signals while rejecting their backgrounds.

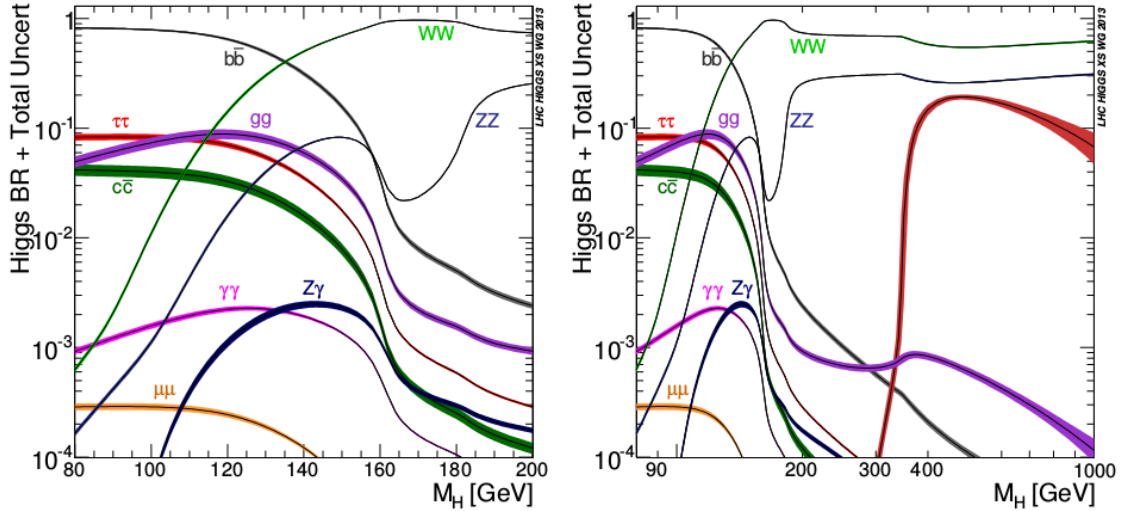


Figure 2.8: Decay branching ratios [95] with uncertainties for the SM Higgs boson as a function of m_H , in a low-mass range (left) and in a range extending to high mass (right).

For values of m_H up to about 135 GeV, the Higgs boson mainly decays into a $b\bar{b}$ pair, but the inclusive signal is overwhelmed by the QCD production of bottom quarks. Therefore, this channel is exploited in boosted regimes, mainly in the VH production mode with the associated W or Z boson decaying leptonically.

As the hypothesized m_H increases, decays to pairs of massive gauge bosons ($H \rightarrow WW^*$ and $H \rightarrow ZZ^*$) open up and their branching ratios grow. The WW channel has the largest branching ratio of all at $m_H = 135$ GeV, and it particularly dominates the ZZ channel around the WW mass threshold. Accurate lepton identification and missing transverse energy reconstruction have made the $WW \rightarrow l\nu l\nu$ decay a sensitive channel for Higgs boson searches at intermediate masses, although the m_H resolution is poor because of the escaping

neutrinos. By contrast, the ZZ channel can offer a complete reconstruction of the final state, with excellent mass resolution. The $H \rightarrow ZZ \rightarrow 4l$ channel, where both Z bosons decay to pairs of either electrons or muons, is one of the two so-called golden channels for the discovery.

The sensitivity at masses around and below 125 GeV is dominated by a channel with low branching fraction, the decay to two photons. Due to the excellent experimental resolution on the diphoton invariant mass, the Higgs boson signal appears as a clear peak on top of backgrounds from QCD production of two photons or jet fragments misidentified as photons. The diphoton decay channel constitutes the other golden channel for the discovery and property measurements, and its characteristics will be discussed in chapter 5.

Three other channels contribute significantly to the decay width at 125 GeV and up to masses near 150 GeV. The decay to a $\tau^+\tau^-$ pair provides intermediate sensitivity and is probed with a variety of experimental strategies, depending on the decays of the τ leptons. The decays to a pair of gluons or to a pair of charm quarks are generally not exploited, because their experimental signatures cannot be distinguished from the overwhelming QCD dijet production. Finally, the decay to a $t\bar{t}$ pair opens up around $2m_t$, and is hence relevant to searches of new high-mass resonances.

The Higgs boson was discovered with at 125 GeV, as described in the next section. The branching ratios with their relative uncertainty for a SM Higgs boson with $m_H = 125$ GeV are summarized in Table 2.2.

Decay channel	Branching ratio	Relative uncertainty
$H \rightarrow \gamma\gamma$	2.27×10^{-3}	+5.0% -4.9%
$H \rightarrow ZZ$	2.62×10^{-2}	+4.3% -4.1%
$H \rightarrow W^+W^-$	2.14×10^{-1}	+4.3% -4.2%
$H \rightarrow \tau^+\tau^-$	6.27×10^{-2}	+5.7% -5.7%
$H \rightarrow b\bar{b}$	5.84×10^{-1}	+3.2% -3.3%
$H \rightarrow Z\gamma$	1.53×10^{-3}	+9.0% -8.9%
$H \rightarrow \mu^+\mu^-$	2.18×10^{-4}	+6.0% -5.9%

Table 2.2: The branching ratios and the relative uncertainty [56, 96] for a SM Higgs boson with $m_H = 125$ GeV.

2.3 The Higgs boson discovery

To guide Higgs boson searches at colliders, the allowed range for the Higgs boson mass m_H has been constrained by theoretical arguments [90, 97], in particular by imposing the energy scale Λ up to which the SM is valid. An upper limit called triviality [98–102] is obtained by requiring that the running quartic coupling λ of the potential V remains finite up to the scale Λ . A lower limit comes from vacuum stability [103–105], requiring that λ remains positive after including radiative corrections, at least up to Λ , which implies that the minimum of the potential is absolute. A looser metastability constraint is found by requiring the minimum to just be local. If Λ were of the order of the Planck scale ($\sim 10^{19}$ GeV), then m_H would be constrained between 130 and 170 GeV. For $\Lambda \sim 1$ TeV, the allowed range goes up to 700 GeV. It should be noted that another long-known upper bound of about 710 GeV arises from the requirement of the unitarity of the scattering matrix.

Direct Higgs boson searches were carried out at the LEP collider exploiting the Higgsstrahlung process $e^+e^- \rightarrow HZ$ (Higgsstrahlung) process, which led to a lower bound of $m_H > 114.4$ GeV at a 95% confidence level (CL) [106]. At the Tevatron proton–antiproton collider, early measurements excluded the mass range 162–166 GeV, exploiting the WW decay channel [107]. Independently from direct searches, precision electroweak measurements using data from different colliders provided an indirect constraint on the SM Higgs boson mass, with an upper limit of 158 GeV at a 95% CL [108], and it is shown in Fig. 2.9.

At the LHC, the first direct Higgs boson searches were based on data from proton–proton collisions collected in 2011 at a centre-of-mass energy of $\sqrt{s} = 7$ TeV, which amounted to an integrated luminosity of 5.1 fb^{-1} . Using the five mentioned decay channels ($H \rightarrow \gamma\gamma$, $H \rightarrow b\bar{b}$, $H \rightarrow \tau\tau$, $H \rightarrow WW$, $H \rightarrow ZZ$), the CMS collaboration excluded a range of masses from 127 to 600 GeV at a 95% CL [111], while the ATLAS collaboration excluded the ranges 111.4–116.6 GeV, 119.4–122.1 GeV, and 129.2–541 GeV at a 95% CL [112]. Within the remaining allowed mass region, both experiments reported an excess of events between 2 and 3 standard deviations near 125 GeV.

In 2012, the proton–proton centre-of-mass energy of the LHC was raised to $\sqrt{s} = 8$ TeV, and an additional data sample of around 5.3 fb^{-1} was collected by the end of June by each of the two experiments, allowing elucidation of the last non-excluded mass region. On 4th July 2012, the ATLAS and CMS collaborations reported the observation of a new boson with mass near 125 GeV, compatible with the SM Higgs boson [113–115]. The largest contributors to the discovery are the $\gamma\gamma$ and ZZ decay modes. They both have very good mass resolution, allowing good localization of the invariant mass of a putative resonance responsible for the excess. Their combined significance reaches 5.0σ in CMS experiment and 6.0σ in ATLAS experiment, as shown in Fig. 2.10.

That same month, the CDF and $D\bar{0}$ experiments published an analysis of the full Tevatron data sample and reported an excess of events of about 3 standard deviations in

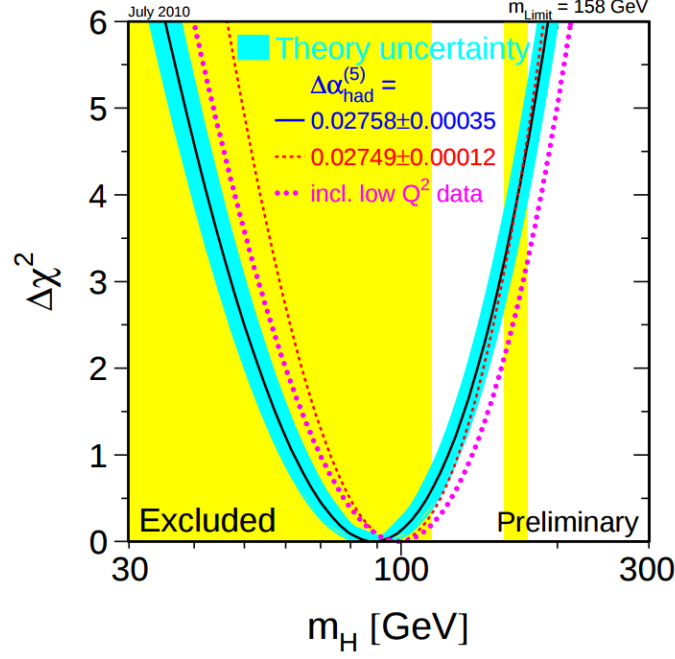


Figure 2.9: Global fit of the SM [109] as a function of the Higgs boson mass. The blue band represents an estimate of the theoretical error due to missing higher order corrections. The vertical yellow band shows the 95% CL exclusion limit on m_H from the direct searches at LEP (up to 114 GeV) and the Tevatron (158 GeV to 175 GeV). The dashed curve is the result obtained using the evaluation of $\Delta\alpha_{had}^{(5)}(m_Z^2)$ from reference [110]. The dotted curve corresponds to a fit including also the low- Q^2 data from [108].

the range 120-135 GeV [116, 117], consistent with the LHC observations. The remainder of the 2012 data taking increased the integrated luminosity of 8 TeV data to approximately 20 fb^{-1} per experiment. The full Run I LHC data sample, both at $\sqrt{s} = 7$ TeV and 8 TeV, was thoroughly analysed in various decay channels and production modes by both collaborations, shedding more light on the newly discovered boson. Improved calibrations allowed for a precise measurement of its mass using the two high-resolution discovery channels [118–121], and ultimately combining data from both experiments [122], leading to a value of $m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})$ GeV. The natural width of a SM Higgs boson with a mass of 125 GeV is about 4 MeV, much smaller than the instrumental mass resolution in the $\gamma\gamma$ and ZZ channels, as shown in Table 2.3. CMS has placed 95% CL bound on the natural width of the observed boson of $\Gamma_H < 2.4$ GeV, using $H \rightarrow \gamma\gamma$ channel [120], and $\Gamma_H < 3.4$ GeV using $H \rightarrow ZZ \rightarrow 4l$ channel [118].

The new boson was soon shown to have spin-parity $J^P = 0^+$ [123], and its production and decay rates and its coupling strengths to SM particles turned out to be consistent with expectations for the SM Higgs boson [121, 124]. Fig. 2.11 illustrates the signal strength of

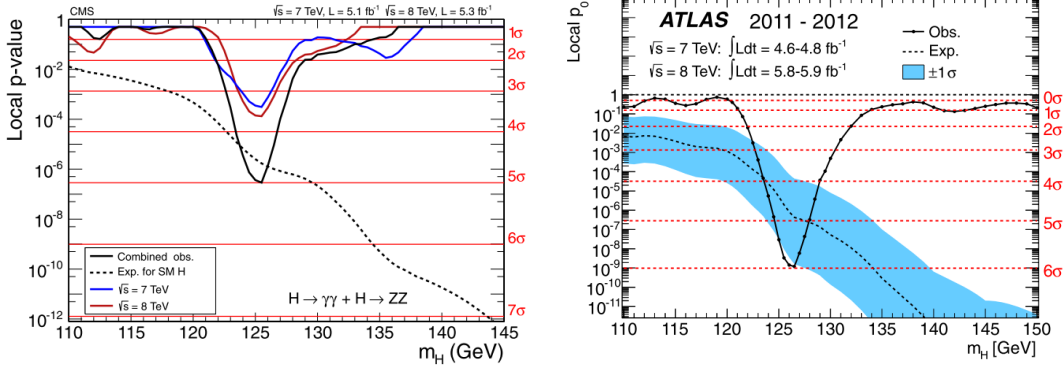


Figure 2.10: Left: The observed local p-value for decay modes with high mass-resolution channels, $\gamma\gamma$ and ZZ , as a function of the SM Higgs boson mass with CMS experiment [113]. The dashed line shows the expected local p-values for a SM Higgs boson with a mass m_H . Right: The observed (solid) local p-value as a function of m_H in the low mass range with ATLAS experiment [114]. The dashed curve shows the expected local p-value under the hypothesis of a SM Higgs boson signal at that mass with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6σ .

Decay	Mass resolution
$H \rightarrow \gamma\gamma$	1-2%
$H \rightarrow ZZ \rightarrow 4l$	1-2%
$H \rightarrow WW \rightarrow 2l2\nu$	20%
$H \rightarrow b\bar{b}$	10%
$H \rightarrow \tau^+\tau^-$	15%

Table 2.3: Mass resolution for five sensitive channels for SM Higgs boson searches at the LHC with $m_H = 125$ GeV, taken from [56].

all Higgs boson production modes for two experiments, CMS and ATLAS, as well as their combination. The combined analyses of ATLAS and CMS data strengthened the fact that the measured properties are consistent across channels and experiments, and with the SM predictions [125].

2.4 The $t\bar{t}H$, $H \rightarrow \gamma\gamma$ channel

The present thesis work shows the inclusive $H \rightarrow \gamma\gamma$ analysis with a specific focus on $t\bar{t}H$ Higgs boson production mode. More details about those channels and the state of the art before the beginning of this thesis are discussed in this section.

The $H \rightarrow \gamma\gamma$ analysis, presented in this thesis, includes the following Higgs boson production modes: ggH , VBF , VH and $t\bar{t}H$. Despite the small branching fraction ($\sim 0.23\%$ at

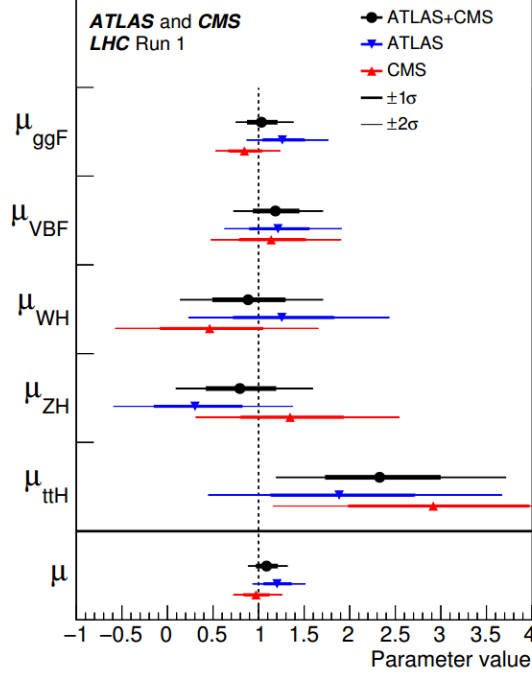


Figure 2.11: Best fit results for the production signal strengths for the combination of ATLAS and CMS data collected during LHC Run I [125]. Also shown are the results from each experiment. The error bars indicate the 1σ (thick lines) and 2σ (thin lines) intervals. The measurements of the global signal strength μ are also shown.

$m_H = 125$ GeV) and the presence of a large diphoton continuum background, the diphoton decay mode provides an expected signal significance for the 125 GeV SM Higgs boson that is one of the highest among all the decay modes. Photon energy is measured using the CMS electromagnetic calorimeter (described in section 3.3) that has an excellent energy resolution. The diphoton invariant mass is computed as:

$$m_{\gamma\gamma} = \sqrt{2E_{\gamma_1}E_{\gamma_2}(1 - \cos\theta_{\gamma\gamma})} \quad (2.2)$$

where E_{γ_1} and E_{γ_2} are photon energies and $\theta_{\gamma\gamma}$ is the angle between photons momenta. A good resolution of the measured photons energies and the angle between their momenta leads to a narrow peak in the invariant mass distribution. One of the Higgs boson discovery plots is shown in Fig. 2.12, where one can see the diphoton invariant mass distribution.

One striking feature of the SM Higgs boson is its strong coupling to the top quark relative to the other SM fermions. Based on the top quark large mass [126] the top-quark Yukawa coupling is expected to be of order one. The coupling dependence on the different particles mass is shown in Fig. 2.13.

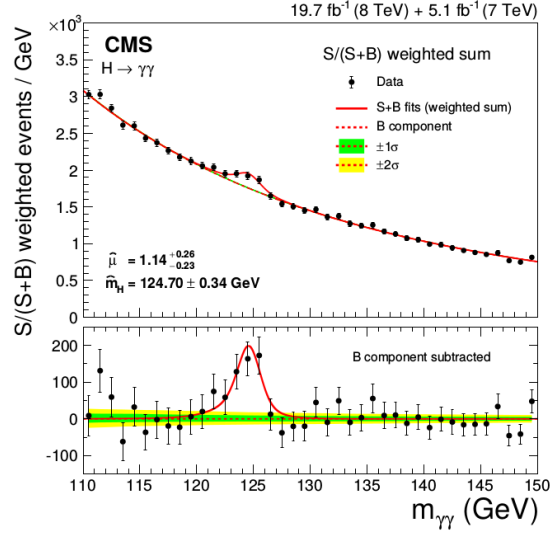


Figure 2.12: The diphoton invariant mass spectrum together with the background subtracted mass spectrum using CMS data collected during the LHC Run I [113].

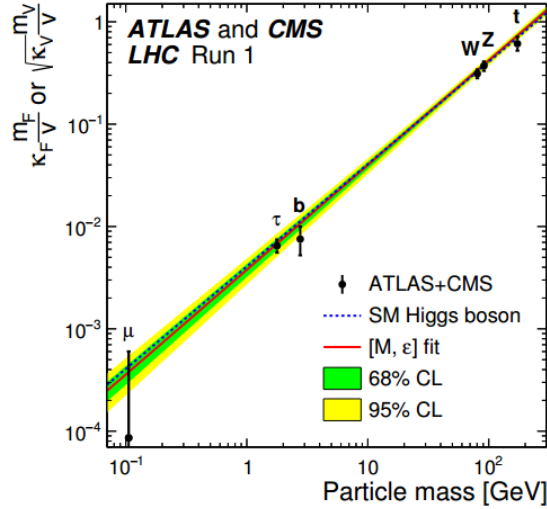


Figure 2.13: Couplings of the Higgs boson to different particles as function of their mass, are computed using data from ATLAS and CMS experiments collected during LHC Run I [125].

Because the top quark is heavier than the Higgs boson, its coupling cannot be assessed by measuring Higgs boson decays to top quarks. However, the Higgs boson coupling to top quarks can be experimentally constrained through measurements involving the gluon fusion production mechanism that proceeds via a fermion loop in which the top quark

provides the dominant contribution (left diagram in Fig. 2.14), assuming there is no physics beyond the Standard Model (BSM) contributing to the loop. Likewise the decay of the Higgs boson to photons involves both a fermion loop diagram dominated by the top-quark contribution (middle diagram in Fig. 2.14), as well as a W boson loop contribution. Current measurements of Higgs boson production via gluon fusion are consistent with the SM expectations for the top-quark Yukawa coupling within the uncertainties [115, 123, 127, 128]. The combined ATLAS and CMS indirect measurements of the top-Yukawa coupling resulted in $\kappa_t = 0.87 \pm 0.15$, assuming the absence of BSM particles in the loops [125].

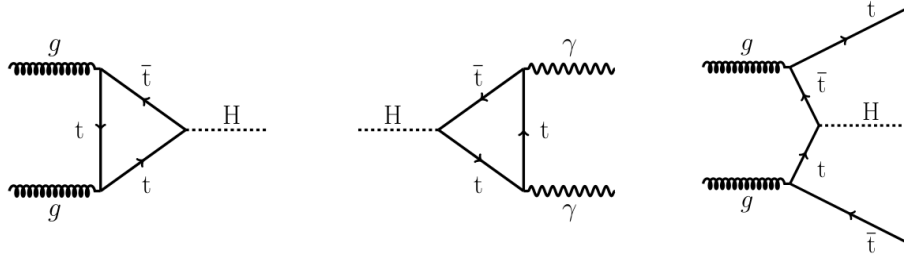


Figure 2.14: Feynman diagrams showing the gluon fusion production of a Higgs boson through a top-quark loop (left), the decay of a Higgs boson to a pair of photons through a top-quark loop (middle), and the production of a Higgs boson in association with a top-quark pair (right). These diagrams are representative of SM processes with sensitivity to the coupling between the top quark and the Higgs boson.

Probing the top-quark Yukawa coupling directly requires a process that results in both a Higgs boson and top quarks explicitly reconstructed via their final-state decay products. The production of the Higgs boson in association with a top-quark pair ($t\bar{t}H$) satisfies this requirement (right diagram in Fig. 2.14). A measurement of the rate of $t\bar{t}H$ production provides a direct test of the coupling between the top quark and the Higgs boson. Furthermore, several new physics scenarios [79–81] predict the existence of heavy top-quark partners, that would decay into a top quark and a Higgs boson. An observation of a significant deviation in the $t\bar{t}H$ production rate with respect to the SM prediction would be an indirect indication of unknown phenomena.

The top quark decays into a W boson and a b quark with $\sim 100\%$ probability, W boson can decay into lepton and neutrino or into a pair of quarks. According to W boson decay, two categories are defined in $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis. If at least one W boson decays leptonically, the event goes into leptonic category, otherwise into hadronic one. The diphoton mass distributions corresponding to two $t\bar{t}H$ categories and produced using CMS data collected during the LHC Run I are shown in Fig. 2.15.

The signal strength relative to the SM prediction ($\mu = \sigma/\sigma_{SM}$) was obtained with full LHC Run I dataset shown in Fig. 2.16 (left), for four main Higgs boson production modes.

The statistics collected during the LHC Run I did not allow an observation of Higgs

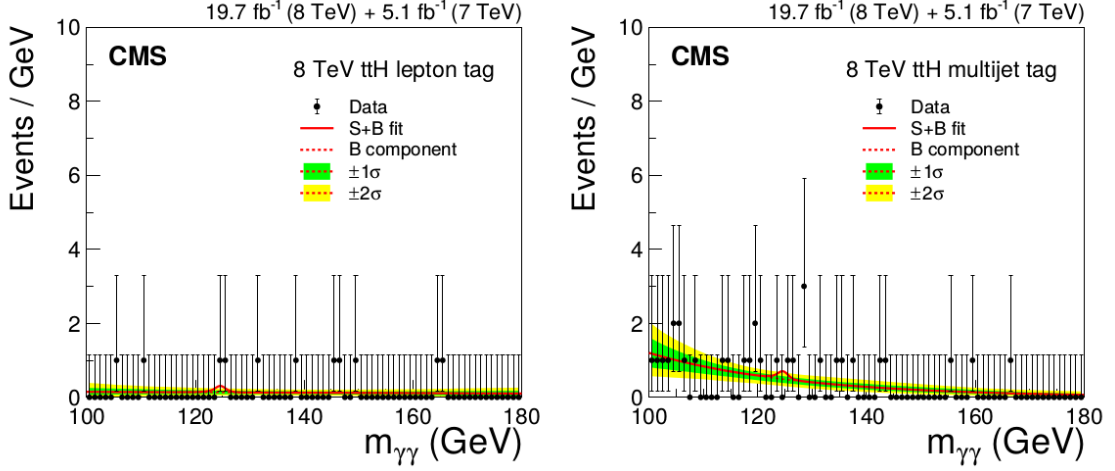


Figure 2.15: Diphoton mass spectrum together with the background subtracted mass spectrum for leptonic (left) and hadronic (right) $t\bar{t}H$ channels using CMS data collected during LHC Run I [120].

boson produced with a pair of top quarks and decaying into two photons. The limits on the $\mu_{t\bar{t}H}$ were set by both CMS (Eq. 2.3) and ATLAS (Eq. 2.4) collaborations at 95% CL [124].

$$\mu_{t\bar{t}H} < 5.4 \text{ (5.3 expected)} \quad (2.3)$$

$$\mu_{t\bar{t}H} < 5.3 \text{ (6.4 expected)} \quad (2.4)$$

A $t\bar{t}H$ combination [129] exploited $H \rightarrow b\bar{b}$, $\tau_h\tau_h$, $\gamma\gamma$, WW and ZZ decay modes. The signal strength was found to be 2.8 ± 1.0 at 68% CL, and shown in Fig. 2.16 (right). This result represented an excess above the background-only expectation of 3.4 standard deviations. Compared to the SM expectation, the observed excess was equivalent to a 2-standard-deviation upward fluctuation. These results were obtained assuming a Higgs boson mass of 125.6 GeV but they did not vary significantly for other choices of the mass in the vicinity of 125 GeV. These results were more consistent with the SM $t\bar{t}H$ expectation than with the background-only hypothesis. A combination with ATLAS collaboration gave an measured (expected) signal strength of $\mu_{t\bar{t}H} = 2.3^{+0.7}_{-0.6}$ ($1^{+0.5}_{-0.5}$) [125], which translates into 4.4 (2.0) σ observed (expected) significance with respect to background-only hypothesis.

In this thesis, the first measurement of the Higgs boson properties at the centre-of-mass energy $\sqrt{s} = 13$ TeV, including the signal strength of $t\bar{t}H$ process, is performed.

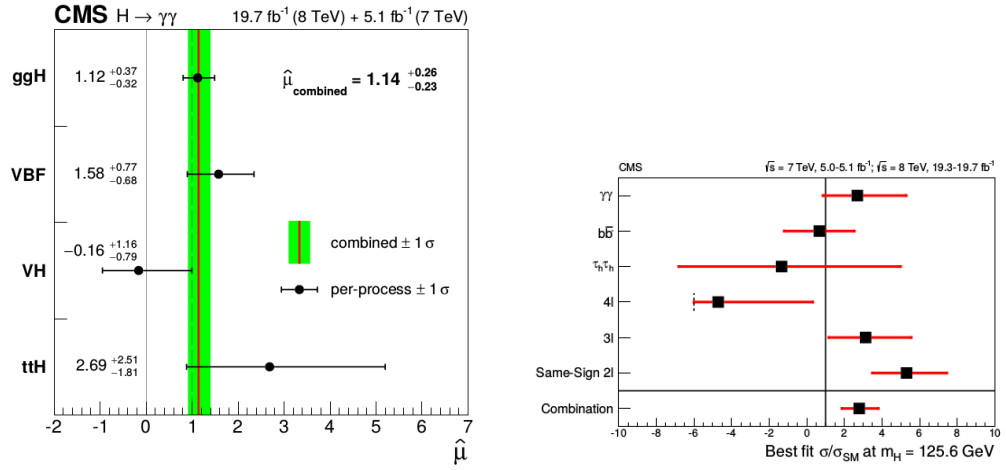


Figure 2.16: Left: The signal strength, $\mu = \sigma/\sigma_{SM}$, measured for each of the Higgs boson production processes and $H \rightarrow \gamma\gamma$ decay mode [120]. The horizontal bars indicate $\pm 1\sigma$ uncertainties in the values for the individual processes. Right: The signal strength for $t\bar{t}H$ channel, while Higgs boson decays to $b\bar{b}$, $\tau\tau$, $H \rightarrow \gamma\gamma$, WW and ZZ, at $m_H = 125.6$ GeV [129]. Both plots exploits CMS data collected during LHC Run I.

Chapter 3

Compact Muon Solenoid experiment at Large Hadron Collider

3.1 Compact Muon Solenoid experiment overview

Although the Higgs boson search is a major physics goal of the Compact Muon Solenoid (CMS) experiment, the CMS detector is designed to be a general purpose detector capable of measuring a wide range of SM and possible new physics processes produced at the LHC. General considerations include excellent momentum and energy resolution, and identification capabilities for muons, electrons and photons, as well as good jet and missing energy resolution. Moreover, the unprecedented collision parameters of the LHC set a strong threefold technical constraint on the design. The detector has to be radiation resistant as the high flux of particles from proton-proton collisions damages detector components on the long term, especially inner tracking and forward calorimetry. The detector has to provide good timing resolution to handle LHC bunch spacing of 25 ns, thus requiring high-performance readout electronics. At the same time it requires high granularity to minimize the probability that particles from pileup¹ interactions be in the same detector element as particles from the main proton-proton interaction, for this reason the detector needs a high number of electronic channels.

The CMS detector layout is organized around a superconducting solenoid magnet of 6 m internal diameter and 12.5 m length, providing a large magnetic field of 3.8 T. The overall apparatus is rather compact: it materializes as a 21.6-metre-long, 14.6-metre-wide, 14 000-tonne cylinder around the LHC beam axis. The magnet coil is large enough to accommodate three major subsystems within its volume: a silicon pixel and strip tracker which measures the trajectories of charged particles, a lead tungstate crystal electromagnetic calorimeter (ECAL) that mainly collects the energies of electrons and photons, and a brass and scintillator hadron calorimeter (HCAL) which stops the more penetrating

¹The average number of concurrent interactions per bunch-crossing.

hadrons. Some forward calorimeters further improve hermeticity. The measurement of muons relies on a combination of inner tracking and information from the muon chambers, which are gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid. A diagram of the CMS detector is shown in Fig. 3.1.

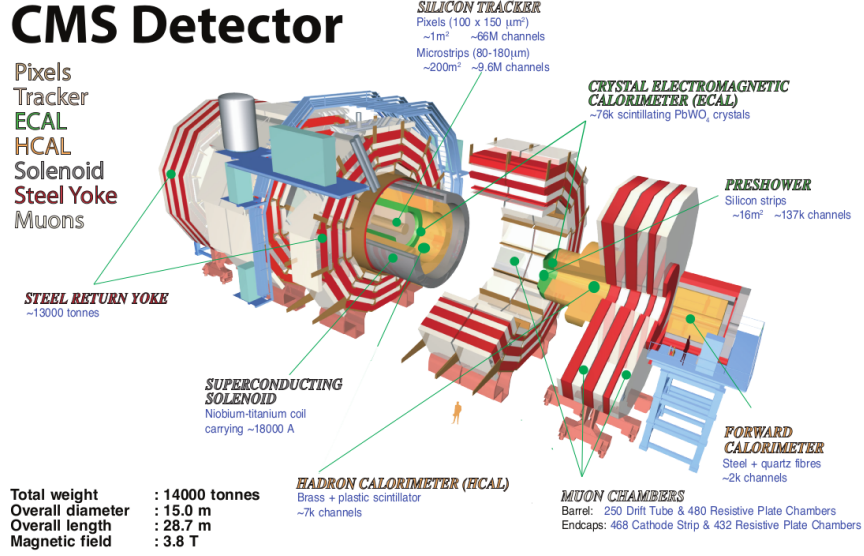


Figure 3.1: Representation of the CMS detector and its major components.

CMS is located in an underground cavern at Point 5 of the LHC, about 100 m below the village of Cessy, in France. The major detector subsystems, mentioned above, will be described in the following sections. A comprehensive description of the whole CMS apparatus can be found in reference [130].

A conventional coordinate system has been used to describe the CMS detector. Its origin is centred at the nominal collision point, and its z axis coincides with the proton beam direction and points toward the Jura mountains from LHC Point 5. The y axis points vertically upwards, while the x axis points radially inwards, toward the centre of the LHC ring. The azimuthal angle ϕ is measured from the x axis in the $x - y$ plane and takes values in $[-\pi, \pi]$. The radial coordinate in this plane is denoted by r . The polar angle θ is measured from the z axis and takes values in $[0, \pi]$. The coordinate system is shown in Fig. 3.2.

The polar angle is usually expressed in terms of the pseudorapidity, defined as $\eta = -\ln[\tan \theta/2]$. The particle momentum and energy transverse to the beam direction, respectively noted as p_T and E_T , are computed from the x and y components of the energy and momentum, which e.g. implies that $E_T = E \sin(\theta) = E/\cosh(\eta)$. The imbalance of the total transverse energy measurement in a collision is referred to as the missing trans-

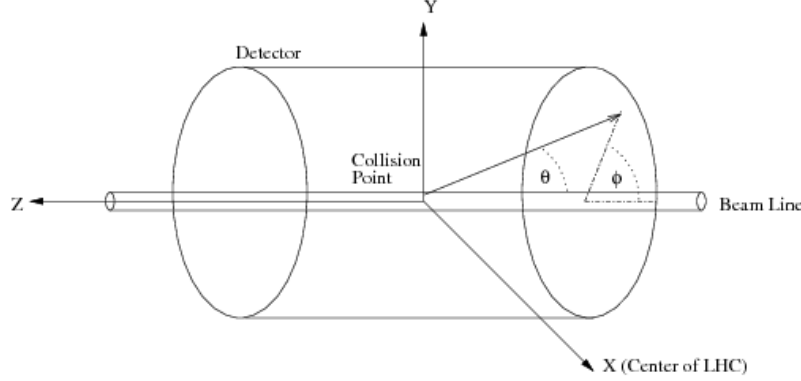


Figure 3.2: The coordinate system of the CMS detector.

verse energy and noted as p_T^{miss} . The angular distance ΔR between two particles i and j is defined as:

$$\Delta R(i, j) = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2} \quad (3.1)$$

Based on the η coordinate, the detector is divided into a central part called the barrel, and two opposite forward parts called the endcaps. The exact boundary depends on the subsystem.

3.2 The tracker

The tracker system [131] is the closest subdetector to the interaction point. The purpose of the tracking system is to measure the trajectory of charged particles, including a precise determination of their momentum as well as their position, extrapolated either to the beamline or to the calorimeter. The inner tracking detectors in CMS consist entirely of solid state silicon based detector. In order to provide the most precise position measurement and separation of charged particles near the interaction point, the inner part of the tracking detector consists of silicon pixel detector which provides precise measurements in three-dimensions.

The pixel detector consists of three barrel layers, referred to as the pixel barrel region, located at radii of 4.4, 7.3, and 10.2 cm and extending from $z = -26.5$ cm to $z = +26.5$ cm. In addition, there are two endcap pixel layers, referred to as the pixel forward region. These are located at $\pm z = 34.5, 46.5$ cm, and cover a region between approximately 6 cm and 15 cm in radius from the beam.

The outer part of the tracking detectors consists of a silicon strip tracker, which provides measurements precisely localized in only two-dimensions, with most strips oriented perpendicular to the ϕ direction. In the barrel, this consists of the Tracker Inner Barrel

(TIB) region, comprised of four layers between 20 and 55 cm in radius, as well as the Tracker Outer Barrel (TOB) region, consisting of an additional 6 layers between 50 and 116 cm in radius. In the endcap region, the strip tracker consists of the Tracker Inner Disk (TID) region of three layers located in $80 \text{ cm} < |z| < 90 \text{ cm}$, plus a Tracker EndCap (TEC) region of nine layers located in $124 \text{ cm} < |z| < 280 \text{ cm}$.

A fraction of the layers include double layered modules, with a second set of strips oriented at an angle of $100 \mu\text{rad}$ with respect to the first. The combination with these stereo measurements can give a position measurement in the third dimension with a precision ranging from 230 to $530 \mu\text{m}$. A schematic view of the tracking detectors, labeled by region, is shown in Fig. 3.3. The combined tracking detector system provides coverage up to $|\eta| = 2.5$, with an average of 13-17 measurements per charged particle, depending on the pseudorapidity region.

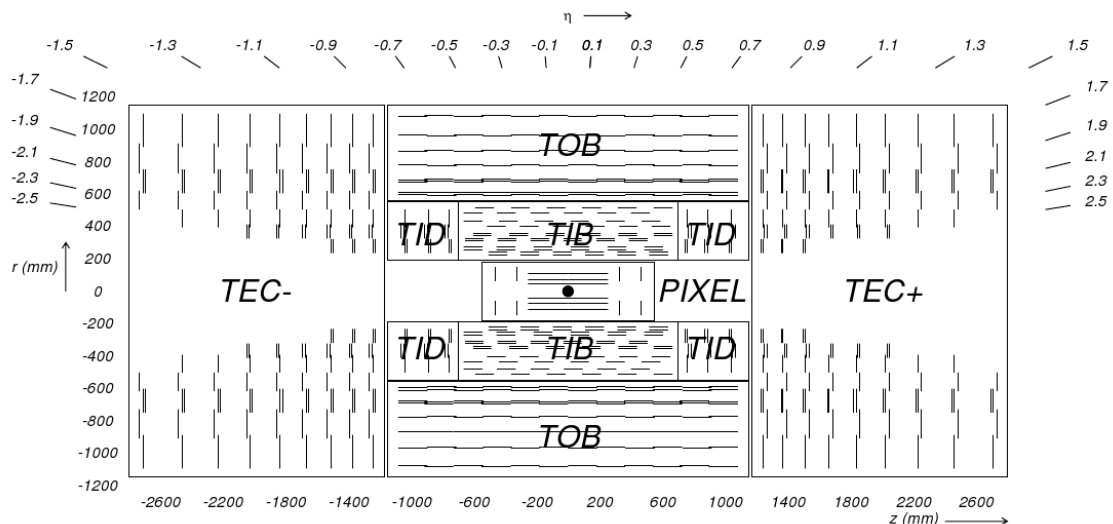


Figure 3.3: Schematic representation of the mechanical layout of the CMS silicon tracker [130].

The pixel detector consists of 66 million pixel elements, each $100 \times 150 \mu\text{m}^2$ in dimension, spread across 1440 modules. Each pixel consists of a p-n semiconductor junction. When a charged particle crosses the junction, it excites electron-hole pairs, and the charge is collected by the readout electronics connected to the junction. In order to keep the data volume reasonable given the very large number of channels, zero suppression is performed by electronics on the sensor modules, in which only pixels with signal above a set threshold are read out. A charged particle crossing the module will generally deposit charge in at least two adjacent pixels, with the amount of charge deposited in each pixel inversely related to the distance between the particle position and the pixel. A measurement of the charge sharing between adjacent pixels therefore allows a single hit position resolution

substantially smaller than the dimensions of a single pixel. In order to exploit the sharing of charge among adjacent pixels, the signal amplitude is digitized with 5 to 8 bits of information, allowing a single hit position resolution of 15-20 μm .

The silicon strip detector consists of about 9.3 million strips across 15148 modules, with strips as well consisting of p-n junctions across which charge carriers are ionized by charged particles as they cross the strip. Depending on the region of the detector, the strip pitch varies between 80 and 184 μm . By exploiting charge sharing between strips, analogous to charge sharing between adjacent pixels, the single hit resolution along the ϕ direction ranges from 23 to 53 μm , smaller than the strip pitch.

The large amount of silicon in the inner tracking detectors, combined with the electronics required leads to a substantial requirement for cabling and cooling services. This results to a relatively large amount of material in the detector. The estimated material budget, as a function of pseudorapidity, is shown in Fig. 3.4. The estimated total material budget ranges from about 0.4 radiation lengths(X_0)² in the very central barrel, to a peak of about 1.8 radiation lengths in the vicinity of $|\eta| = 1.5$, near the barrel-endcap transition region.

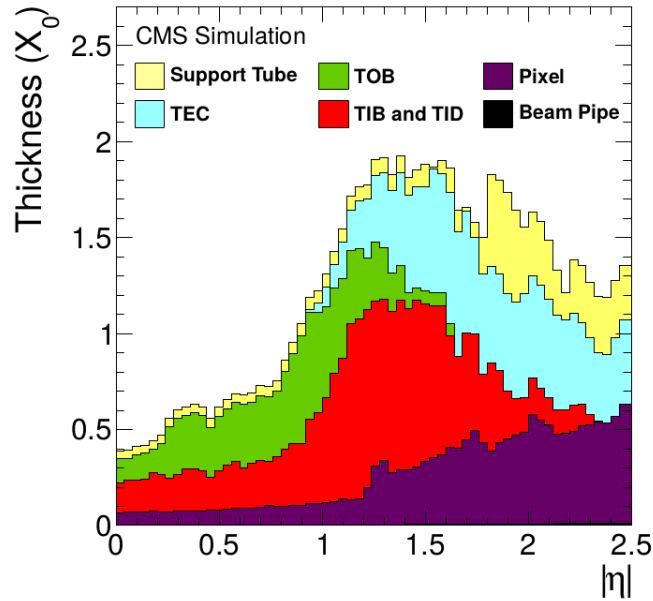


Figure 3.4: Amount of material in the tracker detectors measured in units of radiation lengths, and broken down by detector regions [132].

The tracker system provides a very precise measurement of particle momentum [133].

²The radiation length (X_0) of a material is the mean length to reduce the energy of an electron by the factor $1/e$, e means Euler's number here.

For high p_T track (100 GeV) the p_T resolution is about 1-2% in the central region ($|\eta| < 1.6$) and a bit worse in the endcaps due to the shorter lever arm of these tracks in the $x - y$ plane of the tracker. At this p_T the multiple scattering contribution is about 20-30% and it increases for lower transverse momentum.

3.3 The electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [134] measures the energies of electrons and photons. Its design was driven by the prospect of detecting Higgs boson decays to pair of photons, which called for an excellent energy and position resolution.

In order to achieve the best possible energy resolution, the ECAL is a homogeneous and nearly hermetic calorimeter. It is made of lead tungstate (PbWO_4) crystals, coupled to photodetectors.

Thanks to the short radiation length (0.89 cm) and high density of this material, electromagnetic showers can be absorbed within relatively short crystals, while the small Moliere radius (2.2 cm) allows for a good shower separation. 80% of the scintillation light is emitted in 25 ns, which is fast enough to cope with the LHC bunch spacing. The barrel part of the ECAL (EB) is made of 61 200 crystals of length 230 mm ($25.8 X_0$) and frontal cross section $22 \times 22 \text{ mm}^2$, covering the pseudorapidity range $|\eta| < 1.479$. The endcap parts (EE) involve 7324 crystals each, with length 220 mm ($24.7 X_0$) and frontal cross section $28.62 \times 28.62 \text{ mm}^2$, covering the $1.479 < |\eta| < 3.0$ range, as shown in Fig. 3.5.

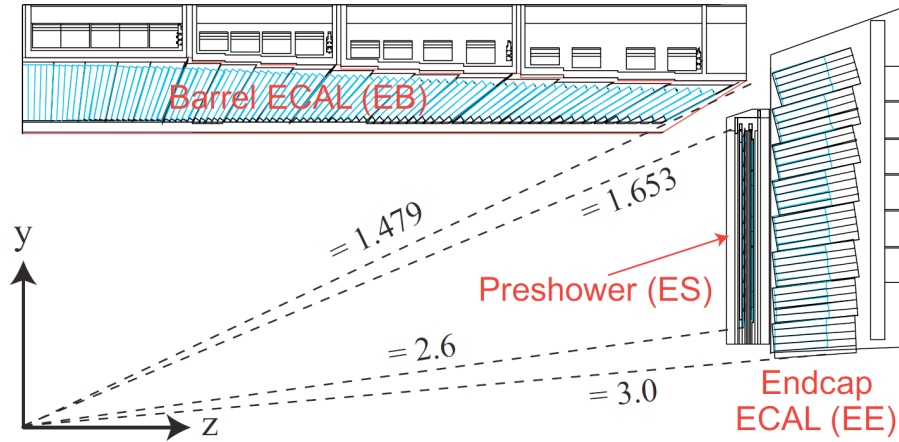


Figure 3.5: Longitudinal sectional view of a quarter of the CMS detector, showing the ECAL barrel and endcap, with the preshower in front.

The structure of the EB relies on 36 supermodules that cover half of the barrel length and 20° in ϕ . Each supermodule is made of four modules that each contain 400 or 500

crystals in an alveolar structure. Each EE is made of two semi-circular dees containing 3662 crystals. This general structure is illustrated in Fig. 3.6. The crystals are mounted in a quasi-projective geometry, so that their axes make a 3° angle with respect to the direction of the nominal interaction point in both the η and ϕ projections, thus avoiding to align inter-crystal gaps with particle trajectories. Still, some gaps (cracks) remain between modules and complicate the energy reconstruction. Larger cracks are present at $\eta = 0$ and at the EB-EE transition.

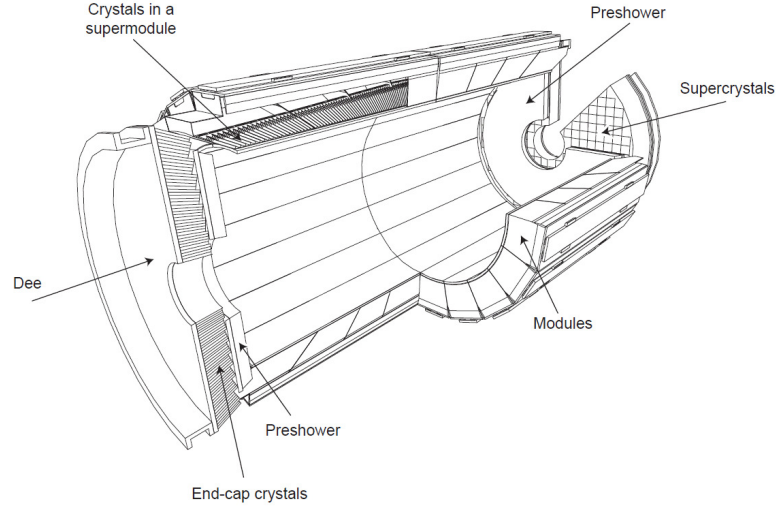


Figure 3.6: Layout of the CMS ECAL, showing the barrel supermodules, the two endcaps and the preshower subdetectors [130].

Two preshower detectors (ES) are installed at each end of the tracker, in front of the EE, covering the $1.653 < \eta < 2.6$ region. These sampling calorimeters are made of a lead radiator layer that initiates electromagnetic showers from incoming particles, followed by silicon strip sensors that measure the deposited energy and transverse shower profiles. The ES helps to distinguish $\pi^0 \rightarrow \gamma\gamma$ decays from single photons, and to identify electrons against minimum ionizing particles.

Scintillation light from the ECAL crystals is read by fast, radiation-tolerant photodetectors, which amplify the small light yield of the particles traversing them. Due to different magnetic field configurations and expected radiation levels, different photodetector technologies are used in the EB and EE, namely avalanche photodiodes (APD) and vacuum phototriodes (VPT), respectively. Since the crystal response is temperature-dependent, a cooling system stabilizes the temperature of both crystals and photodetectors to 18°C with a $\pm 0.05^\circ\text{C}$ accuracy in order to preserve the energy resolution.

APDs are fast (~ 2 ns of rise time), with very good quantum efficiency (70%-80% at $\lambda = 420$ nm), highly insensitive to magnetic field and radiation resistant detectors.

Each APD has $5 \times 5 \text{ mm}^2$ active area and two of them are glued to the back of each crystal. The structure of APD is shown in Fig. 3.7. The light passes through the p^+ layer and it is absorbed by the p layer behind. Electron-hole pairs are produced if the photon energy is higher than the gap energy. A drift in the p-n transition region is followed by an amplification stage in the n volume and by a drift region before the charge is collected by the cathode. Each APD is supplied by reverse voltage to vary the multiplication gain (from 0 to 200), the gain 50 is used as it fits the optimum parameters of the CMS front-end electronics. APD gain depends also on the temperature, so one tenth of all APDs is equipped with temperature sensors.

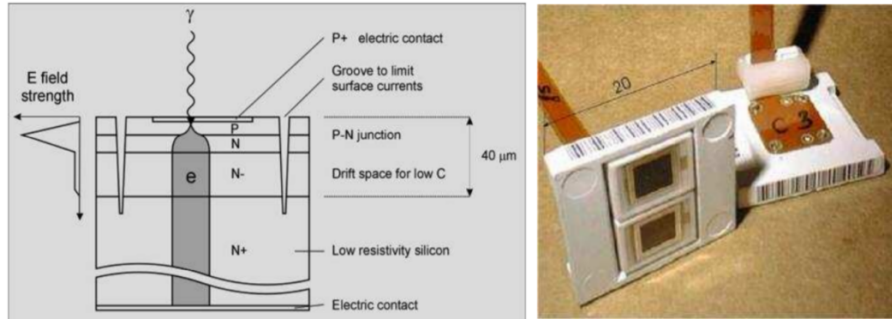


Figure 3.7: Left: APD layout. Right: A photo of APD pair.

APD detectors are insufficiently radiation hard to be used in endcaps, for this reason VPT detectors are used. VPT structure is shown in Fig. 3.8. The photocathode is semitransparent and made of radiation-hard glass. The photoelectrons produced are accelerated by an ultra fine mesh (100 wires/mm) placed 4-5 mm far from the photocathode, and impact on dynode producing secondary electrons (emission factor ~ 20). The secondary electrons are attracted back to the anode mesh where a substantial fraction is captured, leading to a total effective gain of the VPT greater than 8 compensated by a larger effective area $\sim 280 \text{ mm}^2$, so that the total detector response is almost the same for barrel and endcap.

Although lead tungstate crystals are radiation resistant, they are known to undergo loss of optical transmission under irradiation. To measure this effect, laser pulses are injected into the crystals via optical fibres, and the response is normalized by the laser pulse magnitude measured using PN diodes. Time-dependent corrections are then applied to the measured particle energies. The ECAL laser monitoring system is described in chapter 4 along with its performance and possible upgrade for the LHC Phase II.

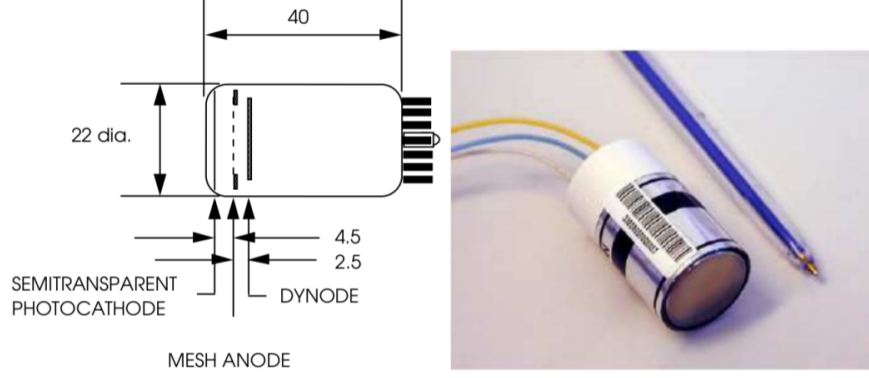


Figure 3.8: Left: VPT layout. Right: A photo of VPT detector.

The ECAL energy resolution consists of a stochastic (S), a noise (N), and a constant (C) contribution:

$$\frac{\sigma_E}{E} = \frac{S}{\sqrt{E[\text{GeV}]}} \oplus \frac{N}{E[\text{GeV}]} \oplus C \quad (3.2)$$

It is discussed in details in chapter 4 as well as the ECAL calibration sequence.

3.4 The hadronic calorimeter

The Hadronic Calorimeter (HCAL) [135] measures the energy of charged and neutral hadrons, critical for the reconstruction and measurement of jets and missing energy. In the forward region beyond the coverage of the ECAL, the HCAL system is also responsible for the measurement of electromagnetic energy. The HCAL consists of a barrel detector (HB) covering the region up to $|\eta| = 1.3$, an endcap detector (HE) covering the region $1.3 < |\eta| < 3.0$, and a forward detector (HF) covering the region $3.0 < |\eta| < 5.2$. In the region covered by the HB, there is an additional outer detector (HO) which is placed outside the magnet solenoid in order to ensure the measurement and containment of the tails of hadronic showers, which may extend beyond the HB and the solenoid material. Including the HO and the solenoid, the total amount of material in the region covered by the HB and HE subdetectors corresponds to at least 11.8 hadronic interaction lengths³, except for a small region at the transition between the HB and HE. The overall layout of the HCAL subdetectors is shown in Fig 3.9.

The HB and HE detectors are both sampling calorimeters. Brass (70% Cu and 30% Zn) is used as the absorber, except for the first and last layers which are made of stainless

³The distance that a hadron can travel before an nuclear interaction occurs is called the hadronic interaction length, $\lambda_n = 35 \text{ g cm}^{-2} \text{ A}^{-1/3}$.

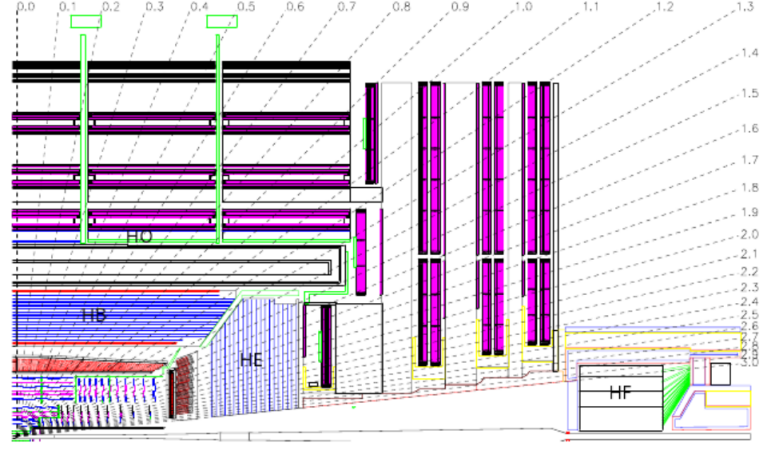


Figure 3.9: Schematic view of the HCAL components in CMS, including the barrel section (HB), the endcap section (HE), the tail catcher outside the solenoid (HO) and the forward section (HF) [130].

steel for structural strength. The active medium uses the tile and wavelength shifting fibre concept [135] to bring out the light, which is then read out by means of hybrid photo diodes (HPDs). Up to $|\eta| < 1.6$, HCAL towers have a size of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$, while for $|\eta| > 1.6$, the size increases to $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$.

The much harsher radiation environment in the forward region necessitates different design for the HF detector, where scintillating quartz fibres, read out by photomultiplier tubes, are used together with steel absorber layers. The quartz scintillators are sensitive to both electrons/photons as well as hadrons, because their longitudinal segmentation into two depths provides limited discrimination between electromagnetic and hadronic deposits.

The HCAL energy resolution is:

- $\frac{\sigma_E}{E} = 65\% \oplus 5\%$ in the barrel
- $\frac{\sigma_E}{E} = 85\% \oplus 5\%$ in the endcaps
- $\frac{\sigma_E}{E} = 100\% \oplus 5\%$ in the forward calorimeters

The energy resolution of the ECAL-HCAL system was evaluated with a combined test beam with high energy pions [136] and it is given by $\frac{\sigma_E}{E} = \frac{84.7\%}{E} \oplus 7.4\%$

3.5 The muon system

The outer muon system of CMS [137] has been designed to achieve high-precision measurement of muon momenta and charge, even without the help of the inner tracker. As

shown in Fig. 3.10, this subsystem is made of muon chambers embedded in the iron return yoke of the CMS magnet, which are in principle not reached by other detectable particles. It is divided into a cylindrical barrel section and two planar endcap regions. Three types of gas-ionization chambers are used, for a total of 25 000 m² of detection planes.

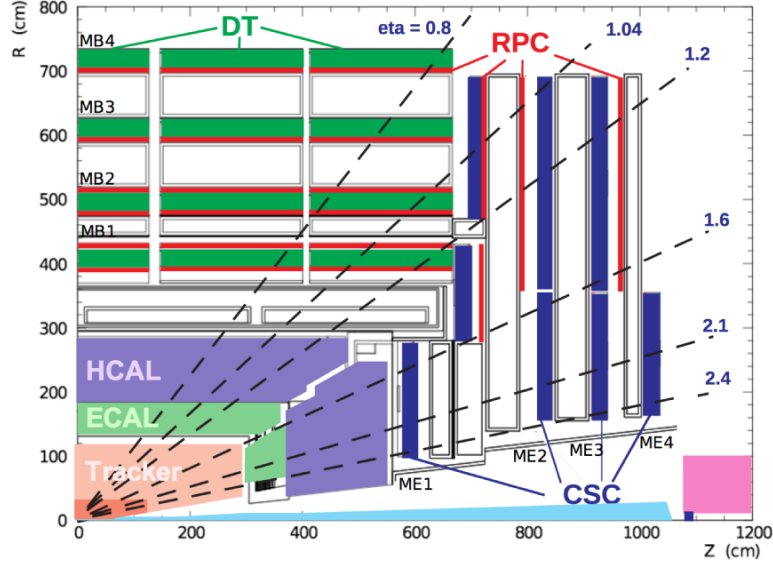


Figure 3.10: Longitudinal sectional view of a quarter of the CMS detector, showing the four DT stations in the barrel (MB1-MB4, green), the four CSC stations in the endcap (ME1-ME4, blue), and the RPC stations (red) [138].

The Drift Tube Chambers (DTs) are located in the barrel region ($|\eta| < 1.2$), where the neutron-induced background is small, the muon rate is low, and the magnetic field is quite uniform. The DTs are organized into four stations interspersed among the layers of the flux return plates. Their basic constituents are rectangular drift cells. These are bounded by two parallel aluminium planes, and aluminium "I"-shaped cathodes, while the anodes are 50 μm stainless steel wires located in the centre of the cells. A muon passing through a cell ionizes the gas mixture that fills the cell volume. The drift time of the resulting electrons is then used to measure the distance between the muon track and the wire. The drift cells of each chamber are offset by a half-cell width with respect to their neighbour, in order to eliminate dead spots in the efficiency. Each chamber has a resolution of 100 μm in the $r - \phi$ plane. The number and orientation of chambers in each station were chosen in a way that helps accurately link muon hits from different stations into a single muon track.

The Cathode Strip Chambers (CSCs) are used in the endcaps ($0.9 < |\eta| < 2.4$), where the muon rates and background levels are high and the magnetic field is large and non-uniform. The CSCs are multiwire proportional chambers, made of 6 anode wire planes

interleaved among 7 cathode panels, with the wires running approximately perpendicular to the strips. A muon passing through a chamber generates an avalanche, inducing a charge on several cathode strips. The ensuing interpolation allows for a fine spatial resolution, of the order of $50\ \mu\text{m}$. There are four stations of CSCs in each endcap; the chambers are positioned perpendicular to the beam line and interspersed between the flux return plates.

The Resistive Plate Chambers (RPCs) are found both in the barrel and in the endcaps ($|\eta| < 1.6$), and are only used at trigger level. While both DTs and CSCs can already trigger muons independently with good efficiency and background rejection, the RPCs provide a complementary trigger system with moderate spatial resolution but excellent time resolution (of the order of 1 ns), thus helping measure the correct beam-crossing time, even at the largest LHC luminosities. The RPCs are double-gap chambers, operated in avalanche mode to ensure reliable operation at high rates; they are arranged in six layers in the barrel and three layers in the endcaps.

The muon momentum resolution [139] varies with pseudorapidity, but overall muon momentum resolution of 5-8% at 10 GeV, 20-40% at 1 TeV is obtained when using only the muon system. The numbers improve significantly to 1-1.5% at 10 GeV, 6-17% at 1 TeV when the combination of the inner tracker and outer muon system is used.

3.6 The trigger system

The CMS trigger system [140] performs online event selection via two successive layers. The first level, called the Level-1 trigger (L1), is implemented on custom designed hardware, and is adjusted to bring the event rate down to about 100 kHz, which is the upper limit imposed by the CMS readout electronics. The second level, called the high-level trigger (HLT), is implemented in software and runs a streamlined version of the CMS reconstruction algorithm to select an average rate of 1 kHz. It then transmits the data to the CMS Tier-0 computing centre for storage and offline processing.

Level-1 trigger of CMS. The hardware-based Level-1 trigger of CMS performs a fast readout of the detector with a limited granularity, selecting events that contain such distinctive detector signals as ionization deposits consistent with a muon, or energy clusters consistent with an electron, photon, tau lepton, or jet. The L1 selection relies on a programmable menu made of 128 algorithms or seeds, each of which selects a particular type of objects and passes them to the HLT for subsequent processing. Every seed is assigned an adjustable prescale value n , meaning that it accept a fraction $1/n$ of events that pass its specific selection criteria. Thresholds and prescales are adjusted to the LHC instantaneous luminosity during data taking, so as to restrict the output rate to the 100 kHz upper limit.

The Level-1 trigger has a fixed latency: it has $4\ \mu\text{s}$ to decide to accept or reject an event, using information from the calorimeters and muon detectors. Trigger primitives are computed from energy deposits in the trigger towers of the ECAL and HCAL calorimeters

in the one hand, and from track segments and hit patterns in the DT, CSC and RPC muon chambers on the other hand. As illustrated in Fig. 3.11, the information is processed through two separate flows: the calorimeter trigger builds the EG (electron or photon), jet, and tau candidates, as well as variables such as p_T^{miss} and H_T (the scalar transverse energy sum of jets above a certain threshold), while the muon trigger builds the muon candidates. The combined event information is finally evaluated in the global trigger, which makes the final decision based on the menu.

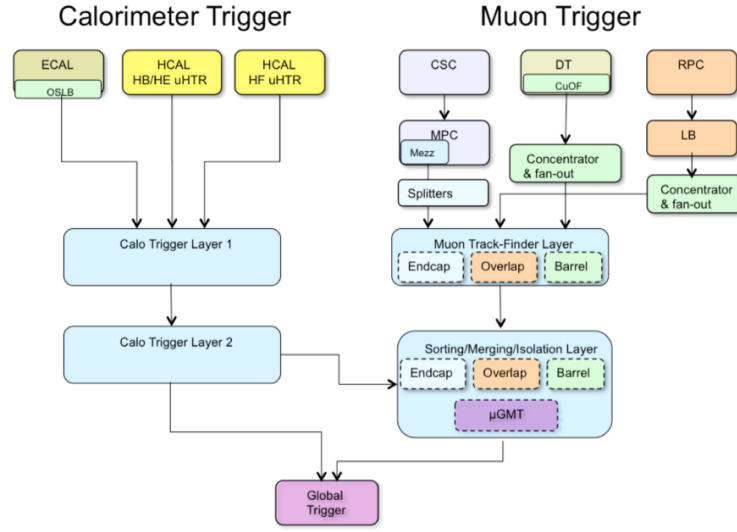


Figure 3.11: Schematic representation of the CMS L1 trigger system.

For Run II of the LHC, with prospects of peak instantaneous luminosities of $1.6 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ or more, and of pileup levels of 50 simultaneous inelastic collisions per crossing, the L1 trigger was known not to be able to keep low trigger thresholds while staying within the 100 kHz bandwidth. Hence, several updates were successively introduced [141] to preserve good performance as the LHC ramps up. First, in the calorimeter trigger, data communication from the ECAL was improved with new optical links, and an earlier sub-system called global calorimeter trigger was replaced with a new data processing card with better data throughput and computational power, allowing the execution of improved algorithms and the inclusion of event-by-event pileup subtraction. Second, in the calorimeter trigger, events are read by two new layers of data processors, with a new time-multiplexed architecture. The position and energy resolution of jet, photons, electrons, and tau candidates were improved, which provided the additional background rejection required to cope with the increased instantaneous luminosity and pileup. Third, the muon trigger was also upgraded, combining all three muon systems to perform an integrated track finding,

allowing for a more elaborate p_T measurement.

HLT of CMS. The high-level trigger is implemented in software and performs a full readout of the CMS detector. Events are reconstructed with the same software as used for offline processing, but in a very optimized configuration which is two orders of magnitude faster. Based on the raw data from events accepted by the L1 trigger, all main classes of physics objects can be reconstructed at HLT, such as electrons, muons, photons, taus, missing transverse energy, and jets, including some more advanced techniques like b tagging or jet substructure study. Specific selection criteria are applied to these objects so as to both keep the rate under control and retain the subset of events most relevant to subsequent data analysis. As opposed to the L1, this HLT filtering procedure is able to exploit the full precision of the data from the CMS detector, with offline-quality algorithms.

The HLT runs on a single dedicated farm of commercial computers. This so-called event filter farm consists of builder units that collect and assemble individual event fragments from the detector, and of filter units that unpack the raw data into detector-specific data structures, and perform event reconstruction and trigger filtering. In total, the farm currently comprises approximately 16 000 CPU cores. Its computing power has been increased by about 50% since Run I to cope with pileup and code complexity. Data processing at the HLT is structured around the concept of a path, which consists in a predefined sequence of algorithmic processing steps of increasing complexity, that both reconstructs a certain type of physics objects and applies a selection to it. Like L1 seeds, HLT paths can be prescaled, and the set of all paths used at a given time is also called a menu. All paths are run in parallel and independently of each other, but the common modules and sequences are shared among different paths. The successive reconstruction modules and selection filters are organized in such a way that the fastest selections such as those relying on information from the calorimeters and muon detectors are run first. This helps reducing the event rate as soon as possible, before considering CPU-expensive steps such as track reconstruction.

Finally, all events that are selected by at least one path are directed to one of various data streams, which are illustrated in Fig. 3.12. The main physics data stream transmits events as full raw detector data, for prompt offline event reconstruction and permanent mass storage. Its maximum average rate has been increased from 400 Hz in Run I to about 1 kHz in Run II. Other physics streams include data parking, i.e. the storage of full event content from special loose HLT paths for a delayed offline processing during LHC long shutdowns, and data scouting, i.e. the storage of reduced, non-reprocessible event content from very loose HLT paths. Some special streams are dedicated to data quality monitoring (online and offline), and to detector alignment and calibration workflows.

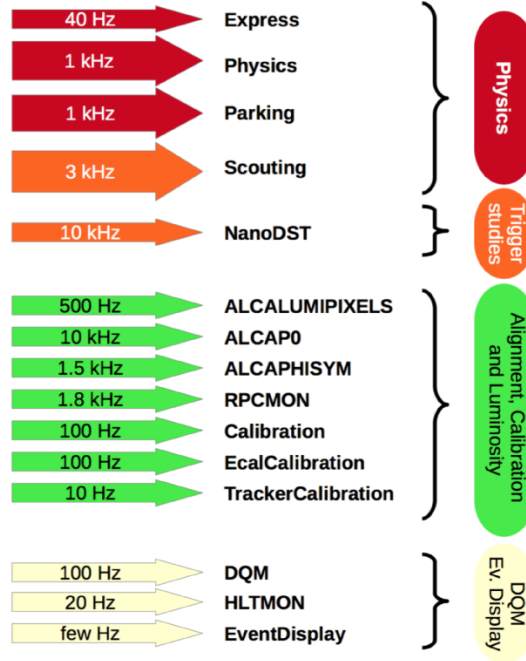


Figure 3.12: Schematic representation of the CMS HLT.

3.7 Event reconstruction

Event reconstruction in CMS relies on the Particle-Flow (PF) algorithm [142, 143]. It aims at reconstructing and identifying each stable particle in the event (electrons, muons, photons, charged hadrons, and neutral hadrons) via a combination of all CMS subdetectors, allowing for an optimal determination of their direction, energy and type. The fundamental elements of PF algorithm are tracks and calorimeter clusters. They are topologically linked into blocks, which are finally interpreted as particles.

Iterative tracking. The CMS tracker provides precise measurements of charged hadrons momentum direction at the production vertex. An iterative tracking strategy [144] is used to achieve both high tracking efficiency and low fake rate. Tracks are first seeded and reconstructed with very tight criteria, ensuring a negligibly small fake rate at the cost of efficiency. The next steps proceed by removing hits that are unambiguously assigned to these already reconstructed tracks, and progressively loosening track seeding criteria (thus increasing the efficiency). In the last iterations, vertex constraints are relaxed to reconstruct charged particles originating from secondary vertex, for example photon conversions or decays of long-live hadrons.

Clustering. The clustering algorithm of calorimeter energy deposits is designed to:

- reconstruct neutral particles such as photons or neutral hadrons,
- measure the energy of photons, electrons and neutral hadrons.

Description of the algorithm. A single particle is expected to give rise to several PF elements: charged-particle track, and/or several calorimeter clusters, and/or muon track. These elements need to be connected to one another by a link algorithm to reconstruct each particle while avoiding possible double counting from different detectors. Link algorithm defines a distance between pair of elements in the event, quantifying the quality of their possible link. It decides whether these elements are linked. The decision is based on a set of criteria that depends on the elements types. This creates PF blocks made of several elements. For each block, the algorithm proceeds through the following sequence:

- Each global muon (defined in section 5.11.1) gives a rise to a PF muon if its combined momentum is compatible with that determined from the inner tracker within 3 standard deviations. Muon tracks are removed from the considered PF block, as well as calorimetric energy deposits of PF muons.
- The algorithm then addresses the electron candidates. To be accepted as PF electrons, these also have to pass identification criteria, defined by an algorithm that exploits tracking and calorimetric variables. The building bricks of PF electrons (tracks and ECAL PF clusters) are removed from further processing of the block.
- The energy deposits in ECAL and HCAL PF clusters have to undergo a calibration procedure, correcting for the non-linear calorimeter response and threshold effects.
- If an ECAL PF cluster and HCAL PF cluster are linked together with a track, the calibrated combined energy of the ECAL and HCAL clusters is compared to the track momentum. If they are compatible, a charged hadron is created, otherwise either neutral hadron or photon is created from the excess of calorimeter energy. Associated tracks and clusters are removed from the block.
- The remaining ECAL PF clusters are associated to photons and HCAL PF clusters (or ECAL and HCAL linked PF clusters) are associated to neutral hadrons.

A final collection of individual PF particles (or PF candidates) of five possible types is obtained for every event.

Chapter 4

Electromagnetic calorimeter laser monitoring system in CMS

The CMS electromagnetic calorimeter (ECAL) is characterized by an excellent energy resolution, which has been one of the main elements in the discovery of the Higgs boson in final states involving electromagnetic particles, in particular in the $H \rightarrow \gamma\gamma$ channel. The laser monitoring system is a key device in the long chain of the ECAL calibration. The excellent energy resolution of the calorimeter relies on a precise and stable laser monitoring system. Ionizing radiation creates color centres in the crystals reducing their transparency and therefore reducing their measured response to the deposited energy. The color centres partially anneal with thermal energy such that the loss in transparency depends on the dose rate, which varies with η . The laser monitoring system permits the computation of the short term intercalibration correction coefficients, needed to stabilize the response of the detector due to radiation-induced crystal transparency variation. The general idea is to inject light through each crystal of the ECAL and compare the signal received on the associated photo-detectors, silicon avalanche photodiodes (APD) in the barrel and vacuum phototriodes (VPT) in the endcaps, with the signal given by a photo-detector (PN diode) that sees directly the light sent to the crystals. The signal ratio APD/PN or VPT/PN is a measurement of the transparency of each crystal.

In this chapter, I introduce the laser monitoring system in the chain of the ECAL calibration in section 4.1. I describe the existing laser monitoring system in section 4.2.1 and show its performance in section 4.2.3. As a service task, I participated in the studies of the possible laser monitoring system upgrade for the LHC Phase II. The system upgrade test bench is described in section 4.3.1, and its installation in CMS underground service cavern is shown in section 4.3.2. The stability study of the test bench with laser data taken in 2015 is demonstrated in section 4.3.3.

4.1 ECAL signal pulse reconstruction and calibration

The ECAL performance was measured in test beams using electrons with energy ranging from 20 to 250 GeV, where there were neither magnetic field, nor material in front of the calorimeter. During the test beams, a supermodule of the ECAL (described in section 3.3) was mounted on a movable, computer controlled table at the H4 beamline at CERN. The supermodule's beam line position was adjusted such that the electron beam was incident at an angle of 3° to the axis of a selected crystal in both transverse directions. The beam could be directed at each crystal in the supermodule. The selected crystal was generally positioned in the beam such that the energy deposit in this crystal was maximized. Due to the 3° off-pointing of the crystal axis, the centre of the crystal front face does not coincide with the beam axis. More details about the test beams setup can be found in reference [145].

The ECAL barrel energy resolution was found to be:

$$\frac{\sigma_E}{E} = \frac{2.8\%}{\sqrt{E[\text{GeV}]}} \oplus \frac{12\%}{E[\text{GeV}]} \oplus 0.3\% \quad (4.1)$$

for electrons impinging on the centre of the crystals. The electron energy is reconstructed as the sum of the energy deposits in a matrix of 3×3 crystals around the impact point. The irreducible constant term, which dominates the energy resolution for high-energy electrons and photons, is affected by the non-uniformity of the longitudinal light collection, energy leakage from the back of the calorimeter, single-channel response uniformity and stability. The stochastic term is affected by the statistical fluctuations in the number of photo-electrons produced in the APDs, the effects of longitudinal non-uniformities of the crystal response and the statistical fluctuations of the containment losses. The electronics noise term was measured by applying the amplitude reconstruction procedure to data taken with a random trigger, when there was no incident electron signal present (pedestal runs).

In contrast to the test beam setup, additional contributions to the energy resolution are present in real CMS environment. Material upstream of the ECAL can cause electron bremsstrahlung and photon conversions that affect all terms of the energy resolution. Moreover, residual miscalibrations of the channel-to-channel response changes with time due to radiation damage of the crystals and environmental instability impact on the constant term of the resolution. These effects have to be controlled to a fraction of a percent to maintain the excellent intrinsic energy resolution of the ECAL.

ECAL signal pulse reconstruction. The electrical signal from the photodetectors is amplified and shaped by a multi-gain preamplifier. The output is digitized by a 12 bit ADC running at 40 MHz, which records ten consecutive samples used to reconstruct the signal amplitude. A template fit with multiple components, named "multi-fit", is used in

the LHC Run II. The multi-fit algorithm estimates the in-time signal amplitude and up to 9 out of time amplitudes by minimization of the χ^2 , given by

$$\chi^2 = \sum_{i=1}^N \frac{(\sum_{j=1}^M A_j p_{ij} - S_i)^2}{\sigma_{S_i}^2} \quad (4.2)$$

where A_j are the amplitudes of up to $M = 10$ interactions. The pulse templates p_{ij} for each bunch crossing j have the same shape, but are shifted in time by multiples of 25 ns within a range of -5 to +4 bunch crossings around the time of the in-time signal. The pulse templates for each crystal are measured from low pileup proton-proton collision data recorded by CMS. The total electronic noise S_i , and its associated covariance matrix, σ_{S_i} are measured from dedicated pedestal runs. The least squares method [146] is used to perform the χ^2 minimization of Eq. 4.2 with the constraint that the fitted amplitudes are all positive. Examples of one fit for signals in the barrel and in the endcaps are shown in Fig. 4.1, for an average pileup of 20 and for 25 ns bunch spacing [147].

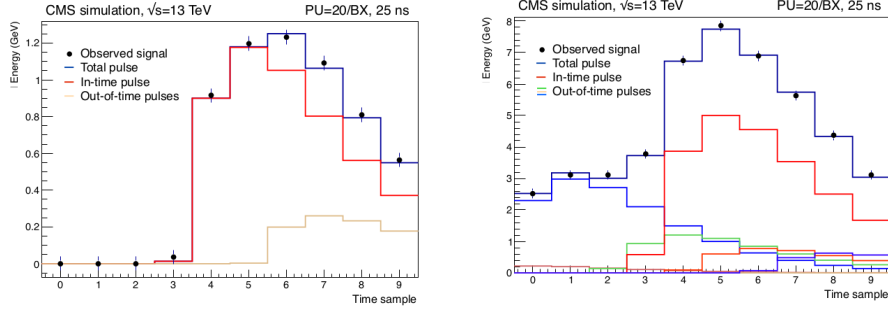


Figure 4.1: Example of fitted pulses for simulated events with 20 average pileup interactions and 25 ns bunch spacing, for a signal in the barrel (left) and in the endcap (right) [147]. Dots represent the 10 digitized samples, the red distributions (other light colors) represent the fitted in-time (out-of time) pulses with positive amplitude. The dark blue histograms represent the sum of all the fitted contributions.

The reconstruction of the laser pulse is different with respect to the ECAL signal pulse, and it is discussed in section 4.2.2.

Clustering algorithms. Electrons and photons deposit their energy over several ECAL channels and the presence of material in front of the ECAL causes conversions of photons and bremsstrahlung from electrons, so that the radiated energy is spread along ϕ by the magnetic field. Clustering algorithms are used to collect the energy deposits in the ECAL, including the contributions from this radiated energy. The electron or photon energy is

estimated as [148]:

$$E_{e,\gamma} = F_{e,\gamma} [G \times \sum_{i=\text{crystal}} (C_i \times S_i(t) \times A_i) + E_{ES}] \quad (4.3)$$

where the sum is performed over all clustered crystals. The amplitude measured in the i -th crystal is labeled by A_i , while $S_i(t)$ is a time dependent correction that accounts for time variation of the channel response due to changes in the crystal transparency. The C_i parameter is a relative calibration constant that takes into account differences among crystals for light yields and photodetector response and G is a scale factor converting the digital scale into GeV. For clusters in the endcap region the corresponding energy in the preshower (E_{ES}) is added. Finally $F_{e,\gamma}$ is a particle dependent correction applied to the clustered energy that accounts for biases in the energy reconstruction related to the geometry of the detector, the upstream material, the electromagnetic shower leakage and the clustering of energy emitted by bremsstrahlung or photon conversions. Residual data-simulation differences are accounted for by in-situ measurements using several physics processes producing electrons and photons in the final state: $\pi^0/\eta \rightarrow \gamma\gamma$, $W \rightarrow e\nu$, $Z \rightarrow e^+e^-$, their contribution is shown below.

Energy calibrations: corrections for time-dependent response changes. As mentioned in the introduction of this chapter, the crystals loose their transparency due to ionizing radiation. The changes in crystal transparency and photodetector response are measured and corrected using a dedicated laser monitoring system which injects laser light into each crystal [149, 150]. The corrections are validated with collisions data, by examining the stability of the reconstructed invariant mass of π^0 decays and using the ratio of E/p for isolated electrons from $W \rightarrow e\nu$ and $Z \rightarrow e^+e^-$ decays, where E is the energy measured in the calorimeter and p is the momentum measured in the tracker.

Energy calibrations: intercalibrations and energy scale. The relative calibration of the crystals (parameter C_i in Eq. 4.3) is obtained from the LHC collision data using several independent methods [148], and the resulting constants are combined to provide one number per crystal. These methods include the use of azimuthal symmetry of the energy flow in minimum bias events (" ϕ symmetry"), the invariant mass of photon pairs from π^0 and η decays, and the E/p ratio of isolated electrons from $W \rightarrow e\nu$ and $Z \rightarrow e^+e^-$ decays. The residual miscalibration in the ECAL barrel after the combination of the three methods is shown in Fig. 4.2 as a function of pseudorapidity [151]. The dataset used is 2.6 fb^{-1} collected in 2015. The red points show the residual miscalibration of the intercalibrations derived with LHC Run1 data, extrapolated to 2015 using the laser monitoring system response, and corrected using the ϕ -symmetry of the low energy deposits in the 2015 dataset. The green points show the residual miscalibration of the intercalibration

constants obtained using photons from $\pi^0 \rightarrow \gamma\gamma$ decays. The blue points show the residual miscalibration of the intercalibration constants obtained using electrons from W and Z decay. The black points represent the residual miscalibration of the combination of the three methods (weighted average).

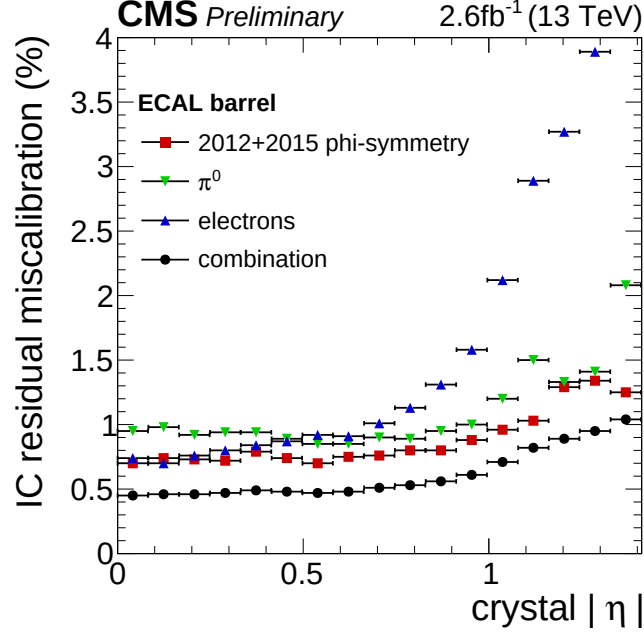


Figure 4.2: Inter-calibration precision using 2015 dataset of 2.6 fb^{-1} [151].

The combined intercalibration precision is 0.5% for central EB crystals ($|\eta| < 1$), and it goes up to 1% for the rest of the EB up to $|\eta| = 1.48$. In EE the precision, measured using 7 TeV data [152], was found to be 1.5% for $1.6 < |\eta| < 2.3$ and better than 2% up to the limit of the electron and photon acceptance at $|\eta| = 2.5$. The variation of the precision with pseudorapidity arises partly from the size of the data sample, and partly from the amount of material in front of the ECAL.

ECAL alignment. Electron identification relies upon matching the measurements in the ECAL and the tracker to better than 0.02 radians in ϕ and 4×10^{-3} units in η [153]. In addition, the accurate position measurement of photons impacting on the calorimeter is used to determine their direction with respect to the collision vertex. The accuracy of the measurement of the opening angle between the two decay photons from the Higgs boson contributes to its reconstructed invariant mass resolution. The precise alignment of the ECAL within CMS is therefore necessary to achieve the required position measurement resolution. The relative alignment of the ECAL with respect to the tracker is performed using electrons from W and Z boson decays.

Energy corrections. Particle-level energy corrections $F_{e,\gamma}$ are applied on top of calibrated clusters to account for the effects of the material upstream of the ECAL as well as local shower containment effects and other geometrical and particle-dependent factors. Corrections have been derived separately for electrons and photons by means of a Monte-Carlo-driven multivariate analysis. Input variables include shower shape information, the shower position within the ECAL and CMS, and global event variables sensitive to pileup. Energy corrections closely follow the distribution of the material budget in front of the ECAL, shown in Fig. 3.4. They are sizable and are up to 6%/10% for low/high showering electrons in the pseudorapidity region $1 < |\eta| < 2$, where the tracker material has a thickness of $\sim 2 X_0$ equivalent.

Energy resolution. The energy resolution achieved with the fully calibrated and corrected clusters has been compared between collision data and Monte Carlo simulation. The ECAL response in the simulation is tuned to match test beams results, including a detailed description of the single channel noise, single-channel response spread corresponding to the estimated residual mis-calibration, and a constant term of 0.3%. The few non-operational channels are also simulated. The electron energy resolution is estimated from the $Z \rightarrow e^+e^-$ peak width using an unbinned maximum likelihood fit to the invariant mass distribution of e^+e^- pairs. The energy scale and resolution of each electron is allowed to float in the fit. The results are obtained in bins of $|\eta|$. Fig. 4.3 shows the energy resolution obtained, using 2015 data sample corresponding to an integrated luminosity of 2.5 fb^{-1} , on the left for low bremsstrahlung electrons ($R_9 > 0.94$) and on the right for high bremsstrahlung electrons ($R_9 < 0.94$)¹ [154]. For low bremsstrahlung electrons, in the central barrel ($|\eta| < 1$) the energy resolution is around 1.5%-1.8% and in the endcaps it varies between 3% and 5%, for high bremsstrahlung electrons, in the central barrel ($|\eta| < 1$) the energy resolution is around 2%-2.5% and in the endcaps it varies between 3.5% and 5%.

4.2 ECAL laser monitoring system

4.2.1 General description

The laser monitoring system is illustrated for the barrel geometry in Fig. 4.4 and consists of the following elements:

- Light injection system, which consist of 3 lasers. One laser produces the light at the nominal wavelength of 447 nm (blue light), to follow the crystals radiation damages. The second laser produces the green light at 495 nm, which allows systematic studies of the calorimeter's evolution. The third laser works at wavelength of 796 nm (near

¹Low bremsstrahlung electrons are electrons with small interaction with the upstream material, and high bremsstrahlung electrons are electrons that have emitted high energetic bremsstrahlung photons.

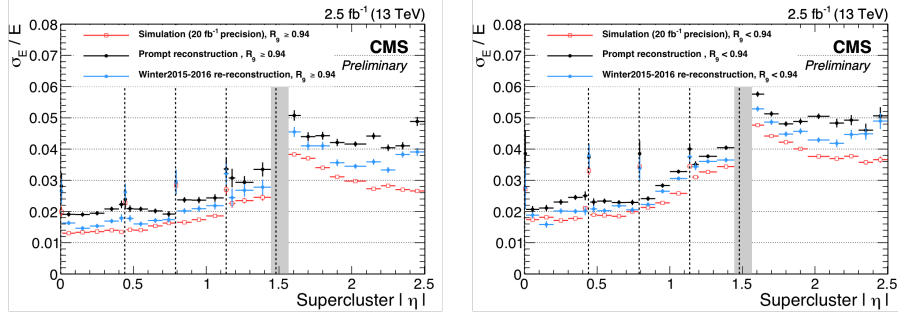


Figure 4.3: The relative electron ECAL energy resolution (σ_E/E) in bins of $|\eta|$ for the barrel and the endcaps [154]. The resolution is shown for low bremsstrahlung electrons on the left and for high bremsstrahlung electrons on the right. The resolution is plotted separately for data (2.5 fb^{-1}) and simulation events.

infra-red), which is less sensitive to radiation damages than the blue and should be able to disentangle the electronics instability from the radiation induced fluctuations.

- One attenuator.
- One switch which directs the pulses to optical fibres distributing the light to the 90 calorimeter elements: 72 half supermodules in the barrel, and 18 quarter dees in the endcaps. Their geometry is shown in Fig 4.5, where one can see one supermodule and one dee.
- A primary optical fibre distribution system.
- A two-level fibre distribution system mounted on the detector.
- PN diodes, two PN diodes per group of 200 crystals.
- APDs, two per each crystal.

The sequence of the laser monitoring system operation is the following:

- A light injection system produces the light.
- The light passes through an attenuator, otherwise the signal would be saturating the PN diodes.
- A switch at the source directs the pulses to optical fibres distributing the light to the selected calorimeter element.
- A primary optical fiber distribution system transports the pulses over a distance of 95 to 130 m to each calorimeter element mounted in CMS, located in the experimental cavern.

- A two-level distribution system mounted on the detector sends the pulses to the individual crystals and to the radiation-hard PN photo-diodes. All crystals have a pair of APDs glued to the rear face. They collect the light passing through the crystals. The design of light distribution in one barrel supermodule is shown in Fig. 4.6. The signal is transferred to the ECAL data acquisition system (DAQ) by the very front-end electronics (VFE). The front-end module (FEM) contains the PN photo-diodes and the front-end electronics (FEE), which is connected to the monitoring electronics read-out module (MEM). The last one sends the signal read-out of the PN diodes to the ECAL DAQ.

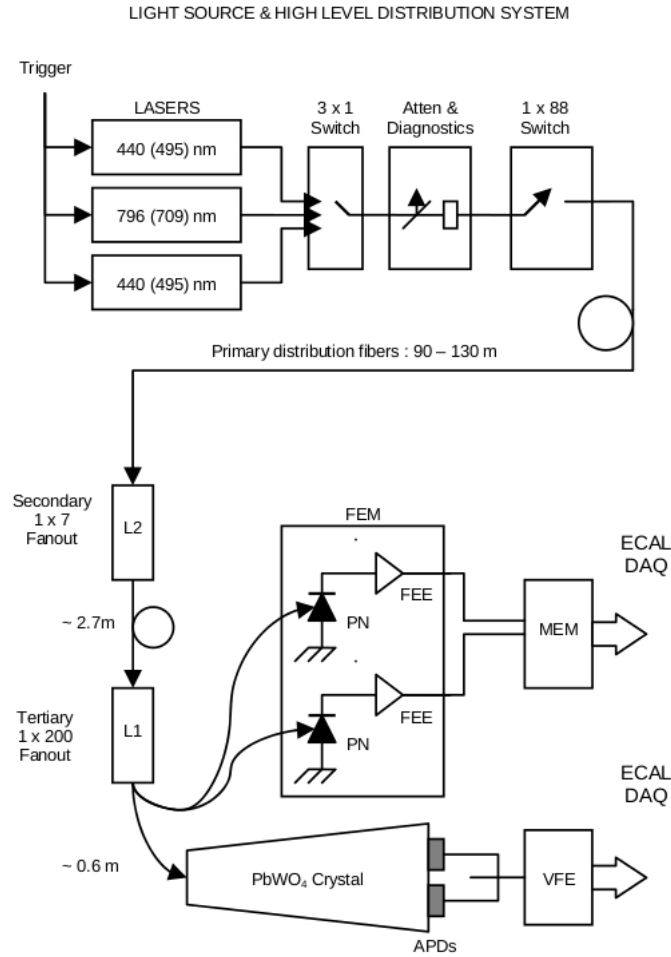


Figure 4.4: The laser monitoring system in the CMS ECAL barrel [149].

The total attenuation of the light distribution system (from the laser source to crystal front face) is measured to be 69 dB.

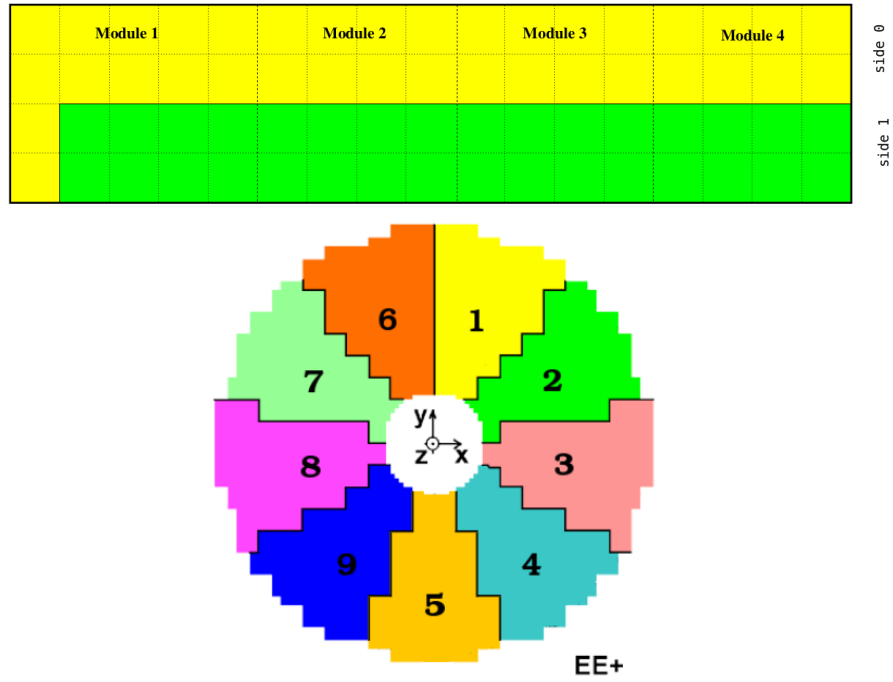


Figure 4.5: Top: Barrel laser monitoring regions in 1 supermodule. The top part (side 0) consists of 900 crystals and the bottom one (side 1) consists of 800 crystals. Bottom: End-cap laser monitoring regions in 1 dee. The regions are separated by black lines and numbered from 1 to 9.

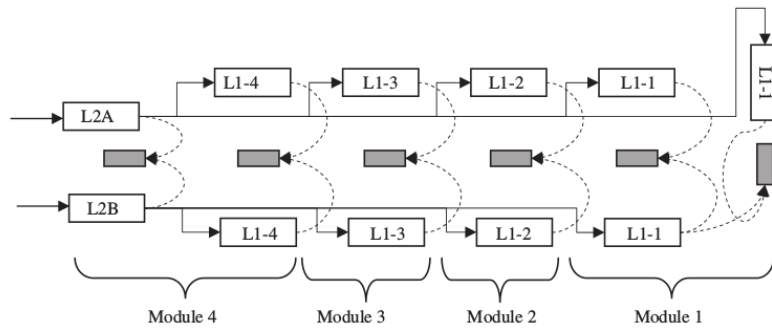


Figure 4.6: Schematic of the ECAL barrel supermodule light distribution [149]. Laser pulses (entering at left) are sent to either one of two secondary fan-outs, and are distributed to crystals along the left or right half of the supermodule's longitudinal axis via four or five tertiary fan-outs located in each of the four modules. The FEM units housing pairs of reference PN photodiodes monitoring each fan-out are shown shaded.

In order not to interfere with the ECAL performance during physics collisions at LHC, the laser pulses are injected during $3 \mu\text{s}$ gaps every $90 \mu\text{s}$ in the LHC beam structure. Running the laser source at 100 Hz means injecting laser pulses in 1% of the available gaps. The laser monitoring system independently measures the injected light for each pulse distributed to a group of typically 200 crystals using pairs of radiation hard PN photodiodes read-out via dedicated front-end electronics. The crystal optical transmission corrections are made using the ratio of the crystal's APD (VPT) response normalized by the associated group PN response. The system continuously cycles over the calorimeter elements, giving a transparency measurement. One transparency measurement is an average of 600 pulses delivered every 40 minutes for every crystal. In 40 minutes cycle blue and green lasers are used to monitor the transparency (the delay of the laser monitoring region change is $\sim 4 \text{ s}$), and also LED light that is needed to stabilize the VPT's response.

4.2.2 Laser pulse reconstruction

To reconstruct the laser signal from APD or PN, it is necessary to know their electronics response, called signal pulse response (SPR). Using it along with the laser signal shape (measured by the MATACQ module), it is possible to prepare the amplitude reconstruction algorithm to any change in the laser behavior. Extracting the SPR of the electronics is done by deconvolution techniques, as described in reference [155]. The dedicated data was taken to obtain SPR for every crystal. The typical SPR is shown in Fig. 4.7.

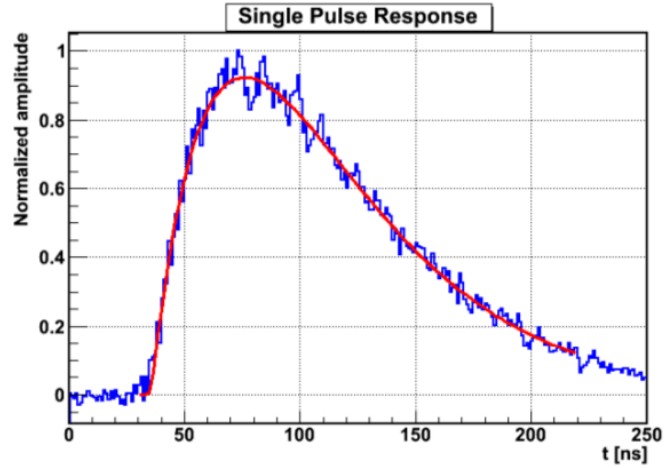


Figure 4.7: Single pulse response of APD obtained by deconvolution [155].

The first step of the laser pulse reconstruction is to build the laser pulse shape, using the MATACQ data. For each event, we compute the peak position and we construct a profile by merging all the individual pulses shifted in time to superimpose the individual

peaks, as shown in Fig. 4.8 for the APD signal. From this profile, we extract the signal part in a time window of 250 ns (we can extend it, assuming a regular exponential decay tail if needed). Then we build the expected shape for each channel by convoluting each SPR with the laser shape. Finally, we fit the expected shape on the time samples (10 samples for APD and 50 samples for PN) and extract the amplitude and the time of the amplitude, as shown in Fig. 4.9 on the left for APD and on the right for PN. The convolution procedure is not the same for APD and PN due to different electronic circuits.

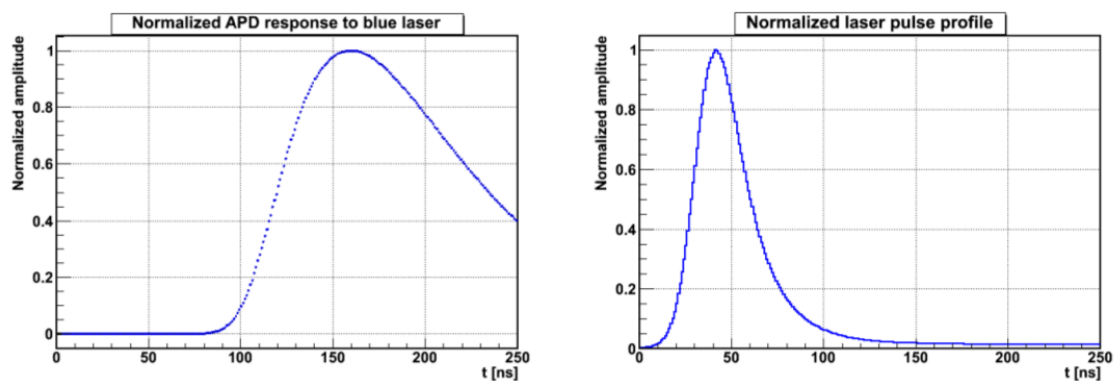


Figure 4.8: Typical pulse shape obtained for APD using the blue laser at P5 (left) and typical laser pulse shape (right) [155].

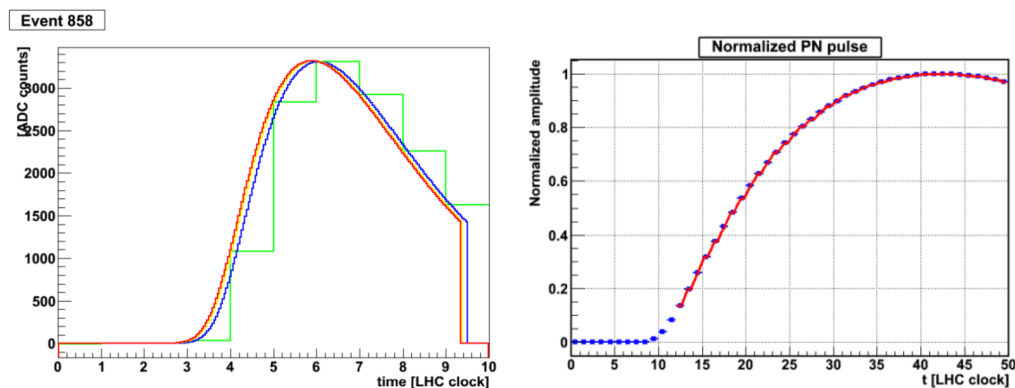


Figure 4.9: Left: Typical 10-sample event superimposed with the best adjustment of the expected APD pulse shape. Right: 50-sample event superimposed with the best fit for the PN pulse shape. Taken from [155].

The normalized laser response APD/PN of each channel is obtained by dividing its APD maximum amplitude by the commonly shared PN maximum amplitude on an event by event basis. Afterwards, an average of the 600 events for the laser run is computed.

Due to the PN shaping time of about $1 \mu\text{s}$, the PN amplitude is almost insensitive to laser pulse width variations, whereas the APD signals are quite sensitive. As a consequence, the ratios APD/PN are dependent on the laser pulse shape variation (width, tails, etc). Fig. 4.10 shows the variation of the expected APD signal (right) when the width of the laser pulse is varying (left). These laser shapes have been obtained by changing the current of the pumping laser. Two effects have to be taken into account: the fact that the shape of the APD is changing in correlation with the laser width means that the pulse amplitude reconstruction algorithm has to be adaptive to avoid bias, and the fact that the measured amplitude, even unbiased, is not directly related to the laser pulse energy and also depends on the laser width. These dependencies must be corrected in the laser monitoring analysis in order to avoid biases in the crystal transparency correction.

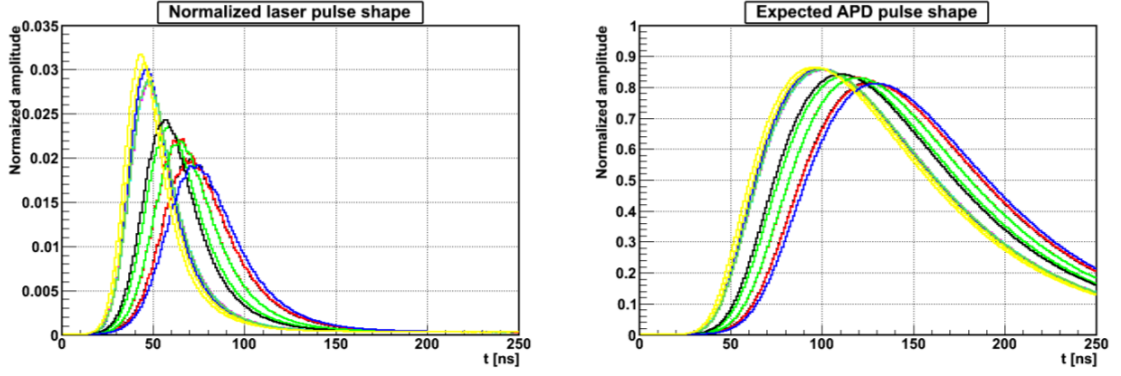


Figure 4.10: Left: Variation of laser signal normalized to a constant pulse energy. Right: Corresponding expected variation of the APD shape. Taken from [155].

Knowing the SPR for APDs and PNs and the laser shape, we can compute what would be the ratios APD/PN for the observed laser signal and give appropriate correction factor to apply to data. Assuming that all SPRs are normalized to 1, the correction factor is equal to:

$$corr = \frac{\max(SPR(APD) * laser)}{\max(SPR(PN) * laser)} \quad (4.4)$$

where $*$ means the convolution product. The correction is applied by dividing the result of the pulse amplitude reconstruction by this factor. This gives what would have been the value of APD/PN if we had used a Dirac delta pulse from the laser. APD/PN before corrections with the variation of $\sim 7\%$ is shown on the left in Fig. 4.11 and after corrections the variation goes down to 1.5×10^{-3} , and it is shown on the right. This method corrects the APD/PN ratios for all changes in the laser behaviour.

Such refined pulse reconstruction and corrections are necessary to get the designed

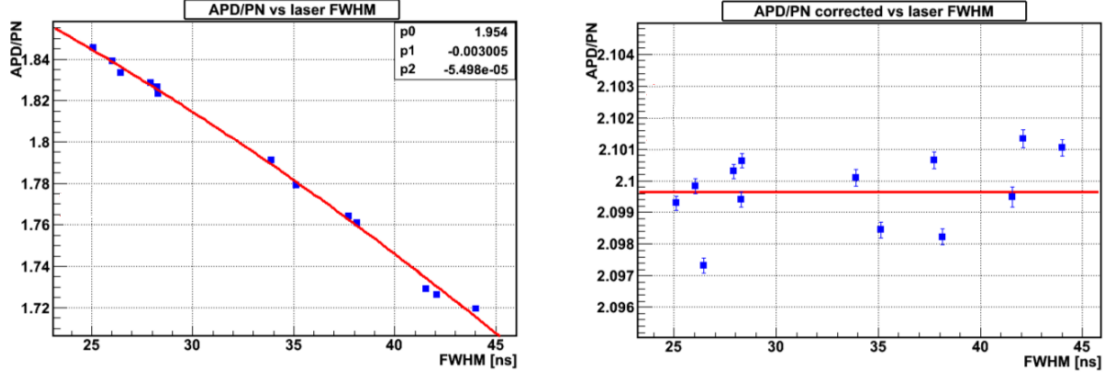


Figure 4.11: Result of the measurement of the ration APD/PN with respect to the laser width: before correction (left) and after correction(right) [155].

precision of $\sim 0.2\%$.

4.2.3 Laser monitoring performance

Changes in the crystal transparency due to radiation damage do not affect the amplitude of the APD signal for an electromagnetic shower in exactly the same way as it affects the APD signal for injected laser pulses. This is principally due to the different mean light paths in the two cases, as shown in Fig. 4.12. The effect can be seen in Fig. 4.13, where single crystal relative responses to both 120 GeV electrons S/S_0 (measured in low intensity electron runs) and to injected laser light from the monitoring system R/R_0 (measured in alternated laser runs) are shown during an irradiation run at 0.15 Gy/h. The measurements were taken both during irradiation phase ending at 15 hrs in the figure, and during the subsequent recovery phase. The initial values S_0 and R_0 correspond to measurements taken prior to the irradiation. The relationship between the two responses can be modelled for small variations as

$$\frac{S}{S_0} = \left\{ \frac{R}{R_0} \right\}^\alpha \quad (4.5)$$

where α is a parameter that varies from crystal to crystal. The excellent agreement between this description and data can be seen in Fig. 4.13, where α is 1.6. Data taken during irradiation and recovery phases follow the same slope. Typical fit precision is 3%, and the intrinsic dispersion for the crystals is deduced to be about 6%. Thus for crystals in the barrel showing a decrease in signal size of 5% a single value of α can be used to correct the loss. This allows to keep the constant term in the energy resolution at 0.3%, within the design specifications.

The variation of the crystal responses has been measured during 2011-2012 (LHC Run

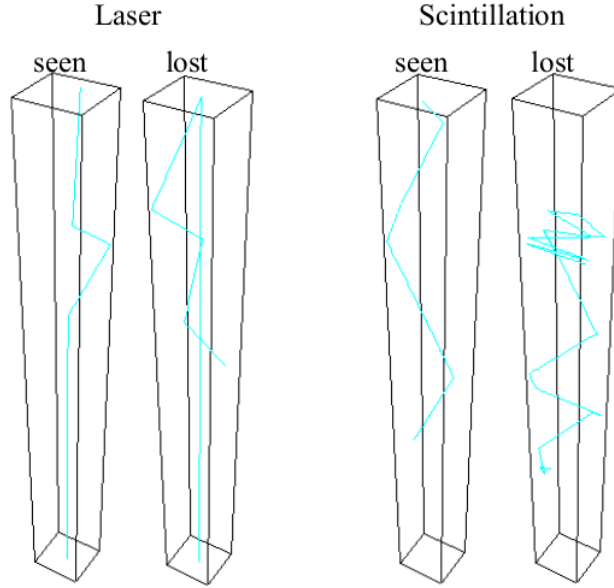


Figure 4.12: Photon path comparison in ECAL crystal, using laser and real data. The seen and lost paths for both cases are shown as well.

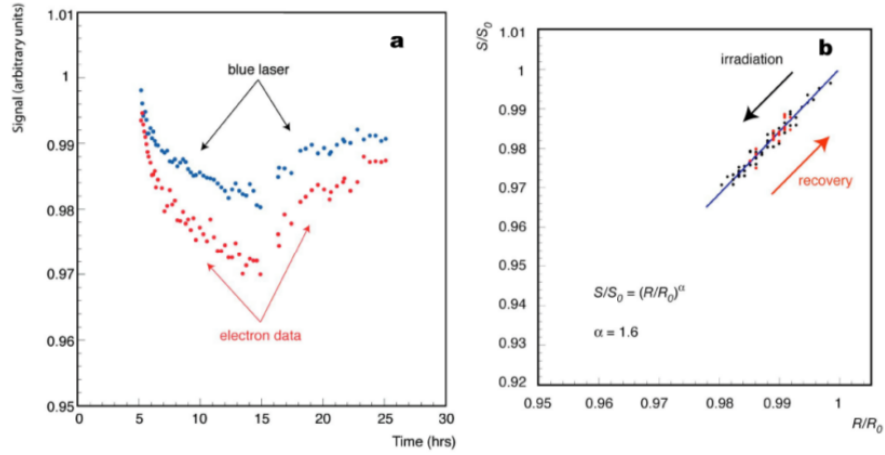


Figure 4.13: Irradiation with 120 GeV electrons and recovery for a single $PbWO_4$ crystal [149]. Left: the upper curve shows the APD response to laser injection at 447 nm (blue laser), and the lower curve shows the response to 120 GeV electrons. Right: the signal response S/S_0 versus the laser response R/R_0 for the same data, the line shows the fit for $\alpha = 1.6$.

I) and 2015-2016 (LHC Run II), and is shown in Fig 4.14. Relative response to laser light injected in the ECAL crystals, is measured by the ECAL laser monitoring system. It is averaged over all crystals in bins of pseudorapidity, for the 2011, 2012, 2015 and 2016 data taking periods, with magnetic field at 3.8 T. The response change observed in the ECAL channels is up to 10% in the barrel and it reaches up to 50% at $\eta \sim 2.5$, the limit of the tracker acceptance. The response change is up to 90% in the region closest to the beam pipe. The recovery of the crystal response during the Long-Shutdown-1 (from 02.2013 to 05.2015) period is visible. The response was not fully recovered, particularly in the region closest to the beam pipe. These measurements are used to correct the physics data. The bottom plot shows the instantaneous LHC luminosity delivered during this time period.

A few independent methods are used to validate the corrections for transparency loss with collisions data: monitoring the stability of the reconstructed invariant mass of $\gamma\gamma$ pairs in $\pi^0 \rightarrow \gamma\gamma$ decays and measuring the stability of the E/p distribution for isolated electrons, where E is the energy measured in the ECAL and p is the momentum measured in the tracker. The stability plot obtained with this second method using 2015 dataset is shown in Fig. 4.15, where the E/p relative scale versus time is shown before and after applying the laser corrections. The reached stability is about 0.14% in the ECAL barrel.

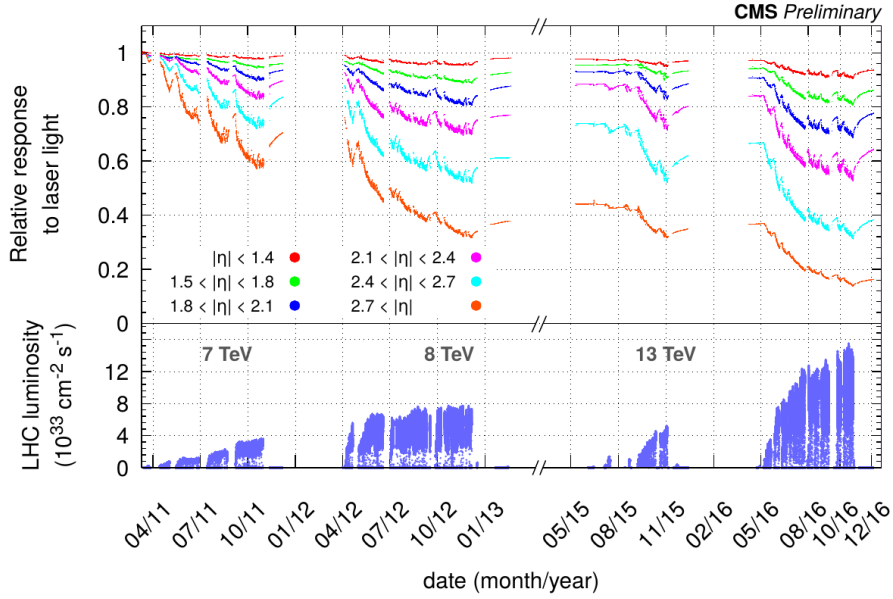


Figure 4.14: Relative response to laser light injected in the ECAL crystals, measured by the ECAL laser monitoring system, averaged over all crystals in bins of pseudorapidity, for the 2011-2012 (LHC Run I) and 2015-2016 (LHC Run II) data taking periods, with magnetic field at 3.8 T. The bottom plot shows the instantaneous LHC luminosity delivered during this time period. Taken from [156].

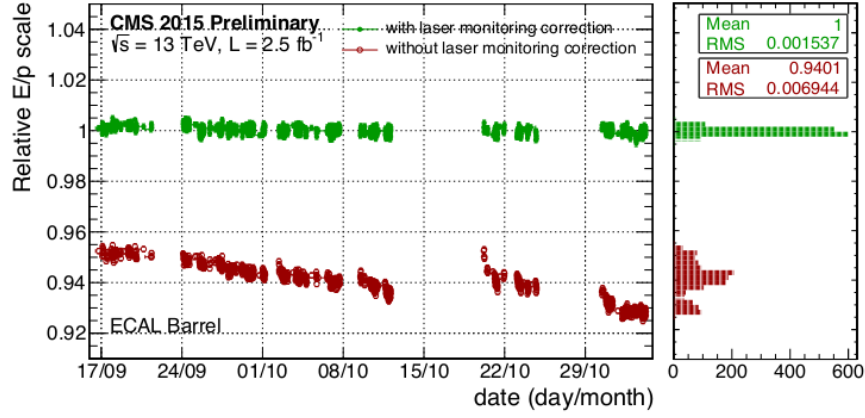


Figure 4.15: The ratio of the electron energy E , measured in the ECAL barrel, to the electron momentum p , measured in the tracker with 2015 data.

4.3 Possible ECAL laser monitoring system upgrade for LHC Phase II

The ECAL detector components were designed and tested to acquire an integrated luminosity of $\sim 350 \text{ fb}^{-1}$ at the LHC in 10 years of data-taking (2011-2021). The LHC Phase II is expected to provide an integrated luminosity of $\sim 3000 \text{ fb}^{-1}$ by 2033. The average number of concurrent interactions per bunch-crossing (pileup) in CMS will be up to 200, and the radiation levels in the detector will be six times higher than for the nominal LHC design. The expected radiation dose after collecting 3000 fb^{-1} is shown in Fig. 4.16.

The Higgs boson mass at 125 GeV constrains the particles produced in its decays to relatively low transverse energies. The detectors must therefore maintain the ability to detect low transverse energy objects. The $H \rightarrow \gamma\gamma$ channel needs a 1% mass resolution to detect the Higgs boson mass peak over a large continuous background. This decay mode is particularly important as it allows a full kinematic reconstruction of the Higgs boson decay, and it will still be of interest at the LHC Phase II, particularly for associated or double Higgs boson production modes ($t\bar{t}H$, VBF, $HH \rightarrow b\bar{b}\gamma\gamma$).

During the LHC Run I, the corrections for response changes have reached a precision of about 0.1%. This requirement may be relaxed at the LHC Phase II, as the conditions will be quite different. In particular the noise levels in the photo-detectors and the fluctuations of the ECAL electromagnetic deposits due to pileup contamination, both contributing to the final energy resolution, will be larger. As a result a requirement at the level of 1% is sufficient.

The main concern for the current monitoring system is its radiation tolerance, which is most important for the light distribution system, the PN diodes, and FEM, placed at the

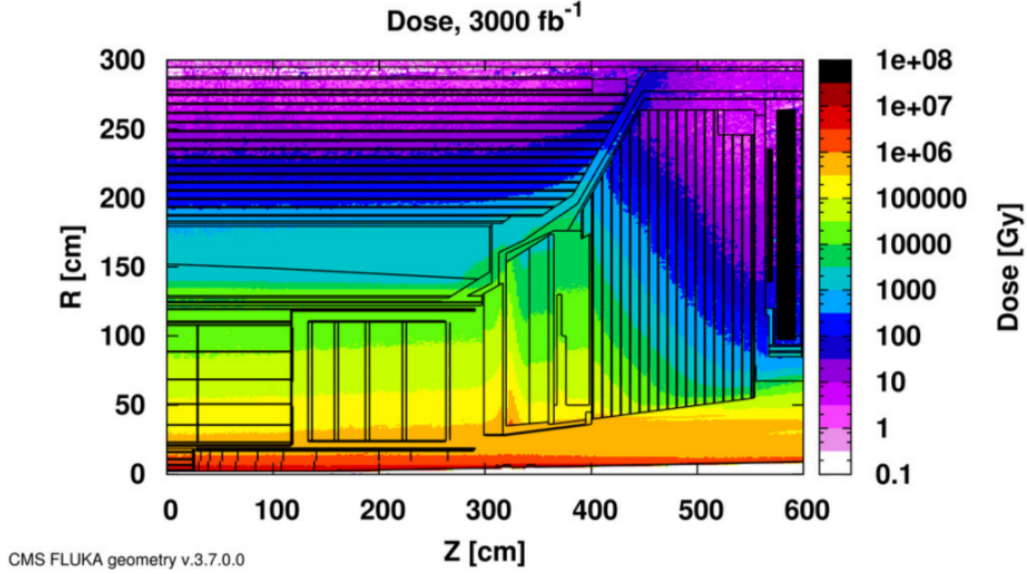


Figure 4.16: Expected radiation dose in the central calorimeters (ECAL Barrel, HCAL Barrel, ECAL Endcap, and HCAL Endcap) after collecting an integrated luminosity of 3000 fb^{-1} .

front of the supermodules. The aim of the upgrade of the ECAL laser monitoring system is to monitor the crystal response changes with the precision required by the running conditions foreseen for the LHC Phase II.

4.3.1 Description

The main change foreseen for the laser monitoring system is to measure the reference value directly at the light injection point in the laser barrack, located in the service cavern, rather than at the last distribution step at the front of the supermodules. The modified part of the laser monitoring system is shown in Fig. 4.17.

A test bench was built in CEA Saclay. It consists of two boxes called "SpyBoxes", with 88 (2x44) optical fibres. Those boxes are added to the nominal laser monitoring system, as shown in Fig. 4.18, to be able to "spy" the laser light sent to the ECAL. To spy the light, the isolation is removed locally from the optical fibre and the PiN diode is seeing the light from the stripped part, as shown in Fig. 4.19. One of the "SpyBoxes" is equipped with 11 Hamamatsu Si PiN photo-diodes [157]. They are connected in parallel to the multiple gain pre-amplifier (MGPA) and powered with 9 V battery. We study the measurements obtained with these 11 PiN diodes, to know the precision of the new laser monitoring system as described in section 4.3.3.

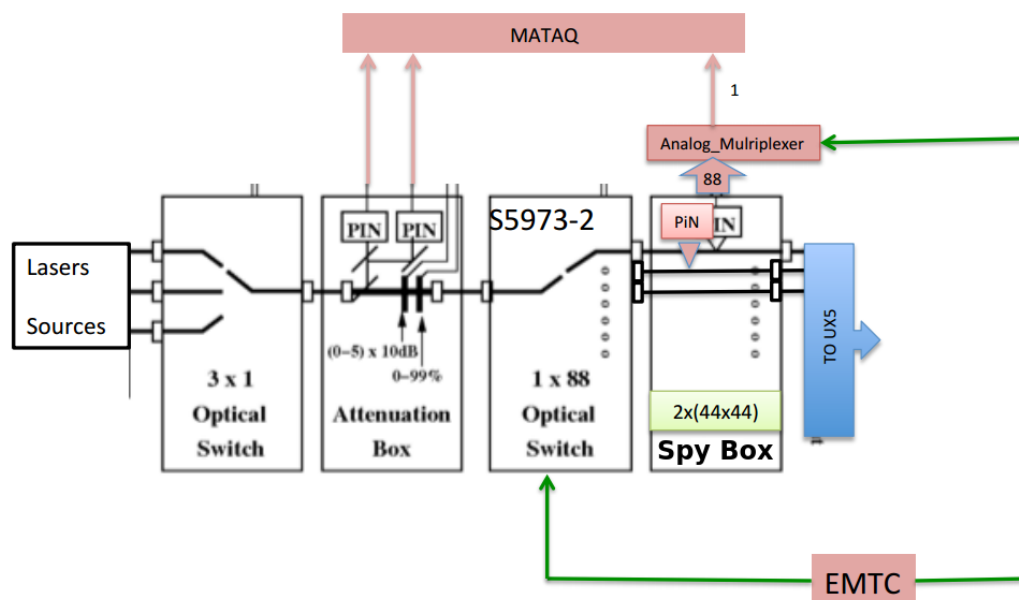


Figure 4.17: The laser monitoring schema with additional "SpyBoxes" connected to the Switch 1x88.

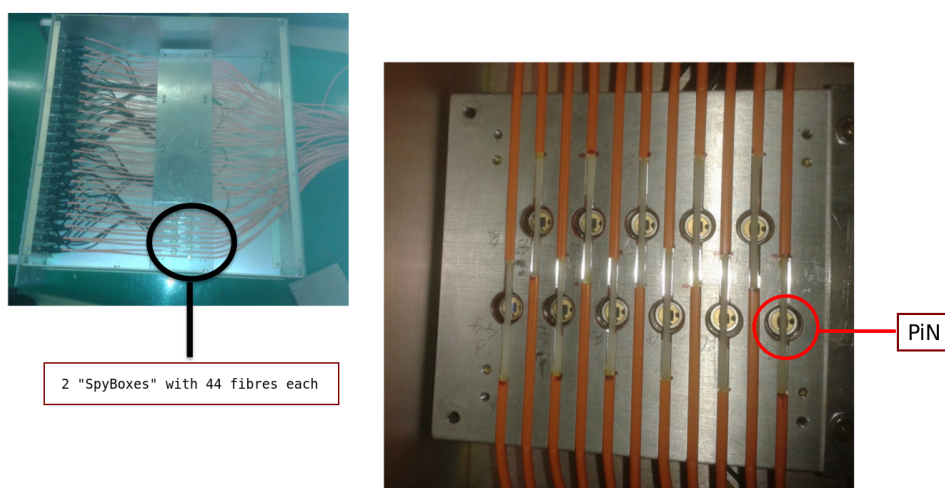


Figure 4.18: Inside the test bench "SpyBox": 11 isolated fibres and 11 PiN diodes, connected to the places where the isolation is stripped.

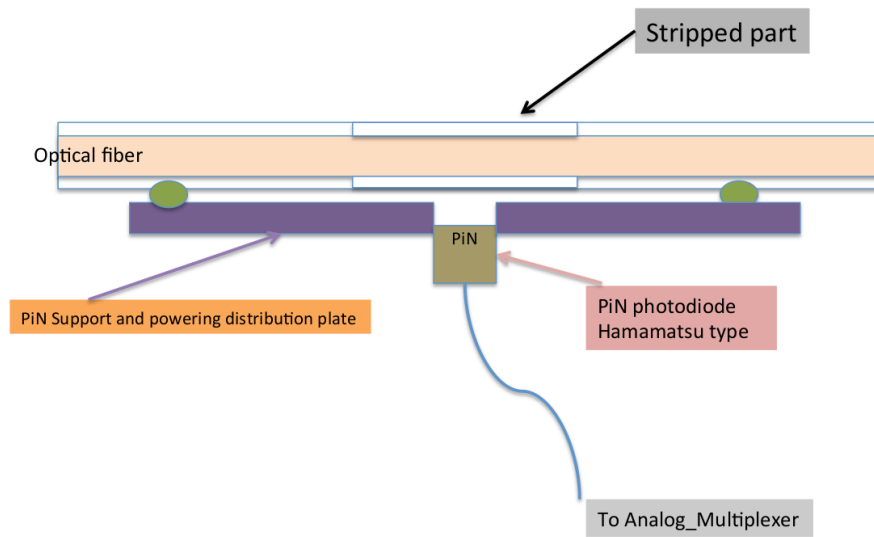


Figure 4.19: Detailed view of 1 fibre with PiN diode, glued to the place with stripped isolation. PiN diode detects the light passing through the fibre.

4.3.2 Installation into the existing laser monitoring system and first tests

I participated to the "SpyBoxes" installation in the CMS laser barrack in December 2014. Before the installation we cleaned, polished and labelled every fibre of the two "SpyBoxes". The work in progress is shown in Fig. 4.20, and all labelled fibres of the "SpyBox" in Fig. 4.21. The cleaning and polishing was also done for the fibres of the Switch 1x88. The view inside the Switch 1x88 is shown in Fig. 4.22.



Figure 4.20: "SpyBox" fibres were cleaned and polished before installation.

The configuration before installation is shown in Fig. 4.23, where the NIM logic (NIM is standard of the module) is used for laser timing and triggering. Firstly, we moved the NIM logic system down, then we installed the first "SpyBox" and L-shaped edges to perform further works inside the "Spy Box" in an easy way, and finally we installed the second "SpyBox". The final configuration is shown in Fig. 4.24. The fibres going down to the ECAL were cleaned as well and connected to the "SpyBoxes", as shown in Fig. 4.25.

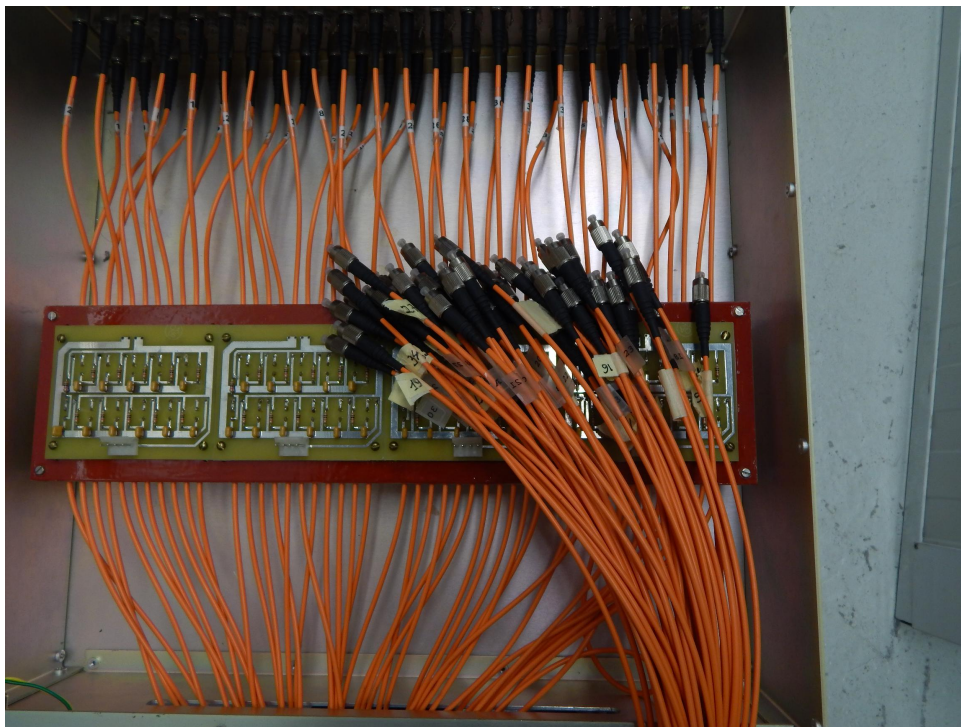


Figure 4.21: The view inside of the "SpyBox". All fibres were labelled during the installation.



Figure 4.22: The view inside of the Switch 1x88. Its fibres were cleaned and polished before installation.



power
NIM logic system
Switch 1x88
Attenuator
Switch 1x5

Figure 4.23: The configuration of the laser monitoring system in laser barrack before "Spy-Boxes" installation.



power
Spy_box (44x44) _ 1
Switch 1x88
Spy_box (44x44) _ 2
Attenuator
Switch 1x5
NIM logic system

Figure 4.24: The configuration of the laser monitoring system in laser barrack after "Spy-Boxes" installation.



Figure 4.25: The general view of the laser monitoring system in the laser barrack after "SpyBoxes" installation.

After the two "SpyBoxes" were installed, it was checked that they were not disturbing the existing laser monitoring system. In the first test, 2 multiplexing schemes were used: in parallel and with analog device ADG904 [158], shown in Fig. 4.26.

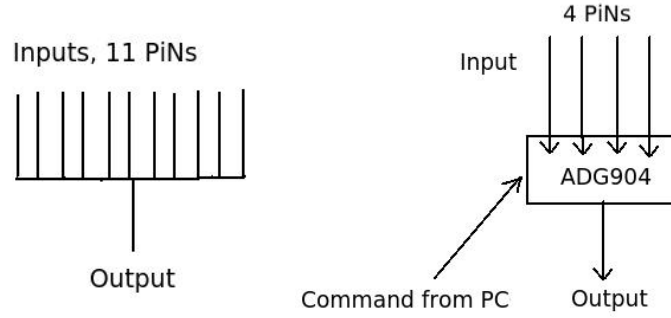


Figure 4.26: Left: Parallel multiplexing. Right: Multiplexing with ADG904.

For different fibres, different configurations were tested:

- Fibre 45 had a PiN diode connected directly to a dedicated diagnostic line (MAT-ACQ) and powered with USB. The USB was not connected during the test, as a result there was no depletion voltage applied to the PiN diode.
- Fibres 46-49 had a multiplexing scheme with ADG904.
- Fibres 50-55 had a parallel multiplexing scheme and were powered with USB.
- Fibre 56 was connected in parallel and used a 4.5 V battery.

The laser power and light attenuation had two configurations, summarized in Table 4.1.

Configuration	Laser	Internal attenuation	Remote attenuation
<i>Nominal ECAL</i>	35A	35%	5%
<i>Full power</i>	35A	no	5%

Table 4.1: The power and attenuation configurations used during first tests of "SpyBoxes". The *full power* configuration cannot be used during usual laser monitoring, as the signal would be saturating PN diodes.

We tested different configurations on two fibres (45 and 56), to check that laser light goes through the "SpyBox" and the PiN diodes work. The tests are listed below and their results are shown in Fig. 4.27:

- Test 1: Laser with blue light and full power configuration.

- Test 2: Laser with blue light and ECAL configuration.
- Test 3: Laser with blue light and full power configuration.
- Test 4: Laser with green light and full power configuration.

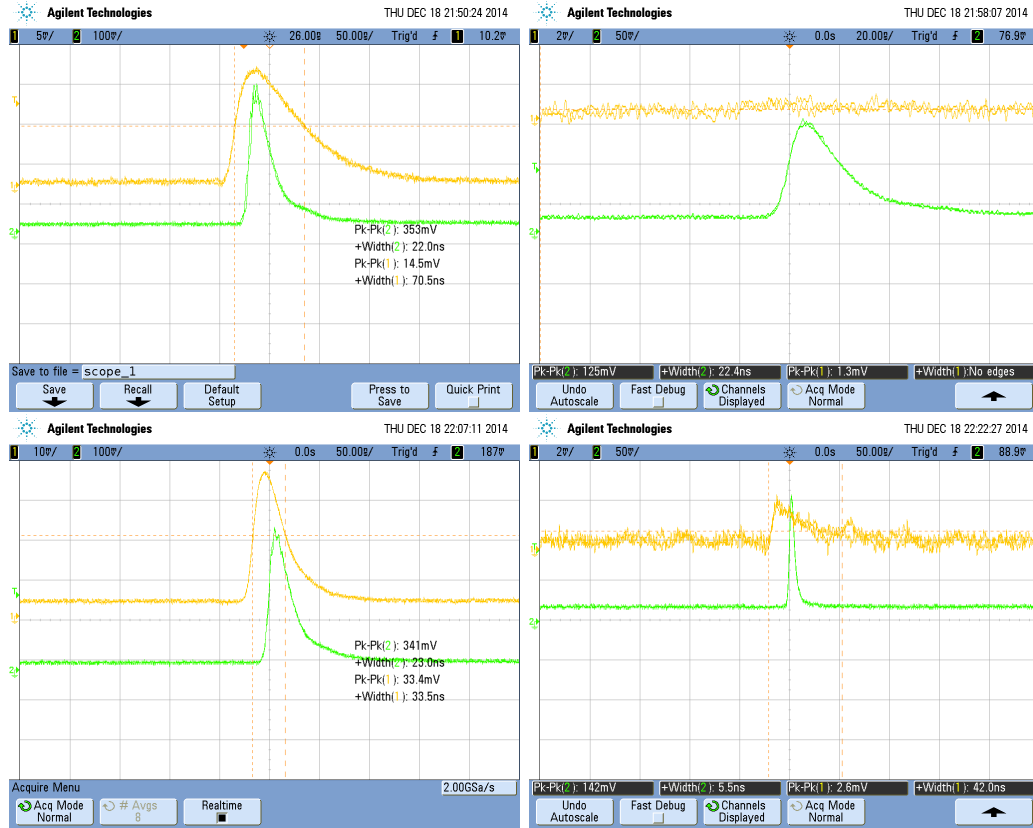


Figure 4.27: Top left: Test 1. The light from the attenuation box: 353mV; 22ns (green). The light from the "spy" fibre 45: 14.5mV; 70.5ns (yellow). Top right: Test 2. The light from the attenuation box: 125mV; 22.4ns (green). The light from the "spy" fibre 45: 1.3mV; no edges (yellow). Bottom left: Test 3. The light from the attenuation box: 341mV; 23ns (green). The light from the "spy" fiber 56: 33.4mV; 33.5ns (yellow). Bottom right: Test 4. The light from the attenuation box: 142mV; 5.5ns (green). The light from the "spy" fibre 45: 2.6mV; 42ns (yellow).

The results show that the light successfully passes via fibres in "SpyBox" and we get a read-out from PiN diodes. To get the proper signal, PiN diodes have to be directly polarized².

²When the diode conducts and it is directly polarized the switch closes the circuit and the current follows the entire circuit, but if the diode is reverse polarized the switch breaks the continuity of the circuit

The light power was checked at the source, and after it passes through one of the fibres in the laser barrack. The latter was measured before and after "SpyBox" installation. In both cases the light power at the source was 96 mW. The light power after passing a fibre in Switch 1x88 was 4.2 mW in November 2014. After the "SpyBoxes" installation and cleaning all fibres, the light power after passing a fibre in the "SpyBox" became 4.96 mW. Despite the introduction of the "SpyBoxes", that adds one meter of fibre and 2 connectors, the net light budget increased because of cleaning and polishing of fibres.

During the LHC Run II the test bench works in parallel with the existing laser monitoring system to check if the precision of such a transparency measurement is sufficient for the LHC Phase II.

4.3.3 "SpyBox" stability and precision study

The ratio of transparency measured with the "SpyBox" and with the nominal laser monitoring system $\left(\frac{APD/PiN(t)}{APD/PN(t)}\right)$ should be stable in time, if the "SpyBox" monitors the transparency well. The RMS/Mean of this ratio gives the precision of the new laser monitoring system.

Eleven laser monitoring regions were studied corresponding to the eleven PiN diodes (shown in section 4.3.1), which are named "feds". Feds numbered from 632 to 636 have two sides 0 and 1. Fed numbered 637 has only one side 0. The sides of the supermodule are shown in Fig. 4.5.

Several datasets with magnetic field on, stable room temperature and stable laser were used to study the stability and precision of the "SpyBox":

- summer data: 04.07.2015 - 27.08.2015, using blue and green lasers
- amplitude scan: a dedicated run with green laser, where the laser amplitude stays the same and attenuation varies, such that the signal amplitude on the PiN varies by ± 25 %.
- full dataset: 04.07.2015 - 30.11.2015, using only blue laser

Summer dataset and blue laser. The transparency ratio $\frac{APD/PiN(t)}{APD/PN(t)}$ and its evolution with time for the feds 632, 635 and 637 are shown in Fig. 4.28. One can see this ratio is stable with time in fed 632, which is not the case for feds 635 and 637.

The precision of the new laser monitoring system is shown by 2D maps in ϕ - η coordinates in the ECAL barrel in Fig 4.29. Feds 632 (side 0) and 634 (side 1) have a precision $< 0.15\%$, 635 (side 0) and 637 (side 0) have precision of $\approx 0.3\%$ and others of $\approx 0.2\%$.

and the current does not pass as the resistance is high. When there is no polarization a part of the current does not pass through the circuit.

To understand the origin of such a precision in feds 635 and 637, 2D plots were created for the laser amplitude (laser qmax), the laser width (FWHM), APD amplitude (APD qmax) and PIN_{Spy} amplitude (PIN qmax) as functions of the transparency ratio $\frac{\text{APD}/\text{PiN}(t)}{\text{APD}/\text{PN}(t)}$. They are shown in Fig. 4.30. One can clearly see two event populations. This effect could come either from the electronics (non-linearity effect) or from the light source. First, the light source was checked, using the green laser.

Summer dataset and green laser. The transparency ratio $\frac{\text{APD}/\text{PiN}(t)}{\text{APD}/\text{PN}(t)}$ and its evolution with time for feds 632, 635 and 637 are shown in Fig. 4.31. One can see the two peaks in feds 635 and 637. It means that the transparency measurement is not stable for those feds also with green laser.

The precision 2D map for the green laser is shown in Fig. 4.32, where one can see the same trend. The comparison of the "SpyBox" precision for blue and green laser data is shown in Table 4.2. The same 2D plots were checked with green laser data. They are shown in Fig. 4.33, where one can see the two populations of events.

green :	fed,side	precision	blue :	fed,side	precision
	632,0 and 634,1	$\approx 0.15 \%$		632,0 and 634,1	$< 0.15\%$
	635 and 637,0	$0.6 - 0.5\%$		635 and 637,0	$\approx 0.3\%$
	others	$\approx 0.25\%$		others	$\approx 0.2\%$

Table 4.2: The precision comparison using the blue and the green lasers.

As the same behaviour is seen in data produced with the green and the blue lasers, we can therefore conclude that the effect is not laser-dependent. To check if the issue originates from the amplitude variations (non-linearity), an amplitude scan with the green laser has been done.

Amplitude scan using the green laser. The dedicated data sample was taken to study potential non-linearities. During the amplitude scan the laser amplitude is constant and the attenuation is variable. The amplitude of the PiN signal varies by 25%, and it is shown in Fig. 4.34.

The transparency ratio in fed 635 side 0 and fed 637 side 0 are shown in Fig. 4.35. There is no evidence for two event populations, and thus we exclude the non-linearity effect.

The precision was checked with the full 2015 dataset, to see if the effect has changed.

Full 2015 dataset and blue laser. The transparency ratio is broken down by modules (from 1 to 4) and it is shown for feds 632, 635 and 636 in Fig. 4.36.

The precision 2D map is shown in Fig. 4.37 (left). The precision degraded comparing to the summer dataset and is also worse in Modules 3 and 4 comparing to Modules 1 and 2, as they are located at higher η , therefore get more radiation.

Full 2016 dataset and blue laser. The precision 2D map using the full 2016 dataset is shown in Fig. 4.37 (right). The precision worsens with respect to the 2015 dataset. In general the precision is worse in Modules 3 and 4 comparing to Modules 1 and 2. The comparison of the laser monitoring system precision with 2015 and 2016 datasets is shown in the Table 4.3 for different modules.

module	fed, side	precision in 2015	fed, side	precision in 2016
1	632, side 0; 634, side 1 others	0.35% 0.25-0.3%	632, side 0 others	1.8% 1.5%
2	632, 633, side 0,1; 634, side 1 636, side 1 others	0.4% 0.9% 0.25-0.3%	632, 633 side 0,1 others	2-2.2% 1.5%
3	632, side 0 635, side 0 others	0.65% 0.3% 0.5%	632, 633 side 0,1 others	2-2.2% 1.8%
4	634, side 1 632, side 0,1; 633, side 0 others	0.9% 0.5% 0.7%	634, side 1; 635, side 0,1; 636, side 0 others	2% 2% 1.8%

Table 4.3: The precision comparison using the blue laser in 2015 and 2016. In general the precision is worse in Modules 3 and 4 comparing to Modules 1 and 2, as they are located at higher η and get higher level of radiation. The precision reaches 0.9% in 2015 data and 2.2% in 2016 data.

4.4 Conclusion

In this chapter, I introduced the ECAL laser monitoring system. I have shown the possible upgrade of the laser monitoring system for the LHC Phase II, its installation and the precision of the transparency measurement. The precision is less than 1% using the 2015 dataset, but it increases to 2.2% with 2016 dataset. There seems to be a fibre ageing effect due to the radiation which is not the same in different modules. This effect has to be modelled and corrected. If after the correction, the precision is not sufficient the alternative is to upgrade the existing laser monitoring system by replacing the existing reference PN diodes (inside the CMS) by the diodes with higher radiation tolerance.

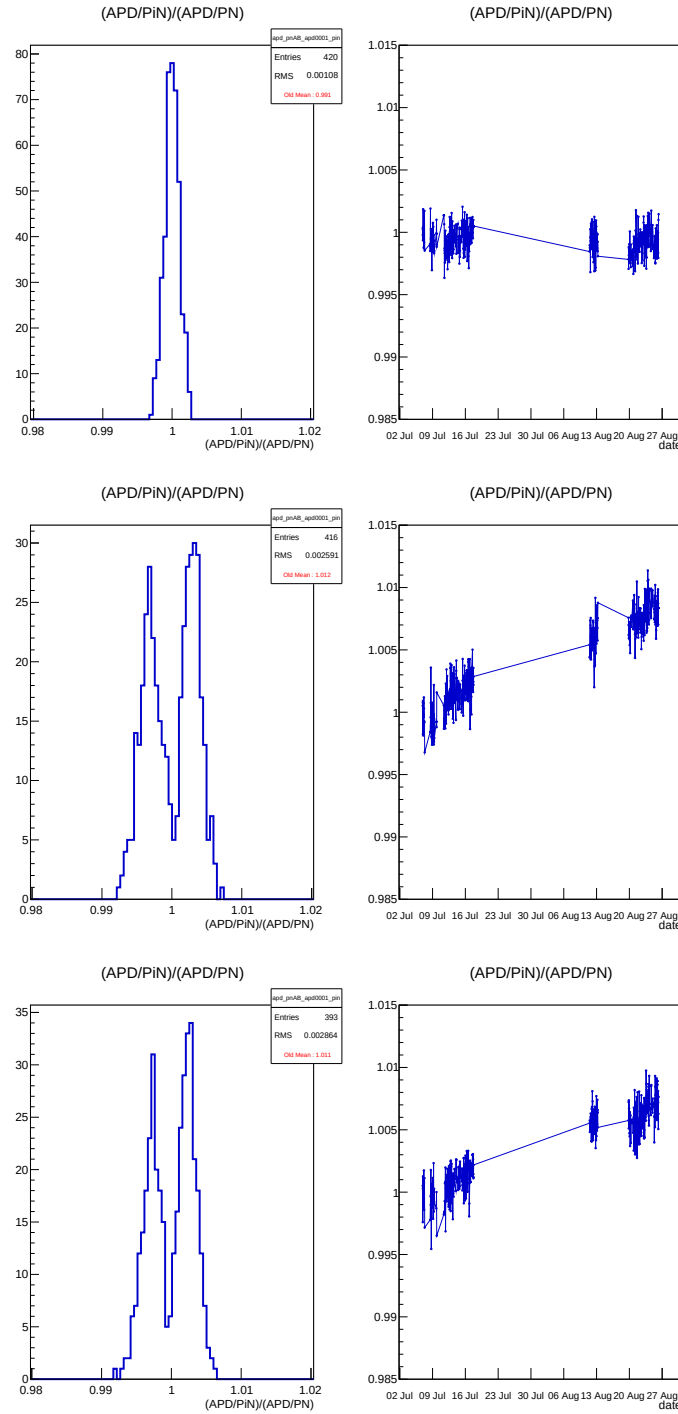


Figure 4.28: Transparency ratio (centred at 1) and its evolution with time in fed 632, side 0 (top), fed 635, side 0 (middle) and fed 637, side 0 (bottom).

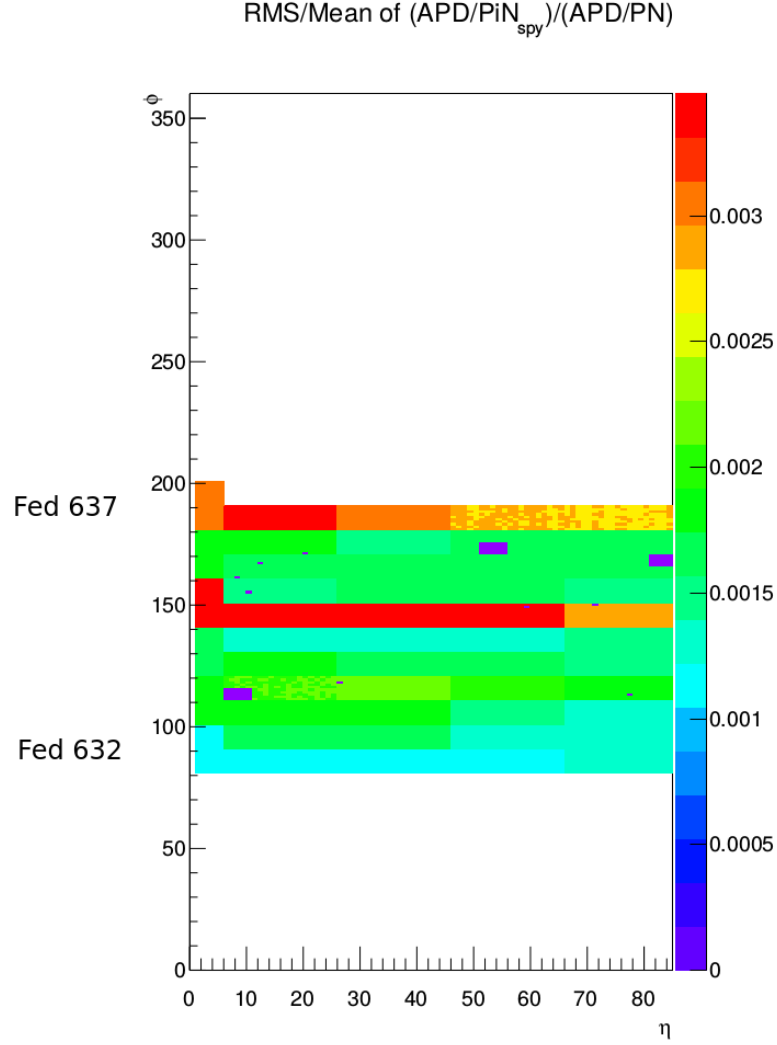


Figure 4.29: "SpyBox" precision map: 2D plot in ϕ - η coordinates in the ECAL barrel. The data are produced with the blue laser and it corresponds to the summer dataset. Studied feds correspond to the following ϕ coordinates: 632: 80° - 100° , 633: 100° - 120° , 634: 120° - 140° , 635: 140° - 160° , 636: 160° - 180° , 637: 180° - 200° .

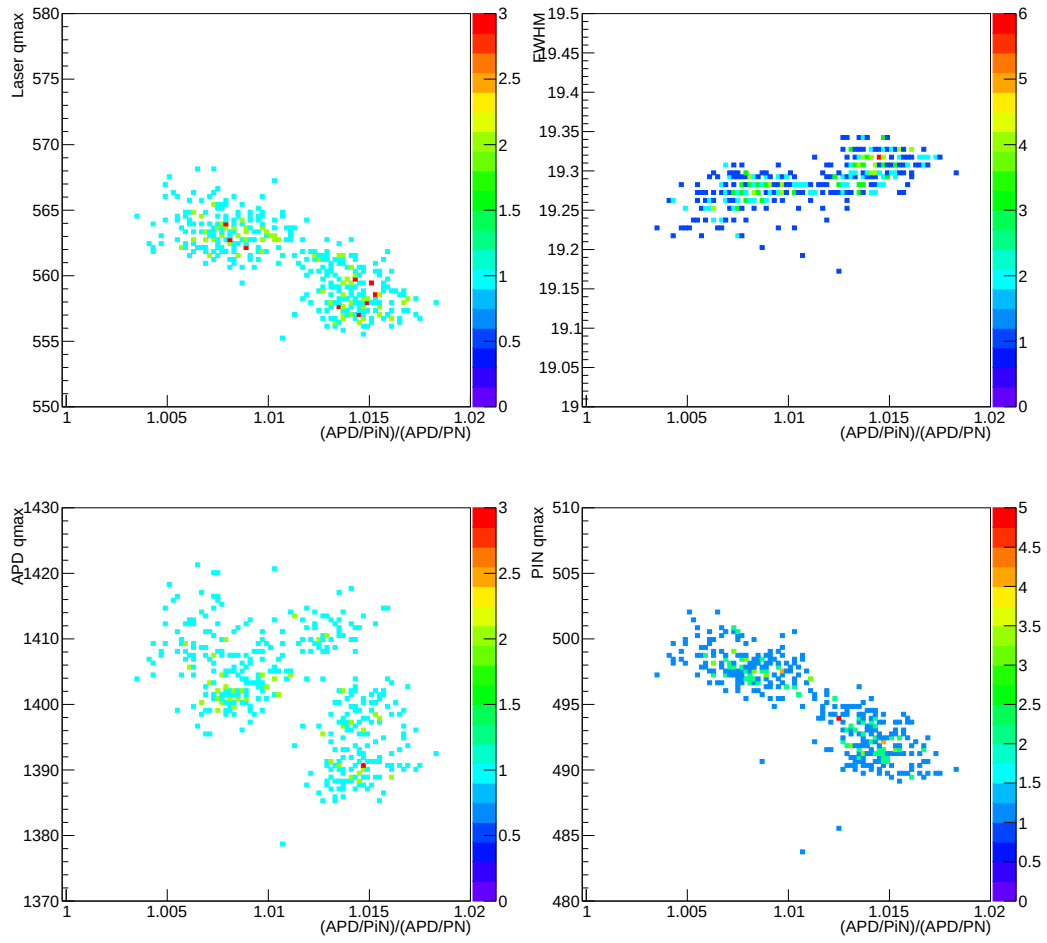


Figure 4.30: Fed 635 (both sides together). Top left: The transparency ratio as a function of the laser amplitude (laser qmax). Top right: The transparency ratio as function of the laser width. Bottom left: The transparency ratio as a function of the APD amplitude (APD qmax). Bottom right: The transparency ratio as a function of the PiN amplitude (PiN qmax).

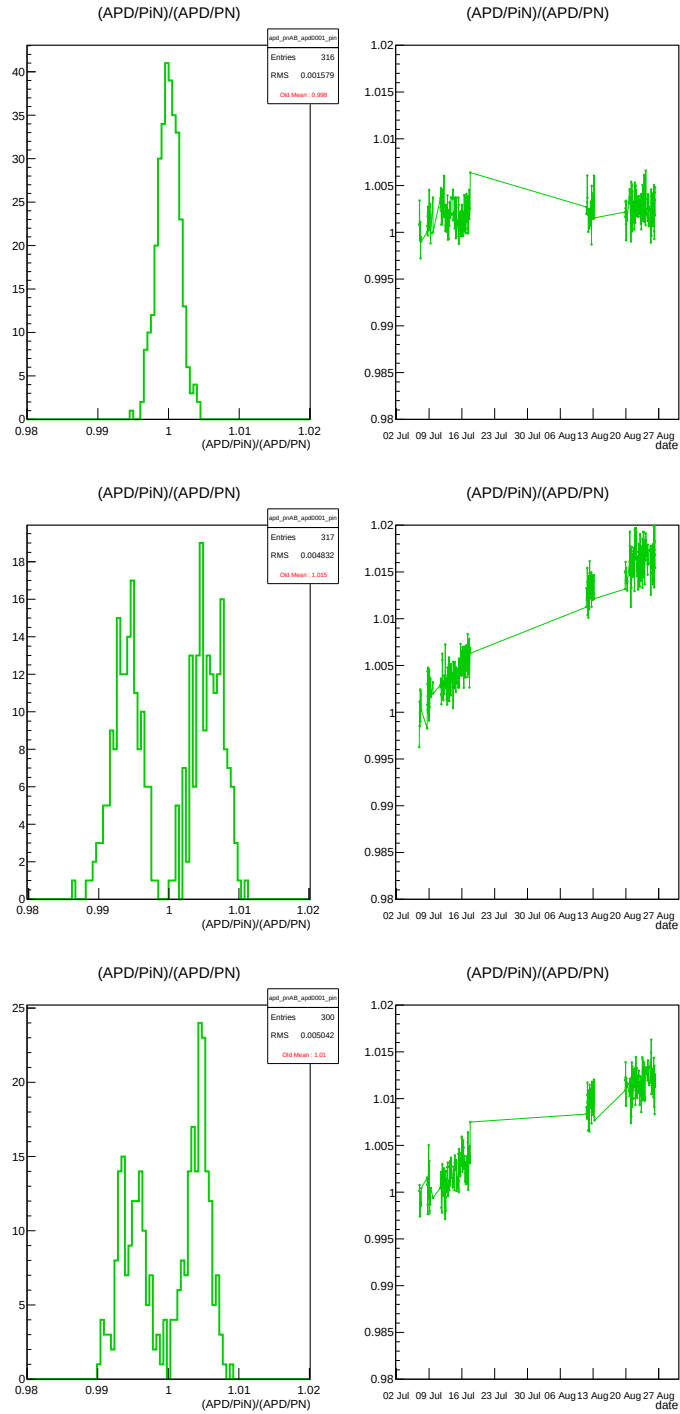


Figure 4.31: Transparency ratio, centred at 1, (left) and its evolution (right) in fed 632 side 0 (top), fed 635, side 0 (middle), and fed 637, side 0 (bottom).

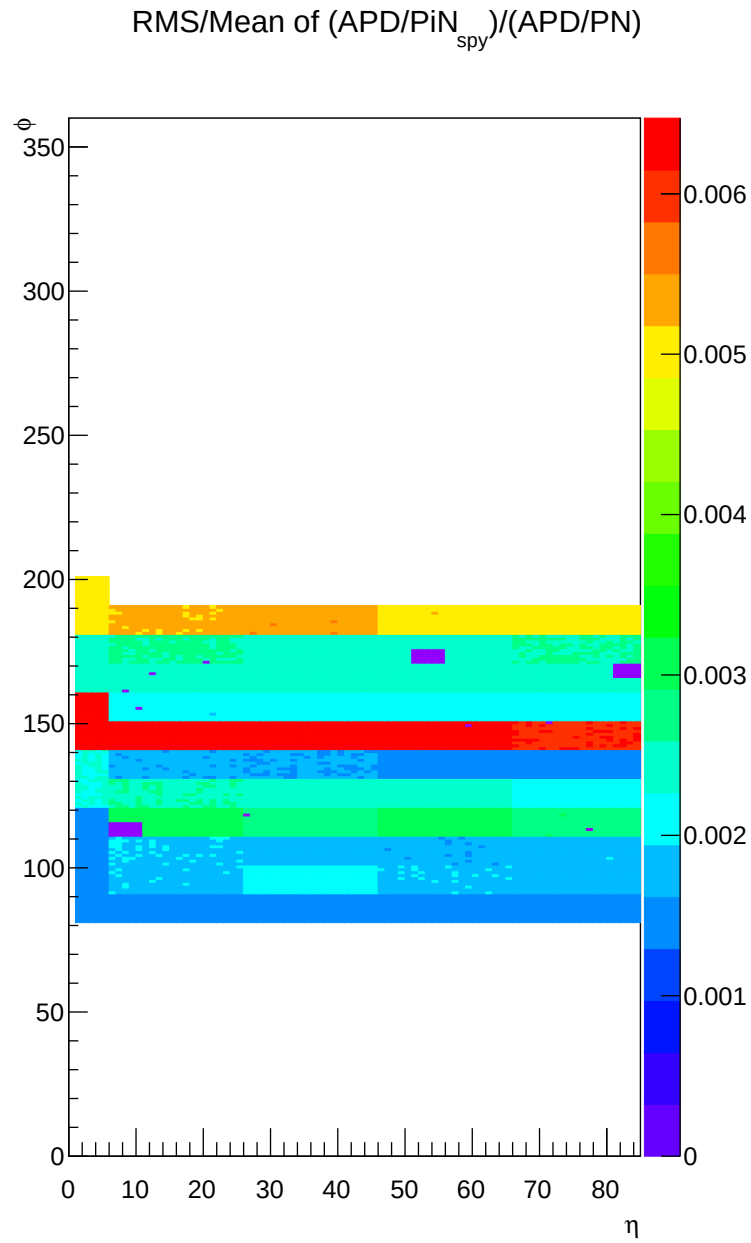


Figure 4.32: "SpyBox" precision map: 2D plot in ϕ - η coordinates in the ECAL barrel. The data are produced with the green laser and it corresponds to the summer dataset.

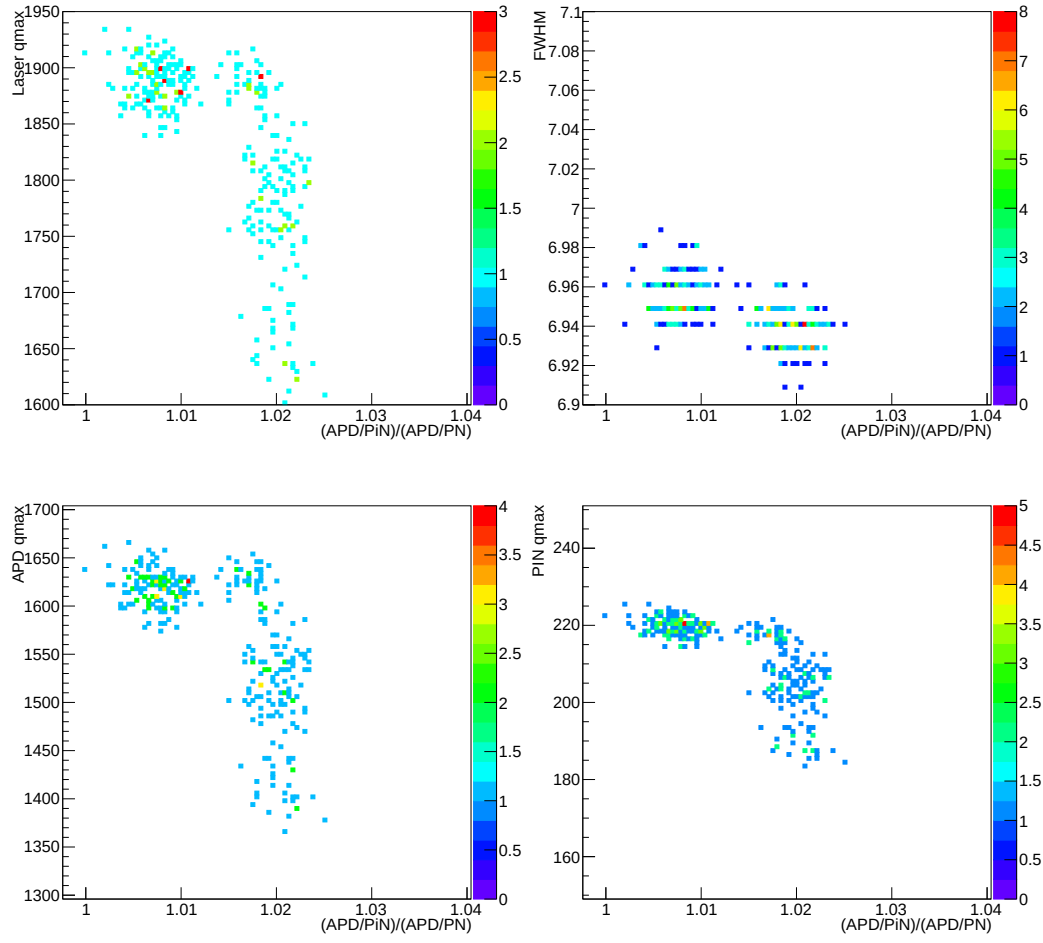


Figure 4.33: Fed 635 (both sides together). Top left: The transparency ratio as a function of the laser amplitude (laser qmax). Top right: The transparency ratio as function of the laser width. Bottom left: The transparency ratio as a function of the APD amplitude (APD qmax). Bottom right: The transparency ratio as a function of the PiN amplitude (PIN qmax).

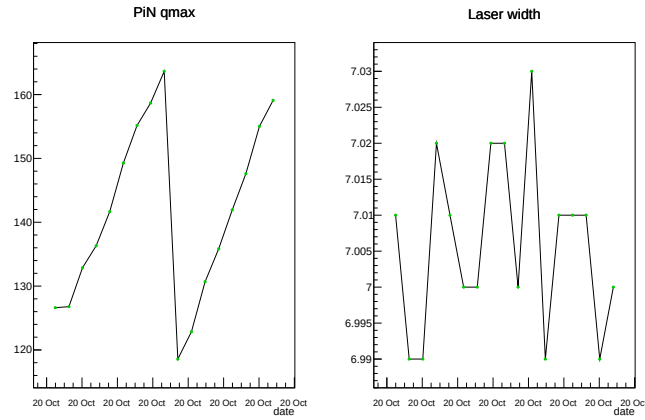


Figure 4.34: The PiN current and green laser width during the amplitude scan.

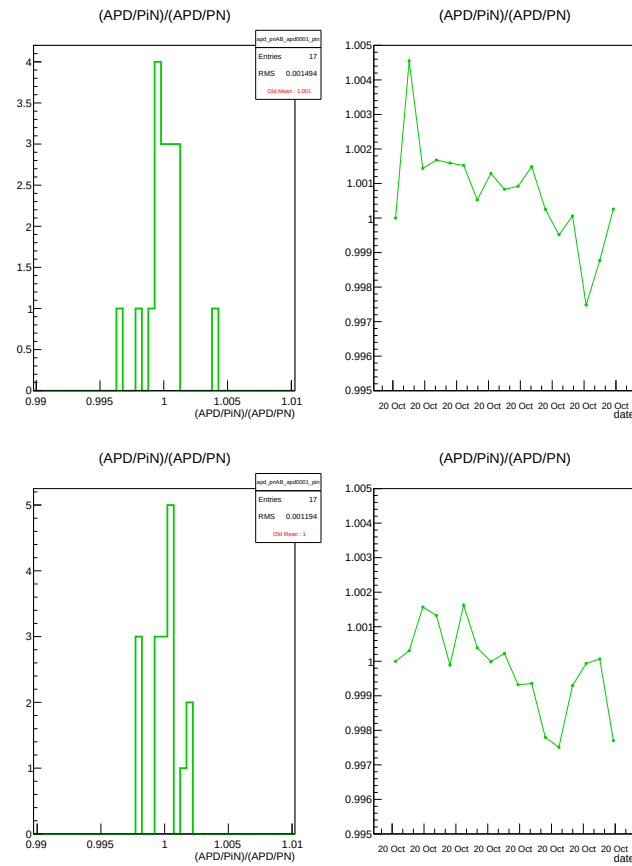


Figure 4.35: Transparency ratio, centred at 1, (left) and its evolution (right) in fed 635, side 0 (top) and fed 637, side 0 (bottom).

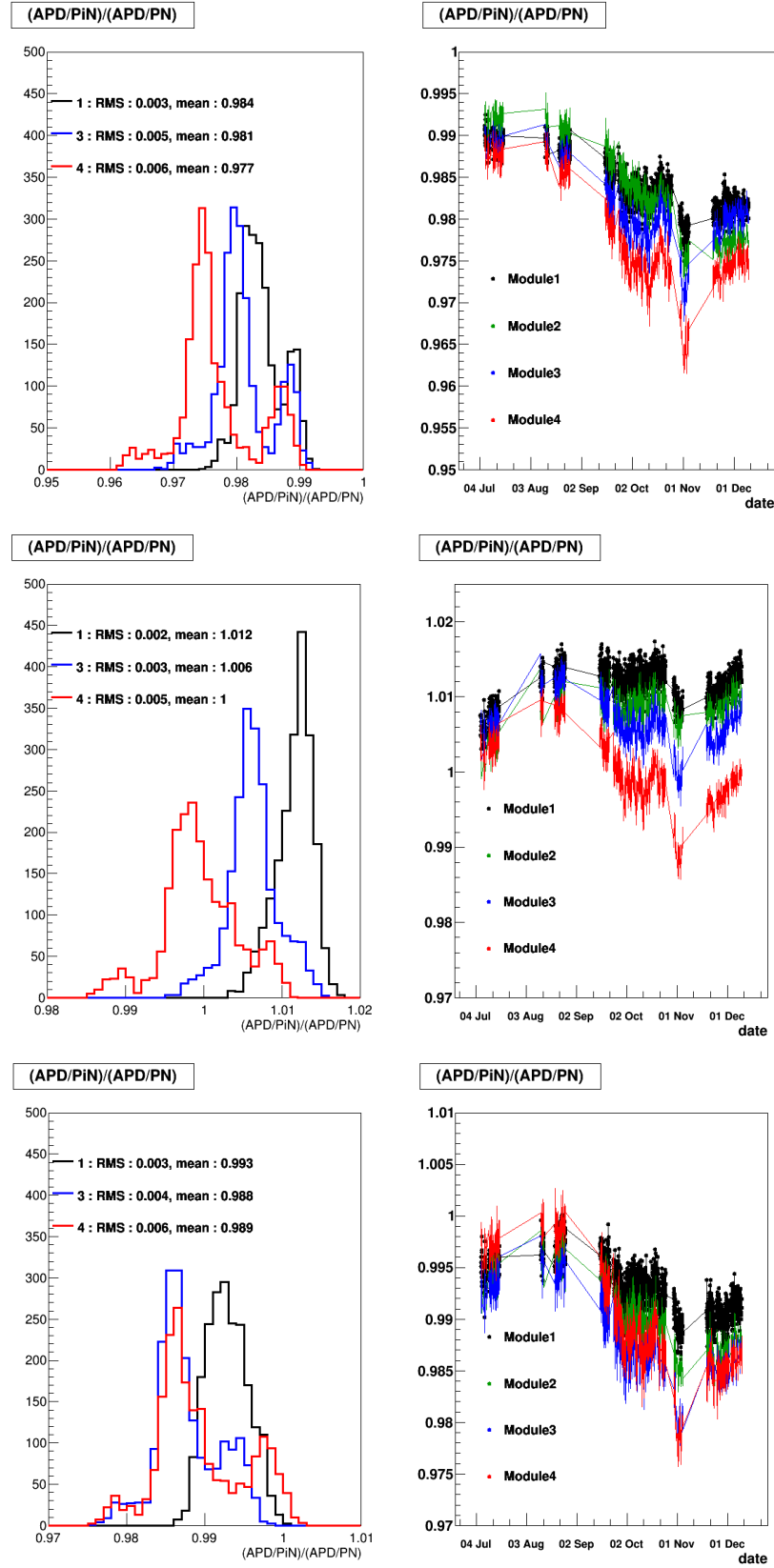


Figure 4.36: The transparency ratio (left) and its evolution with time (right) for the full 2015 dataset broken down by modules. Fed 632 side 0 on top, fed 635 side 0, in the middle and fed 636, side 0 on the bottom. Module 2 is not shown on the left plots as it has the same tendency as Module 1 or Module 3.

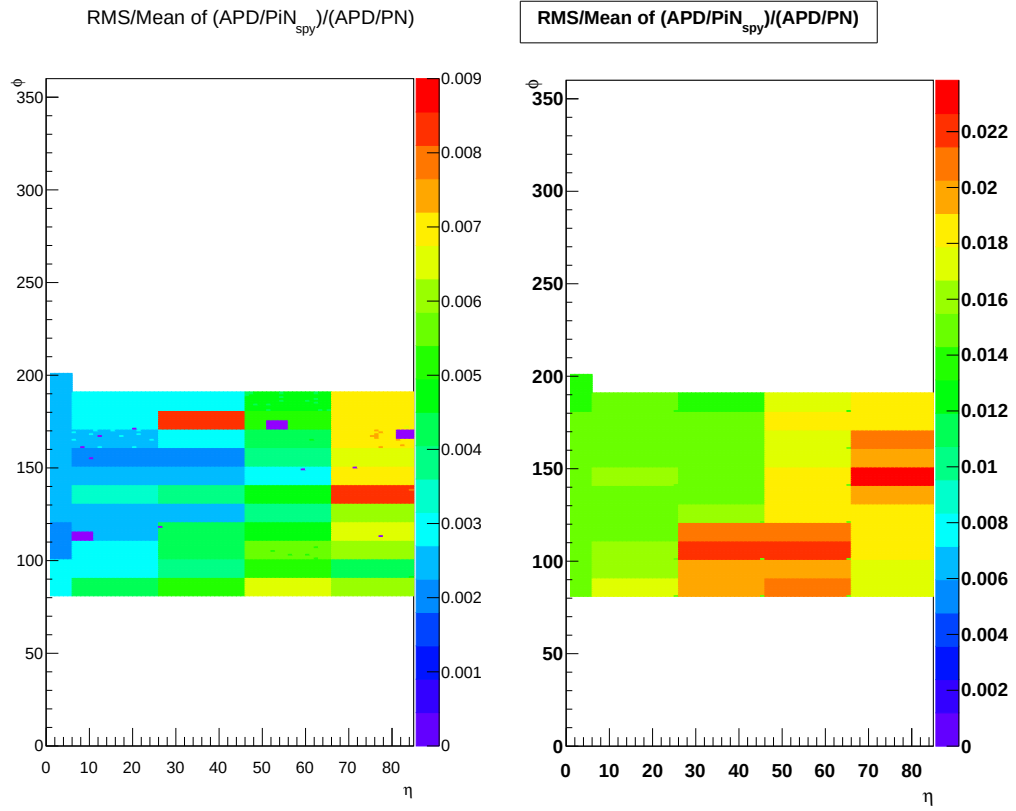


Figure 4.37: "SpyBox" precision map: 2D plot in ϕ - η coordinates in the ECAL barrel. The data are produced with the blue laser and it corresponds to the full 2015 (left) dataset and 2016 (right) dataset. In general the precision is worse in Modules 3 and 4 comparing to Modules 1 and 2, as they are located at higher η and get higher level of radiation. The precision reaches 0.9% in 2015 data and 2.2% in 2016 data.

Chapter 5

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

Introduction

In this chapter, the data analysis of the Higgs boson decaying into two photons search is presented. Since the discovery of a new boson at the LHC in 2012 [113–115], experimental studies have focused on determining the consistency of this particle’s properties with the expectations from the SM Higgs boson. Despite its small branching ratio, the $H \rightarrow \gamma\gamma$ decay channel provides a clean final-state topology with a fully reconstructed final state with an excellent invariant mass resolution. As a consequence, the $H \rightarrow \gamma\gamma$ channel is one of the most important channels for observing and investigating the properties of the Higgs boson. It is one of the channels contributing into the Higgs boson mass measurement. The dominant sources of background for $H \rightarrow \gamma\gamma$ channel comes from the irreducible direct diphoton production, and the reducible $pp \rightarrow \gamma + \text{jet}$ and $pp \rightarrow \text{jet} + \text{jet}$. The expected Higgs boson signal rate is at least an order of magnitude smaller than the SM background rate.

I participated in several iterations of the $H \rightarrow \gamma\gamma$ analysis during the LHC Run II. In this chapter, I describe the general strategy and mention my contributions. The results which correspond to the full 2016 dataset are presented in chapter 9. Previous results, shown at Moriond’16 and ICHEP’16 conferences can be found in corresponding public documents [159, 160]. The general analysis strategy is similar to the one in Run I, described in reference [120].

The data and simulation samples used in the analysis are listed in section 5.1. The trigger description, along with its efficiency, is shown in section 5.3. Photon reconstruction and photon energy corrections are discussed in sections 5.2 and 5.4. Preselections used in the analysis are shown in section 5.5. To get a good diphoton mass resolution in a high pileup environment, it is important to select the correct $H \rightarrow \gamma\gamma$ vertex, which is not trivial in this analysis. I am one of the main authors of the vertex identification algorithm, and I show my work in a dedicated chapter 6. Photon identification is very

important, since we need to distinguish the photons coming from hard interactions, such as our signal, from the ones coming from jets fragmentation. I also contributed to the photon identification algorithm described in chapter 7. After photons are identified the diphoton object has to be identified as well, that is presented in section 5.9. Once all objects are identified, the events corresponding to the Higgs boson different production modes have to be selected. The sensitivity of the analysis is enhanced by creating different categories, corresponding to different production modes: vector boson fusion (in section 5.10), the associated production with vector bosons (W or Z) (in section 5.11) or the associated production with a pair of top-antitop quarks (since I am focusing on this production mode in my thesis, it is described in a dedicated chapter 8). Events are classified according to additional objects present in the event, such as leptons, jets and missing transverse energy (MET). Events which are not selected in any of the categories mentioned above go into the "Untagged" categories populated mostly by the events produced by gluon-gluon fusion. Those events are categorized based on photon kinematics, mass resolution and other inputs related to signal to background ratio. The diphoton mass distribution is created for each event category, described above. The signal model is obtained from simulation, while the background one is data-driven, they are described in sections 5.13.1 and 5.13.2. The Higgs boson signal strength is finally extracted based on simultaneous likelihood fits to the diphoton mass spectra over all events classes, as described in section 5.13.3.

5.1 Samples

5.1.1 Data samples

The data sample used in the analysis corresponds to an integrated luminosity of 35.9 fb^{-1} with 25 ns bunch spacing and magnetic field of 3.8 T, recorded in 2016, unless otherwise specified. The analysis is performed in the invariant mass region $100 < m_{\gamma\gamma} < 180 \text{ GeV}$, keeping the signal region $115 < m_{\gamma\gamma} < 135 \text{ GeV}$ blinded.

5.1.2 Simulation samples

Monte-Carlo based simulations are used for signal modelling. Monte-Carlo based simulations for background processes are used only for selection optimizations and to build multivariate discriminators, but they are not used for background modelling in the analysis.

The signal samples used for four Higgs boson production modes and seven mass points 120 GeV, 123 GeV, 124 GeV, 125 GeV, 126 GeV, 127 GeV, 130 GeV, are listed in Table 5.1. Signal samples for associated production with single top quark and $b\bar{b}$ are used at 125 GeV mass point. Different mass points are used to construct a parametric signal model. The signal samples for the four Higgs boson production modes are generated with MADGRAPH_aMC@NLO [161]. The parton level samples are interfaced to

PYTHIA8 [162] for parton showering and hadronization. The cross-sections recommended by the LHC Cross Section Working Group [163] are used and mentioned in Table 5.1.

Process	Matrix Element Generator	Cross-section at 13 TeV (pb)
<i>Higgs boson production modes</i>		$m_H = 125$ GeV
Gluon-gluon fusion	MADGRAPH_aMC@NLO	48.5800
Vector boson fusion	MADGRAPH_aMC@NLO	3.7820
Associative production with W/Z boson	MADGRAPH_aMC@NLO	2.2569
Associative production with $t\bar{t}$	MADGRAPH_aMC@NLO	0.5071
Associative production with top quark tHq	MADGRAPH PYTHIA	0.0742
Associative production with top quark tHW	MADGRAPH PYTHIA	0.0151
Associative production with $b\bar{b}$	MADGRAPH_aMC@NLO	0.5329
<i>Background</i>		
Drell-Yan di-lepton + 0-2 jets	MADGRAPH_aMC@NLO	5765.4
Diphoton+jets (Born and box diagrams) M80-Inf	Sherpa	84.4
Diphoton+jets (Born and box diagrams) M40-80	Sherpa	303.2
Photon + jet Pt20-40 Mass80-Inf	PYTHIA	220.0
Photon + jet Pt40-Inf Mass80-Inf	PYTHIA	850.8
Photon + jet Pt20-Inf Mass40-80	PYTHIA	3216.0
Dijet Pt30-40	PYTHIA	22110.0
Dijet Pt40-Inf	PYTHIA	113400.0

Table 5.1: List of Monte-Carlo based simulations, which are used in $H \rightarrow \gamma\gamma$ analysis, their corresponding matrix element generators and cross-sections at 13 TeV.

A large irreducible background is coming from diphoton+jets production in both quark and gluon initial states. Leading order histograms for the quark-initiated born diphoton production and gluon-initiated box diphoton production are depicted in Fig. 5.1. Reducible background is coming from QCD dijet or γ +jet productions, where jets fake photons. Two Feynman diagrams for such processes are shown in Fig. 5.2.

Pileup conditions are simulated such that the running conditions of the 2016 runs are reproduced. The average number of pileup events in data is 23. The simulated samples have to be reweighed a posteriori to match the actual pileup profile in data. The average number of additional pileup interactions is either computed from the number of reconstructed primary vertices or from the measured instantaneous luminosity per bunch crossing, and its distributions in simulation and data are used to compute pileup weights for a particular data taking period, and in a region of the phase space similar to that of the analysis. The distributions for number of vertices and per-event measure of the amount of transverse energy from pileup interactions (ρ) are shown for data and simulation in Fig 5.3 after the pileup-reweighing procedure. We do not have any correction for the difference between data and simulation in number of vertices. There are corrections and systematic uncertainty assigned for the primary vertex choice, as described in section 6.3.

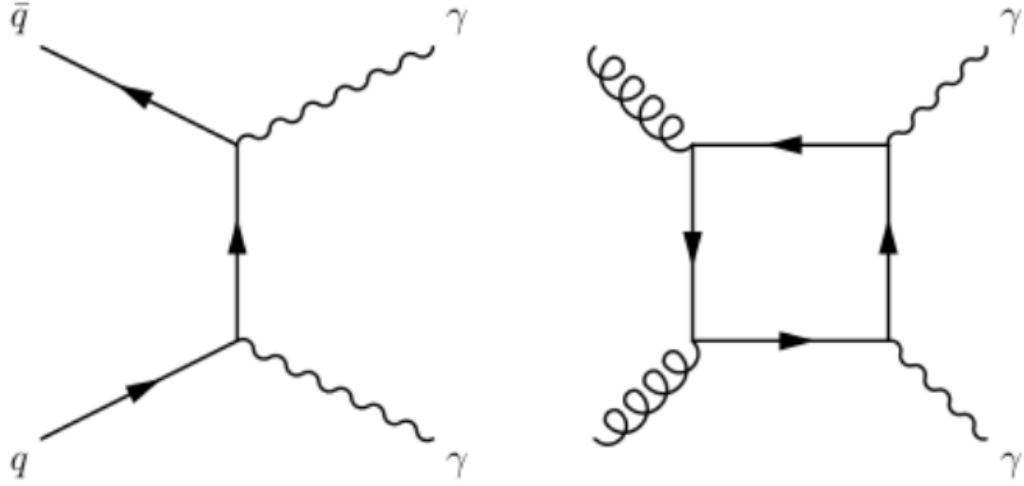


Figure 5.1: Born and box Feynman diagrams for diphoton+jet production originating from quark (left) and gluon (right).

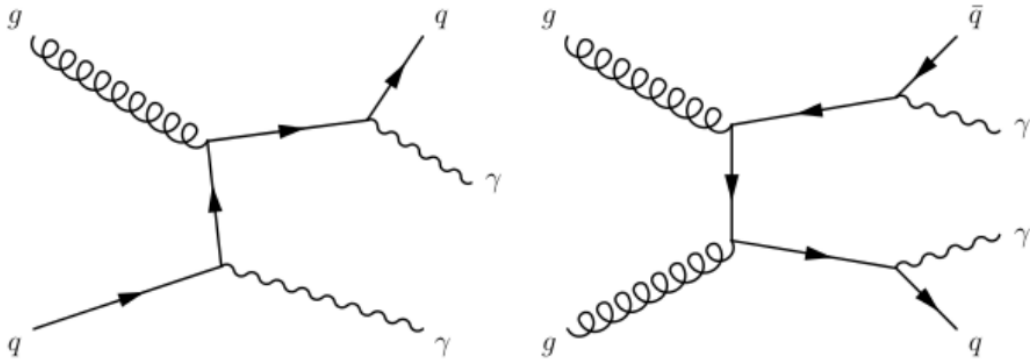


Figure 5.2: Tree-level Feynman diagrams for diphoton production in association with one or two jets.

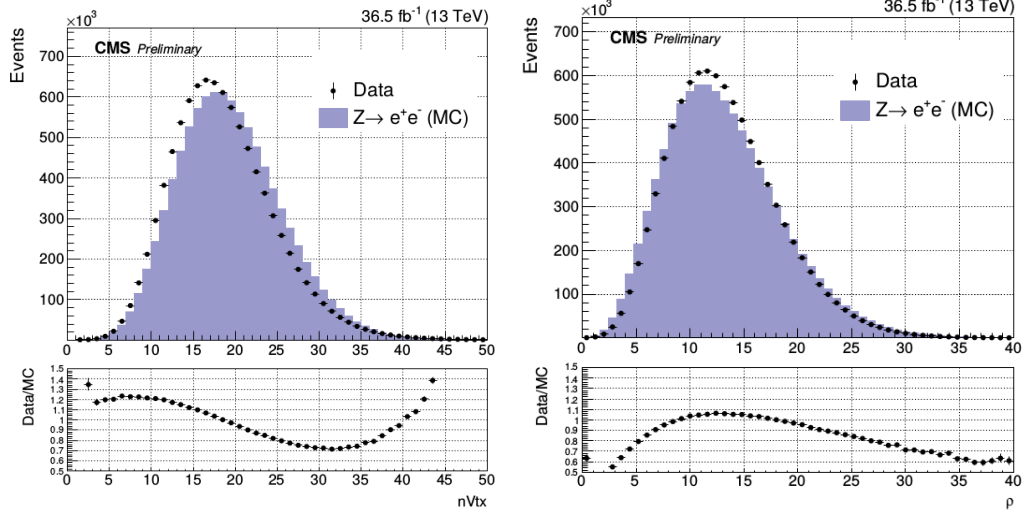


Figure 5.3: Distributions for number of vertices (left) and per-event measurement of the amount of transverse energy from pileup interactions (right) for $Z \rightarrow e^+e^-$ data events (black dots) and simulation events after applying pileup reweighting using minimum bias cross-section (filled histogram).

5.2 Photon reconstruction

A photon produced at the proton-proton interaction point passes through the tracker and deposits its energy into the ECAL via an electromagnetic shower, which can spread into several neighbouring crystals. In most cases, photons do not interact with the tracker and deposits $\sim 97\%$ of their energy in a 5×5 crystal matrix in the ECAL. In the remaining cases photons interact with the tracker material and convert into an electron-positron pair before entering the ECAL. Since electron and positron are charged particles, their trajectories bend in ϕ coordinate due to the magnetic field, as a result they deposit energy into large spread of crystals. To include all energy deposits, photon candidates are reconstructed from clusters of their energy deposits in the ECAL (as explained in section 4.1) and merged into superclusters [164, 165]. The reconstruction algorithm of photon clusters allows almost a complete recovery of the energy of photons that convert due to the material in front of the ECAL. The reconstruction algorithm consists of three steps:

1. Cluster seeds are identified as a local calorimeter cell (ECAL crystal) energy maximum above a given energy.
2. Topological clusters are formed from the seeds by aggregating crystals which have at least one side in common with a cell already in the cluster, and with an energy greater than a given threshold. These thresholds represent about two standard deviations of

the electronic noise in the ECAL (80 MeV in the barrel and up to 300 MeV in the endcaps).

3. Clusters are dynamically merged into superclusters. Dynamic superclustering allows for good energy containment, robustness against pileup (as size of the supercluster depends on its position in η and ϕ , and also on supercluster p_T , as shown in Fig. 5.4) and automatically takes the detector geometrical variations with η (e.g. EE crystal size) into account. Clusters lying in the area between two p_T - η dependent parabolas, which are centred around the most energetic cluster, are dynamically gathered giving to the supercluster a mustache-like shape. This is particularly important when moving to higher $|\eta|$ regions as the shape of the shower extends in η . In Fig. 5.5 the $\Delta\eta$ and $\Delta\phi$ distances between each cluster and the most energetic cluster are shown for $0 < |\eta| < 3$ with a step of $\eta = 0.25$. It can be easily noticed that going to higher η the shower shape becomes different and extends not only in ϕ , but also in η .

Photon reconstruction efficiency is defined as $\epsilon_{reco} = \frac{N_{\gamma \text{ matched}}}{N_{\gamma \text{ gen}}}$, where $N_{\gamma \text{ gen}}$ is number of photons generated with $p_T > 18$ GeV and $|\eta| < 1.442$ or $|\eta| > 1.566$, and $N_{\gamma \text{ matched}}$ is number of reconstructed photons with $p_T > 18$ GeV and $|\eta| < 1.442$ or $|\eta| > 1.566$, associated to generated photons within a cone of $\Delta R \leq 0.1$. It is computed using ggH simulation sample, and found to be $98.4\% \pm 0.01\%$.

A supercluster is promoted to a photon candidate if its reconstructed transverse energy is greater than 10 GeV. For the analysis only photons passing the following cuts are considered:

- $p_T > 14$ GeV and ratio of HCAL energy deposits in a cone $\Delta R < 0.14$ around the supercluster and the energy deposited in ECAL supercluster $H/E < 0.15$
- $10 \text{ GeV} < p_T < 14 \text{ GeV}$, $H/E < 0.15$ and $Iso_{CH_had} < 10 \text{ GeV}$, where Iso_{CH_had} is the scalar sum of transverse energy from HCAL deposits in a cone of $\Delta R < 0.3$ around the photon supercluster, excluding an inner-veto cone of $\Delta R < 0.15$.

5.2.1 Reconstruction of photon conversions

Reconstruction of photons converting into the tracker material is used in this analysis to help with the identification of the correct Higgs boson production vertex as explained in chapter 6. About 38% of the events have at least one of the photons reconstructed and selected as a conversion with two or one legs (tracks).

Although converted photons are clustered in the ECAL using the method mentioned in the previous section for photon reconstruction and tagged with good approximation by the R_9 shower-shape variable, additional information is gained by reconstructing the associated e^+e^- track pairs. The R_9 variable is the energy sum of the 3×3 crystals centred on the most energetic crystal (named "seed") in the supercluster divided by the raw energy of the

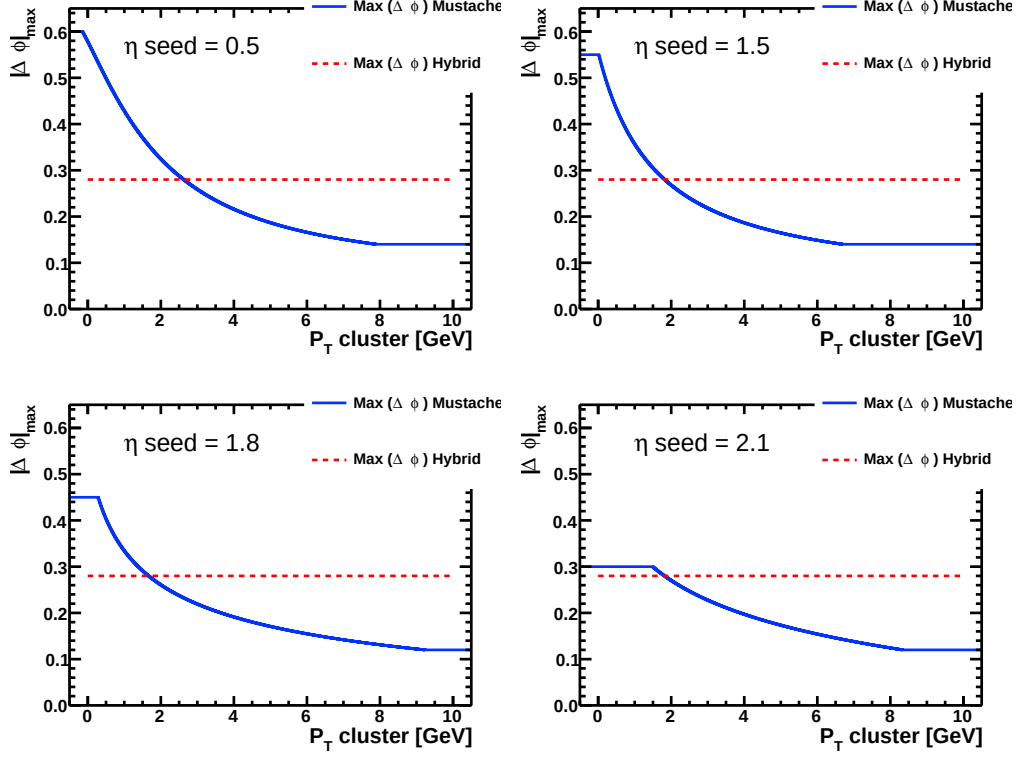


Figure 5.4: $\Delta\phi_{max}$ of the supercluster as function of the p_T of the supercluster for 4 different η bins. The red line shows the $\Delta\phi$ of the fixed-size window reconstruction (called Hybrid). Blue line shows the $\Delta\phi$ of the new algorithm used in Run 2 (called Mustache). $\Delta\phi_{max}$ decreases with η and p_T of the supercluster. On the bottom right one can see that at high eta (where more PU is expected) the supersluster's size is smaller using Mustache algorithm, so it includes less contamination from PU.

supercluster. The R_9 variable distribution is shown in Fig. 5.6. Unconverted photons have high R_9 (> 0.9), whereas converted photons have a much wider R_9 distribution.

Track reconstruction is based on an iterative tracking procedure. The first iteration finds tracks originating from the interaction vertex while subsequent iterations find tracks from displaced (secondary) vertices at increasing distance from the primary vertex. In addition, tracks starting from clusters in the ECAL and propagated backward to the tracker volume are sought, so as to reconstruct late-occurring conversions [166].

All tracks associated to the electron are refitted with the Gaussian sum filter method [132, 167]. Electron tracks are selected with basic quality requirements on the minimum number of hits and goodness of the track fit. Tracks are then required to have a positive charged-signed transverse impact parameter, so that the primary vertex lies outside the trajectory helix. Track-pairs of opposite charge are then filtered to remove tracks that might have

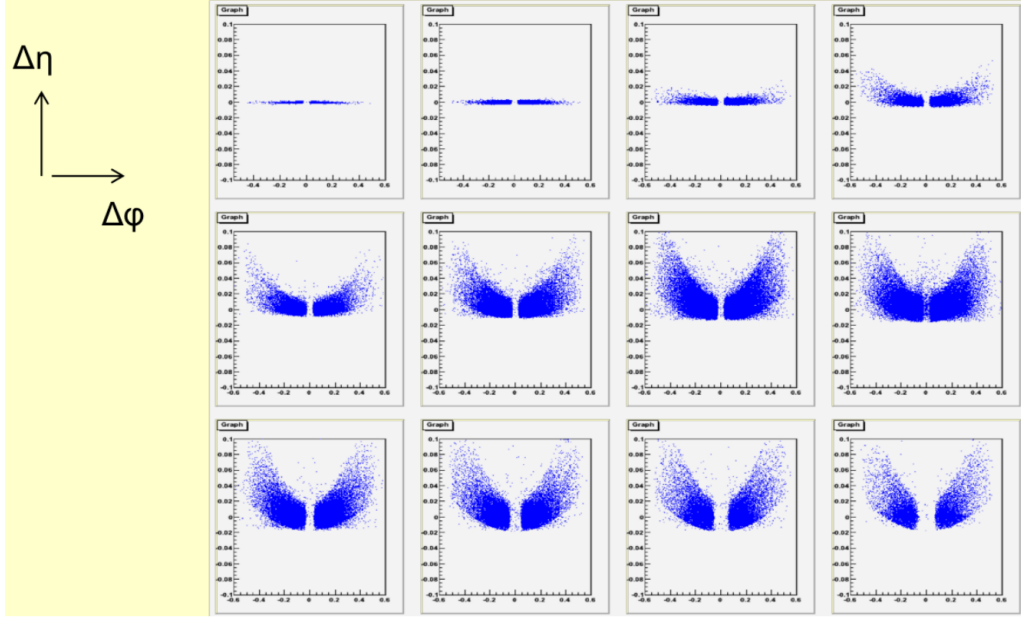


Figure 5.5: Plots of the $\Delta\eta$ and $\Delta\phi$ distances between each cluster and the most energetic cluster are shown for $0 < |\eta| < 3$ with a step of $\eta = 0.25$. They are produced using simulated photons with $10 < p_T < 100$ GeV.

resulted from conversions in the beam pipe, or could possibly consist of electrons originating from the primary vertex. Photon conversion candidates can be distinguished from massive meson decays, nuclear interactions or vertices from misreconstructed tracks by exploiting the fact that the momenta of the conversion electrons are approximately parallel since the photon is massless. For this purpose, the angular separation of the track pair in the longitudinal plane, measured in terms of $\Delta \cot \theta$, is required to be less than 0.1. Also, the two-dimensional distance of minimum approach between the two tracks is required to be positive to remove intersecting helices. Finally, the point in which the two tracks are tangent is required to be well contained in the tracker volume.

Track pairs passing the selection are fitted to a common vertex with a 3D-constrained kinematic vertex fit. The 3D constraint imposes the tracks to be parallel in both transverse and longitudinal planes. The pair is retained if the vertex fit converges and the χ^2 probability is greater than a given threshold. The transverse momentum of the pair is finally refitted with the vertex constraint. Reconstructed conversions are required to satisfy a minimum transverse momentum threshold, meant to reduce accidental or poorly reconstructed pairs. The threshold on the converted photon p_T as measured by the tracks is 10 GeV. Sometimes more than one conversion track-pair candidate can be reconstructed for the same supercluster. When such a case occurs, the optimal conversion is chosen by finding the best directional match between the momentum direction of the track pair

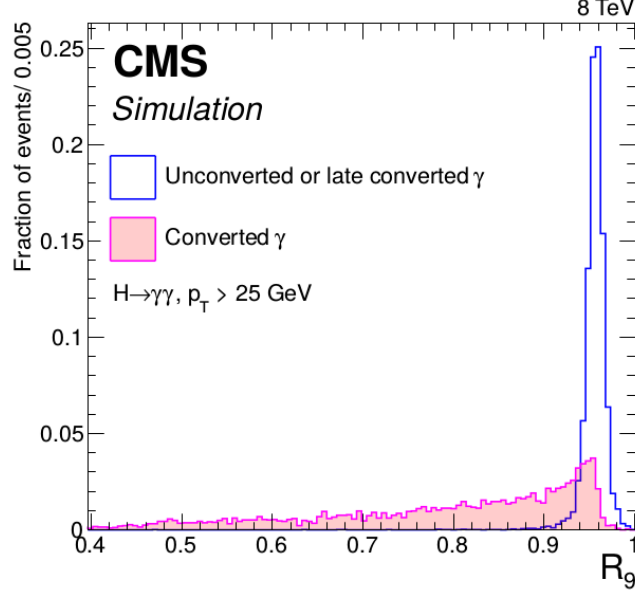


Figure 5.6: Distributions of the R_9 variable for photons in the ECAL barrel that convert in the material of the tracker before a radius of 85 cm (in red), and those that convert later, or do not convert at all before reaching the ECAL (in blue).

and the position of the supercluster. The matching criterion is expressed in terms of the $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ distance between the supercluster direction and the conversion direction. The conversion candidate with minimum ΔR is retained if ΔR is less than 0.1. Both the conversion and supercluster directions are redefined with respect to the fitted conversion vertex position.

Only one leg conversion is reconstructed, when photons convert at a small radius and the e^+e^- tracks have a very unbalanced p_T sharing (implying that one lepton carries most of the photon energy, leaving the second lepton with a too low momentum to be reconstructed) or when photons convert at large radius and tracks cannot be separated because they are short and collinear.

5.3 Trigger

As mentioned in section 3.6, there are two levels of triggers in CMS: the Level-1 Trigger (L1) and the High Level Trigger (HLT). Events entering the analysis have to pass a diphoton trigger decision. The trigger efficiency is measured using data collected by CMS. The diphoton trigger and the measurement of its efficiency are described below.

5.3.1 Level 1 trigger

Each high level trigger (HLT) diphoton trigger path is seeded by at least one hardware level 1 (L1) electromagnetic candidate. Since any photon candidate in the event can produce an L1 seed, the overall efficiency for a diphoton path seeded by a single L1 seed (singleEG) is higher than requiring two L1 seeds(doubleEG). Due to bandwidth limitations at L1, the lowest transverse momentum of single electromagnetic L1 candidates is 40 GeV (during lowest instantaneous luminosity periods this is lowered to 25 GeV). This creates an inefficiency at lower transverse energy. To mitigate this, the HLT paths are also seeded by a suite of L1 pairs with varying transverse momenta. At the highest instantaneous luminosity, the requirement on the transverse momenta are 22 GeV and 15 GeV for a diphoton pair. L1 efficiency is measured on data using the tag and probe method [168] (described below) with $Z \rightarrow e^+e^-$ events, where electrons are reconstructed as photons.

It is shown as function of the probe p_T and η in Fig. 5.7 for "photons" passing different R_9 selections.

5.3.2 High Level Trigger

In the CMS experiment, reconstruction of electrons and photons (e/γ) at the HLT level [169] is done in a sequence of steps summarized in Fig. 5.8. The diphoton HLT trigger is seeded at L1 by either one electromagnetic cluster (singleEG) with $p_T > 40$ GeV or by two electromagnetic clusters (doubleEG) with $p_T > 22$ GeV on the first candidate and $p_T > 15$ GeV on the second one, as mentioned above. Events passing L1 trigger, are processed through a streamlined version of the CMS offline reconstruction algorithm. The algorithm starts by clustering the hot crystals inside the ECAL to collect the full energy deposit of the candidates. The e/γ candidates are then selected with identification criteria involving ECAL cluster shape and isolation observables, which are defined as the sum of transverse energy deposit around the candidate. For electrons only, the presence of a well reconstructed track matched to the supercluster is also required. Finally, a selection on the tracker isolation is applied on the e/γ candidates.

The HLT selection uses general variables, isolation and shower shapes calorimeter variables, and R_9 :

- General variables: transverse energy (E_T) of both photons, diphoton invariant mass ($m_{\gamma\gamma}$), H/E;
- Shower shape + isolation variables: the lateral extension of the shower, computed as the energy weighted spread within the 5×5 crystal matrix centred on the seed crystal ($\sigma_{i\eta i\eta}$), corrected for the pileup ECAL particle flow cluster isolation, tracker isolation in a hollow cone (the scalar sum of the tracks transverse momenta in a cone of $\Delta R < 0.29$ around the photon direction, the tracks falling into the inner cone $\Delta R < 0.06$ are not included);

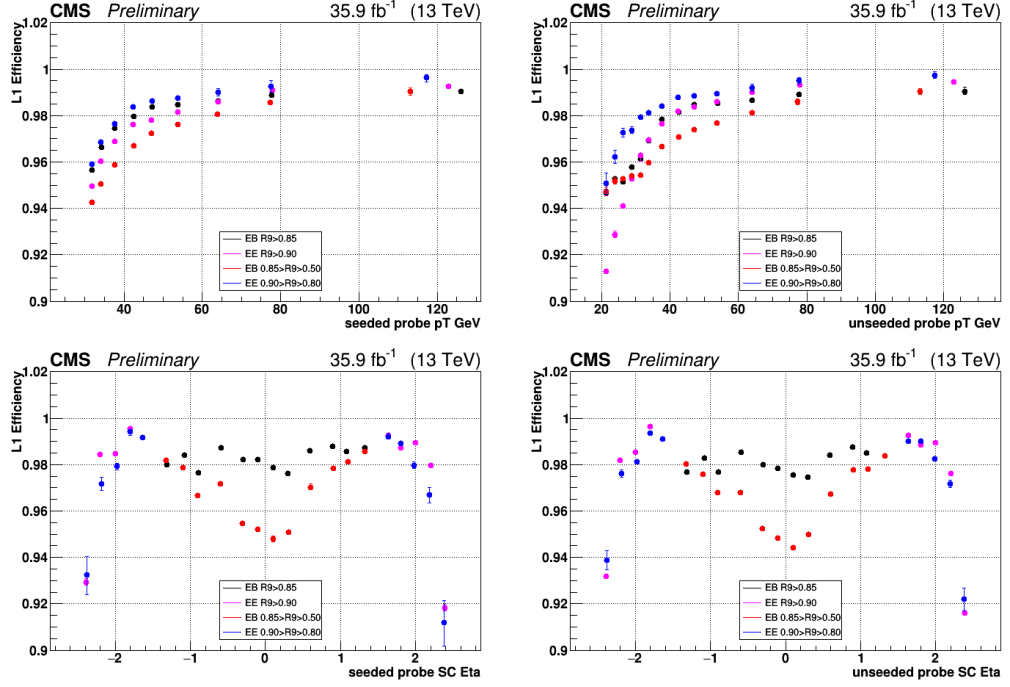


Figure 5.7: Top: trigger efficiency measured on data for $Z \rightarrow e^+e^-$ events using the tag and probe method. Efficiency with respect to offline probe p_T . Efficiency of photons in 4 R_9 categories are shown. The plots correspond to L1 objects from EG 22 (on the left) and EG 15 (on the right). Bottom : Trigger efficiency measured on data for $Z \rightarrow e^+e^-$ events using the tag and probe method. Efficiency with respect to offline probe η . Curves correspond to L1 objects matched to offline "photons" in 4 R_9 categories. L1 objects EG 22 (on the left) and EG 15 (on the right) are the seeds for diphoton HLT. An additional selection of leading "photon" ("photon" with the highest E_T in the event) $E_T > 41.67$ GeV and subleading "photon" ("photon" with the second highest E_T in the event) $E_T > 31.25$ GeV for "diphotons" with $m_{\gamma\gamma} = 125$ GeV was applied.

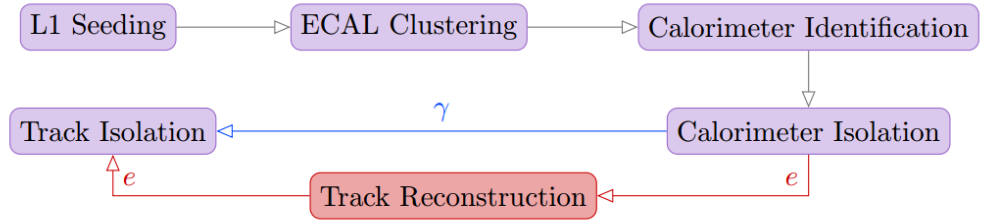


Figure 5.8: Flowchart of the e/γ reconstruction chain at HLT level.

- R_9 .

The L1-seeded leg of the HLT is required to have $E_T > 30$ GeV. The requirements for clusters are the following:

- $|\eta| < 2.5$,
- $R_9 > 0.5$ (0.8) in the ECAL barrel EB (the ECAL endcap (EE)),
- $H/E < 0.12$ (0.1) in EB (EE).

After the clusters are filtered with "high- R_9 " requirement: $R_9 > 0.85$ (0.9) in EB (EE). If the clusters pass those criteria, they continue to the unseeded requirements (described below), and they are not required to pass the shower shape and isolation cuts. If the clusters fail the "high- R_9 " requirement, they have to pass $\sigma_{i\eta i\eta} < 0.015$ (0.035) in EB (EE) and ECAL isolation $< 6.0 + 0.012 \times E_T$, in order to continue to the unseeded requirements.

Two clusters are then required to pass the unseeded requirements (at least one of them being seeded as described above). Those requirements are the same as for seeded leg, except $E_T > 18$ GeV. Both legs seeded and unseeded have to pass track isolation $< 6.0 + 0.002 \times E_T$, and the diphoton object invariant mass is required to be above 90 GeV.

5.3.3 Trigger performance

The requirements of the HLT were chosen to keep the bandwidth of the HLT at sustainable levels while still accepting as many signal events as possible. In order to measure the efficiency the tag and probe method is used. The efficiency of both the seeded and unseeded leg of the diphoton HLT path is measured on $Z \rightarrow e^+e^-$ data events, where electrons reconstructed as photons. Events are required to pass unscaled single electron trigger with $p_T > 27$ GeV and the tight identification working point (described in section 5.11.1). Since the mentioned trigger only requires a single electron, the second electron remains an unbiased probe to measure the efficiency of the diphoton path.

The events which pass the tag and probe trigger are reconstructed. "Photons" are required to pass the analysis preselection, as described in section 5.5. Normally, a conversion-safe electron veto is required for each photon. It is used to reject electrons other than the ones originating from photon conversion. A typical converted photon will have charged-particle tracks associated to a reconstructed conversion vertex, pointing to an ECAL cluster. This veto reject the photon candidate if its supercluster is matched to an electron track with no missing hits in the innermost tracker layers. For trigger efficiency estimation, it is inverted to select electrons.

If both "photons" of the "diphoton" pair pass the "tag" selection they are both used also in the "probe" analysis. The tag must match an HLT object which passed the tag and probe trigger. The tag must also pass a photon identification criteria of photon

identification $MVA > -0.6$ (a photon identification method is explained in chapter 7). The remaining "photon" is required to pass looser photon identification criteria of photon identification $MVA > -0.9$, and defined as the probe. The tag and probe pair must have an invariant mass corresponding to the Z boson ($70 < m_{\gamma\gamma} < 110$ GeV). The collection of probes passing this selection becomes the denominator for the seeded leg of the diphoton HLT. If the probe can be matched within $\Delta R < 0.3$ to the seeded leg of the diphoton HLT path an entry is added to the numerator. The same analysis is done separately for the unseeded leg of the diphoton HLT path, with the lower preselection transverse energy requirement (20 GeV) applied as well as the requirement that the tag matches a seeded leg diphoton HLT object. This procedure is duplicated for all combinations of "photons" in the "diphoton" collection which can pass the tag and probe requirements.

Due to the material upstream of ECAL, electrons and photons have different shower shapes in ECAL. To account for this (as well as the different η distributions for $Z \rightarrow e^+e^-$ and $H \rightarrow \gamma\gamma$) the "photons" are weighted in R_9 and η from the respective simulated samples to match those of the gluon fusion $H \rightarrow \gamma\gamma$ distributions.

Since photons have a harder R_9 distribution than electrons, this gives an overall increase of the efficiency on the $Z \rightarrow e^+e^-$ events. The HLT efficiency of both the seeded and unseeded legs of the diphoton trigger with respect to the offline photon p_T can be found on top of Fig 5.9 and the efficiency with respect to η can be found on bottom of Fig 5.9.

The analysis does not require the trigger in simulations, since the preselection criteria (described in section 5.5) are applied, and they are tighter than trigger requirements. The trigger efficiency measured on data, binned in E_T , R_9 , and η , is used to scale the simulation. The corrections are a combination of the HLT and L1 efficiency to properly simulate the seeding of the HLT by the L1. These corrections are applied to the simulation to replicate the effect of the trigger.

5.4 Photon energy resolution

A good understanding of the expected signal shape is crucial as an input to the final statistical analysis. The width of the reconstructed Higgs boson mass peak is driven by the detector resolution whenever the correct vertex is chosen.

It is thus necessary to correct the photon energy scale in data and resolution in simulation, in order for the simulation to reproduce properly the detector performance. In order to achieve the best energy resolution, one has to apply different sets of corrections to the ECAL reconstructed hits and photon energy. The first set consists of crystal-level corrections in order to equilibrate the channel to channel response variations [170]. Then a high-level correction method, the photon energy regression (called "regression" in the remainder of this chapter), is applied in order to take into account finer effects, such as the local containment of the photon shower, in particular the energy losses in the gaps and cracks of the ECAL, the global containment of the converted photon showers and

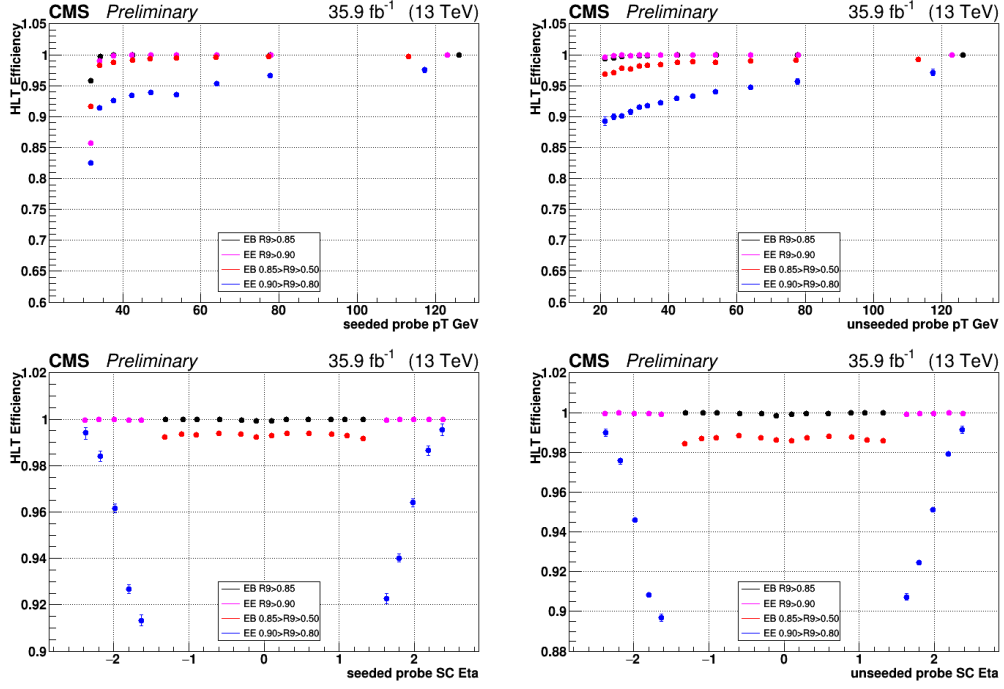


Figure 5.9: Top: trigger efficiency measured on data for $Z \rightarrow e^+e^-$ events using the tag and probe method. Efficiency with respect to the offline probe p_T . Efficiency of photons in 4 R_9 categories are shown. The plots correspond to seeded (left) and unseeded (right) HLT leg of the diphoton trigger. Bottom : Trigger efficiency measured on data for $Z \rightarrow e^+e^-$ events using the tag and probe method. Efficiency with respect to the offline probe η . Curves correspond to L1 objects matched to "photons" in 4 R_9 categories. Seeded (left) and unseeded (right) HLT leg of the diphoton trigger. An additional selection of leading "photon" $E_T > 41.67$ GeV and subleading "photon" $E_T > 31.25$ GeV based on scaling p_T cuts for "diphotons" with $m_{\gamma\gamma} = 125$ GeV was applied.

the pileup. Starting from the raw supercluster and (in EE) the preshower energy, the energy regression method aims to make the best prediction of the true photon energy. It is based on a multivariate approach. The input variables are the shower shape, the energy ratios within the seed cluster, the absolute position and the pileup information. The semi-parametric regression technique, which has the advantage of being able to estimate simultaneously the energy of the photon and its resolution, is used. The training is performed on prompt photons from simulated events and its result is shown in Fig. 5.10. The distributions are fitted with a double Crystal Ball function¹ [171].

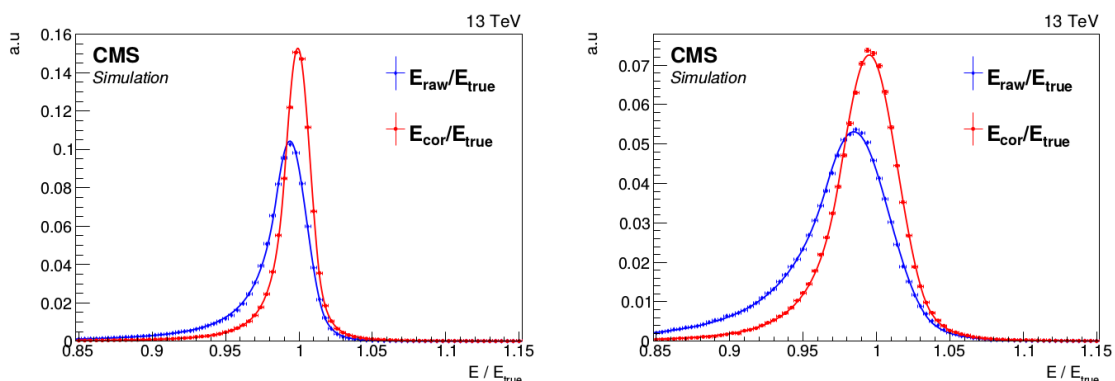


Figure 5.10: Ratio of photon raw energy to its true energy compared to the ratio of photon energy corrected by the regression to its true energy. Photons from simulated $H \rightarrow \gamma\gamma$ ($m_H = 125$ GeV) events are used in the barrel (left) and in the endcaps (right).

Data-driven corrections of the photon energy scale and resolution are derived using $Z \rightarrow e^+e^-$ events. The measurement technique, which is described in more details in [172], consists in 3 main steps:

- The purpose of the first step is to fix the drift of the energy scale in data as a function of time (run number). Both data and simulation invariant mass distributions are fitted with a Breit-Wigner function convoluted with a Crystal Ball. The mean parameter are obtained for data and simulation in run bins. The outcome of the first step provide run dependent energy corrections which are applied to data to correct for possible time dependent energy scale variations during data-taking.
- In the second iteration, scale corrections are derived in η - R_9 categories on top of the time-dependent corrections found in the first step.
- In the third iteration, resolution corrections are derived once applied the final set of scale corrections. The result of this step is a set of smearing corrections in 8 cate-

¹The Crystal Ball function consists of a power law tail stitched to a Gaussian core such that the function and its first derivative are continuous.

gories ($4\eta \times 2R_9$). They are derived minimizing the likelihood between the smeared simulation and data.

The results of the energy scale and resolution adjustment are shown in Fig. 5.11 for 2 categories: 2 electrons in EB with $R_9 > 0.94$, and 2 electrons in EE with $R_9 > 0.94$. The result for the full R_9 range for 2 electrons in the barrel are shown in on the left in Fig. 5.12, for one electron in the barrel, another one in the endcaps in the middle, and for both electrons in the endcaps on the right.

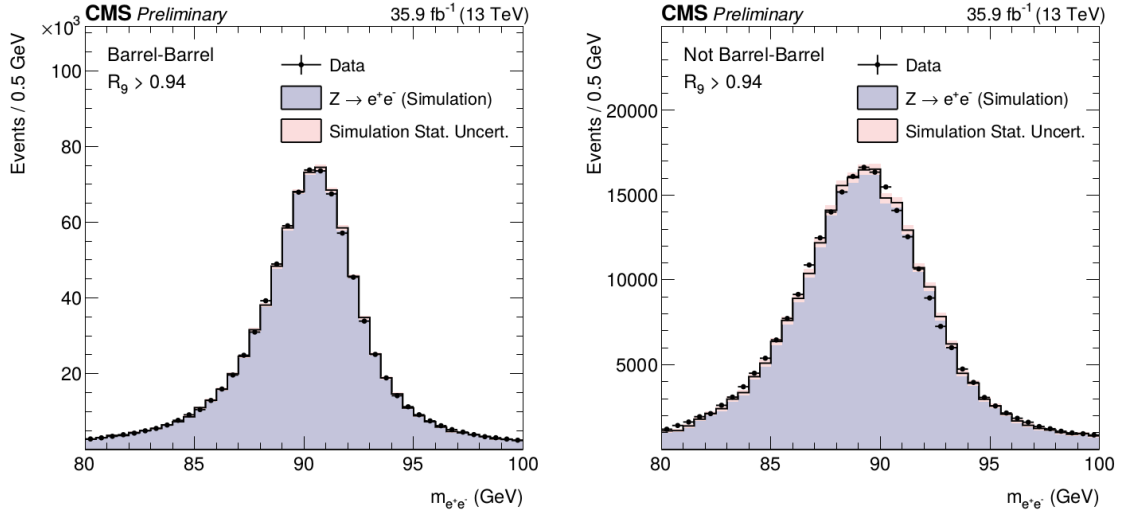


Figure 5.11: Comparison of the dielectron invariant mass distributions in data and simulation (after energy smearing) for $Z \rightarrow e^+e^-$ events where electrons are reconstructed as photons. The comparison is shown requiring $R_9 > 0.94$ for both electrons and for (left) events with both showers in the barrel, and (right) the remaining events. The simulated distribution is normalized to the integral of the data distribution in the range $87 \text{ GeV} < m_{e^+e^-} < 93 \text{ GeV}$.

5.5 Event preselection

All photons within geometrical acceptance ($|\eta| < 2.5$), and not in the EB-EE gap, ($1.442 < |\eta| < 1.566$) are required to pass the preselection criteria. They are similar to, but slightly more stringent than, the trigger requirements.

Several loose cuts are applied on the ECAL shower shape and isolation variables: H/E , $\sigma_{i\eta i\eta}$, photon isolation and tracker isolation (both with a ρ correction applied to match the HLT), and R_9 . Photons are classified according to the photon location in the ECAL (barrel or endcap) and the R_9 value (> 0.85 or < 0.85 for the barrel and > 0.9 and < 0.9 for the endcap). All requirements are listed in Table 5.2. The leading photon is required

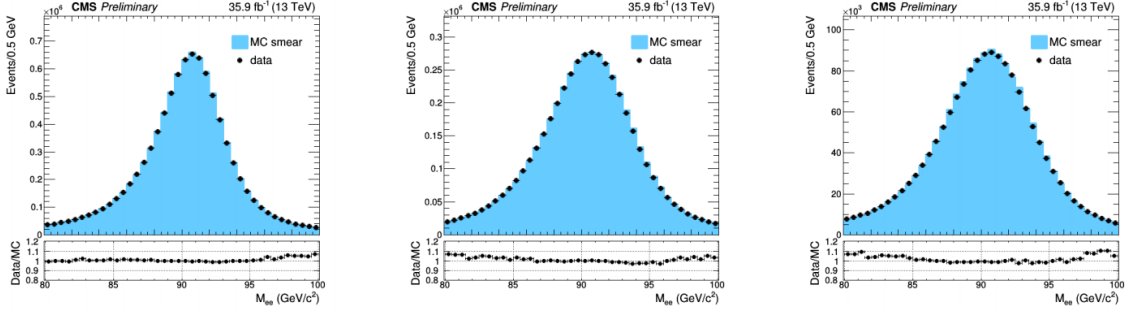


Figure 5.12: Comparison of the dielectron invariant mass distributions in data and simulation (after energy smearing) for $Z \rightarrow e^+e^-$ events where electrons are reconstructed as photons. The comparison is shown for both electrons in the barrel (left), for one electron in the barrel and another one in the endcaps (middle), and both electrons in the endcaps (right). The simulated distribution is normalized to the integral of the data distribution in the range $87 \text{ GeV} < m_{e^+e^-} < 93 \text{ GeV}$.

to have $p_T > 30 \text{ GeV}$ and the sub-leading one is required to have $p_T > 20 \text{ GeV}$ ². The selected pair of photons is required to satisfy $m_{\gamma\gamma} > 100 \text{ GeV}$.

	H/E	$\sigma_{i\eta i\eta}$	R_9	photon isolation	tracker isolation
EB; $R_9 > 0.85$	< 0.08	—	> 0.5	—	—
EB; $R_9 \leq 0.85$	< 0.08	< 0.015	> 0.5	< 4.0	< 6.0
EE; $R_9 > 0.90$	< 0.08	—	> 0.8	—	—
EE; $R_9 \leq 0.90$	< 0.08	< 0.035	> 0.8	< 4.0	< 6.0

Table 5.2: List of photon preselection requirements.

As was already mentioned, to avoid misidentifying an electron as a photon, a conversion-safe electron veto is applied. It is used to reject electrons other than the ones originating from photon conversion. A typical converted photon will have charged-particle tracks associated to a reconstructed conversion vertex, pointing to an ECAL cluster. This veto reject the photon candidate if its supercluster is matched to an electron track with no missing hits in the innermost tracker layers.

Preselection efficiencies are determined from data using $Z \rightarrow e^+e^-$ events with the tag and probe method, which is then compared to simulation to calculate the ratio of preselection efficiencies measured in data and simulation. Those correction are then applied to simulation.

In the computation of the corrections for simulation, the photon and electron R_9 difference is taken into account by weighting the electron R_9 distribution to match the photon

²There is no trigger for simulation applied, so these p_T cuts mimic the trigger for the simulation. The efficiency was checked, and it was found very close to 1.

one. Table 5.3 shows the efficiencies with statistical uncertainties measured in data and simulation along with their ratio in four η , R_9 categories. The systematic uncertainty has been estimated taking into account the choice of the parton distribution function (PDF) used for the signal modeling, the choice of the PDF for the background modeling, p_T and R_9 differences (reweighing the probe spectra to match the one of photons from Higgs boson decay) and varying the tightness of the tag selection.

By construction the tag and probe technique using $Z \rightarrow e^+e^-$ does not allow to measure the electron-veto efficiency. The effect on the signal photon efficiency of the electron veto requirement has been evaluated from data in a sample of $Z \rightarrow \mu\mu\gamma$ events with high photon purity and compared with events selected in a simulated sample of Drell-Yan events. The electron veto efficiency is measured as the ratio of the number of photons passing the electron veto divided by the total number of preselected photons. Table 5.4 lists the electron veto efficiencies with statistical uncertainties measured in data and simulations, and their ratio.

	Data		Simulation		Ratio	
	Eff.	Unc.	Eff.	Unc.	Scale	Unc.
Barrel; $R_9 > 0.85$	0.9488	0.0001	0.9499	0.0001	0.9988	0.0001
Barrel; $R_9 < 0.85$	0.8471	0.0001	0.8423	0.0002	1.0057	0.0002
Endcap; $R_9 > 0.90$	0.9207	0.0004	0.9256	0.0002	0.9947	0.0004
Endcap; $R_9 < 0.90$	0.5309	0.0001	0.5622	0.0003	0.9443	0.0005

Table 5.3: Preselection efficiencies measured in the 4 photon categories using tag and probe with $Z \rightarrow e^+e^-$ events (for all preselection criterion except electron veto).

	Data		Simulation		Ratio	
	Eff.	Unc.	Eff.	Unc.	Scale.	Unc.
Barrel; $R_9 > 0.85$	0.9928	0.0003	0.9970	0.0005	0.9958	0.0006
Barrel; $R_9 < 0.85$	0.9741	0.0010	0.9795	0.0020	0.9945	0.0023
Endcap; $R_9 > 0.90$	0.9789	0.0009	0.9863	0.0015	0.9924	0.0017
Endcap; $R_9 < 0.90$	0.9360	0.0033	0.9574	0.0055	0.9777	0.0065

Table 5.4: Efficiency of the conversion-safe electron veto, measured in the four photon categories using tag and probe with $Z \rightarrow \mu\mu\gamma$ events. The data to simulation ratio is also shown with its uncertainty. The efficiency is calculated using for the denominator the number of photons passing all preselection criterion except the electron veto, and for the numerator the number of photons passing all selection requirements including the electron veto.

5.6 Diphoton vertex identification

Since I am one of the main authors of the vertex identification in the $H \rightarrow \gamma\gamma$ analysis, it is described in details in chapter 6.

5.7 Photon identification

Since I am one of the main authors of the photon identification in the $H \rightarrow \gamma\gamma$ analysis, it is described in details in chapter 7.

5.8 Event classification

The event classification is done to enhance the analysis sensitivity and to measure the signal strengths of different Higgs boson production modes. Higgs boson production mechanisms, different from the ggH, can be identified by selecting final state objects in addition to the diphoton pair. In case the Higgs boson is produced in vector-boson fusion, it is accompanied by a pair of jets separated by a large rapidity gap. If the Higgs boson is produced in association with vector bosons (W or Z), it can be accompanied by jets coming from W, or by charged leptons or missing energy coming from W or Z. If it is produced in association with a top-antitop pair, it will be accompanied by b-quark jets, and may be accompanied by charged leptons or additional jets coming from W decay.

Tagging of dijet events, for VBF, significantly increases the overall sensitivity of the analysis and precision on the measured signal strength, and allows to measure the coupling to vector bosons. Tagging the associated production processes, WH, ZH and $t\bar{t}H$, increases the sensitivity of the measurement of the coupling to vector bosons and to the top quark, and further probes the compatibility of the observed signal with the SM Higgs boson. The events with additional objects are tagged as exclusive categories, while those that remain untagged are classified to the inclusive categories mostly populated by the gluon-gluon fusion production mechanism. The inclusive categories are divided into groups of events with similar mass resolution. If more than one diphoton candidate within the analysis mass window is present, the candidate with the highest priority tag is selected. In the case of multiple diphotons ($0.9\% \pm 0.01\%$ of events, estimated using a ggH simulation sample) with equal tag priority, the diphoton with the highest p_T is selected. The exclusive categories are defined below and listed by order of priority:

- events with leptons from the leptonic or semi-leptonic top decays, this selection is described in details in chapter 8;
- events with leptons from the ZH and WH productions, as described in section 5.11.2;
- events with jets from hadronic top decays, as described in chapter 8;

- events with two forward jets (three categories) to target the VBF process, this selection is described in section 5.10;
- events with missing transverse energy to tag VH events where the Z boson has decayed to two neutrinos and the W boson into lepton and neutrino as described in section 5.11.3;
- events with two jets to pick up VH where the vector boson decays to quarks as described in section 5.11.4.

The above categories provide a thorough classification of final states. The remaining inclusive events are categorized using a multivariate classifier, called diphoton identification classifier (diphoton MVA). It is described in section 5.9. It creates four categories based on photon kinematics, mass-resolution, as well as other inputs to indicate the signal-to-background ratio.

5.9 Diphoton multivariate classifier

As mentioned above, in order to increase the sensitivity of the analysis, the events are divided into categories, such that "high-performance" categories are characterised by a higher signal-to-background ratio and better mass resolution.

5.9.1 Setup and performance

Events with two photons passing the preselection criteria described in section 5.5 are selected. In addition a loose selection on the photon identification output³ (photon identification MVA > - 0.9, 99% signal efficiency) is applied. The photons are also required to satisfy a diphoton mass-dependent requirement on the transverse momentum, set to $p_T > m_{\gamma\gamma}/3$ ($m_{\gamma\gamma}/4$) for the leading (sub-leading) photon.

The event classifier includes the kinematic properties of the diphoton system, a per-event estimate of the diphoton mass resolution and a per-photon score given by the identification output described in chapter 7. The input variables are chosen such that the algorithm is blinded to the diphoton invariant mass. For this purpose, dimensional variables are rescaled by the mass of the diphoton system. The following variables are used as input to the event classifier:

- the transverse momenta for both photons, rescaled for the diphoton mass, $p_T^{1,(2)}/m_{\gamma\gamma}$,
- the pseudorapidities of both photons, $\eta^{1,(2)}$,
- the cosine of the angle between the two photons in the transverse plane, $\cos(\Delta\phi)$,

³see chapter 7

- the identification BDT score for both photons,
- the per-event relative mass resolution estimate, under the hypothesis that the mass has been reconstructed using the correct primary vertex, σ_{rv} ,
- the per-event relative mass resolution estimate, under the hypothesis that the mass has been reconstructed using an incorrect primary vertex σ_{wv} ,
- the per-event probability estimate that the correct primary vertex has been selected, it is described in chapter 6, $prob_{vtx}$

The classifier is implemented using a Boosted Decision Tree (BDT) from the TMVA package [173]. The per-event relative mass resolution estimate is computed from the propagation of the photon energy resolution estimates, assuming Gaussian resolution functions:

$$\sigma_{rv} = \sigma_m^{right}/m_{\gamma\gamma} = \frac{1}{2}\sqrt{\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2} \quad (5.1)$$

The energy estimate and its resolution are extracted from regression, described in section 5.4. The energy resolution σ_E of each photon is corrected to match the estimate of the energy resolution in data by adding in quadrature the additional smearing applied to the single photon energy in the simulation. Under the assumption that the correct vertex has been selected, no additional contributions are considered for the computation of the mass resolution, since in this case the energy measurement of the photons is the dominant contribution. In order to correctly estimate the mass resolution for the events where an incorrect primary vertex is selected, a second estimator of the mass resolution is computed. It includes an additional term σ_m^{vtx} given by the displacement between the correct and the selected primary vertex. The distance between the true and a randomly selected vertex is distributed as a Gaussian with width $\sqrt{2}\sigma_Z^{beamspot}$ and the term σ_m^{vtx} can be computed analytically given the impact positions of the two photons in the calorimeter. Since the simulation does not model the variation of the beamspot length over the course of a fill, we use an estimate of the average beamspot length measured in 13 TeV data (3.5 cm) to compute σ_{wv} for both data and simulation in order to construct consistent inputs for the event classifier. The relative mass resolution under the incorrect vertex hypothesis is then computed as:

$$\sigma_{wv} = \sigma_m^{wrong}/m_{\gamma\gamma} = \sqrt{(\sigma_m^{right}/m_{\gamma\gamma})^2 + (\sigma_m^{vtx}/m_{\gamma\gamma})^2} \quad (5.2)$$

In the training of the BDT, information needs to be provided so that signal-to-background is inversely proportional to the mass resolution. This is achieved by weighting the signal events used to train the BDT:

$$w_{sig} = \frac{prob_{vtx}}{\sigma_{rv}} + \frac{1 - prob_{vtx}}{\sigma_{wv}} \quad (5.3)$$

Events with a better mass resolution will thus tend to have a higher score assigned by the BDT.

The BDT is trained on samples simulating signal and background processes at 13 TeV. The signal sample is the cross-section weighted mixture of the four main production mechanisms of a SM Higgs boson (ggH, VBF, VH and $t\bar{t}H$) with a mass of 125 GeV.

The background sample is the cross-section weighted mixture of SM backgrounds including prompt-prompt (from diphoton+jets sample), prompt-fake (from photon+jet and QCD dijet samples) and fake-fake (from QCD dijet samples) contributions as described in sections 5.1. Signal and background samples are split based on the event number in a training sample, containing 3/4 of the available events, and a test sample, containing the remaining 1/4 of the events and used in later stages of the analysis. The split is necessary in order to avoid any bias due to overtraining in later category extraction and optimisation.

The performance in terms of discriminating power between signal and background events of the BDT classifier is summarised in Fig. 5.13 where the signal efficiency as a function of background rejection (ROC) and the normalised shapes of the BDT output variable for signal and background are shown. Both ROC and shapes are extracted using the test sample, in order to prevent any bias in the results.

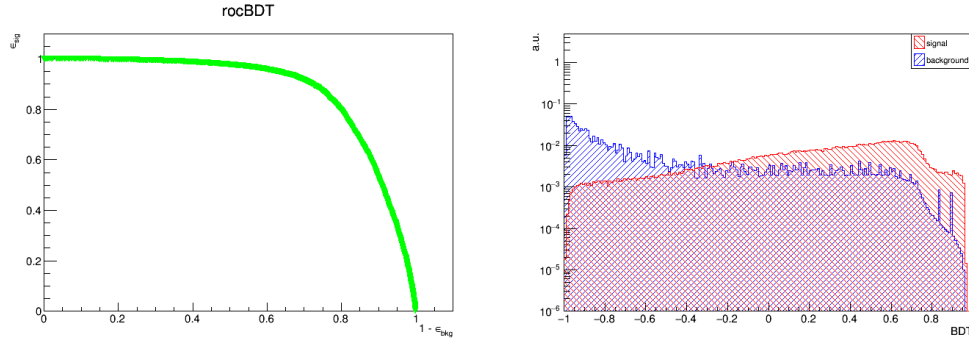


Figure 5.13: The diphoton BDT classifier ROC curve (left) and the shapes normalised to unity of the BDT output variable for signal and simulated background events (right).

The diphoton invariant mass distribution shape for background events, normalised to 1 in different signal efficiency (ϵ_{sig}) ranges, is shown in Fig. 5.14 (on the top). It can be seen that the distribution is smooth in all efficiency bins, confirming the mass-blindness of the BDT classifier. The same distributions for signal events, in Fig. 5.14 (on the bottom), on the other hand, show how events with narrower mass resolution tend to have higher BDT scores.

Data-simulation comparison for the transformed diphoton MVA output (to range between 0 and 1) distribution in the mass region $100 < m_{\gamma\gamma} < 180$ GeV is shown in Fig. 5.15. The diphoton MVA was transformed such that the sum of signal events from all production

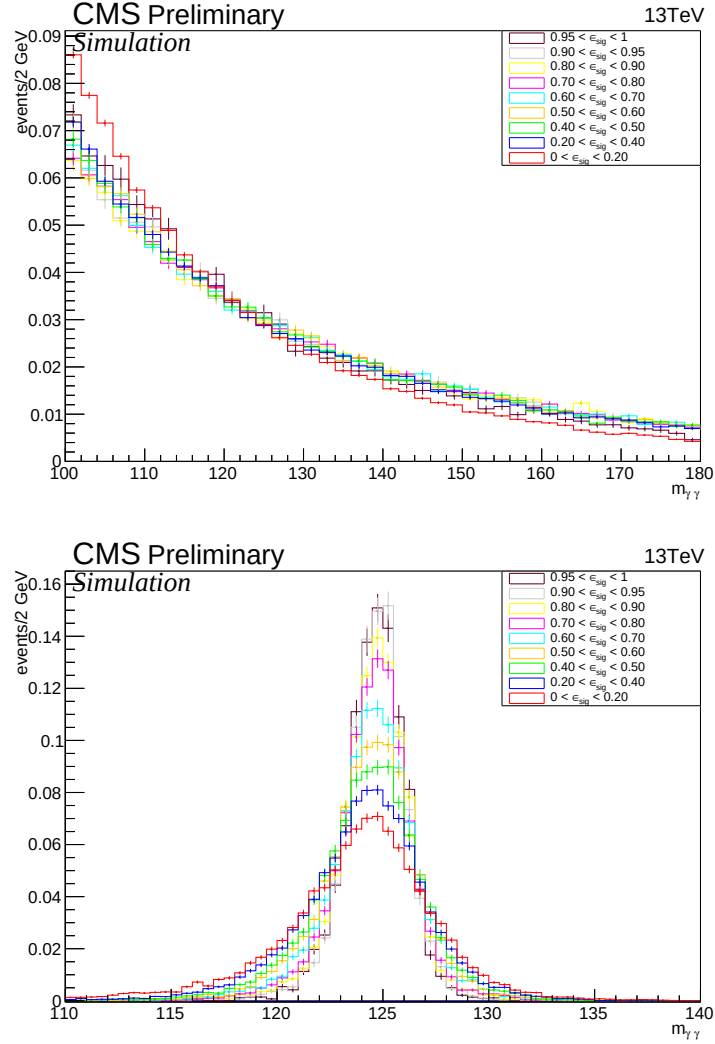


Figure 5.14: Diphoton mass distributions in every signal efficiency bin, normalised to unity, for signal (top) and background (bottom) events.

modes has a uniform, flat, distribution.

5.9.2 Systematic uncertainties

Since the background is modelled in a fully data-driven manner (described in section 5.13.2), the result does not rely on the simulation prediction for the event classifier output shape of the various background components. On the opposite, the signal modelling is based on simulation prediction with the relevant correction from data applied to simulation. The two main sources of systematic uncertainty on the diphoton MVA output

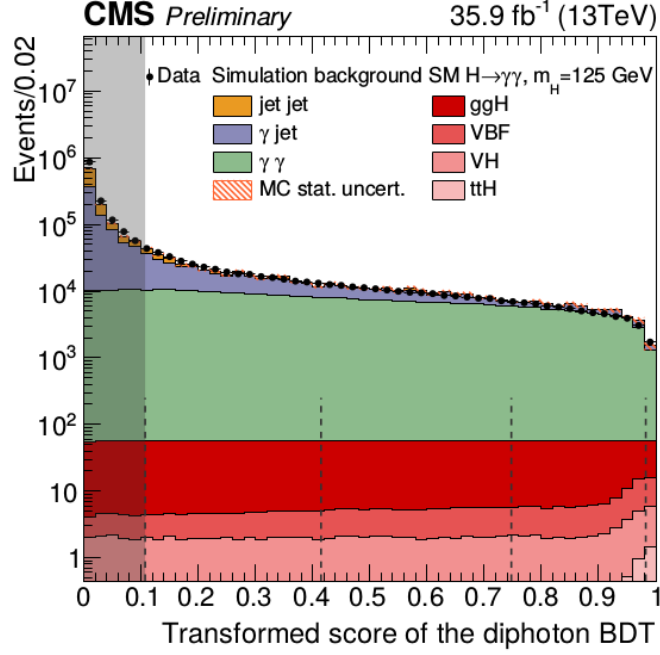


Figure 5.15: The shape of the transformed diphoton MVA variable is compared between simulation (stacked histograms) and data (black points).

are arising from the photon identification output and the per-photon energy resolution estimate from the regression.

The photon identification output is a continuous discriminator with output ranging from -1 to 1 , with prompt photons tending to have values close to 1 . In order to assign a systematic uncertainty to this variable, we shift its value for every photon in the simulation according to a transformation combining a shift of ± 0.03 with a linear increase of the uncertainty for events with a low score from the photon identification output, as described in section 7.2.3. Since the value of the diphoton identification output tends to be larger for larger values of the photon identification output, the simultaneous translation of the photon identification MVA for both photons leads to maximal migration of events in the diphoton identification output, which is then propagated as a migration of signal yield among the final event classes.

The per-photon resolution estimate is also affected by an imperfect modelling of the electromagnetic shower shape in the simulation. This quantity affects the diphoton identification output through its impact on the mass resolution estimate, both under the right and wrong vertex hypotheses (σ_{rv} and σ_{wv}). The impact is diluted since it competes with the simulated energy resolution smearing term. The systematic uncertainty from modelling of σ_E/E is assigned by shifting its value by $\pm 5\%$ for each photon, as mentioned in

section 5.13.4. The diphoton identification output is trained to get lower values for larger energy resolution estimate.

The impact of these two sources of systematic uncertainty on the diphoton MVA output variable is shown in Fig. 5.16. It is constructed using $Z \rightarrow e^+e^-$ events. One can see that the difference between data and simulation is covered by statistical and systematic uncertainties.

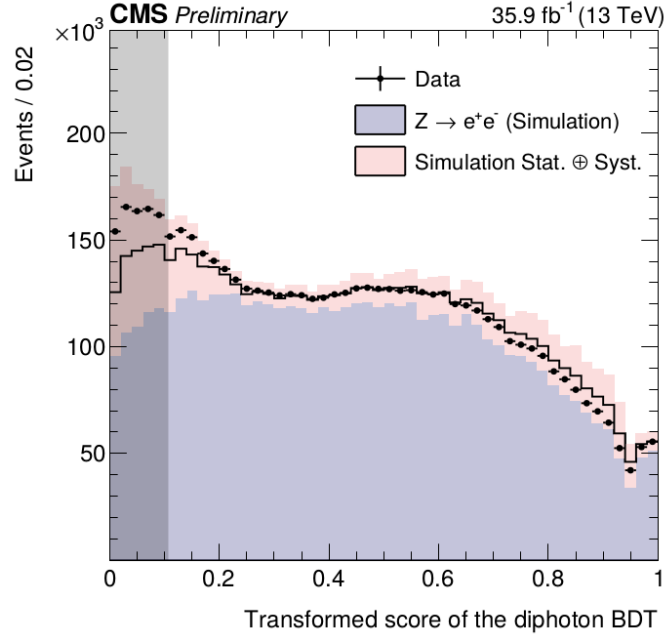


Figure 5.16: Data-simulation comparison for transformed diphoton MVA output for electrons reconstructed as photons from the decay of the Z boson. The pink band indicates the impact of the photon identification output and σ_E/E systematic uncertainties on the diphoton MVA output variable.

5.9.3 Event categorization with diphoton MVA output

Splitting the events into four categories increases the expected significance of the analysis. The category boundaries in the event classifier output spectrum are chosen as a result of an optimisation method, described below.

For an n -category analysis, n boundaries are initially placed at equal distances in signal quantiles from each other, given the two fixed boundaries at the edges of the spectrum -1 and 1 . Using n boundaries $n + 1$ categories are defined, of which only n are selected for the analysis and for the calculation of the expected significance, while the events falling in the lowest-BDT score category are discarded. The positions of the boundaries are free

to float and are adjusted minimising the combined p-value⁴. In each category, signal and background models are extracted from simulation. The background model is extracted through an exponential shape fit to the diphoton invariant mass distribution in simulated background samples. The sum of simulated background components is scaled by a factor $\simeq 1.38$ in order to match the normalisation obtained in data. The signal model is extracted through the fit to the same variable in simulated signal samples. The shape used to extract the signal model is the sum of two gaussian functions, one fitting the signal component where the mass resolution is completely dominated by the photon energy resolution (i.e. the distance dz between the selected vertex for mass reconstruction and the correct vertex is smaller than 1 cm along the z coordinate) and the other gaussian fitting the worse-resolution signal component given by the non-negligible term related to the wrong choice of the vertex ($dz > 1$ cm). The p-value is then extracted from a fit to an Asimov toy dataset⁵ [174] created from the signal+background model in each category.

Since the p-value has to be computed many times during the optimisation, no systematic uncertainty on the signal is included in the procedure and the simplified models for signal and background described above are chosen.

The procedure is done from $n = 2$ up to $n = 6$. Given the small gain achievable going from $n = 4$ to 5, as shown in Fig. 5.17, the untagged events are split into 4 categories. The optimised positions of boundaries defining these categories are found to be $-0.405, 0.204, 0.564, 0.864$, expressed in terms of the diphoton MVA output variable.

5.10 VBF selection

5.10.1 Analysis strategy

In order to be tagged as VBF-like, the events are required to have at least 2 jets, separated by a large rapidity gap. Jet reconstruction and selection are shown in section 5.10.2. In order to reduce background contributions, jets are required to meet selection criteria aimed at rejecting fake jets due to the clustering of several low energy emissions from pileup interactions, as described in section 5.10.2. Signal events are selected using a multivariate analysis, which combines two additional variables with the diphoton MVA output described in section 5.9. First an additional multivariate discriminant (dijet MVA) exploiting the kinematic properties of the jets (section 5.10.3) is trained. This variable is combined with the diphoton MVA and diphoton p_T divided by $m_{\gamma\gamma}$ to form the combined dijet MVA.

⁴The p-value is the probability, under assumption of a null hypothesis H_0 , of obtaining a result as compatible with H_0 . If H_0 is background-only hypothesis, the p-value is the probability that the background fluctuates to the observed value. In case of an observed excess above the expected background, the p-value is used to estimate the significance of the excess.

⁵Estimation of the median significance by replacing the ensemble of simulated data sets by a single representative one, referred to here as the "Asimov" dataset. The name of the Asimov data set is inspired by the short story Franchise, by Isaac Asimov). In it, elections are held by selecting the single most representative voter to replace the entire electorate.

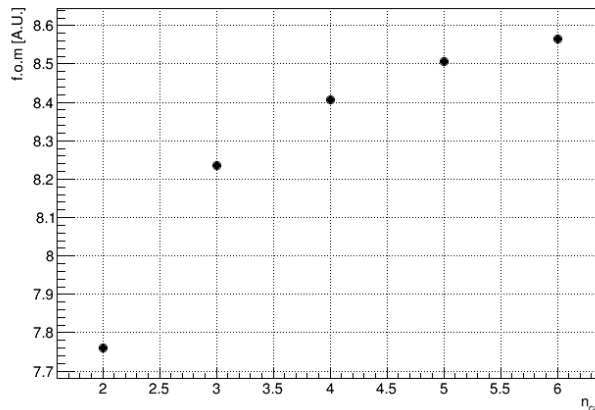


Figure 5.17: Expected significance as a function of the number of analysis categories. The estimate of the expected significance is approximated since it is extracted from a simplified signal+background fit. The relative gain in expected significance is of the order of 1% after $n = 4$.

The combined dijet MVA is then used to define three categories which are optimised to maximise the signal significance. The procedure for the categorization is described in section 5.10.4. Events failing the exclusive VBF selection, either at the preselection level or because their combined dijet MVA score is too low, are subsequently tested for inclusion in the untagged categories.

5.10.2 Jet definition

While quarks and gluons are abundantly produced in high-energy processes such as hard scattering of partons in pp collisions, they cannot be observed directly as they immediately fragment and hadronize, giving rise to collimated showers of particles referred to as jets. Obtaining information on the original partons thus requires to reconstruct jets, i.e. to accurately combine the reconstructed particles and to determine the total jet momentum. It is also crucial to reach a good understanding of the jet energy scale and resolution, since they are usually an important component of systematic uncertainties.

Jets are reconstructed from the aforementioned collection of PF particles. They are grouped using a clustering algorithm that consists in defining some measure of distance between pairs of particles, and iteratively combining the closest ones. More precisely, CMS uses the anti- k_T algorithm [175], with a radius parameter of 0.4, from Particle Flow (PF) candidates (described in section 3.7). Charged candidates associated with a vertex other than the selected vertex for the event are excluded, and this procedure is called Charged Hadron Subtraction (CHS) [142]. Jets are required to have $|\eta| < 4.7$.

The particles produced in the pileup interactions are sometimes clustered by the jet

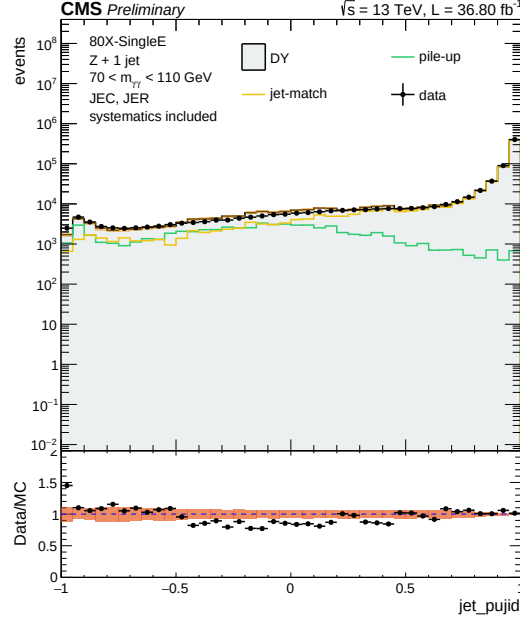


Figure 5.18: Data and simulation distributions of the per-jet PUJID variable for events containing a Z boson balanced with one jet.

clustering algorithm into objects of relatively large p_T . The resulting "pileup jets" are removed using selection criteria based on the jet shape and the compatibility of the jets tracks with the primary vertex. An algorithm called, Pileup-Jet Identification (PUJID) exploits these characteristics [176] and this algorithm returns a PUJID jet score, which is shown in Fig 5.18. The pileup contribution comes mainly from jets with $|\eta| > 2.5$, where the tracker is no longer present and therefore pileup jets can no longer be mitigated by the CHS technique. A PUJID working point which ensures 80% of quark-jet efficiency is used [176].

5.10.3 Kinematic dijet MVA

Events passing the following criteria are pre-selected as potential VBF candidates:

- Two photons passing the diphoton preselection described in section 5.5,
- Two photons passing the photon identification MVA > -0.2 ,
- $p_T^{\gamma_1}/m_{\gamma\gamma} > 1/3$ and $p_T^{\gamma_2}/m_{\gamma\gamma} > 1/4$ ⁶,

⁶Those cuts were optimized for different categories to reject more background in each individual category.

- Two jets passing the pileup jet identification described in the previous section and with $p_T^{j_1} > 30$ GeV, $p_T^{j_2} > 20$ GeV,
- dijet mass $m_{j_1 j_2} > 250$ GeV.

In order to increase the number of simulated background events during the training procedure, the preselection for jets and photons is relaxed. The minimum thresholds for $p_T^\gamma/m_{\gamma\gamma}$ on the leading and sub-leading photons are reduced to 1/4 and 1/5, respectively. The threshold on the dijet invariant mass is reduced to 100 GeV. The diphoton invariant mass is restricted to the range 100 GeV to 180 GeV.

The kinematic dijet MVA is built using the following variables:

- $p_T^{\gamma_1}/m_{\gamma\gamma}$ and $p_T^{\gamma_2}/m_{\gamma\gamma}$,
- $p_T^{j_1}$ and $p_T^{j_2}$,
- the dijet invariant mass, $m_{j_1 j_2}$,
- the difference in pseudo-rapidity between the two jets, $\Delta\eta_{j_1 j_2}$,
- the centrality defined as,

$$C_{\gamma\gamma} = \exp\left(-\frac{4}{(\eta_1 - \eta_2)^2} \left(\eta_{\gamma\gamma} - \frac{\eta_1 + \eta_2}{2}\right)^2\right) \quad (5.4)$$

where η_1 and η_2 are the pseudo-rapidities of the two leading jets. This variable quantifies the η position of the diphoton system with respect to the two leading jets in the event. It takes a value of 1 when the diphoton is halfway in η between the two jets, $1/e$ when the diphoton is aligned with one of the jets, and $< 1/e$ when the object is not between the jets in η .

- the difference in azimuthal angle between the two leading jets $\Delta\phi_{j_1 j_2}$,
- the minimum distance between a leading or subleading jet and leading or subleading photon $\min \Delta R(\gamma, jet)$. This variable is larger for the VBF signal.

The kinematic variables are combined by means of a Boosted Decision Tree using a Gradient Boosting (BDTG) algorithm. The BDTG is trained on simulated events: VBF sample with $m_H = 125$ GeV is used as signal and all main background samples (diphoton+jets, $\gamma + jet$ and dijet) are used in the training. Gluon fusion $H \rightarrow \gamma\gamma$ with $m_H = 125$ GeV is used as an additional source of background in the training in order to improve the purity of the dijet signal.

5.10.4 Combined dijet MVA and categorization

The Combined dijet MVA is built using the kinematic dijet MVA, the diphoton MVA, and $p_T^{\gamma_1\gamma_2}/m_{\gamma\gamma}$ as inputs. The purpose of this variable is to discriminate the VBF dijet signal from background utilising the information from all relevant objects tagged in the event. The ratio of $p_T^{\gamma_1\gamma_2}/m_{\gamma\gamma}$ is included as an input because of its significant correlation to both the dijet MVA and the diphoton MVA. The information is combined by means of a Gradient boosting algorithm. The method is trained on simulated events: VBF sample with $m_H = 125$ GeV is used as signal. All standard backgrounds are used in the training. Fig. 5.19 left distribution shows the BDT output for the VBF, ggH simulation and the background estimated from the data sidebands. The Combined dijet MVA provides good separation between the signal and background. A validation of the Combined dijet MVA obtained in $Z \rightarrow e^+e^-$ jets events, where the electrons are reconstructed as photons and at least two jets is shown on the right in Fig. 5.19 for data and simulation. Data and simulation agree within the uncertainties.

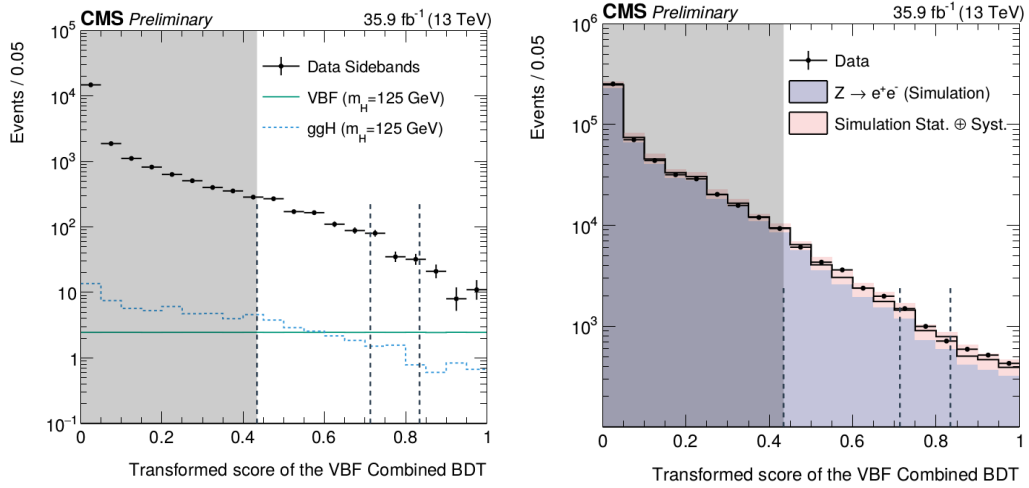


Figure 5.19: Left: Score of the combined dijet MVA of the ggH and VBF signal distributions, compared to background taken from data in the mass sideband region. The signal region selection is applied to the expected simulated ggH and VBF events; these are compared to points representing the background, as determined from data using the signal region selection in mass sidebands. Right: Score of the combined dijet MVA with $Z \rightarrow e^+e^-$ jets events. The signal selection is applied with electrons reconstructed as photons, with points showing the score for data and the histogram showing the score for simulated Drell-Yan events, including statistical and systematic uncertainties (pink band). Dotted lines indicate the three VBF categories, while the grey region is discarded from the analysis.

Three VBF categories are defined by simultaneously optimising three cuts on the combined MVA score, these constitute the category boundaries. For a given number of categories N points in an N -dimensional volume where each dimension corresponds to a choice

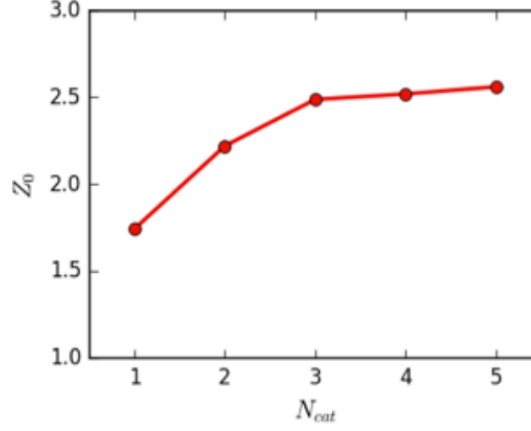


Figure 5.20: The effect of increasing number of categories, there is little gain beyond three.

of boundary are evaluated. The point with the highest overall significance is then chosen. In more details, a double Gaussian model to fit the signal and extract the peak position and the width is used. An exponential function to model the background is used. This is then used to estimate the significance in a $\pm 2\sigma_{eff}$ ⁷ window around the peak [177]:

$$Z_0 = \sqrt{2 \left((s + b + b_{reg}) \ln \left(1 + \frac{s}{b + b_{reg}} \right) \right)} \quad (5.5)$$

where s is the number of signal events, b is the number of background, and b_{reg} is regularisation term set to 5 as it maximizes the significance.

As shown in Fig. 5.20 the significance is maximal starting from 3 categories, and there is no gain in creating more.

Signal events accepted by the VBF-0 category consist of 77% VBF and 22% ggH events. For the other categories, these two signals each make up 65% and 34% for VBF-1 and 47% and 50% for VBF-2, of the total accepted signal. The remainder in each case is from the other production modes.

5.10.5 Systematic uncertainties

Many of the corrections and uncertainties impacting the VBF channels are common with other channels in the analysis and covered elsewhere. In this section, systematics related to jets and to the VBF topology are described. The jet energy corrections and jet energy resolution [178] are applied:

1. Jet energy corrections (JEC) are applied to data and simulation. The uncertainties on

⁷ σ_{eff} is the smallest interval in the distribution which contains 68.3% of the entries

the corrections are applied to simulation and used to derive the magnitudes of shifts between VBF categories, as well as between the VBF categories and the untagged categories. Nuisance parameters⁸ are defined such that migrations between the two VBF tags can be adjusted independently from those to untagged events in the final fit. JEC account for a 8 to 18% migration within VBF categories and 11% from VBF to untagged categories.

2. Jet energy resolution corrections are applied to simulation, and they are derived as explained in reference [178]. This shifts imply event migrations between categories similarly to the jet energy corrections. The jet energy resolution has an impact on the event migration of less than 3%.
3. Uncertainties on the Pileup jet identification are obtained by comparing the PUJID score value distributions in data and simulation for a sample of $Z + 1$ jet events and finding the shift required in order to cover the data/MC disagreement. The shift is typically around 0.01 but can go up to 0.06 for $30 \text{ GeV} < p_T < 50 \text{ GeV}$, $2.75 < |\eta| < 3.0$. The discrepancy between data and simulation is propagated to estimate the event migration. The migration is found to be the order of 1%.

5.11 VH selection

The Higgs Strahlung production mechanism, where a V is produced in association with the Higgs boson, can be selected by requiring the presence of the decay products of the V boson. In VH production mode $V=W/Z$, the possible final states are:

- $W \rightarrow l\nu$ and $Z \rightarrow l^+l^-$ which are tagged by requiring at least one lepton as described in section 5.11.2,
- $Z \rightarrow \nu\nu$ and $W \rightarrow l\nu$ (when lepton is not well reconstructed or out of acceptance), which is tagged with the presence of high missing transverse energy as described in section 5.11.3,
- $Z/W \rightarrow q\bar{q}$, which are tagged by the presence of jets as described in section 5.11.4.

5.11.1 Objects definition

Muons. Muon tracks are first reconstructed independently in the inner tracker and in the outer muon system. They are called tracker tracks and standalone-muon tracks, respectively. The muon reconstruction algorithm then combines the information from both subsystems using two approaches [138].

⁸In the statistical analysis each independent source of systematic uncertainty is assigned a nuisance parameter.

- Global muons are formed by propagating standalone-muon tracks inwards to the inner tracker. If a matching tracker track is found, a global-muon track is fitted combining the hits from the tracker and standalone-muon track, using the Kalman Filter technique [166].
- Tracker muons are formed by extrapolating tracker tracks outwards to the outer muon system, requiring that at least one muon track segment made of DT or CSC hits matches the extrapolated track. The possibility for tracker muons to have one single matched segment in the muon system makes this algorithm more efficient than global-muon reconstruction at low transverse momentum ($p_T < 5$ GeV).

About 99% of muons produced in pp collisions with high enough momentum are reconstructed by either algorithm. Muon candidates found by both methods (sharing the same tracker track) are merged into one single candidate. To be selected for the analysis, muons with $p_T > 20$ GeV and $|\eta| < 2.4$ have to pass the following criteria:

- χ^2/n_{dof} of the global muon track fit < 10 ,
- transverse impact parameter⁹ (d_0) of its tracker track with respect to the muon vertex < 0.2 cm,
- the longitudinal impact parameter (d_z) of the track with respect to the muon vertex < 0.5 cm,
- number of pixel hits > 0 ,
- number of tracker layers with hits > 5 ,
- number of hits in the muon chamber that are included in the global muon track fit > 0 ,
- the number of muon segments in different muon stations matching the muon track > 1 .

An important quantity for distinguishing signal and background leptons is isolation. Muon isolation (see Eq 5.6) is defined as the sum of the transverse energy of the particles (photons, charged and neutral hadrons) in a cone $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.4$ around the muon direction, excluding the muon itself, where $\Delta\eta$ and $\Delta\phi$ are the angular differences between the muon and the other particles in η and ϕ . To correct for the effects of the pileup, the charged particles contributions not originating from the primary vertex are explicitly removed from the isolation sum, and the neutral particles contribution is corrected assuming a ratio of 0.5 for the contribution of neutral to charged objects to the pileup activity. The ratio of the corrected isolation sum to the muon p_T has to be < 0.25 .

⁹The transverse (longitudinal) impact parameter of a track is defined as the transverse (longitudinal) distance of closest approach of the track to the primary vertex.

$$\begin{aligned}
Isol(\mu) = & \frac{1}{p_T(\mu)} \Sigma p_T(\text{ChgHad}_{PV}) + \frac{1}{p_T(\mu)} \max(0, \Sigma E_T(\text{NeutralHad}) \\
& + \Sigma E_T(\text{Photon}) - \frac{1}{2} \Sigma p_T(\text{ChgHad}_{\text{pileup}}))
\end{aligned} \tag{5.6}$$

Electrons. Electrons are reconstructed within geometrical acceptance of the tracker $|\eta| < 2.5$. The reconstruction combines information from cluster energy deposit in the ECAL and the electron tracks reconstructed in the tracker. Trajectories in the tracker are reconstructed using a dedicated modelling of the electron energy loss and fitted with a Gaussian sum filter [132, 167]. The electron momentum is determined from the combination of the ECAL and the tracker measurements. Electron identification relies on cut-based approach with a separate set of criteria in the barrel (EB) and endcap (EE) regions:

- ECAL crystal based shower covariance in the η direction ($\sigma_{i\eta i\eta}$, defined in section 7.1.1) < 0.0103 (EB) and < 0.0301 (EE),
- ratio of the energy measured in the HCAL to that in the ECAL (H/E) < 0.104 (EB) and < 0.0897 (EE),
- pseudorapidity distance between the energy weighted position of the supercluster and the track position of closest approach to the supercluster, extrapolated from the innermost track position and direction: $\Delta\eta_{in} < 0.0105$ (EB) and < 0.00814 (EE),
- distance in ϕ between the energy weighted position of the supercluster and the track position of closest approach to the supercluster, extrapolated from the innermost track position and direction: $\Delta\phi_{in} < 0.115$ (EB) and < 0.182 (EE),
- difference between the inverse electron energy measured in the ECAL and the inverse momentum measured in the tracker ($1/E - 1/p$) < 0.102 (EB) and < 0.126 (EE)
- number of missing inner layer hits in the electron track < 2 (EB) and < 1 (EE),
- transverse impact parameter (d_0) < 0.0261 (EB) and < 0.118 (EE),
- transverse impact parameter (d_z) of the electron track with respect to the vertex < 0.41 (EB) and < 0.822 (EE),
- relative combined isolation corrected for pileup using effective areas (see Eq. 5.7) in a cone of $\Delta R = 0.3$ < 0.0893 (EB) and < 0.121 (EE).

$$\begin{aligned}
Isol(e) = & \frac{1}{p_T(e)} \Sigma p_T(\text{ChgHad}_{PV}) + \frac{1}{p_T(\mu)} \max(0, \Sigma E_T(\text{NeutralHad}) \\
& + \Sigma E_T(\text{Photon}) - \rho \times EA)
\end{aligned} \tag{5.7}$$

- electrons are vetoed if they are found to match electrons coming from converted photons. An electron is matched to the conversion if its track is found to be the same as the conversion one.

Selected electrons are required to have $p_T > 20$ GeV and pseudorapidity $|\eta| < 1.4442$ or $1.566 < |\eta| < 2.5$.

Jets. Jets are reconstructed as described in section 5.10.2. The selection is also similar, but for the pseudorapidity $|\eta| < 2.4$ and the PUJID selection, which is looser and designed to have 99% jet efficiency.

Missing transverse energy (MET). The $\vec{p}_T^{miss}$ vector is defined as the projection onto the plane perpendicular to the beam axis of the negative vector sum of the momenta of all reconstructed particle-flow objects in an event. Its magnitude is referred to as p_T^{miss} . In VH production mode, the p_T^{miss} originates from the neutrino(s) in the final state.

5.11.2 Leptonic channel

An exclusive selection of diphoton events produced in association with at least one high p_T lepton originating from the leptonic decay of vector bosons in the VH mechanism is done in the analysis. Requiring a lepton suppresses QCD multijet backgrounds. The main remaining background comes from electroweak processes $Z\gamma/W\gamma$, where Z or W bosons decay leptonically.

5.11.2.1 Selections

Most signal events come from WH production due to the cross section ($\sigma_{WH} = 1.373$ pb and $\sigma_{ZH} = 0.884$ pb) and branching fraction to leptons ($\mathcal{B}(Z \rightarrow \ell\ell) \sim 7\%$, $\mathcal{B}(W \rightarrow \ell\nu) \sim 20\%$). Most signal events have a single lepton and a neutrino, giving rise to high missing- E_T . By requiring two leptons, one can also select events from ZH. Three sub-categories are defined with at least one lepton tagged.

To be selected in one of the sub-categories, events have to satisfy the following requirements:

- leading photon $p_T / m_{\gamma\gamma} > 0.375$,
- sub-leading photon $p_T / m_{\gamma\gamma} > 0.25$,
- $\Delta R(\text{photon}, \text{lepton}) > 1.0$,
- for channels involving electron $\Delta R(\text{photon}, \text{GSFtrack}^{10}) > 0.4$ and $|m_{e,\gamma} - m_Z| > 20$ GeV.

¹⁰GSFtrack is the reconstructed electron track.

For **ZH leptonic** channel events have to pass additional cuts:

- at least two opposite sign, same flavour leptons,
- two leptons invariant mass $70 \text{ GeV} < M_{\ell\ell} < 100 \text{ GeV}$,
- diphoton MVA > -0.405 . The chosen value corresponds to the lowest boundary of the untagged categories and a loose selection is sufficient given the high purity of this category.

For **WH leptonic** category events have to pass additional cuts:

- $p_T^{miss} > 45 \text{ GeV}$,
- diphoton MVA > 0 . It allows to reject about 50% of the background and to retain 90% of the signal.

For **VH loose** category events have to pass additional cuts:

- $p_T^{miss} < 45 \text{ GeV}$,
- diphoton MVA > 0 . It allows to reject about 50% of the background and to retain 85% of the signal.

An event is rejected from WH and VH loose categories if it has more than 2 jets, to reduce the contamination from $t\bar{t}H$ events, with the following jets:

- $p_T > 20 \text{ GeV}$,
- $|\eta| < 2.4$,
- $\Delta R(\text{jet}, \text{lepton}) > 0.5$,
- $\Delta R(\text{jet}, \text{photon}) > 0.5$.

5.11.2.2 Systematic uncertainties

Lepton identification efficiency. For both electrons and muons, the uncertainty on the identification efficiency is computed varying the correction for simulation (computed as a ratio of efficiency measured in data and simulation) by its uncertainty. The resulting difference in the signal efficiency estimated from the simulation is taken as systematic uncertainty ($\sim 1\%$). The systematic uncertainties are propagated as signal migration to the inclusive categories, in WH leptonic category it is 1.5% and in ZH leptonic one is 3.0%.

5.11.3 Missing transverse energy channel

Signal events come from ZH production mode, when $Z \rightarrow \nu\nu$ or from WH production mode, when $W \rightarrow l\nu$ and the lepton is either badly reconstructed or is out of detector acceptance.

5.11.3.1 Selection

The requirements for an event to be selected in this channel are the presence of two photons passing preselection criteria, large missing transverse energy and further kinematic requirements both to enhance the signal over background ratio and to reduce the contamination of the ggH events, as well as the instrumental effects.

Events have to satisfy the following requirements:

- leading photon $p_T / m_{\gamma\gamma} > 0.375$,
- subleading photon $p_T / m_{\gamma\gamma} > 0.25$,
- diphoton MVA > 0.6 ,
- $p_T^{miss} > 85 \text{ GeV}$,
- $\Delta\Phi(\gamma\gamma, p_T^{miss}) > 2.4$.

5.11.3.2 Systematic uncertainties

p_T^{miss} measurement. The systematic uncertainty is accounted for by shifting the p_T^{miss} by $\pm 1\sigma$ in reconstructed p_T of the particle flow candidates and propagating that through to the p_T^{miss} . The effect of this shift is event migration to the VH leptonic categories, in $W \rightarrow l\nu$ case it is 1.3% and in $Z \rightarrow \nu\nu$ one is 1.5%.

5.11.4 Hadronic channel

Events where the Higgs boson is produced in association with a hadronically decaying vector boson $V \rightarrow jj$ (with $V = W/Z$) are selected for this category. The full final state is reconstructed requiring the presence of two photons and two jets. The full reconstruction of the vector boson decay chain ($V^* \rightarrow VH$) provides strong means to reduce the background, mainly due to diphoton events produced in association with two hard jets arising from pileup, radiation or underlying event activity.

Since the energy scale of the event is set by the masses of the decaying bosons (H and V), signal objects in the final state will result, on average, into a harder spectrum than objects arising from background. As the two jets arise from the decay of a massive boson, their invariant mass peaks around the boson mass, which is not the case for background

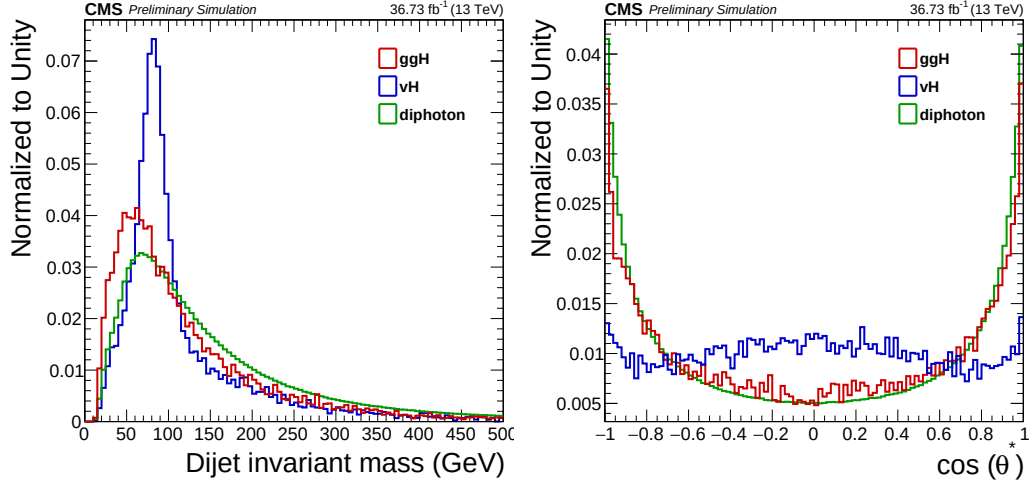


Figure 5.21: Dijet invariant mass (left) and $\cos\theta^*$ (right) distributions. The expected shape of VH events (blue line) is compared to the shape of gluon fusion Higgs boson production (red line) and the major background: diphoton continuous production (green line). All distributions are normalized to unity.

events. The invariant mass of the jets is shown for VH, ggH and background events in Fig. 5.21 (left).

Higgs boson production in association with an electroweak vector boson occurs via $V^* \rightarrow VH$. The angle between the direction of the V^* in the laboratory frame and its decay products in the V^* rest frame, named θ^* , provides further discrimination between signal and background. The distribution of $\cos\theta^*$ is shown in Fig. 5.21 (right) for VH, ggH and diphoton+jets background simulated events.

5.11.4.1 Selection

Events are selected if they satisfy the following requirements:

- leading (sub-leading) photon $p_T > m_{\gamma\gamma}/2$ ($m_{\gamma\gamma}/4$),
- $p_T^{\gamma\gamma}/m_{\gamma\gamma} > 1$,
- at least two jets with $p_T > 40$ GeV and $|\eta| < 2.4$,
- $60 < m_{jj} < 120$ GeV,
- $|\cos\theta^*| < 0.5$,
- diphoton MVA > 0.7 .

5.11.4.2 Systematic uncertainties

Jet energy scale and resolution. The systematic error on the jet energy scale is estimated by shifting the scale by $\pm 1\sigma$, where σ is the full jet energy scale uncertainty. The systematic error on jet energy resolution is estimated by varying the resolution according to the level of disagreement between the resolution measured in data and in simulation. It causes event migration into the inclusive categories, for WH category it is 1.2% and for ZH category 0.7%.

5.12 $t\bar{t}H$ selection

In my thesis I focus on the $t\bar{t}H$ production mode, and it is described in details in chapter 8.

5.13 Statistical analysis

Events are classified into fourteen exclusive categories, corresponding to four Higgs boson production modes. Three categories are used to classify events coming from vector boson fusion (VBF). These are noted "VBFTag 0" and "VBFTag 1" and "VBFTag 2", with the first being the most sensitive. Two further categories are reserved for events likely to have been produced in association with top quarks, decaying either leptonically or hadronically: "TTHLeptonicTag" and "TTHHadronicTag". There are 3 categories corresponding to the VH production mode with V decaying leptonically, they are noted "ZHLeptonic", "WHLeptonic" and "VHLeptonicLoose". There is one category for the VH production mode, where the Z boson decays into two neutrinos and the W boson decays leptonically, but the lepton is out of acceptance, it is noted "VH MET". The category corresponding to the VH production mode with vector boson decaying hadronically is noted "VHHadronic". In the latest analysis iteration, three new categories were added: "BBH" corresponding to $b\bar{b}H$ production mode, "TQH" and "THW", corresponding to the associative production of the Higgs boson with a single top quark, to check their contributions into "TTHLeptonicTag" and "TTHHadronicTag" categories. The remaining events are categorized into four inclusive categories, noted "UntaggedTag 0", "UntaggedTag 1", "UntaggedTag 2", "UntaggedTag 3", ordered from the most sensitive to the least sensitive category. Events which are not categorized into any of the above categories are discarded.

5.13.1 Signal model

Interpreting data requires a description of the expected signal for a SM Higgs boson. The shape of the $m_{\gamma\gamma}$ distribution is parametrised separately for each event category.

Since the mass of the Higgs boson is not precisely known, parametrisations from simulated signal samples assuming different m_H values are combined to form a parametric model. A simultaneous fit of all the different m_H samples is performed (simultaneous signal fitting (SSF) method), where the individual parameters of the functional form are themselves polynomials of m_H . The floating parameters in the fit are then the coefficients of these polynomials. The SSF method is applied separately for each process, category and right vertex (RV) or wrong vertex (WV) case, using seven different mass points: 120 GeV, 123 GeV, 124 GeV, 125 GeV, 126 GeV, 127 GeV, 130 GeV, that is illustrated in Fig 5.22 for "Untagged0" category. The parametrization of signal model with SSF method is shown in Fig. 5.23.

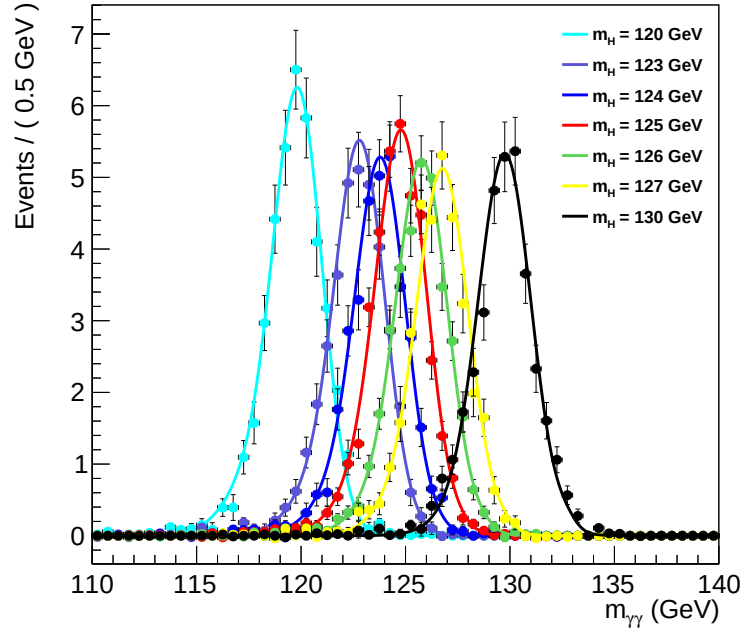


Figure 5.22: The diphoton invariant mass parametrisations in "Untagged0" category from simulated signal samples assuming seven different mass points.

For each process, category and RV/WV scenario, the $m_{\gamma\gamma}$ distributions are fitted using a sum of at most five Gaussians. The width, mean and relative size of each Gaussian used in the fits are allowed to float. If a process/category has too few simulated events to be able to make a meaningful fit, the analytic function from a higher-statistical reference process (for example, "TTHLeptonicTag" and "TTHhadronicTag" categories are merged to perform the fit) is used, with the normalisation taken from the event count in the original category.

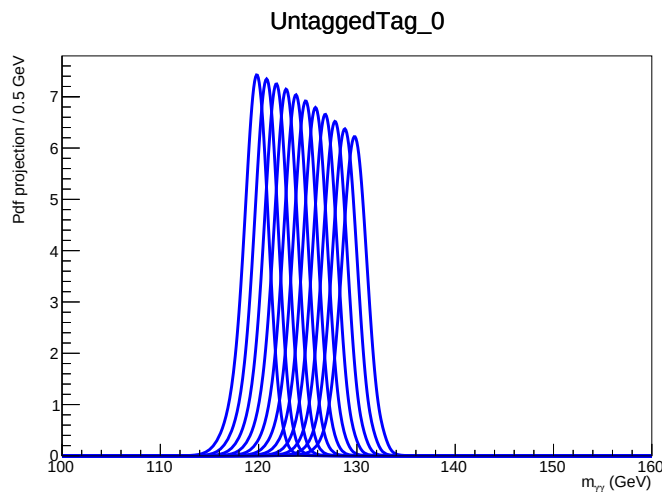


Figure 5.23: The parametrization of signal model using SSF method for "Untagged0" category.

The distribution of $m_{\gamma\gamma}$ changes considerably depending on whether the vertex associated with the candidate diphoton was correctly identified. In the cases where the correct vertex was identified, the shape of the $m_{\gamma\gamma}$ distribution is chiefly determined by the detector resolution and reconstruction. If the incorrect vertex was chosen, the shape of the $m_{\gamma\gamma}$ distribution is substantially smeared. As explained in chapter 6, if the distance between the "true" Higgs boson production location and the chosen vertex is less than 1 cm, the impact on the mass resolution is negligible and one can therefore consider that the correct vertex was chosen.

Since the vertex choice affects the shape of the $m_{\gamma\gamma}$ distribution, the modelling is performed separately for cases where the RV or WV was selected. The signal $m_{\gamma\gamma}$ distributions are all parametrised using a sum of up to five Gaussian functions. The fits for the RV and WV scenarios are summed to produce the signal model in each category. The relative size of the RV and WV shapes in the summed fit is given by the vertex selection efficiency determined in the simulation.

The final parametric model for the "Untagged0" category is depicted on the left in Fig. 5.24 and for all categories together on the right in Fig. 5.24.

5.13.2 Background model

The $m_{\gamma\gamma}$ distribution is fitted to data in the range of 100-180 GeV. The method used to treat the background in this analysis is called discrete profiling or "envelope" method [179]. The discrete profiling method was designed as a way to determine the systematic uncertainty associated with choosing a particular analytic function to fit the

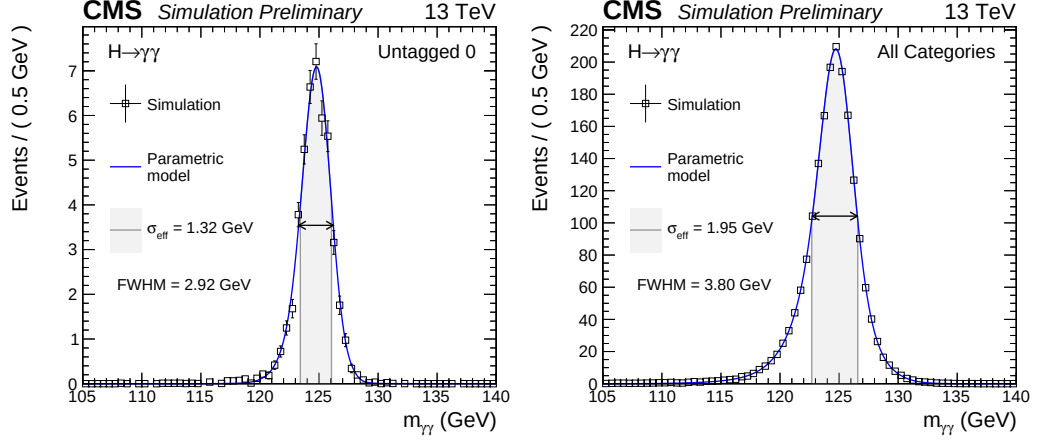


Figure 5.24: Final parametric signal model for Untagged0 category (left) and for all categories together (right)

background $m_{\gamma\gamma}$ distribution. The method treats the choice of the background function as a discrete parameter in the likelihood fit. The choice of the background function enters the statistic uncertainty instead of systematic one. For this method to be valid, a complete set of candidate function families should be considered. Four families of analytic functions are considered:

- sums of exponentials:

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

- sums of polynomials (in the Bernstein basis):

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

- Laurent series:

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

- sums of power-law functions:

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

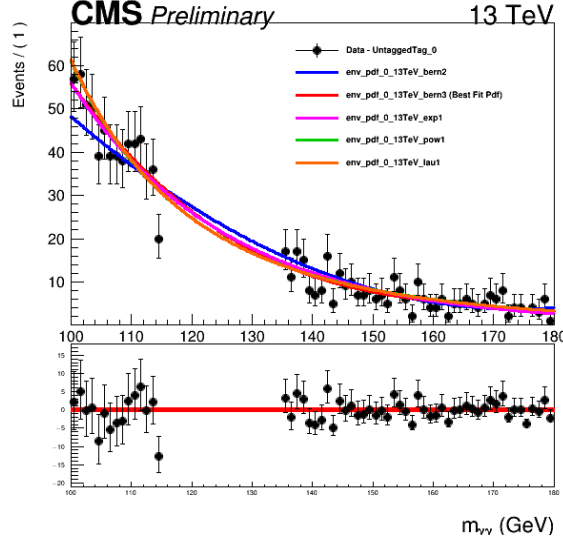


Figure 5.25: Set of functions chosen to fit the background using the discrete profiling method for each of the inclusive "Untagged" categories considered in this analysis. Four families of functions are considered: sums of exponentials, polynomials in the Bernstein basis, Lauren series and sums of power laws.

where for all i , p_i are a set of floating parameters in the fits. Example of background fits with these functions are shown in Fig. 5.25 for the "Untagged0" category.

When fitting these functions to the background $m_{\gamma\gamma}$ distribution, the value of twice the negative logarithm of the likelihood $-2\text{Log}\lambda$ is minimized, and λ is defined in the appendix A. A penalty is added to $-2\text{Log}\lambda$ to take into account the number of floating parameters in each candidate function. When making a measurement of some parameter of interest, the discrete profiling method determines the envelope of the lowest values of $-2\text{Log}\lambda$ (with appropriate penalties) profiled as a function of the parameter of interest. The envelope obtained through this method will yield a broader curve than the $-2\text{Log}\lambda$ curve obtained from a single choice of function. The 1σ uncertainty is then obtained by taking the width of the 68% range, or equivalently the points where $-2\text{Log}\lambda = 1$.

5.13.3 Results of event selection

5.13.3.1 Event yields

Table 5.5 shows the expected number of signal events for each category. The total number is broken down by percentage contribution of each production mode to any particular event category. The percentage of the processes which each category was designed to collect are in bold. σ_{eff} and σ_{HM} are also listed in the table. The former represents the smallest interval in the distribution which contains 68.3% of the entries. The latter repre-

sents the width of the distribution at half of its maximum, divided by 2.35. The expected number of background events per GeV in the corresponding σ_{eff} window around 125 GeV, which is taken from the $m_{\gamma\gamma}$ fit to data, is also listed in the table. VH processes are additionally divided between leptonically decaying Z boson (ZH lep) and W boson (WH lep). Hadronically decaying Z and W bosons are labeled as ZH had and WH had respectively. Fig. 5.26 displays the same information, with an additional plot of the number of signal events divided by the total number of signal plus background events.

Event Categories		SM 125GeV Higgs boson expected signal												Bkg (GeV ⁻¹)	
		Total	ggH	VBF	ttH	bbH	tHq	tHW	WH lep	ZH lep	WH had	ZH had	σ_{eff}		
Untagged 0	45.83	80.19 %	11.75 %	1.83 %	0.40 %	0.47 %	0.22 %	0.41 %	0.19 %	2.96 %	1.58 %	1.32	1.24	21.92	
Untagged 1	480.56	86.81 %	7.73 %	0.56 %	1.15 %	0.13 %	0.02 %	0.47 %	0.27 %	1.81 %	1.04 %	1.47	1.32	924.21	
Untagged 2	670.45	89.76 %	5.48 %	0.44 %	1.18 %	0.08 %	0.01 %	0.51 %	0.34 %	1.40 %	0.81 %	1.94	1.68	2419.53	
Untagged 3	610.07	91.13 %	4.51 %	0.48 %	1.07 %	0.07 %	0.01 %	0.55 %	0.30 %	1.21 %	0.69 %	2.62	2.28	4855.00	
VBF 0	10.01	21.69 %	77.09 %	0.34 %	0.35 %	0.29 %	0.03 %	0.03 %	0.00 %	0.18 %	0.00 %	1.51	1.30	1.60	
VBF 1	8.64	33.58 %	64.64 %	0.39 %	0.52 %	0.36 %	0.04 %	0.13 %	0.03 %	0.24 %	0.07 %	1.66	1.38	3.25	
VBF 2	27.76	50.14 %	46.46 %	0.81 %	0.73 %	0.53 %	0.07 %	0.20 %	0.06 %	0.71 %	0.27 %	1.61	1.36	18.89	
ttH Hadronic	5.85	10.99 %	0.70 %	77.54 %	2.02 %	4.13 %	2.02 %	0.09 %	0.05 %	0.63 %	1.82 %	1.48	1.30	2.40	
ttH Leptonic	3.81	1.90 %	0.05 %	87.48 %	0.08 %	4.73 %	3.04 %	1.53 %	1.15 %	0.02 %	0.02 %	1.60	1.35	1.50	
ZH Leptonic	0.49	0.00 %	0.00 %	2.56 %	0.00 %	0.02 %	0.13 %	0.00 %	97.30 %	0.00 %	0.00 %	1.65	1.43	0.12	
WH Leptonic	3.61	1.26 %	0.59 %	5.18 %	0.18 %	3.03 %	0.73 %	84.48 %	4.33 %	0.12 %	0.09 %	1.64	1.43	2.09	
VH LeptonicLoose	2.75	9.16 %	2.70 %	2.34 %	0.57 %	1.81 %	0.13 %	63.62 %	18.87 %	0.56 %	0.23 %	1.67	1.56	3.50	
VH Hadronic	9.69	57.38 %	3.68 %	3.61 %	0.35 %	1.39 %	0.27 %	0.17 %	0.42 %	20.47 %	12.26 %	1.38	1.31	7.22	
VH Met	4.25	23.63 %	2.46 %	14.45 %	0.41 %	2.00 %	1.14 %	25.17 %	28.60 %	1.32 %	0.82 %	1.55	1.38	3.49	
Total	1883.77	86.96 %	7.09 %	1.00 %	1.09 %	0.15 %	0.04 %	0.81 %	0.42 %	1.55 %	0.89 %	1.95	1.62	8264.73	

Table 5.5: The expected number of signal events per category and the percentage breakdown per production mode. The effective σ is provided as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events per GeV around 125 GeV is also shown.

5.13.3.2 Event distributions

Fig. 5.27, 5.28 and 5.29 show the final models for the 125 GeV Higgs boson mass scenario with all production processes summed as described previously for each category individually. Fig. 5.30 shows the signal efficiency $\times$ acceptance of the event selection as a function of the mass hypothesis.

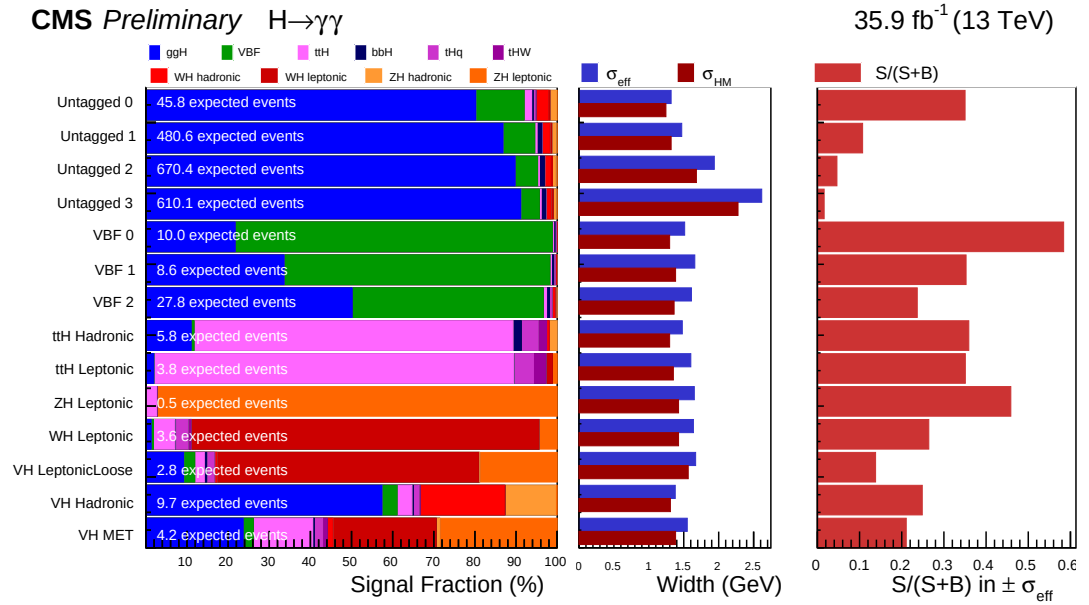


Figure 5.26: The expected number of signal events per category and the percentage breakdown per production mode. The effective σ is also provided as an estimate of the $m_{\gamma\gamma}$ resolution in that category and compared directly to the full-width at half maximum. The ratio of the number of signal events (S) to the number of signal plus background events (S+B) is shown on the right hand side.

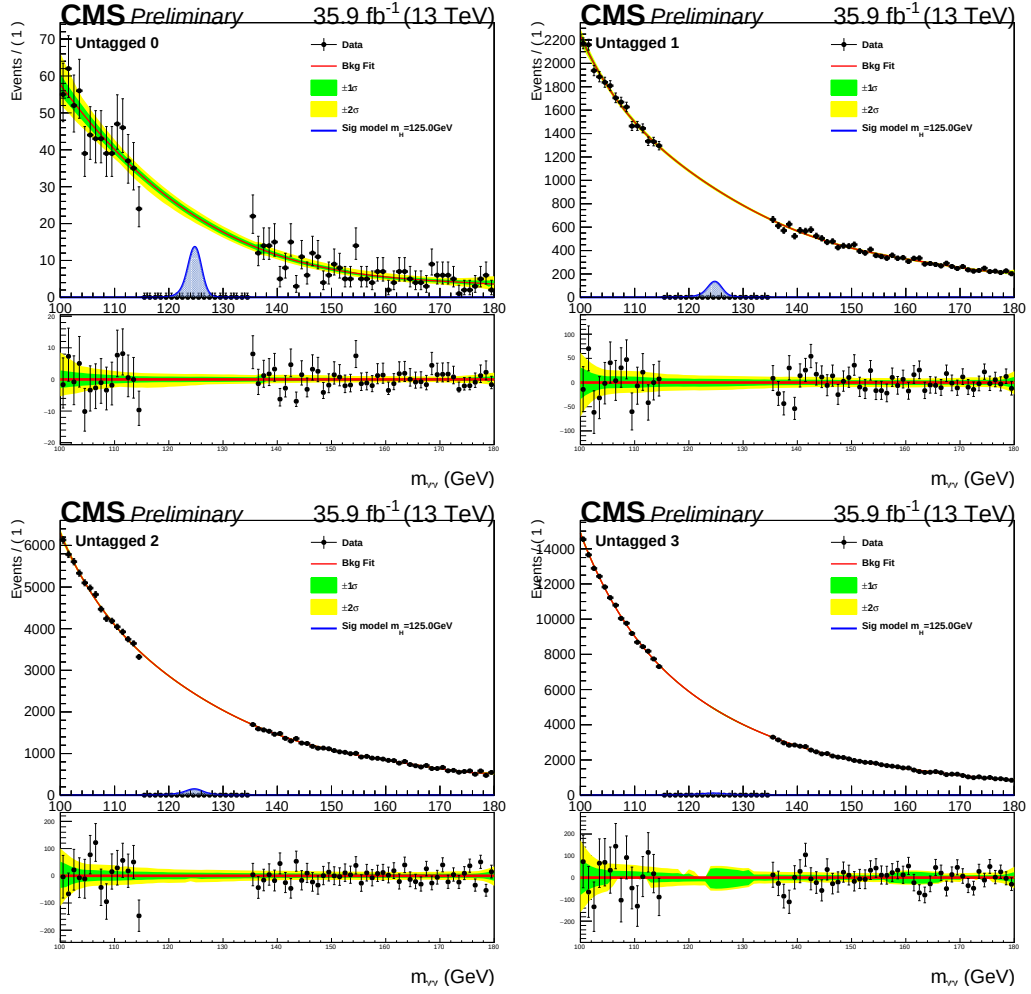


Figure 5.27: The best fit background parametrisation plotted alongside the data in each inclusive "Untagged" category. The green ($\pm 1\sigma$) and yellow ($\pm 2\sigma$) bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

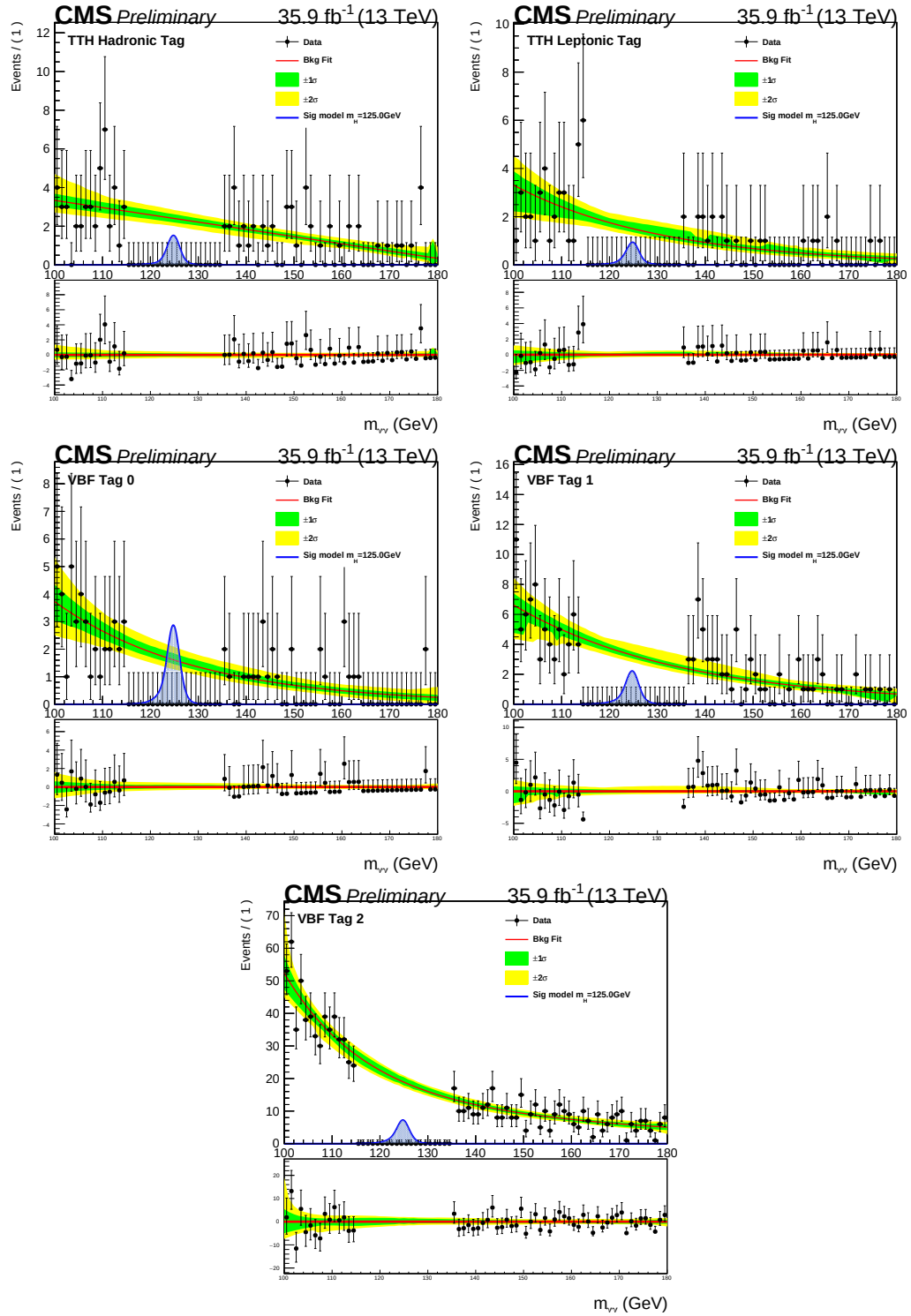


Figure 5.28: The best fit background parametrization plotted alongside the data in the $t\bar{t}H$ and the VBF categories. The green ($\pm 1\sigma$) and yellow ($\pm 2\sigma$) bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

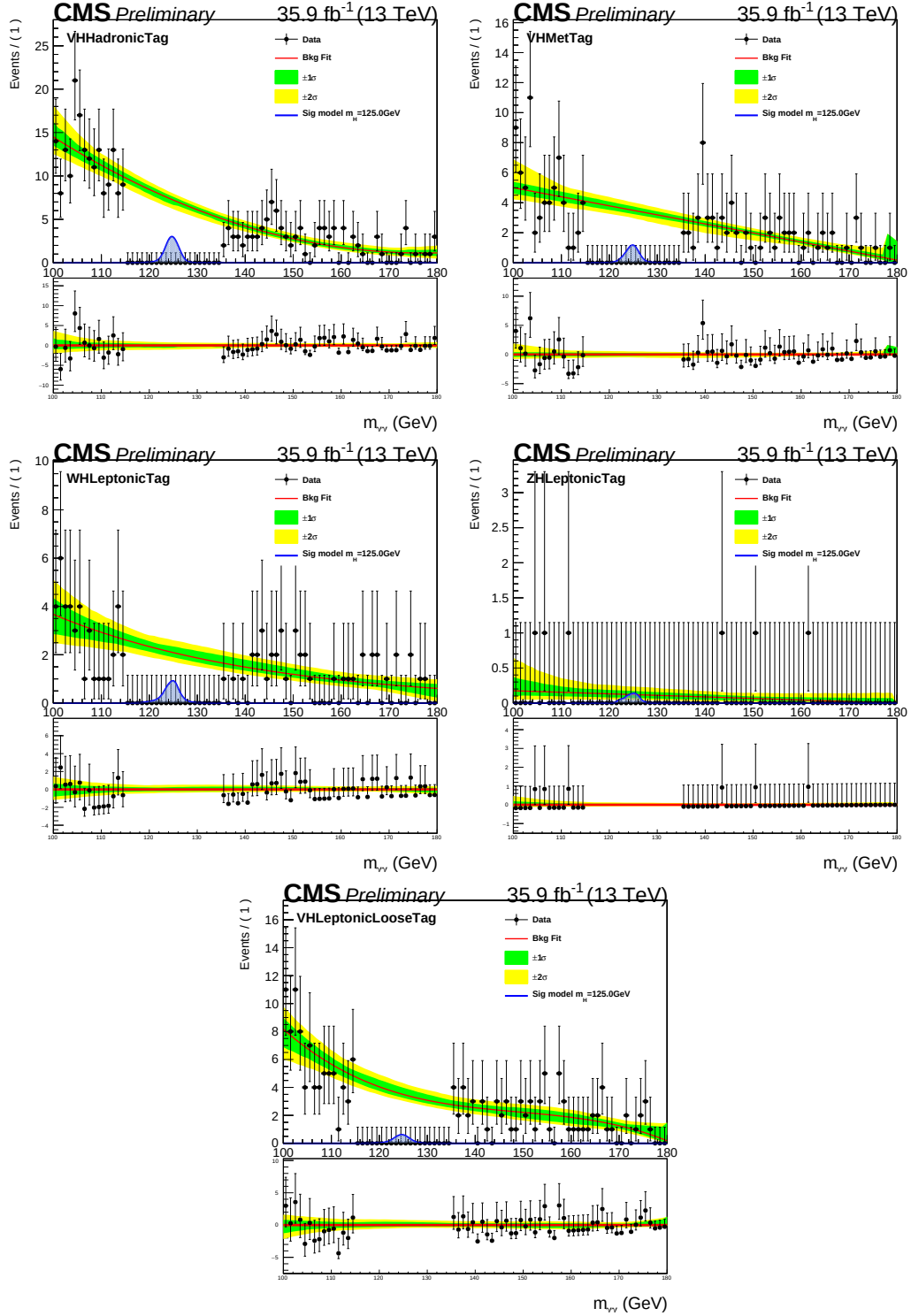


Figure 5.29: The best fit background parametrization plotted alongside the data in all VH categories. The green ($\pm 1\sigma$) and yellow ($\pm 2\sigma$) bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

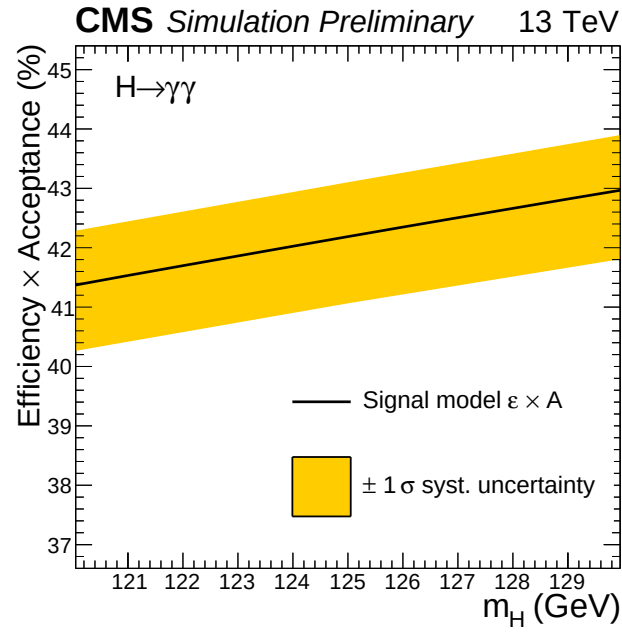


Figure 5.30: The efficiency $\times$ acceptance of the signal model as a function of m_H for all categories combined. The yellow bands indicate the effect of the systematic uncertainties.

5.13.4 Systematic uncertainties

This section lists the main sources of systematic uncertainty. They affect mainly the signal model, since the background model is derived using a data driven technique: the discrete profiling method is used to account for the uncertainty associated with the choice of background fit function. Three types of signal systematic uncertainties are considered, they affect signal shape, overall yield per process, or category migration.

Systematic uncertainties which modify the shape of the $m_{\gamma\gamma}$ distribution are generally built directly into the signal model as parametric nuisance parameters. In cases where the shape of the $m_{\gamma\gamma}$ distribution is largely unaffected, the systematic uncertainty affects signal yields following a log-normal distributions. For cases where the systematic has an effect on the input of the classification MVAs, the variation takes the form of a correlated log-normal uncertainty on the category yields, i.e as a category migration systematic.

The systematic uncertainties considered in this analysis are the following:

- Theoretical systematics:
 - *parton density functions (PDF) uncertainties*: the effect of the uncertainty from the choice of PDF is assessed by estimating the relative yield variation in each process and category, after re-weighting the events of the simulated signal sample. The re-weighting is done according to the PDF4LHC15 combined PDF set and NNPDF30 [180] using the MC2hessian procedure [181]. The effect on overall yield and the effect on categorization of events are handled separately. The category migrations are found to be less than 1%. The overall normalization variation is taken from [182].
 - α_s *uncertainty*: the uncertainty on the value of the strong force coupling constant α_s is evaluated following the PDF4LHC prescription [183]. The effect on the overall yield and the effect on the categorization of events are handled separately. The overall variation in the relative event yield due to the α_s uncertainty has been computed to be 2.6%.
 - *Underlying event and parton shower uncertainty*, which is obtained using samples where the choice and tuning of the generator has been modified. This systematic uncertainty is treated as an event migration systematic as it will chiefly affect the jets in the analysis. The possibility that an event could move from one VBF Tag to another or from either VBF Tag to an inclusive category is assigned a systematic uncertainty of 7% and 9% respectively.
 - *QCD scale uncertainty*, related to varying the renormalization and factorization scales. The effect on the overall yield and the effect on the categorization of events are handled separately. For the overall yields, separate scale factors are considered for each production mechanism. Category migrations arise from the

three scenarii with up/down variations of the renormalization and factorization scales. These effects are factorised from the yields uncertainty by dividing through the overall effect on signal normalisation. The overall effect on the normalisation is taken from [95].

- *Uncertainty on the $H \rightarrow \gamma\gamma$ branching fraction*: is estimated to be about 2% [96].
- *Gluon fusion contamination in VBF and $t\bar{t}H$ categories*: the theoretical predictions for gluon fusion are not reliable in a regime where the Higgs boson is produced in association with a large number of jets. The uncertainty on the yield of gluon fusion events in the VBF tagged classes has been estimated using the Stewart-Tackmann procedure following the recommendation of the LHC Higgs Cross Section Working Group [96]. The overall normalization has been found to vary by 29% while migrations between the two VBF categories are about 7% and 3% respectively. The systematic uncertainty on the gluon fusion contamination in the $t\bar{t}H$ tagged classes has been estimated taking into account several contributions:
 - * uncertainty due to the limited size of the simulated sample 10%
 - * uncertainty from the parton shower modelling. This uncertainty is estimated as the observed difference in the jet multiplicity between MADGRAPH5_aMC@NLO predictions and data in $t\bar{t} + \text{jets}$ events (which are dominated by gluon fusion production $gg \rightarrow t\bar{t}$), with fully leptonic $t\bar{t}$ decays. This uncertainty is about 15% in the leptonic category selections and about 35% in the hadronic category [184]
 - * uncertainty on the gluon splitting modelling. It is estimated by scaling the fraction of events from gluon fusion with real b-quark jets by the observed difference between data and simulation in the ratio $\sigma(t\bar{t}b\bar{b})/\sigma(t\bar{t}jj)$ at 13 TeV. A measurement of the scale factor is available in [185] and found to be 0.7. It gives an uncertainty on the ggH yield in $t\bar{t}H$ categories of 50%.
- *Integrated luminosity* is estimated from data, and amounts to a 2.5% uncertainty on the signal yield [186].
- *Trigger efficiency*: the trigger efficiency is measured on $Z \rightarrow e^+e^-$ events using the tag-and-probe technique as described in section 5.3.3. The size of the uncertainty varies between the photon categories, which are binned in R_9 , η and p_T . The size of the effect on the event yields is of at most 0.1%.
- *Photon preselection*: the systematic uncertainty is taken as the uncertainty on the ratio between the efficiency measured in data and in simulation, as described in section 5.5. It ranges from 0.1% to 0.7% according to the photon category and results in an event yield variation from 0.2% to 0.5% depending on the category.

- *Vertex finding efficiency:* the largest contribution to the uncertainty comes from the modelling of the underlying event, plus the uncertainty on the ratio of data and simulation obtained using $Z \rightarrow \mu\mu$ events, and from the uncertainty on the Higgs boson p_T . It is handled as an additional nuisance parameter built into the signal model which allows the fraction of events in the right vertex/wrong vertex scenario to change. The size of the uncertainty of the vertex selection efficiency is 2% and it is described in section 6.3.
- *Energy scale and resolution:* scale and resolutions are studied with electrons from $Z \rightarrow e^+e^-$ and then applied to photons. The main source of systematic uncertainty comes from the different interactions of electrons and photons with the material upstream the ECAL. Uncertainties are assessed by changing the R_9 distribution, the regression training (using electrons instead of photons) and the electron selection used to derive the corrections. The uncertainty on the additional energy smearing is assigned propagating the uncertainties on the various $|\eta|$ and R_9 bins to the Higgs boson signal phase space. In both cases dedicated nuisance parameters are included as additional systematic terms in the signal model and amount to a 0.15% to 0.5% effect on the photon energy depending on the photon category. The effect of the measurement of the inclusive signal strength is found to be of about 2.5%.
- *Non-uniformity of the light collection:* The uncertainty on the response of the ECAL crystals. The uncertainty has been slightly amplified with respect to Run 1 to account for the effect of larger transparency loss of the ECAL crystals. The size of the effect on the photon energy scale for 2016 data is estimated to be 0.07%.
- *Non-linearity:* The uncertainty associated with the non-linearity of the photon energy. This effect is estimated using Z boson decays to electron-positron pairs. The effect is found to be 0.1% on the photon energy in all categories, except in the untagged category with highest signal-to-background ratio, for which it is 0.2%.
- *Electromagnetic shower modelling:* A further small uncertainty is added to account for imperfect electromagnetic shower simulation in GEANT 4 [20]. A simulation made with an improved shower description, changes the energy scale for both electrons and photons. Although mostly consistent with zero, the variation is interpreted as a limitation on our knowledge of the correct simulation of the showers, leading to a further uncertainty of 0.05% on the photon energy.
- *Modeling of the material budget:* The uncertainty on material budget between the interaction point and the vertex, which affects the behaviour of electrons and photons showers, is estimated with special simulation samples where the material budget is uniformly varied by $\pm 5\%$. The effect on the energy scale is at most 0.24%.

- *Shower Shape Corrections*: The uncertainty deriving from the imperfect shower shape modelling in simulation. It is estimated using simulation with and without the corrections to derive a double ratio of response variation in different photon categories. This effect is at most 0.01-0.15% on the energy scale, depending on the photon category.
- *Photon identification BDT score*: in order to cover the observed discrepancies between data and simulation, the uncertainty on the signal yields in the different categories of the analysis is estimated conservatively by propagating the uncertainty as explained in section 5.9.2.
- *Per photon energy resolution estimate*: it is parametrized as a rescaling of the resolution estimate by $\pm 5\%$ about its nominal value, in order to cover all data/simulation differences in the distribution of the estimator output.
- *Jet energy scale and smearing corrections*: this uncertainty is implemented as migration within VBF categories, within $t\bar{t}H$ categories and from tagged to untagged categories. Jet energy scale corrections (JEC) account for a 8 to 18% migration within VBF categories and 11% from VBF to untagged categories. Migration due to energy scale in $t\bar{t}H$ categories is about 5%. The jet energy resolution has an impact on the event migration of lower than 3%.
- *Missing transverse energy*: this uncertainty is computed by shifting the reconstructed p_T of the particle candidates entering the computation of p_T^{miss} , within the momentum scale and resolution uncertainties appropriate to each type of reconstructed object, as described in [187]. Its effect is the event migration between the leptonic categories targeting the associated production of the Higgs boson with a vector boson. It was found to be 1.3% for $W \rightarrow l\nu$ category and 1.5% for $Z \rightarrow \nu\nu$ one.
- *Pileup jet identification*: this uncertainty is estimated by comparing in data and simulation the identification score of jets in events with a Z boson and one balanced jet. The full discrepancy between data and simulation is propagated to estimate the event migration. The migration is of the order of 1% or below.
- *b-tagging efficiency*: it is evaluated by varying the ratio between the measured b-tagging efficiency in data and simulation within their uncertainty [188]. For the $t\bar{t}H$ hadronic category, the uncertainty is evaluated by modifying the shape of the b-tagging discriminant in the simulation. The uncertainties accounted for in both lepton-tagged and hadronic-tagged categories include the statistical component on the data and simulation estimate of the fraction of heavy- and light-flavour jets, and corresponding mutual contaminations. The resulting uncertainty on the signal yield is about 2% in the lepton-tagged category and 5% in the hadronic-tagged category.

- *Lepton identification*: for both electrons and muons, the uncertainty is computed by varying the ratio of the efficiency measured in data and simulation by its uncertainty. The resulting differences in the selection efficiency for the $t\bar{t}H$ category tagged by leptons, is less than 1%.
- *Lepton isolation*: a systematic uncertainty of 0.5% associated to the special isolation criterion used to select events produced in association with a top quark pair is computed using the tag and probe technique on Z events. An additional 1% accounts for the different topology of Z and $t\bar{t}H$ events. The resulting changes in the yields of the $t\bar{t}H$ categories is at most 1%.
- *Background modelling*: the choice of background parametrization is handled using the discrete profiling method. This automatically leads to an uncertainty on the choice of background function as described in section 5.13.2.

Table 5.6 quantifies the size of the effect each systematic uncertainty has on the analysis. Uncertainties on the overall and per-process signal strength modifiers are shown. The contribution from uncertainties of similar origin are combined.

Expected relative uncertainty for SM Higgs boson ($m_H = 125$ GeV)					
Systematic	μ	μ_{ggH}	μ_{VBF}	μ_{VH}	$\mu_{t\bar{t}H}$
Photon identification	3.40 %	4.59 %	5.39 %	2.05 %	2.77 %
Per photon energy resolution estimate	2.68 %	4.20 %	4.85 %	1.47 %	0.80 %
Trigger efficiency	0.25 %	0.04 %	0.11 %	0.12 %	0.10 %
Diphoton preselection	0.53 %	0.23 %	0.24 %	0.40 %	0.28 %
Electron veto	0.46 %	0.12 %	0.21 %	0.23 %	0.22 %
Vertex finding efficiency	0.56 %	1.22 %	0.28 %	0.52 %	0.42 %
Photon energy scale and smearing	2.55 %	3.02 %	6.62 %	3.21 %	3.19 %
Nonlinearity of detector response	0.40 %	0.27 %	0.26 %	0.33 %	0.15 %
Nonuniformity of light collection	0.31 %	0.20 %	0.03 %	0.16 %	0.26 %
Shower shape corrections	0.44 %	0.36 %	0.14 %	0.55 %	0.27 %
Modelling of material budget	0.25 %	0.08 %	0.12 %	0.13 %	0.29 %
Modelling of detector response in GEANT4	1.17 %	0.32 %	0.11 %	0.17 %	0.08 %
Jet energy scale and resolution	1.70 %	2.53 %	25.82 %	2.09 %	2.17 %
ggF contamination in VBF categories	0.59 %	1.12 %	13.58 %	1.27 %	0.05 %
UE and PS	0.53 %	1.87 %	8.88 %	0.51 %	0.32 %
Lepton reconstruction and btag efficiencies	0.18 %	0.07 %	0.06 %	1.08 %	2.56 %
MET	0.08 %	0.08 %	0.11 %	0.66 %	0.11 %
ggF contamination in ttH categories	0.40 %	0.09 %	0.10 %	0.23 %	5.83 %
Integrated luminosity	2.13 %	2.83 %	5.77 %	3.24 %	2.67 %
Branching ratio	1.65 %	1.96 %	5.12 %	2.69 %	1.97 %
QCD scale yield	2.63 %	4.42 %	0.29 %	1.41 %	10.38 %
PDF and alphaS yield	2.13 %	3.29 %	5.13 %	2.19 %	4.08 %
QCD scale migrations	1.75 %	1.53 %	8.21 %	6.76 %	1.45 %
PDF migrations	0.46 %	0.22 %	0.73 %	0.40 %	0.79 %
AlphaS migrations	0.62 %	0.05 %	1.03 %	0.33 %	0.17 %
Total	7.35 %	10.27 %	34.38 %	9.80 %	14.24 %

Table 5.6: The size of systematic uncertainties contributed by various groups, for the overall and per-process signal strength modifier measurements. Values are calculated by omitting the appropriate uncertainties in the measurement and comparing to the overall uncertainty, then combining in quadrature. The background modelling enters the statistic uncertainty instead of systematic one, as explained in section 5.13.2.

Chapter 6

Diphoton vertex identification in the $H \rightarrow \gamma\gamma$ analysis

Introduction

In the decay of the Higgs boson into two photons, the final state unconverted photons are not detected in the tracker, so in the presence of pileup, the determination of the primary vertex associated with the signal is not trivial. The ECAL detector alone cannot be used for pointing back to the primary vertex since it is not segmented longitudinally as discussed in chapter 3. The information from the recoiling tracks can be used to determine the vertex position, and, when at least one of the photons is converted in the tracker, the conversion tracks can be used in addition. The principle of the vertex identification used for the $H \rightarrow \gamma\gamma$ analysis is described in section 6.1.1.

For the LHC Run II, the tracking algorithm has changed [189]. Moreover, the energy and pileup conditions are different. From a more technical point of view, the tools used for the Run I have evolved and the parameters of the method used in the past cannot be used in the Run II. Thus the vertex identification algorithm for the $H \rightarrow \gamma\gamma$ analysis had to be revised. Globally the same strategy is used as for the Run I algorithm. In section 6.1.2, the vertex identification for the Run II is described, with an emphasis on the changes with regard to the Run I analysis. Its performance is shown in section 6.1.3.

The vertex choice has a direct impact on the diphoton mass resolution, as a wrong choice would worsen the resolution by about 1 GeV on average. The resolution on the photon opening angle makes a negligible contribution to the diphoton mass resolution when the chosen vertex lies within 1 cm of the true vertex. It is illustrated in Fig. 6.1 for all photons and in Fig. 6.2 for the high- R_9 photons in the barrel. The effective mass resolution (on the left) is represented as a function of the z difference between the true and the chosen vertex (in red), and the mass extra smearing from a wrong vertex choice as a function of the z difference between the toy vertex and the true vertex (in blue). The first one (in red)

is produced with selected events from ggH simulation, and selected vertices (most events are located in the first bins). The second (in blue) uses also the same simulation, only for events where the true vertex is chosen, and the over-smearing is computed using toy values of the z vertex with a spread from 0 to 15 cm along z . The mass resolution worsening as function of the z difference between the toy vertex and the true vertex is represented on the right. One can see that the impact of the vertex position is negligible below 1 cm for both all photons (Fig. 6.1) and photons in the barrel with high- R_9 (Fig. 6.2).

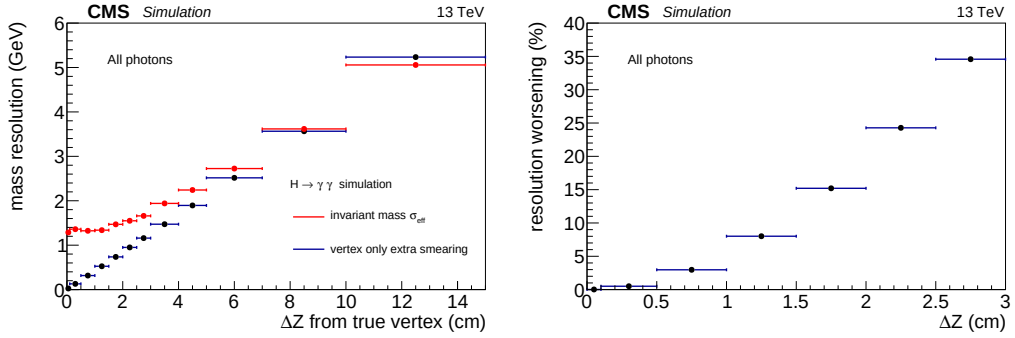


Figure 6.1: On the left: effective mass resolution as a function of the z difference between the true and the chosen vertex (in red), and the mass extra smearing from a wrong vertex choice as a function of the z difference between the toy vertex and the true vertex (in blue). On the right: mass resolution worsening as a function of the z difference between the toyvertex and the true vertex. Both are done using $H \rightarrow \gamma\gamma$ simulation with ggH production mode and for all photons.

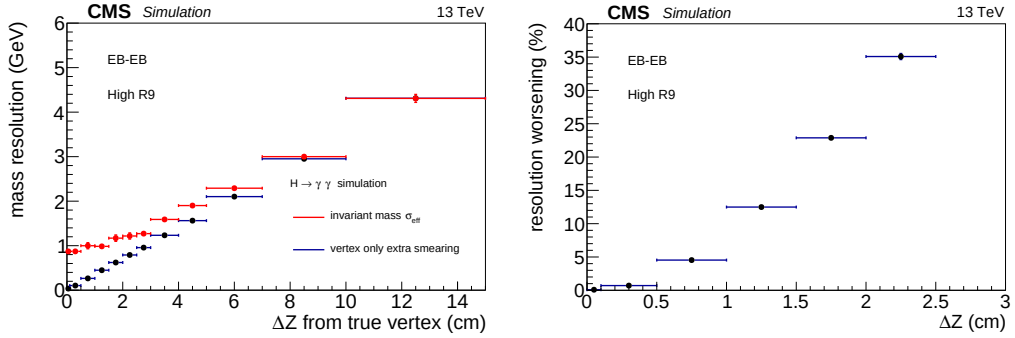


Figure 6.2: On the left: effective mass resolution as a function of the z difference between the true and the chosen vertex (in red), and the mass extra smearing from a wrong vertex choice as a function of the z difference between the toy vertex and the true vertex (in blue). On the right: mass resolution worsening as a function of the z difference between the toy vertex and the true vertex. Both are done using $H \rightarrow \gamma\gamma$ simulation with ggH production mode and for the photons located in the barrel and with high values of R_9 .

A per-event probability to choose the right vertex is determined allowing to fully benefit of the excellent ECAL resolution in the event categorisation procedure. This probability determination is summarized in section 6.2. As the per-event probability to choose the right vertex obviously depends on the vertex identification algorithm, it also has to be re-examined. The performance of the updated algorithm is shown in section 6.2.2.

Both algorithms were studied and optimized on $H \rightarrow \gamma\gamma$ simulation. As the simulation does not describe data perfectly, a validation on data is necessary. The validation of the vertex identification algorithm is described in sections 6.1.4 and 6.1.5 and the validation of the per-event probability in sections 6.2.3 and 6.2.4.

6.1 Vertex identification algorithm

6.1.1 Principle

The vertex identification algorithm exploits the correlation between the recoiling tracks attached to the $H \rightarrow \gamma\gamma$ vertex and the diphoton system. In events where at least one of the two photons converts into an e^+e^- pair, the conversion tracks can be reconstructed and linked to the photon supercluster. This information is used to build an optimal discriminating variable.

To identify the primary vertex, a multivariate analysis (MVA) is used. It allows to exploit several discriminating variables in order to get one single variable with the best separation between signal and background, taking into account correlations between variables. In this study, the signal is defined as the closest (in z coordinate) reconstructed vertex to the true generated one, and background as any of the other vertices.

Multivariate analysis method : BDT

The study was implemented using Toolkit for Multivariate Data Analysis (TMVA) [173] with the Boosted Decision Trees (BDT) method.

A decision tree is a sequence of binary splits of the data as sketched in Fig 6.3.

Repeated "yes/no" decisions are taken on one single variable at a time until a stop criterion is fulfilled. The phase space is split this way into many regions that are eventually classified as signal or background, depending on the majority of training events that end up in the final leaf node. Each output node represents a specific value of the target variable. A decision tree needs to be trained on a dataset which already provides the outcome (signal, background). The data are split into two parts: a training and a testing samples, the testing sample being used to validate the training, and to check that there is no overtraining (when the performance is significantly better with the training sample than with the testing one). The first step is to train the BDT, with a set of known training events. The second step is evaluate the results using a separate set of known testing events. The output of the BDT is a discriminating variable, called in the following "MVA", distributed between -1 and +1,

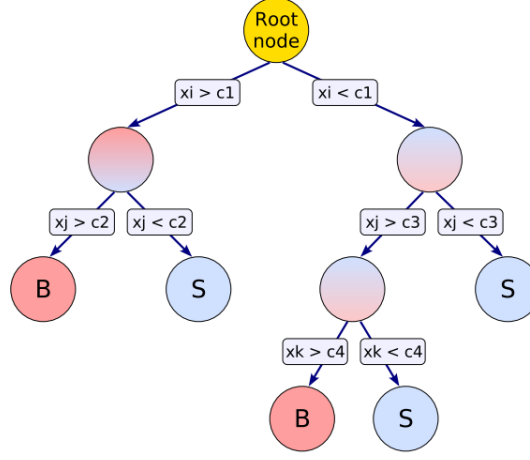


Figure 6.3: Schematic view of a decision tree.

peaking around $+1$ value for signal and -1 value for background. The vertex chosen in the analysis for each event is the one with the most "signal-like" value of this discriminating variable (i.e. with the maximal MVA value).

Simulated samples

The simulated samples used in this study are $H \rightarrow \gamma\gamma$ samples for production modes including: gluon fusion, vector boson fusion, associated production with vector boson (W/Z) and associated production with a pair of top quarks. They are generated at a center-of-mass energy of 13 TeV and with a Higgs boson mass $m_H = 125$ GeV. Each production mode is weighted according to its cross-section. The software, used for generating the events is MadGraph5_aMC@NLO [161] and the parton level samples are interfaced to Pythia 8 [162] for parton showering and hadronization. Half of the events are used for the BDT method training and the other half for the BDT output evaluation.

Discriminating variables

Two unconverted photons. In the Run I analysis, in presence of unconverted photons, the three most discriminating variables for the vertex identification were found to be [120]:

1. $sumPt2 = \sum_i |\vec{p}_T^i|^2$: the sum over the tracks of the square of the transverse momentum of i -th track associated with a given vertex. As the true primary vertex is the hard vertex, it tends to take higher values for the true vertex than for the wrong ones.

2. $ptBal = -\sum_i (\vec{p}_T^i) \cdot \frac{\vec{p}_T^{\gamma\gamma}}{|\vec{p}_T^{\gamma\gamma}|}$: the negative sum of projections of the $\vec{p}_T$ of the tracks on the diphoton $\vec{p}_T^{\gamma\gamma}$. It tends to be positive for the true vertex, as tracks recoil against the diphoton pair. It tends to be centred at 0 for wrong vertices.
3. $ptAsym = (|\sum_i \vec{p}_T^i| - p_T^{\gamma\gamma}) / (|\sum_i \vec{p}_T^i| + p_T^{\gamma\gamma})$: the asymmetry between the total momentum of the tracks attached to a given vertex and the modulus of the diphoton p_T . This quantity tends to have higher values for the true vertex and to peak at -1 for wrong vertices, as tracks recoil against the diphoton pair.

For all variables, $\vec{p}_T^i$ is the momentum of the i -th track associated with a given vertex and $\vec{p}_T^{\gamma\gamma}$ is the momentum of the diphoton pair. The sum runs over all well reconstructed Particle Flow candidates (described in section 3.7) with an electric charge and associated with the given vertex.

The diphoton p_T spectrum (measured in $H \rightarrow \gamma\gamma$ simulation) is harder at 13 TeV compared to the one at 8 TeV, which is expected and it is shown in Fig 6.4.

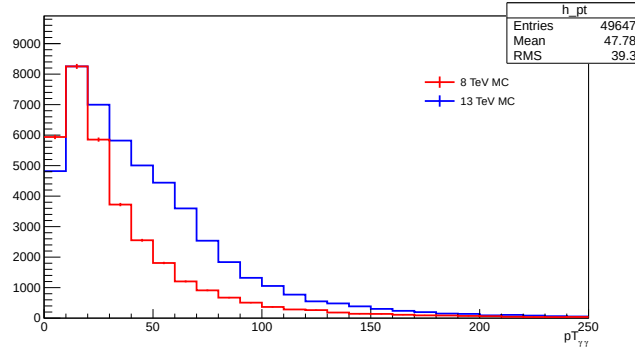


Figure 6.4: Comparison of the diphoton p_T spectrum using $H \rightarrow \gamma\gamma$ simulation sample at 13 TeV (blue) and 8 TeV (red).

The distributions of the discriminating variables for 8 TeV and 13 TeV $H \rightarrow \gamma\gamma$ Monte-Carlo simulation samples are shown for signal and background in Fig. 6.5, where 13 TeV MC is compared to 8 TeV MC, and in Fig. 6.6, where signal is compared to background.

At least one converted photon. Electron-positron pairs from reconstructed converted photons can be exploited to determine the longitudinal coordinate of the primary interaction vertex where the Higgs boson is produced.

Two methods have been developed which give different performances depending on where the photon conversion occurs. Both methods exploit the knowledge of the converted photon direction extracted from the conversion reconstruction. Once the direction is known, it is extrapolated back to the beam line to obtain an estimate of the z position of the primary interaction vertex, z_{PV}^{conv} .

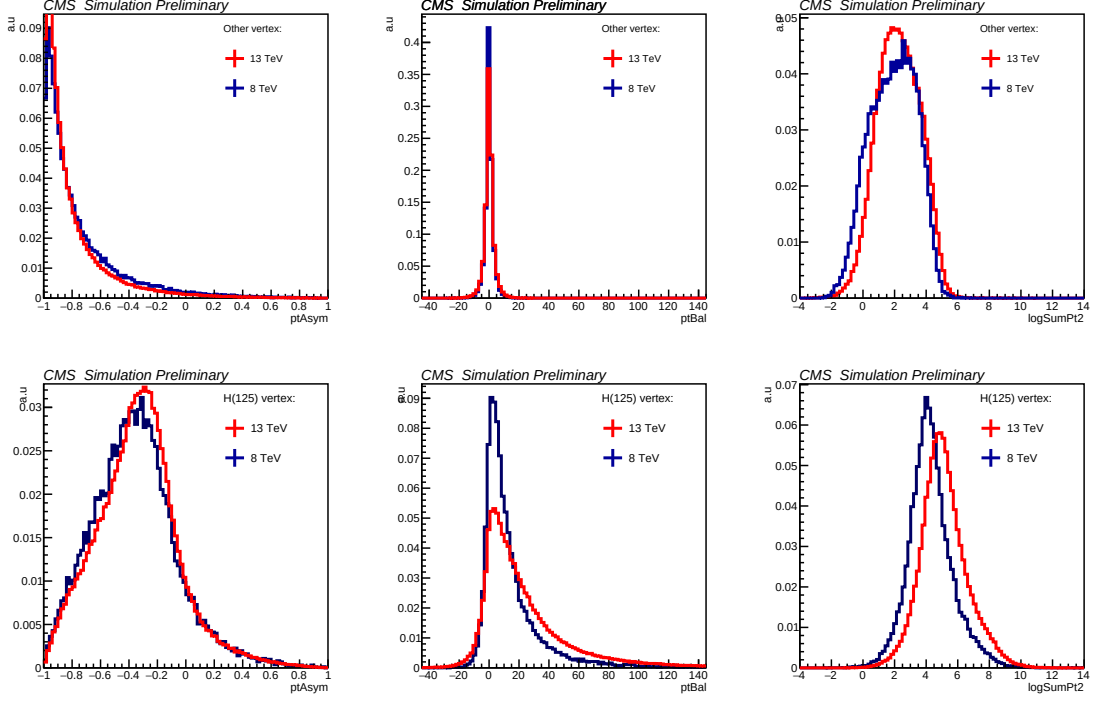


Figure 6.5: Comparison of the tracks input variables for 13 TeV and 8 TeV MC. At the top are shown distributions for the background (random wrong vertex) and at the bottom distributions for the signal (reconstructed vertex closest to the true vertex).

In the first method, called "CO" for "conversion only" in the following, the photon direction is calculated using the angle, α_{conv} , between the conversion momentum and the z -axis. The conversion momentum is evaluated from the track pair refitted with the vertex constraint. The longitudinal coordinate of the primary interaction vertex is then calculated as:

$$z_{\text{PV CO}}^{\text{conv}} = z_{\text{conv}} - r_{\text{conv}} \times \cot(\alpha_{\text{conv}}), \quad (6.1)$$

where z_{conv} and r_{conv} are the z coordinate and the distance to z axis of the fitted conversion vertex.

In the second method, called "SC" for "supercluster" in the following, the direction of the converted photon is instead determined by combining the information on the conversion vertex position and the position of the ECAL supercluster. The longitudinal coordinate of the primary interaction vertex is calculated in this method as:

$$z_{\text{PV SC}}^{\text{conv}} = z_{\text{conv}} - r_{\text{conv}} \frac{z_{\text{SC}} - z_{\text{conv}}}{r_{\text{SC}} - r_{\text{conv}}}, \quad (6.2)$$

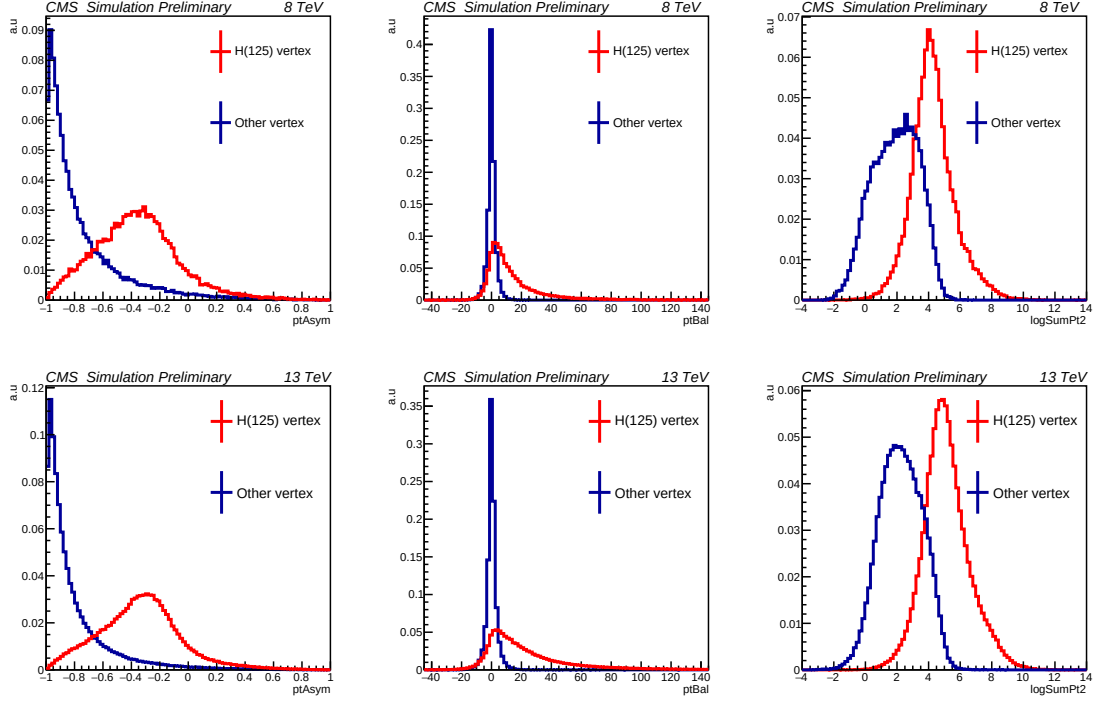


Figure 6.6: Tracks input variables distributions from MC at 8 TeV (top) and 13 TeV (bottom) for signal (in red) and background (in blue).

where z_{conv} and r_{conv} are the z and the distance to the z axis of the fitted conversion vertex as in the previous method, and z_{SC} and r_{SC} are the z position of the ECAL supercluster and its distance to the beam axis. Fig. 6.7 illustrates the two methods.

The two methods are complementary and the interplay of conversion vertex position resolution, conversion momentum direction resolution and, for the second method, resolution on the position in the ECAL will define the best method for every converted photon.

The variables used in the analysis in presence of converted photons are the number of converted photons ($N_{conv} = 0, 1, 2$) and $Pull_{conv}$:

$$Pull_{conv} = \frac{|z_{PV} - z_{PV}^{conv}|}{\sigma_{conv}} \quad (6.3)$$

where z_{PV} is the z position of the tested primary vertex, z_{PV}^{conv} is the z position of the primary vertex estimated from one of the 2 algorithms described above, and σ_{conv} is the associated z effective resolution, for the tracker part where the conversion vertex is reconstructed. As $z_{PV} - z_{PV}^{conv}$ distributions are non Gaussian, the z effective resolution is computed as the half width around maximum where 68% of the distribution is located.

The most precise of the 2 methods is chosen on an event-by-event basis, based on the

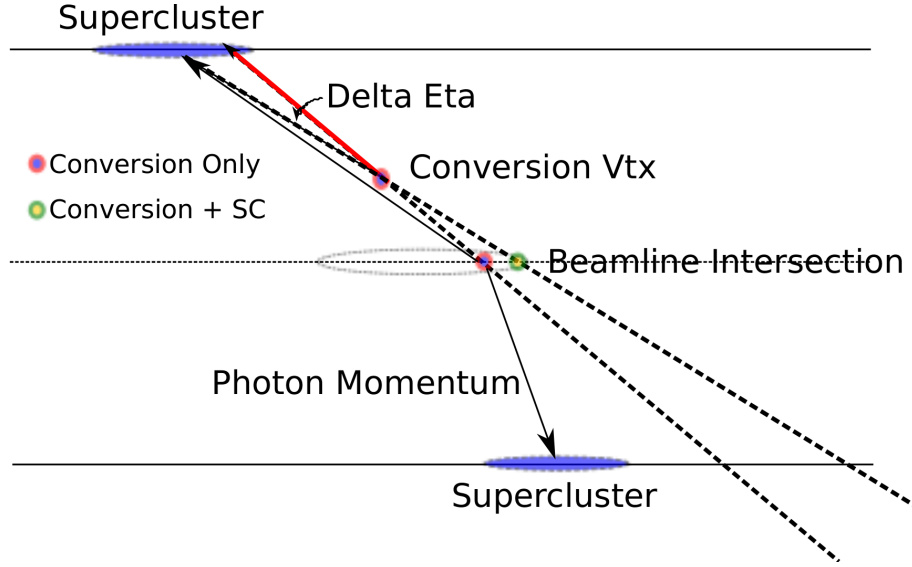


Figure 6.7: Illustration of the two methods considered to determine the z position of the diphoton vertex in case at least one of the photons is a conversion. One method extrapolates the conversion track ("conversion only (CO)"), whereas the other takes into account the photon position in the ECAL ("conversion + super-cluster (SC)").

z_{PV}^{conv} resolution in each sub-detector, where the conversion vertex is reconstructed. A $r-z$ slice of the CMS tracker structure is depicted in the Fig. 6.8, showing different tracker sub-detectors: pixel inner barrel (PIB), pixel inner forward (PIF), tracker inner barrel (TIB), tracker outer barrel (TOB), tracker endcap (TEC), and tracker inner disk (TID).

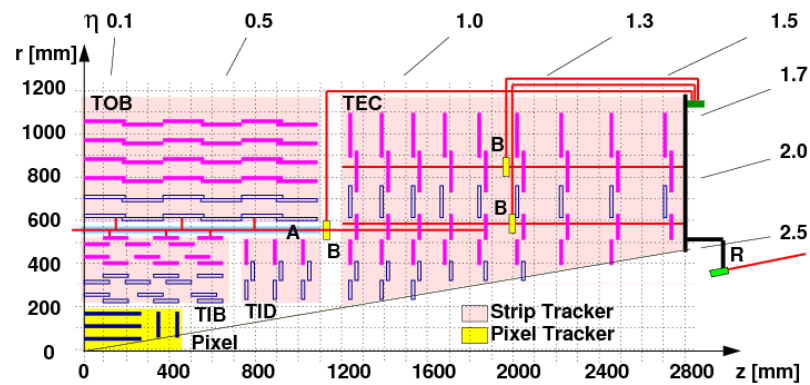


Figure 6.8: Quarter of the $r-z$ slice of the CMS tracker, where the center of the tracker is at the leftbottom corner, the horizontal axis, to which the detector has a cylindrical symmetry, points along z , and the vertical axis points along the radius r . The LHC beams are along the z axis. Various pseudorapidity values are shown at the ends of the black lines.

The comparison of the $z_{PV} - z_{PV}^{\text{conv}}$ distribution with 8 TeV and 13 TeV simulation for CO algorithm is shown in Fig 6.9 and the one for SC is in Fig. 6.10. The effective resolutions of the $z_{PV} - z_{PV}^{\text{conv}}$ are listed in Table 6.1 for two and single leg conversions. Single leg conversion occurs when the only one from two electron tracks are reconstructed. The CO algorithm performs better in subdetectors which are close to the interaction point: PIB, PIF and TID. For the outer regions of the tracker, TIB, TOB and TEC, the addition of information from the ECAL helps to improve the effective resolution, so the SC algorithm performs better.

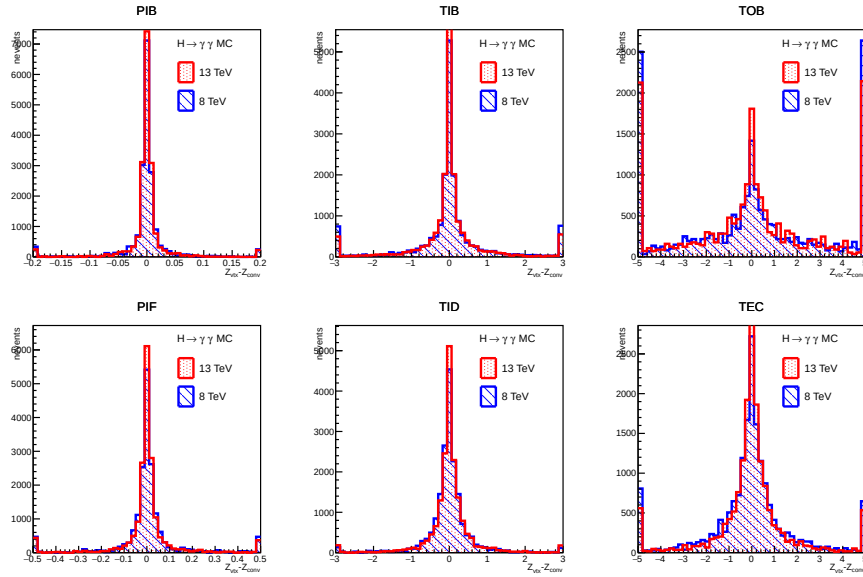


Figure 6.9: Distributions of $z_{PV} - z_{PV}^{\text{conv}}$ for 13 TeV and 8 TeV simulation with CO algorithm per tracker region.

The distributions of the $Pull_{\text{conv}}$ variable for 8 TeV and 13 TeV $H \rightarrow \gamma\gamma$ Monte-Carlo samples are shown for signal and background in Fig. 6.11, where 13 TeV MC is compared to 8 TeV MC, and in Fig. 6.12, where signal is compared to background.

In case of 2 converted photons, the weighted average is taken to calculate z_{PV}^{conv} and its resolution:

$$z_{PV}^{\text{conv}} = \frac{z_{PV}^{\text{conv}1}/\sigma_1^2 + z_{PV}^{\text{conv}2}/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \quad (6.4)$$

and

$$\sigma_{\text{conv}} = \sqrt{\frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}}, \quad (6.5)$$

where 1 and 2 refer to the 2 converted photons.

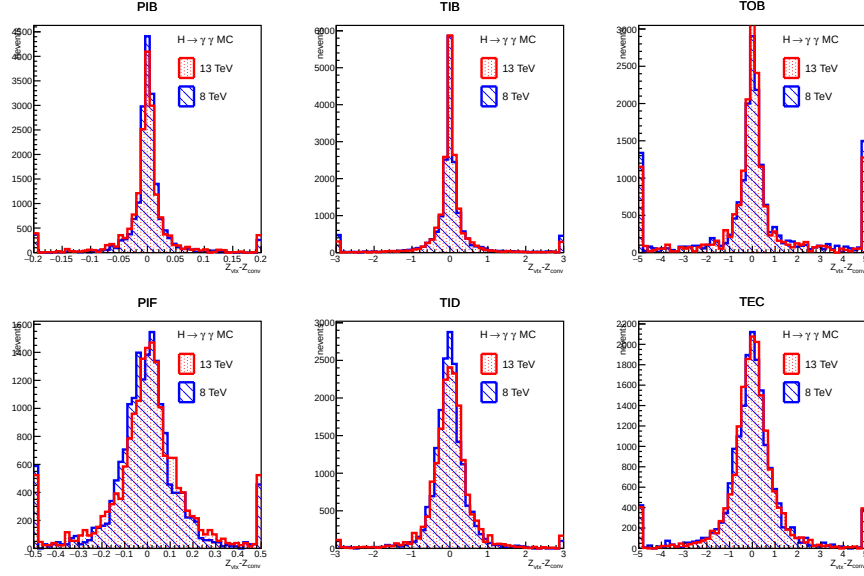


Figure 6.10: Distributions of $z_{PV} - z_{PV}^{\text{conv}}$ for 13 TeV and 8 TeV simulation with SC algorithm per tracker region.

Subdetector	$\sigma_{\text{conv CO}} \text{ (cm)}$				$\sigma_{\text{conv SC}} \text{ (cm)}$			
	Two legs		Single leg		Two legs		Single leg	
	8 TeV	13 TeV	8 TeV	13 TeV	8 TeV	13 TeV	8 TeV	13 TeV
PIB	0.01	0.01	0.01	0.02	0.02	0.03	0.05	0.09
TIB	0.50	0.72	1.2	1.3	0.30	0.41	0.60	0.76
TOB	4.4	3.2	2.1	2.2	1.7	1.1	0.48	0.62
PIF	0.05	0.06	0.07	0.15	0.15	0.18	0.28	0.58
TID	0.32	0.38	0.38	0.70	0.39	0.49	0.50	0.89
TEC	1.1	1.7	1.9	2.5	1.0	1.2	1.0	1.6

Table 6.1: z_{PV}^{conv} effective resolutions (13 TeV) obtained for the two different algorithms (CO and SC) and compared to the Run I effective resolutions (8 TeV), for two legs conversions and one leg conversions.

6.1.2 Investigation for possible improvements

The investigations described in this section were done with the very first simulation and reconstruction available in the Run II. Several possibilities were investigated to improve the separation between signal and background. Firstly, the parameters of the BDT were re-optimized as the Run I parameters did not exist anymore in the new version of TMVA and the default TMVA parameters might not be optimal. It has been decided to keep the same

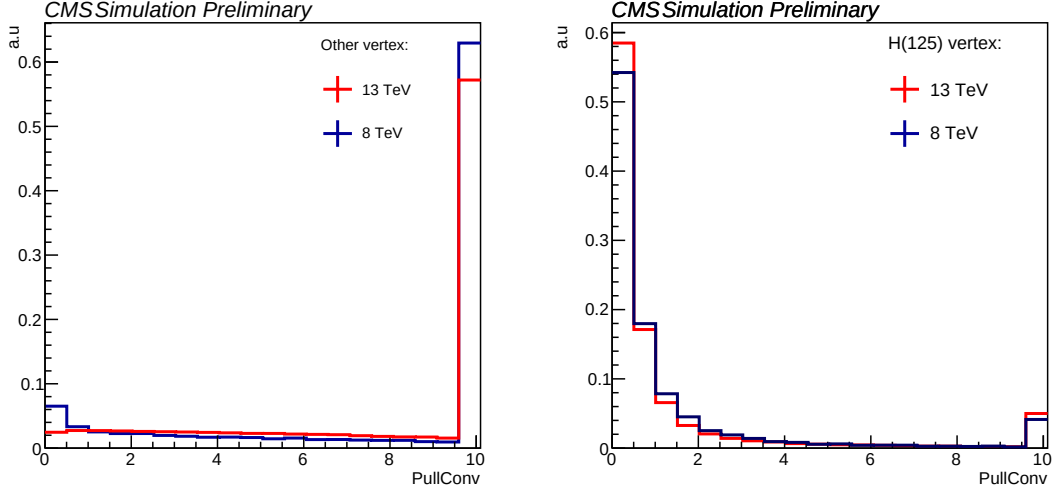


Figure 6.11: Comparison of the $Pull_{conv}$ variable for 13 TeV and 8 TeV MC. On the left it is shown for the background (random wrong vertex) and on the right for the signal (the closest vertex to the true one).

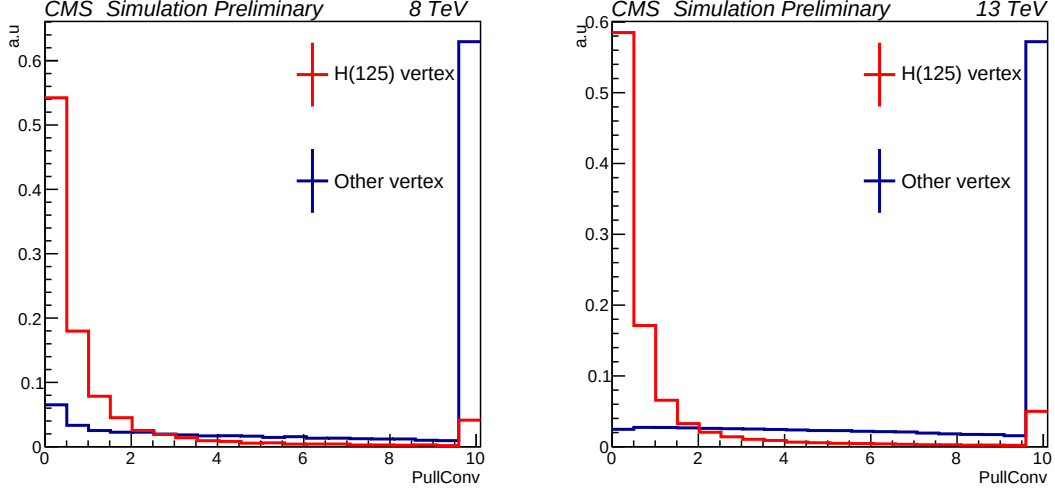


Figure 6.12: $Pull_{conv}$ variable distributions from MC at 8 TeV on the left, and 13 TeV on the right for signal (the closest vertex to the true one) in red and background (random wrong vertex) in blue.

input variables as during the Run I for the BDT parameters optimization. Secondly, further tracks and conversions variables were added in order to test if the discriminating power

could be improved significantly. Finally, as the primary vertex identification is significantly improved when there is at least one converted photon, because of the presence of conversion tracks, the single leg conversions were also added to the analysis.

At the time of the study shown in this section, the criteria to decide if a photon is converted or not were much tighter than the ones in the latest version of the analysis, where it was purely geometrical. Conversion is geometrically matched to photon if the angular distance between conversion pair direction and photon direction is less than 0.1. The angular distance is $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, where $\Delta\eta$ and $\Delta\phi$ are the distances in the η and ϕ directions between the photon direction and the conversion direction. In the first iteration an additional requirement has been applied. It checks if the photon has any conversions explicitly referenced from the Particle Flow algorithm. This criteria resulted into a drop in the number of conversions by almost a factor 2. All plots and numbers, presented in this section are done with all criteria including the last requirement, unless otherwise specified.

The figure of merit used in these studies is the probability to choose the correct primary vertex, computed from the signal and background distributions of the MVA output. The vertex chosen in the analysis for each event is the one with the most "signal-like" output MVA value (i.e with the maximal output MVA value). Let v be the output MVA value. This probability can be computed from the probability densities for the signal $S(v)$ and the background $B(v)$. An example of these distributions for signal and background is shown in Fig. 6.13. The probability $b(x)$ for a background vertex to take a value v lower than x is:

$$b(x) = \int_{-1}^x B(v)dv \quad (6.6)$$

The probability to choose the right vertex among a total number of vertices N is:

$$p_{\text{good}}^N = \int_{-1}^{+1} b(x)^{N-1} S(x) dx \quad (6.7)$$

This probability has to be maximal for the BDT to be the most efficient.

MVA method parameters optimization. The main parameters used in the Boosted Decision Trees method are listed below. Further explanation about BDT parameters can be found in TMVA user guide [173]. The boosting method used is "Gradient Boost".

- **NTrees** : the number of decision trees,
- **MaxDepth** : the maximum depth of a decision tree,
- **Shrinkage** : the learning rate for Gradient Boost algorithm,
- **MinNodeSize** : the minimum percentage of training events required in a leaf node,

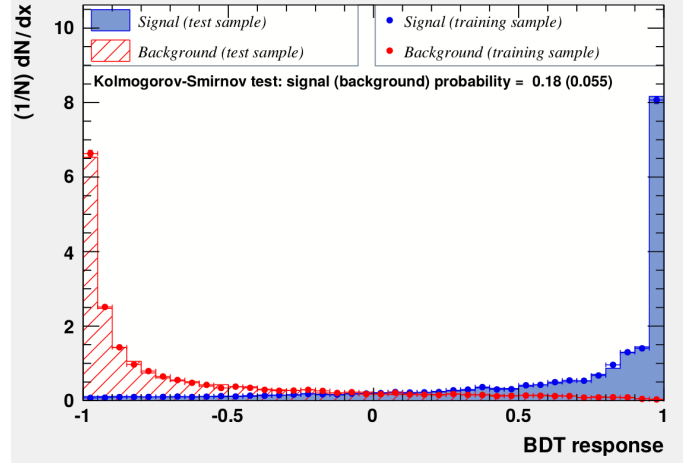


Figure 6.13: Example of BDT output distributions for signal (in blue) and background (in red).

- **nCuts** : number of grid points in variable range used in finding optimal cut in node splitting.

An automatic parameter optimization exists in the TMVA, but this procedure has several drawbacks: first, it does not discard overtrainings, secondly, the parameters' ranges used in the optimization can not be modified, and finally, the figure of merit is not the efficiency of interest. First, this automatic optimization was used, followed by a home-made optimization of the parameters, using the efficiency p_{good}^N as a figure of merit, and discarding overtrainings. The ranges of parameters scanned by the automatic parameter optimization and our optimization are listed in Table 6.2.

Parameter	Range Autoopt	Range Home-made opt
NTrees	10-1000	1000
MaxDepth	2-3	3
MinNodeSize	1-30	1-10
Shrinkage	0.05-0.5	0.01-0.1

Table 6.2: Parameters for BDT auto and home-made optimizations and their ranges.

For the training, the sample was split into two categories, with and without converted photons, to profit from conversion tracks when they exist. Therefore, two types of efficiencies were obtained with different parameters configurations. As about 79% (63% for geometrical matching) of events do not have converted photons, the BDT parameters were optimized with only unconverted photons.

The efficiency for the events with conversions was also computed and it was checked that the overall efficiency was the best. The efficiency to choose the correct vertex is defined as the fraction of events where the z coordinate of the chosen vertex is located within 1 cm from the true $H \rightarrow \gamma\gamma$ vertex. In Fig. 6.14, one can see the efficiencies p_{good}^N versus the number of additional vertices $N - 1$, for several different sets of parameters, for the events with only unconverted photons on the top and for the events with converted photons on the bottom.

The following set of parameters gives the highest global efficiency: **MaxDepth** = **3**, **MinNodes** = **2**, **NTrees** = **1000**, **Shrinkage** = **0.05**, **nCuts** = **20** and was thus chosen for the vertex identification algorithm for the Run II. In Fig. 6.14, these parameters correspond to the efficiency depicted in magenta.

Adding more track-related variables. Several track-related variables were also added to the training to see if a sizable improvement could be obtained:

- **PtAsymZ**: $\frac{|\sum_i p_z^i| - p_z^{\gamma\gamma}}{|\sum_i p_z^i| + p_z^{\gamma\gamma}}$,
- **MaxPt**: the largest p_T of the tracks associated with the vertex,
- **MaxPt3**: the scalar sum $|\sum_i p_T^i|$, where i runs over the 3 tracks with the largest p_T ,
- **SumPt**: $\sum_i |\vec{p}_T^i|^2$, the sum of the p_T^2 of all tracks associated to the vertex,
- **Ntracks**: the number of tracks associated to the vertex,
- **DPhi**: the $\Delta\phi$ angle between $\sum_i \vec{p}_T^i$ and $\vec{p}_T^{\gamma\gamma}$,
- **PtBalScalSumPt**: the ratio between PtBal and SumPt,
- **SumPtDiPhoPt**: the ratio between SumPt and the diphoton p_T .

All these variables are strongly correlated to the three variables used during the Run I. The distributions for the signal (in blue) and the background (in red) for each variable in the list are shown in Fig. 6.15.

The BDT was retrained including all this variables and the efficiency was compared to the efficiency of the nominal BDT. The resulting efficiency as function of number of vertices is shown in Fig. 6.16 in black, whereas the nominal BDT efficiency is shown in green. The improvement is marginal given the additional complexity, so it was decided to keep the same 3 variables as for the Run I :

- $sumPt2 = \sum_i |\vec{p}_T^i|^2$,
- $ptBal = -\sum_i (\vec{p}_T^i) \cdot \frac{\vec{p}_T^{\gamma\gamma}}{|\vec{p}_T^{\gamma\gamma}|}$,
- $ptAsym = (|\sum_i \vec{p}_T^i| - p_T^{\gamma\gamma}) / (|\sum_i \vec{p}_T^i| + p_T^{\gamma\gamma})$.

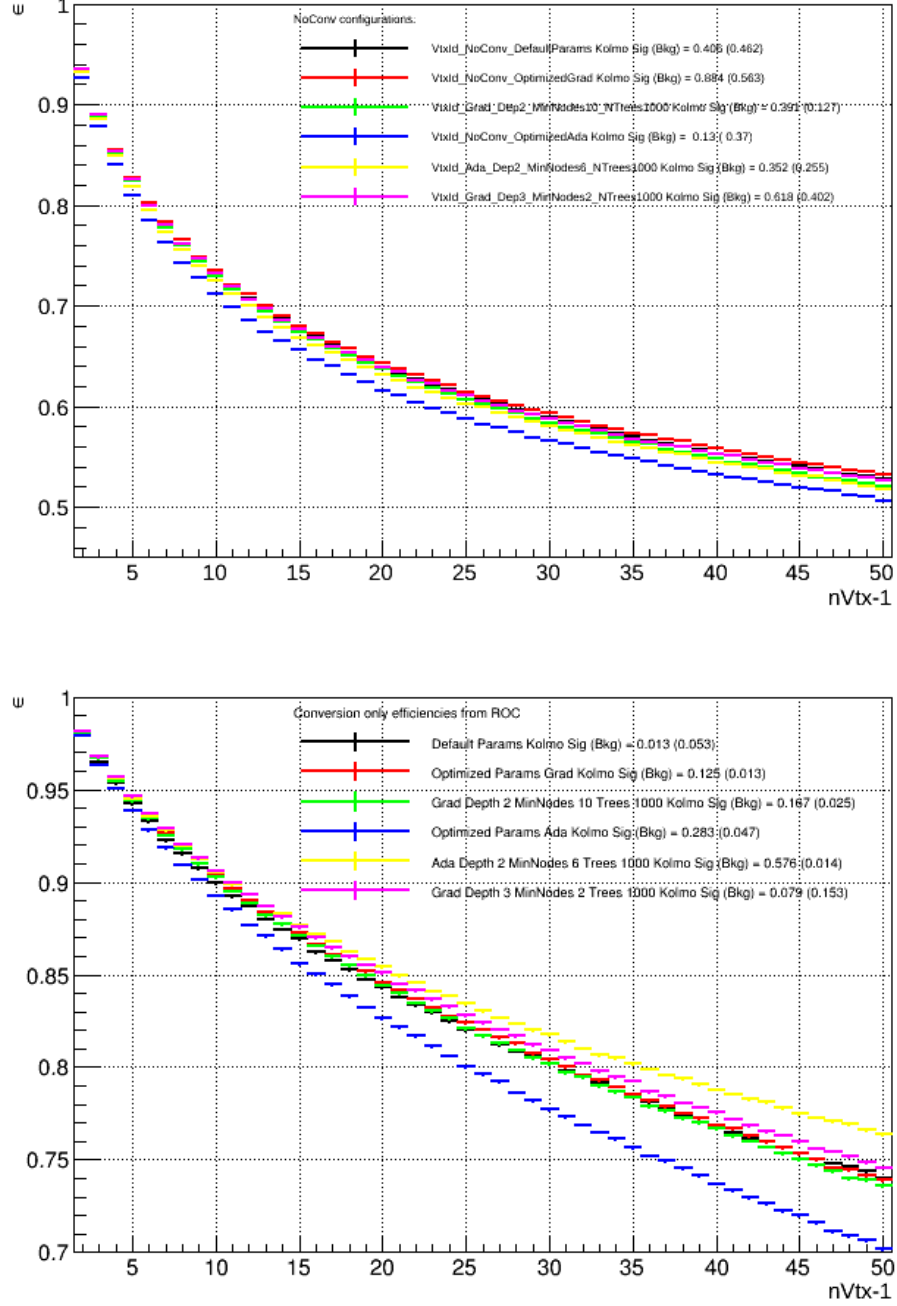


Figure 6.14: Expected efficiencies p_{good}^N computed from the BDT output as a function of the number of additional vertices $N - 1$ for different BDT parameters configurations, for the sample with only unconverted photons (top), and for the sample with converted photons (bottom). The parameters corresponding to the magenta curves were chosen.

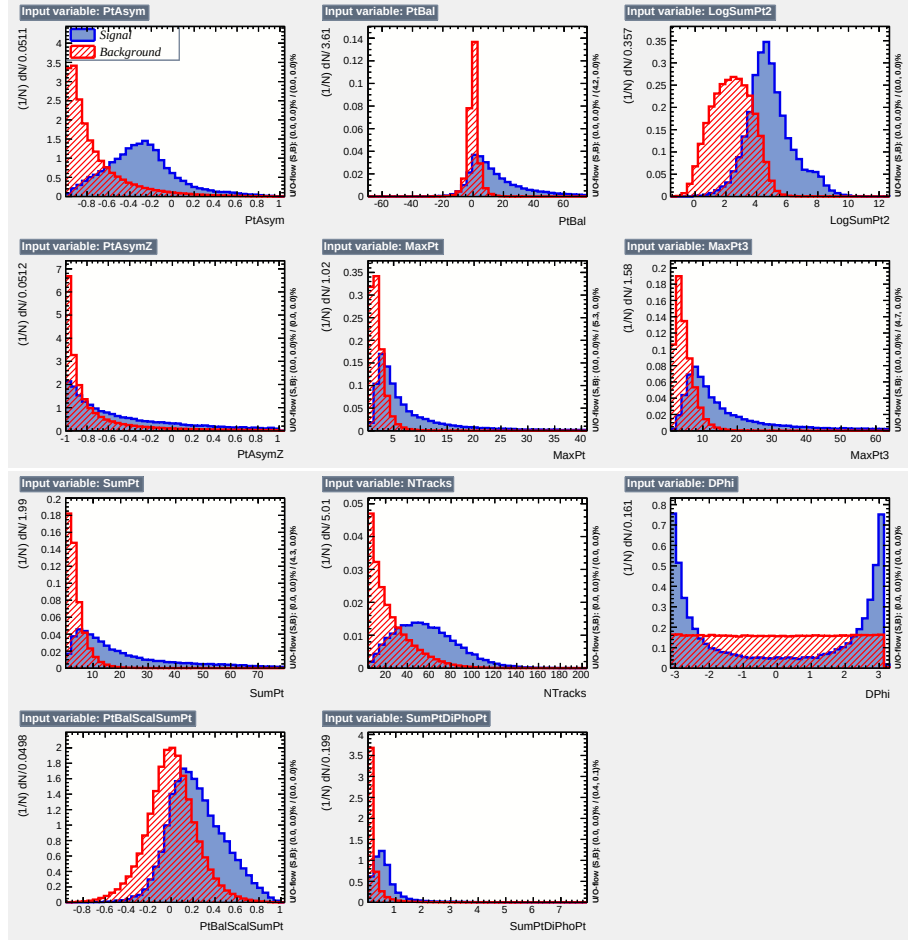


Figure 6.15: Signal and background distributions for tracker input variables included in this study.

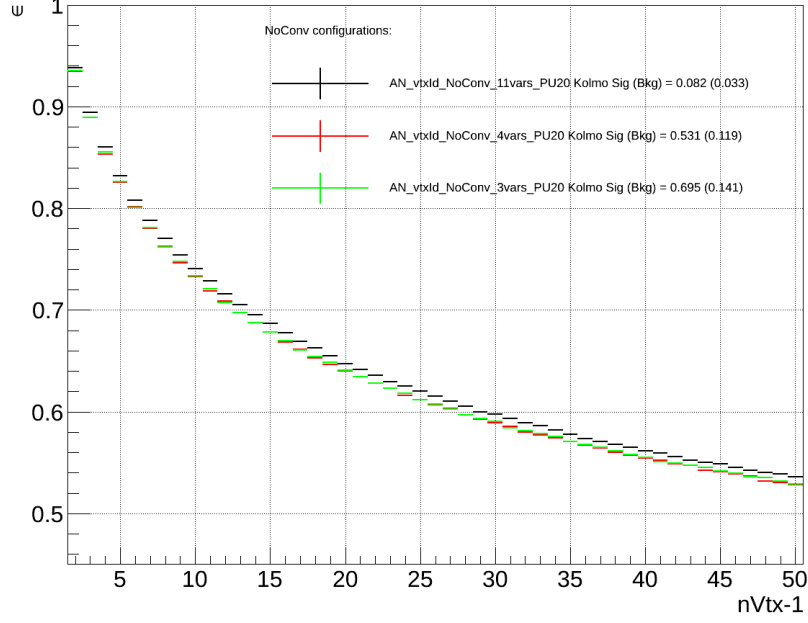


Figure 6.16: Efficiency to find the correct primary vertex as a function of the number of vertices for different sets of input variables. The efficiency obtained using 11 discriminating variables is shown in black and the nominal BDT efficiency is shown in green.

Adding more conversion variables. As described in section 6.1.1, when at least one of the photons is converted, the variable $Pull_{conv}$ is added to improve the discriminating power between the wrong and correct vertices. When at least one photon is converted, the average efficiency to find the correct vertex is $\sim 85\%$ (for an average pileup¹ 20), while without any conversion the efficiency is only $\sim 65\%$. Depending on where the conversion happens in the tracker, this variable is either $Pull_{conv CO}$ or $Pull_{conv SC}$, depending on which of the two algorithms has the best resolution in the given tracker part. In this section, we describe the performances that one can obtain by using all the available information. In the different configurations studied, $Pull_{conv}$ was replaced by:

1. Both $Pull_{conv CO}$ and $Pull_{conv SC}$,
2. Their average weighted with their effective resolutions,
3. $Pull_{conv CO}$, $Pull_{conv SC}$ and the tracker part where the conversion happened (a discrete number between 1 and 6, describing the 6 parts mentioned before).

¹Pileup is the average number of interactions per bunch crossing.

For these different configurations, the efficiencies estimated from the BDT output as described above are depicted in Fig. 6.17.

In this figure, the nominal BDT efficiency is shown in black. One can see that the efficiency for the third configuration is better by about 3% for a number of vertices of about 25. This improvement is substantial for events with conversions, but it is diluted to a total improvement on the efficiency of only 0.6% because only 21% of events have converted photons. It was considered not worth adding two more variables to the algorithm for such a small global improvement.

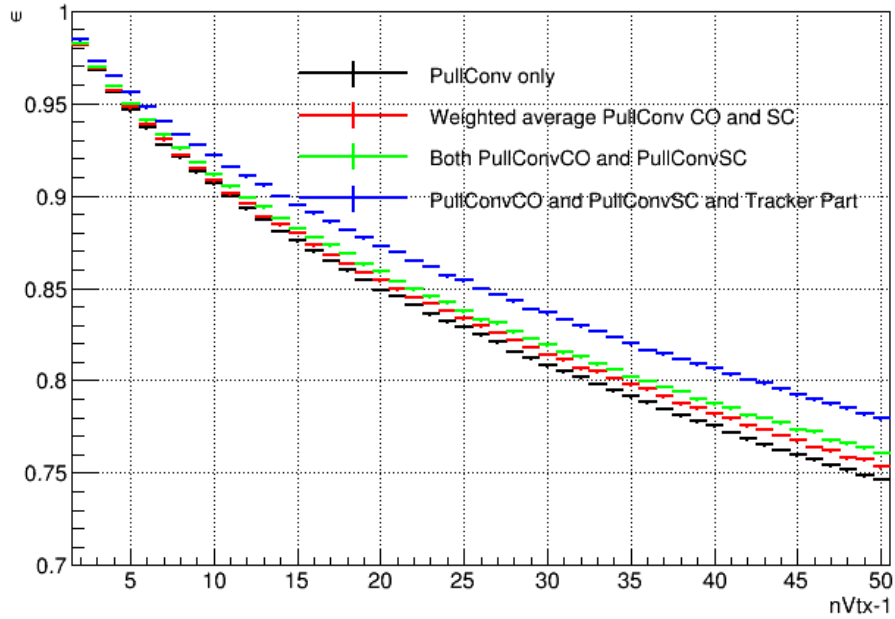


Figure 6.17: Efficiencies to find the correct primary vertex as a function of the number of vertices in the event for different sets of conversion variables.

Adding single leg conversions. Single leg conversions, where only one of the two tracks is reconstructed, were added to the vertex selection algorithm, to try to increase the vertex identification efficiency. From a technical point of view, it is first checked if a double leg conversion is matched to the photons, and if this is not the case, it is checked whether a single leg one matches. As a result, instead of 21% of events with converted photons, 26.5% of events have converted photons, i.e. an increase of a 5.5%. The distributions of the $Pull_{conv}$ and N_{conv} including single leg conversions are shown in Fig. 6.18. In this case, the conversion momentum is taken as the track momentum measured at the innermost hit, and the conversion vertex as the innermost hit and the effective resolutions for SC and

CO algorithms are derived independently. Adding single leg conversions into the study increases the fraction of events with at least one converted photon by 5.5% which leads to a 1% improvement in the global efficiency.

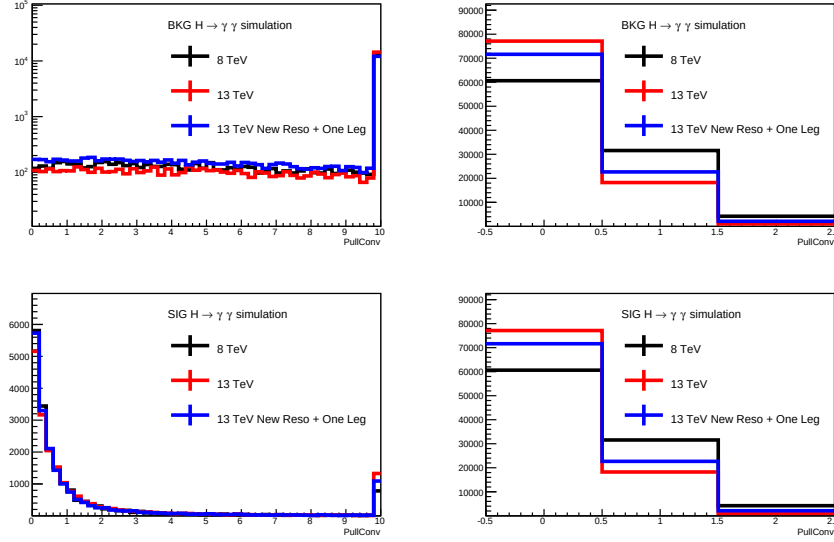


Figure 6.18: Conversion variables ($Pull_{conv}$ on the left and N_{conv} on the right) comparison for 13 TeV and 8 TeV $H \rightarrow \gamma\gamma$ MC, including single leg conversions (in blue). At the top are distributions for the background (random wrong vertex), and at the bottom for the signal (the closest vertex to the true one).

The efficiency as a function of p_T is depicted in Fig. 6.19. The configuration including single leg conversions gives the best efficiency and shown in green.

Releasing the criteria for conversion matching, the percentage of events with at least one converted photon (single leg conversion are included) goes up to 38%. This conversion matching is used in the final analysis.

6.1.3 Performance in $H \rightarrow \gamma\gamma$ MC simulated events

All performance and validation studies were done for 3 iterations of the analysis: Moriond'16 conference, ICHEP'16 conference and for the LHCP conference in 2017 using the full 2016 dataset of 35.9 fb^{-1} . The latest iteration is presented in this chapter. Two previous ones can be found in references [159, 160].

It was found that the beam spot size was very different between data and simulation. This is illustrated in Fig. 6.20, where one can see the difference Δz between the beam spot z coordinate and the muon impact on the z axis for data and simulation, for 2015 and 2016. In simulation, the beam spot spread is $\sigma_{BS} = 5.1 \text{ cm}$ both in 2015 and 2016, whereas

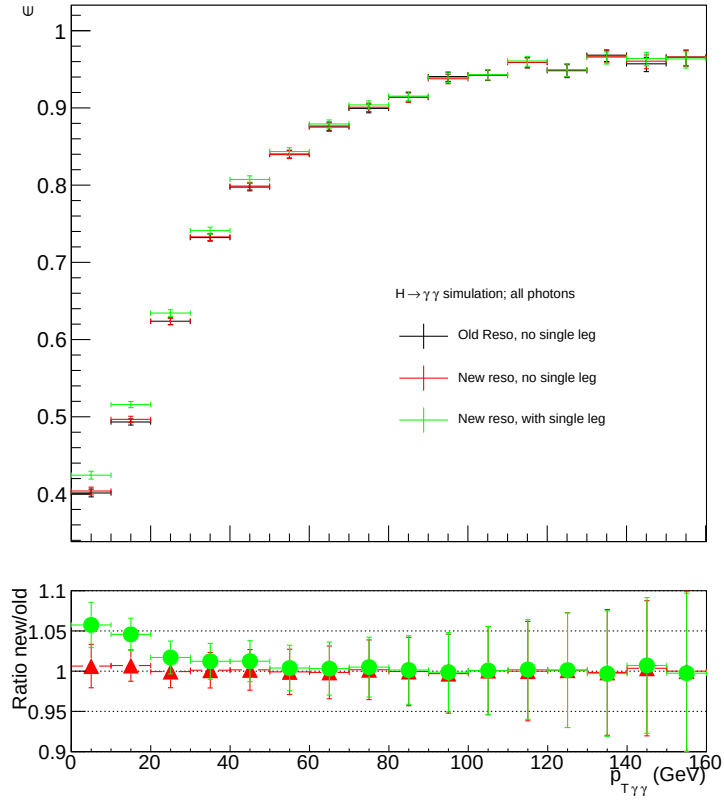


Figure 6.19: Efficiency as a function of p_T for different vertex identification BDT configurations: the effective resolutions taken from the Run I and without single leg conversion in black, the effective resolutions recomputed for the Run II and without single leg conversion in red, the effective resolutions recomputed for the Run II and single leg conversions included in green.

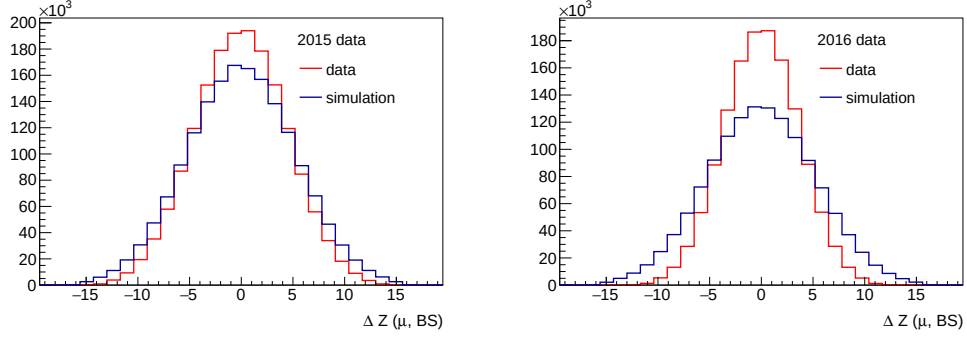


Figure 6.20: Illustration of the beam spot differences between data (red) and simulation (blue) in 2015 (left) and 2016 (right). The X-axis is the Δz between the beam spot z coordinate and the muon impact on the z axis.

in data it was $\sigma_{BS} = 4.2$ cm in 2015 and is about $\sigma_{BS} = 3.6$ cm starting from 2016. To cope with this difference, the $\Delta z = z_{vtx} - z_{true}$ distribution in simulation for strictly wrong vertices is reweighed using a ratio of normalized Gaussians evaluated at Δz :

$$weight = gauss(0, \sqrt{2} \times \sigma_{BS, DATA})[\Delta z] / gauss(0, \sqrt{2} \times \sigma_{BS, MC})[\Delta z], \quad (6.8)$$

where the $\sqrt{2}$ factor provides the conversion from beam spot Gaussian to Δz Gaussian. Δz distribution before and after the beam spot reweighing is shown in Fig. 6.21.

The overall efficiency to find the correct vertex in selected events from simulation reweighted to data pileup conditions ($PU = 23$) and beam spot in data using the full 2016 dataset is $\sim 81\%$. This efficiency is shown as function of diphoton p_T and number of vertices in Fig. 6.22.

6.1.4 Validation with $Z \rightarrow \mu\mu$ events for unconverted photons

As the vertex identification algorithm was optimized on simulation, which does not describe data perfectly, a validation on data events is needed. For unconverted photons, it is performed on $Z \rightarrow \mu\mu$ events, where the muon tracks are removed in order to mimic the diphoton system.

Firstly, the muon tracks are removed and the vertices are refitted without those tracks. Doing so, the z vertex position resolution goes from about $30 \mu m$ with the muon tracks to about $60 \mu m$ without the muon tracks, which is depicted in Fig 6.23. This resolution is well below the 1 cm criteria that defines a "good" vertex and has thus has a negligible impact.

Secondly, the muons are removed from the tracks entering the vertex identification algorithm. The vertex is chosen by this algorithm within the vertex collection obtained

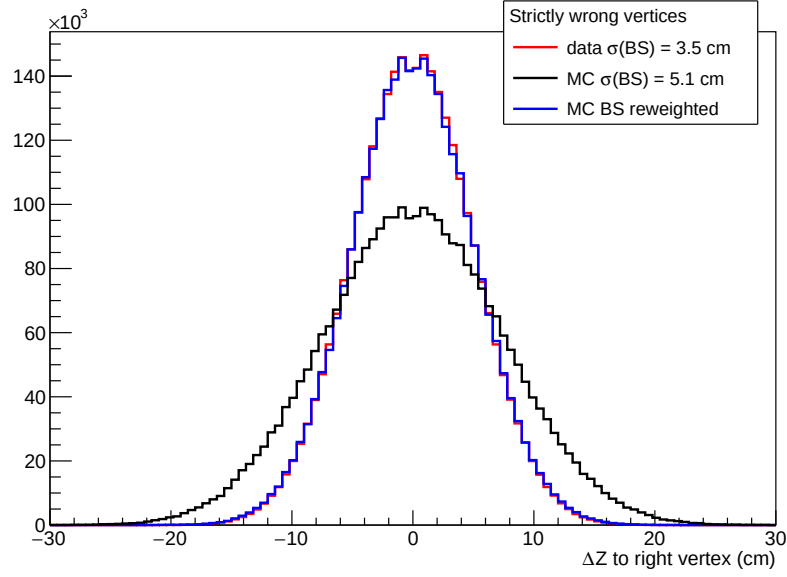


Figure 6.21: Δz distribution for strictly wrongly chosen vertices in data and in simulation before and after the beam spot reweighing.

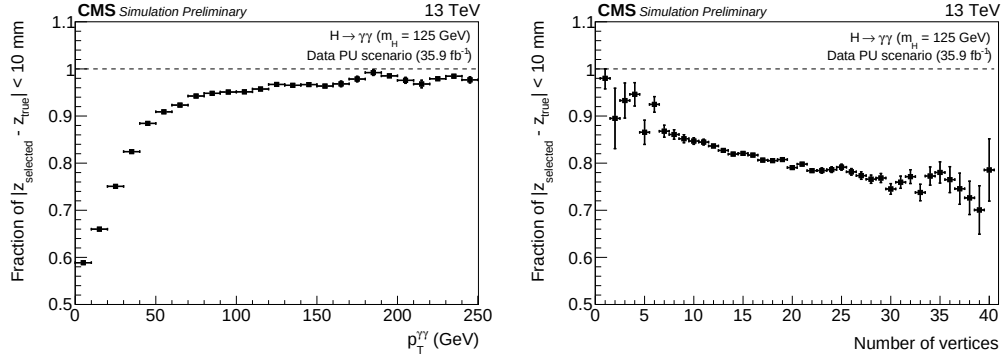


Figure 6.22: The vertex identification efficiency as a function of the reconstructed diphoton p_T (left) and number of vertices (right) in simulated $H \rightarrow \gamma\gamma$ events with $m_H = 125$ GeV. The distributions are reweighted to match the final dataset PU and from beam spot size differences between data and simulation.

without the muon tracks. The true $H \rightarrow \gamma\gamma$ vertex is determined from the muons tracks in the collection of vertex including the muon tracks. It was checked using simulation truth information that this choice is always the relevant one. The efficiency of finding the good vertex is estimated both in data and simulation.

The datasets used come from the proton-proton collisions with 25 ns bunch spacing

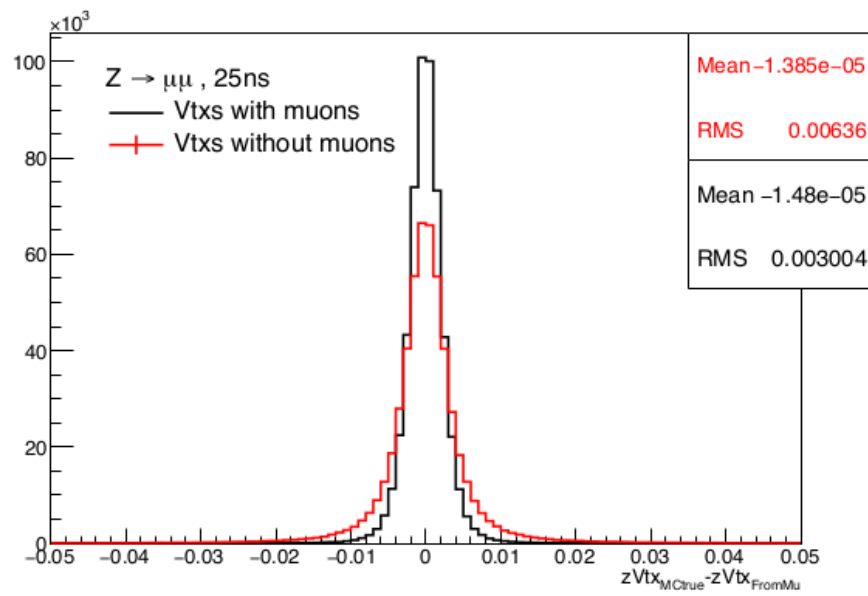


Figure 6.23: Distribution of the difference between the z position of the vertex from Monte-Carlo truth and the z position of the reconstructed vertex matching the best two muon tracks (in cm) in simulated Drell-Yan events. The vertex position computed with muon tracks is shown in black, and without muon tracks in red, with the vertex fitting algorithm. The resolution on the vertex position is 2 times larger without the muon tracks ($60 \mu\text{m}$ instead of $30 \mu\text{m}$).

at center of mass energy of 13 TeV. The events selected for this study are required to be triggered by the HLT_IsoTkMu27 trigger. The integrated luminosity of the dataset analysed is 35.9fb^{-1} . The average number of pileup vertices in the datasets used is 23. The Monte-Carlo simulated samples used come from the Drell-Yan process, which is simulated with MADGRAPH5_aMC@NLO at next-to-leading order with up to two jets from the matrix element level.

Event is selected if two muons pass all tight identification criteria (as listed in section 5.11.1), except the ones involving the vertex position. The dimuon mass is required to be within 70 GeV and 110 GeV.

The distribution of input variables for data and simulation, together with their ratios are shown in Fig. 6.24 respectively for signal (the right vertex, matched to the muons) and background (a wrong vertex, not matching the muons). The distributions for simulation are reweighted to match the final dataset PU and for beam spot size differences between data and simulation.

The efficiency to select the vertex within 1 cm from the true vertex in z by the $\text{H} \rightarrow \gamma\gamma$ vertex identification algorithm in $\text{Z} \rightarrow \mu\mu$ events in data and MC simulation, along with their ratio, are shown in Fig. 6.25 as a function of the p_T of the dimuon system and also as a function of the number of vertices in the event. The simulation is reweighted for pileup as in the global $\text{H} \rightarrow \gamma\gamma$ analysis, as shown in section 5.1.

The difference in efficiency is coming from the different beam spot in data and simulation. The efficiency in data and simulation are in good agreement after the beam spot reweighing, which is illustrated in Fig. 6.26. The data/simulation efficiencies ratio versus p_T and number of vertices are used to correct the efficiencies in the $\text{H} \rightarrow \gamma\gamma$ simulation, and varied within uncertainties to estimate the associated systematic uncertainty.

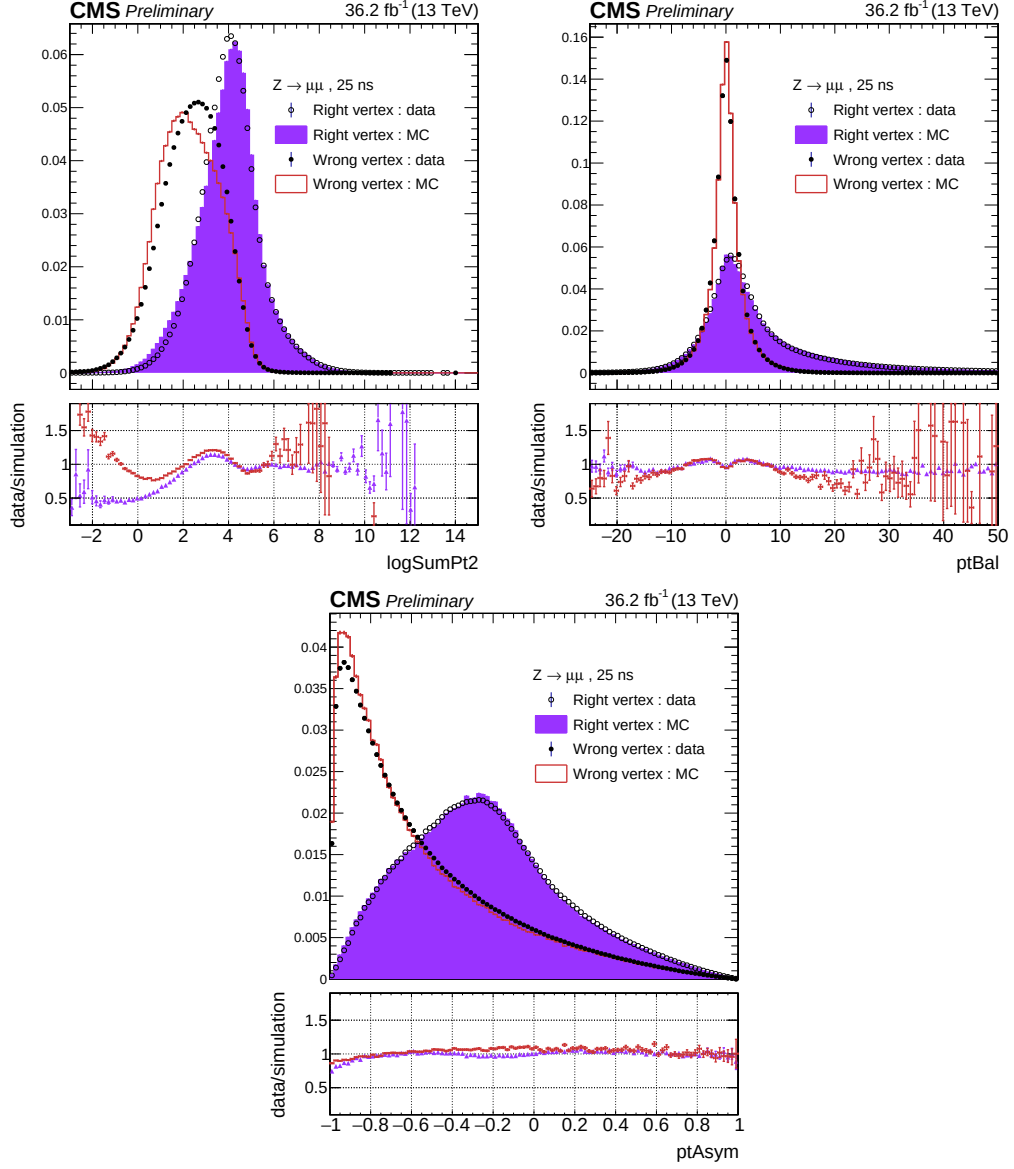


Figure 6.24: Distributions of input variables for data and simulation, together with their ratios for signal vertices (the right vertex, matched to the muons) and background vertices (a wrong vertex, not matching the muons) reweighted to match data pileup and beam spot.

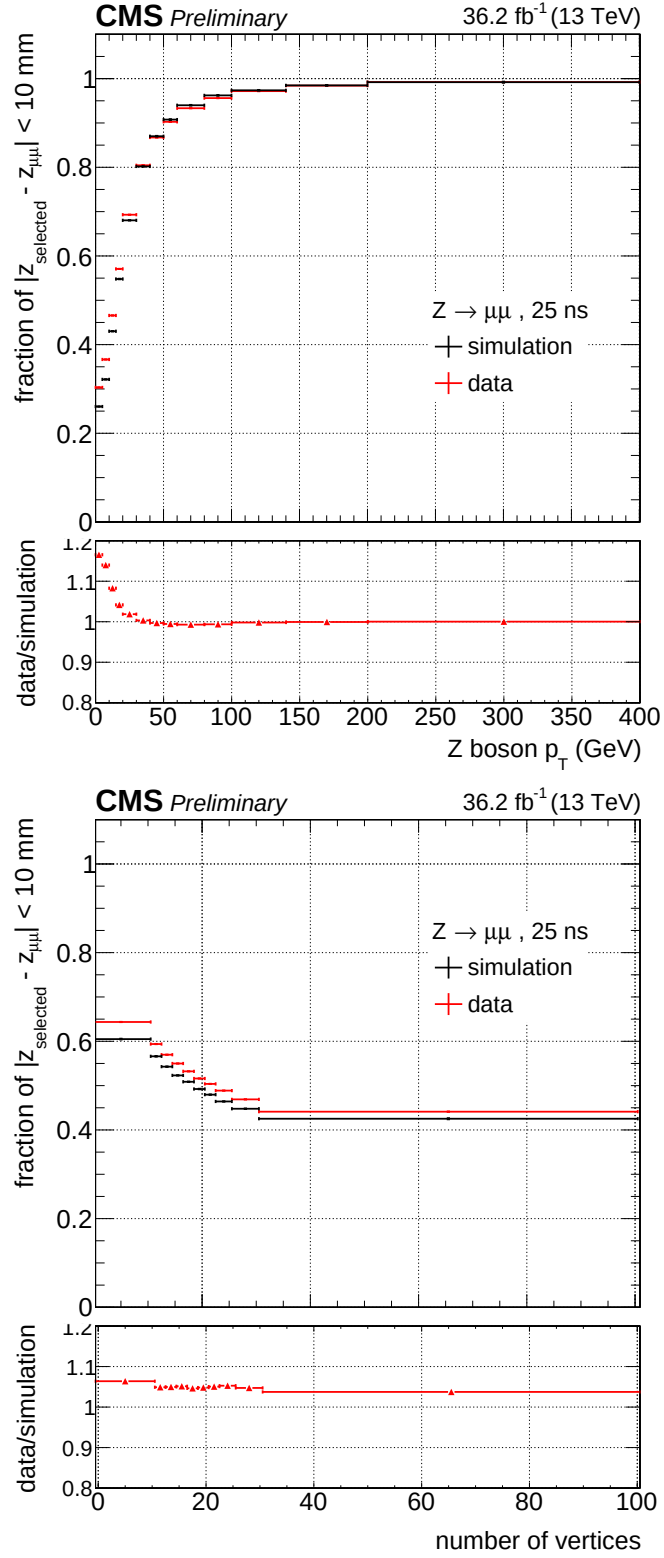


Figure 6.25: Efficiency as a function of p_T (top) and as a function of the number of vertices (bottom)), using $Z \rightarrow \mu\mu$ event for data and MC before the beam spot reweighing.

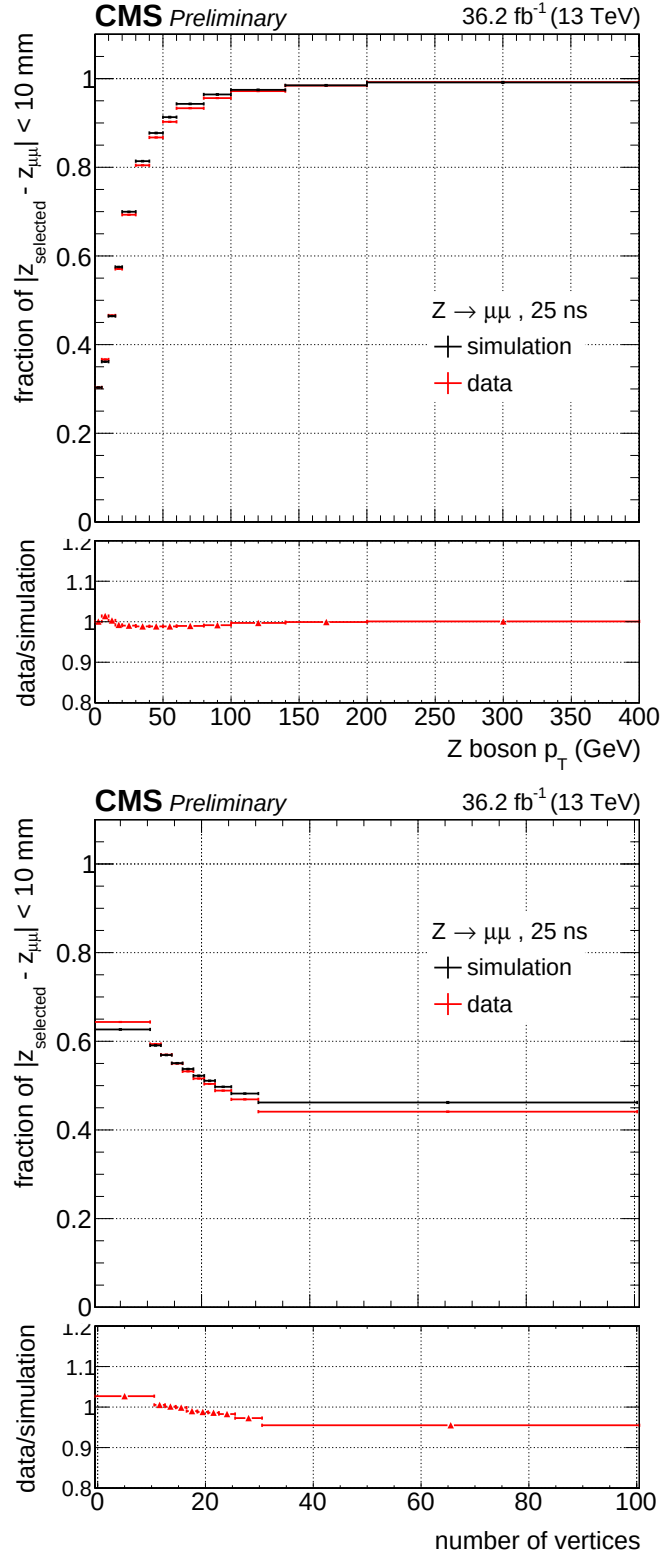


Figure 6.26: Efficiency as a function of p_T (top) and as a function of the number of vertices (bottom)), using $Z \rightarrow \mu\mu$ event for data and MC after the beam spot reweighing.

6.1.5 Validation with $\gamma + \text{jet}$ events for converted photons

The performance of the $H \rightarrow \gamma\gamma$ vertex identification algorithm is validated in the case of converted photons using $\gamma + \text{jet}$ events. The basic principle behind using $\gamma + \text{jet}$ events is to create a photon-jet system by pairing a photon and a jet while removing the tracks associated with the jet during the process of vertex identification, in order to mimic a diphoton system, and subsequently to compare the vertex selected by the $H \rightarrow \gamma\gamma$ vertex identification algorithm with the vertex that is tagged by the jet in order to calculate the efficiency of selection of the correct vertex.

The datasets used for the study come from the proton-proton collisions with 25 ns bunch spacing at center of mass energy of 13 TeV. The events selected for the study are required to be triggered by the HLT_Photon50 trigger, which was prescaled during data taking. The integrated luminosity of the dataset analysed, which correspond to the full 2016 dataset, after accounting for the trigger prescale, is 31.8 pb^{-1} . The average number of pileup vertices in this dataset is 23.

The MC simulated samples used are $\gamma + \text{jet}$ samples. Those samples generally contain one photon and one electromagnetically enriched object originating from a jet. They are generated with Pythia8 and a "double EM-enriched"² filter was applied during the production in order to select events which are likely to pass the final diphoton selection of the analysis, and in this way improve the production efficiency.

Events containing a converted photon with p_T greater than 55 GeV passing the cut based photon identification criteria, designed to have 90% of efficiency and at least one jet with p_T greater than 30 GeV are selected. For the jets, the sum of the transverse momenta of the constituent charged particles is also required to be greater than 30 GeV. A photon and a jet are paired to form a photon-jet system if they do not overlap ($\Delta R_{\gamma, \text{jet}} > 0.4$). The events in the MC samples are weighted such that the distribution of the reconstructed number of vertices in the MC matches the one in data. Since the trigger is prescaled, the simulation cannot be reweighted to the data pileup as in the global $H \rightarrow \gamma\gamma$ analysis and $Z \rightarrow \mu\mu$ validation. The normalized distributions of the number of vertices per event, the p_T of the photon-jet system, the p_T of the converted photons and the p_T of the associated jets in the $\gamma + \text{jet}$ selected events in data and MC simulation are shown in Fig. 6.27.

The $H \rightarrow \gamma\gamma$ vertex identification algorithm is used to determine the vertex associated with the photon-jet system, with the tracks associated with the jet, i.e. all tracks lying within $\Delta R < 0.4$ with respect to the four momentum of the jet, removed from the input of the vertex finding BDT. The tracks associated with the jet are removed so that the

²The "Double EM-enriched" filter is created for the production of the QCD dijet and $\gamma + \text{jet}$ samples. This filter requires an electromagnetic activity, coming from photons, electrons or neutral hadrons, to pass a p_T threshold of 15 GeV. It also requires no more than two charged particles in a cone $\Delta R < 0.2$, mimicking a tracker isolation. These charged particles can be electrons, muons, taus, pions and kaons, and they are required to have tracks $p_T > 1.6 \text{ GeV}$, $|\eta| < 2.2$. For each event, the potential photon signals are coupled together to form double EM enriched object.

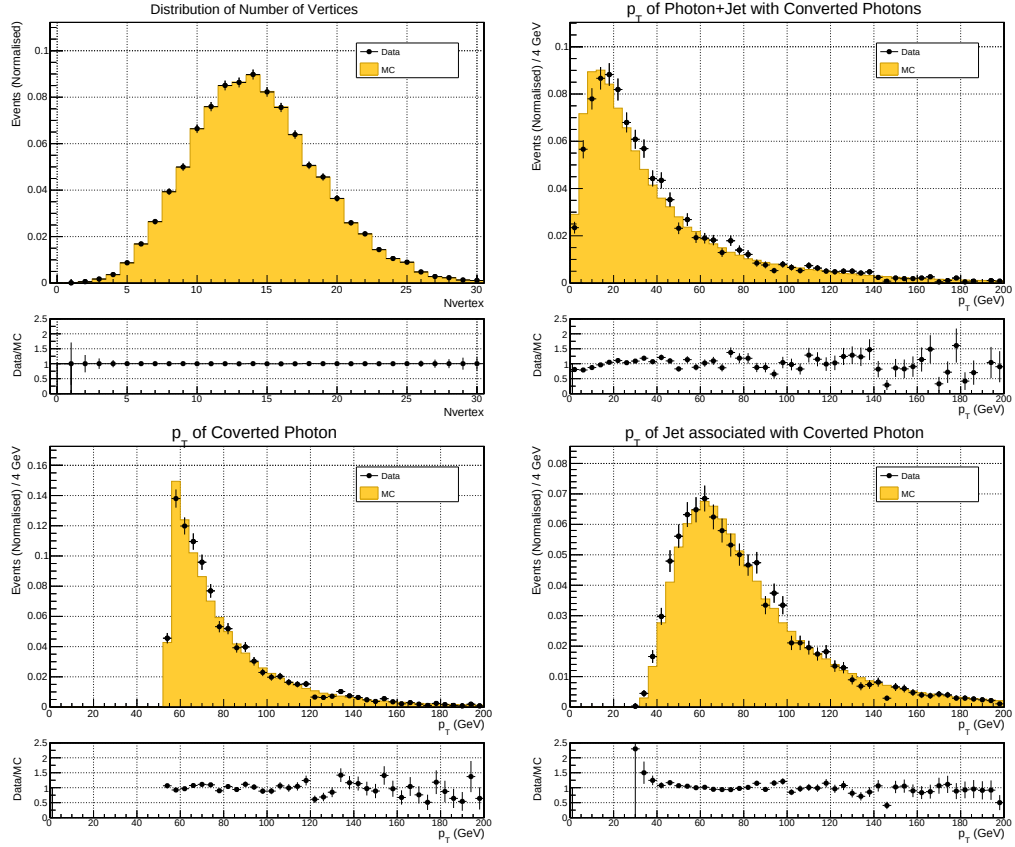


Figure 6.27: Distributions of the number of vertices in the event (top left), p_T of the photon-jet system (top right), p_T of the converted photon (bottom left) and p_T of the associated jet (bottom right) in $\gamma + \text{jet}$ events with converted photons in data (black dots) and MC simulated events (yellow) and their ratio. The distributions are normalized to the area.

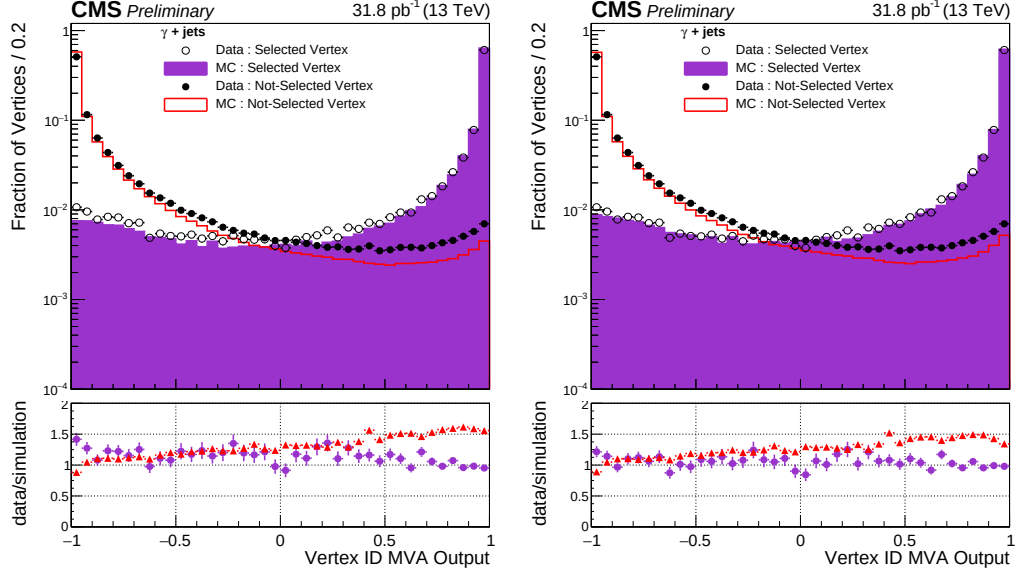


Figure 6.28: Normalized distributions of the $H \rightarrow \gamma\gamma$ vertex identification BDT output for the selected vertices in data (open black circles) and MC (purple histogram) and for vertices that are not selected in data (closed black circles) and MC (open red histogram) in $\gamma + \text{jet}$ events with converted photons. The distributions on the left are not reweighted for the beam spot size differences between data and simulation, while the distributions on the right are reweighted.

photon-jet system can mimic a diphoton system. The vertex selected by the $H \rightarrow \gamma\gamma$ vertex identification algorithm is compared with the vertex that is tagged by the jet in order to calculate the efficiency of finding the correct vertex. If the z coordinate of the vertex selected by the $H \rightarrow \gamma\gamma$ vertex identification algorithm lies within 1 cm of the z coordinate of the jet-tagged vertex, then the vertex is taken to be correctly identified. The normalized distribution of the BDT score of the $H \rightarrow \gamma\gamma$ vertex identification MVA is shown in Fig. 6.28 separately for the vertices that are selected by the algorithm and those that are not in $\gamma + \text{jet}$ events in data and MC simulation. The distributions are reweighted for the beam spot size differences between data and simulation are also shown.

The efficiency of the selection of the correct vertex by the $H \rightarrow \gamma\gamma$ vertex identification algorithm in $\gamma + \text{jet}$ events with converted photon in data and MC simulation, along with their ratio, are shown in Fig. 6.29 as a function of the p_T of the photon-jet system and also as a function of the number of vertices in the event. The vertex identification algorithm is found to be highly efficient for the case of converted photons, and the efficiency agreement between data and simulation is at the same level as for $Z \rightarrow \mu\mu$ validation.

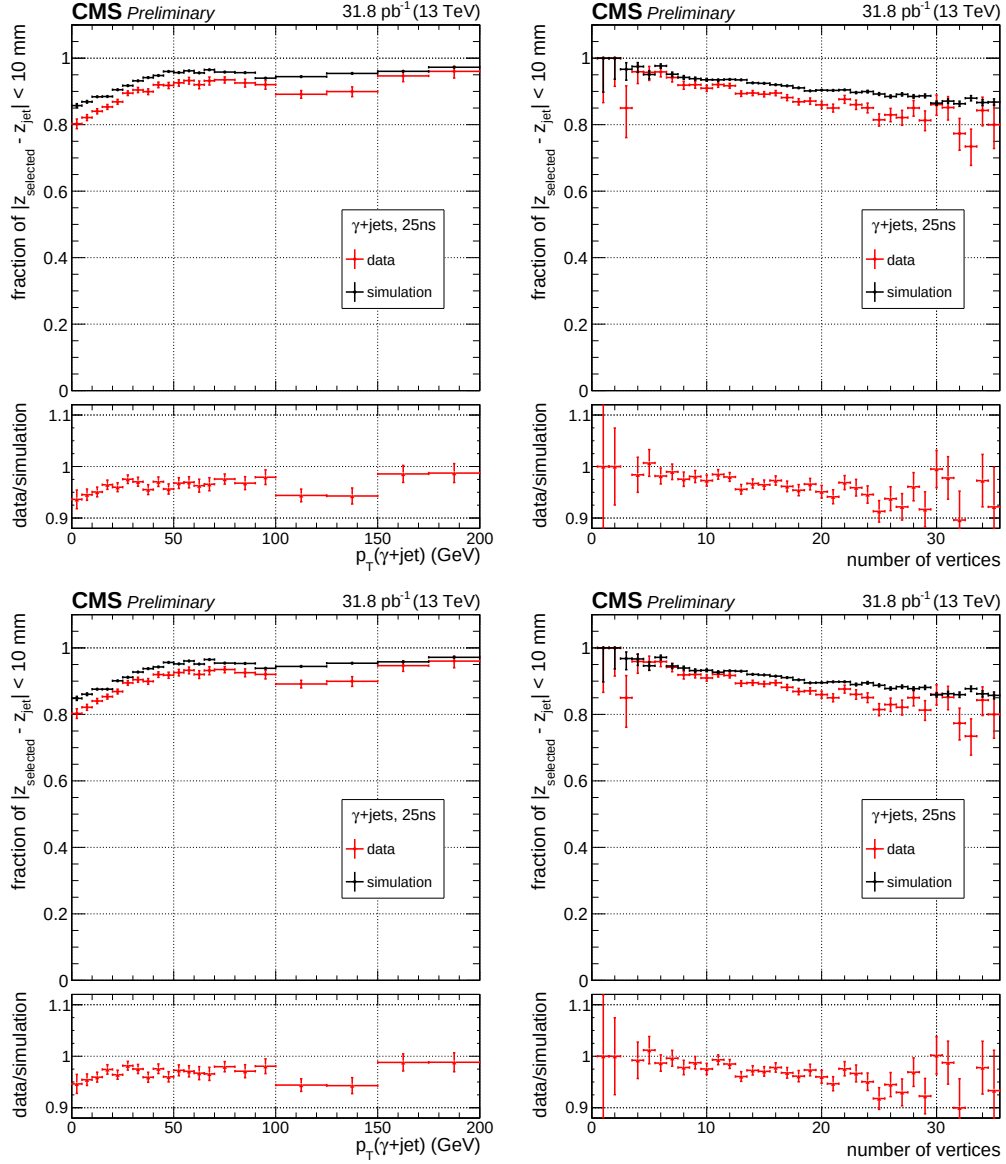


Figure 6.29: The efficiency of the selection of the correct vertex by the $H \rightarrow \gamma\gamma$ vertex identification algorithm in $\gamma + \text{jet}$ events with converted photons as a function of the p_T of the photon-jet system (left) and as a function of the number of vertices in the event (right) in data (red) and MC simulated events (black) and their ratio. The distributions on the top are not reweighted for the beam spot size differences between data and simulation, while the distributions at the bottom are reweighted.

6.2 Per-event probability of correct diphoton vertex choice

6.2.1 Principle

The probability to choose the vertex within 1 cm of the true vertex is used in the rest of the analysis to fully benefit from the excellent ECAL resolution. This probability is computed from the output of a second MVA (called "MVAprob") obtained using a BDT method, as for vertex identification. The simulation samples used are the same as described in section 6.1.1.

The input variables are the following:

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

The distributions of the input variables for the second MVA in ggH production mode $H \rightarrow \gamma\gamma$ Monte-Carlo samples at 13 TeV are shown in Fig. 6.30. In this case, vertices located further than 1 cm from the true vertex are taken as background and the one within 1 cm, as signal.

The per-event probability to choose the vertex within 1 cm of the true one is taken from a parametrisation of the efficiency as a 4th order polynomial function of the output of this MVA with a constraint at MVAprob value = -1, where the probability has to be 1, separately for converted and unconverted photons. This parametrization is shown in Fig. 6.31.

6.2.2 Performance in $H \rightarrow \gamma\gamma$ MC simulated events

The comparison between the true vertex identification efficiency and the average estimated vertex probability is shown in Fig. 6.32 as a function of the reconstructed diphoton p_T (top) and the number of vertices (bottom). All production modes are included with a weight corresponding to their expected cross-sections. The simulation events have been reweighted such that the pileup distribution in simulation matches that in data for full 2016 dataset (35.9 fb⁻¹) and also such that beam spot size matches the one in data.

6.2.3 Validation with $Z \rightarrow \mu\mu$ events for unconverted photons

The performance of the BDT for the calculation of the per event probability of correct diphoton vertex choice is validated using $Z \rightarrow \mu\mu$ events. The datasets, MC simulated

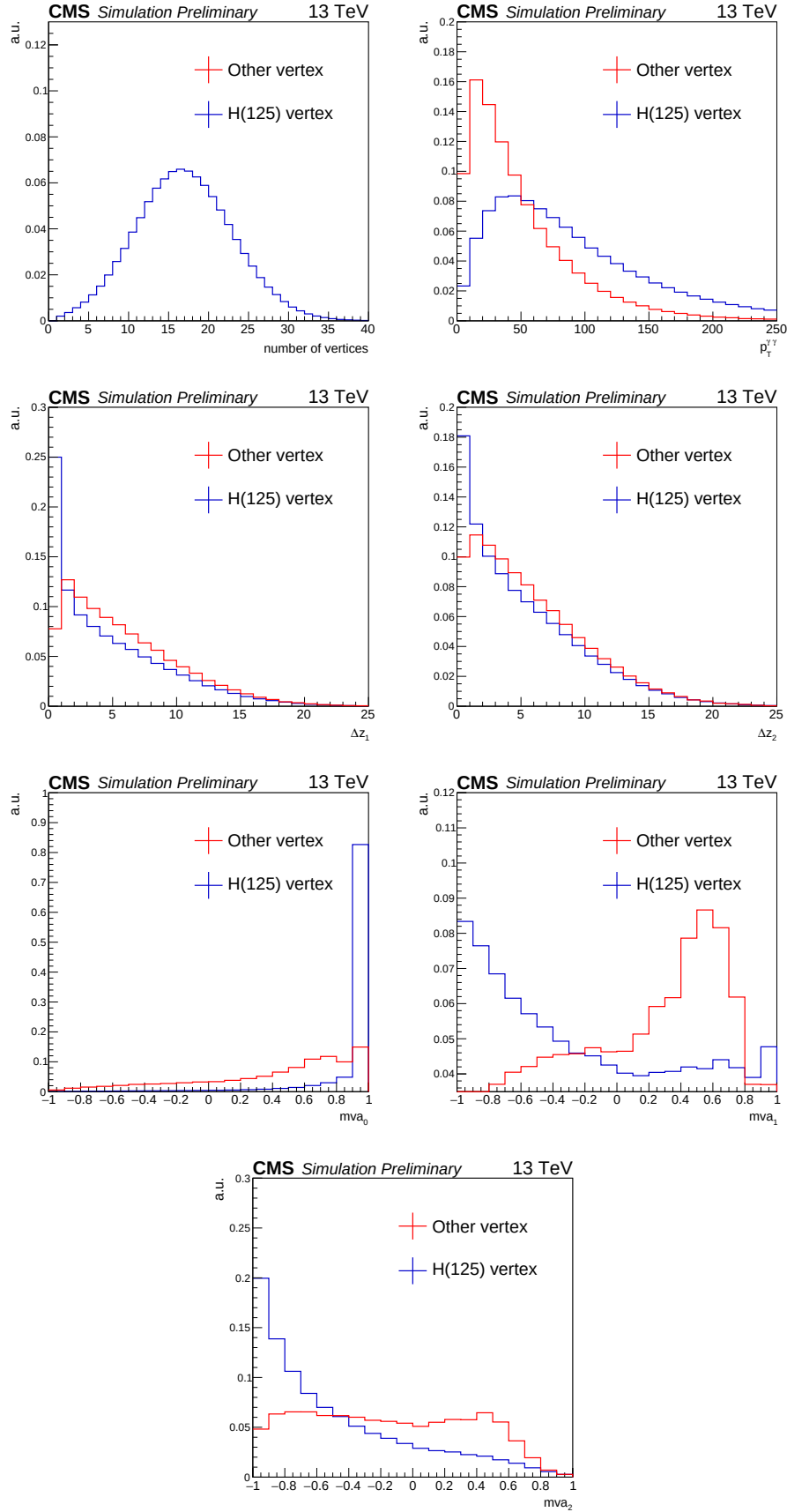


Figure 6.30: Signal (in red) and background (in blue) distributions of input variables for the second MVA.

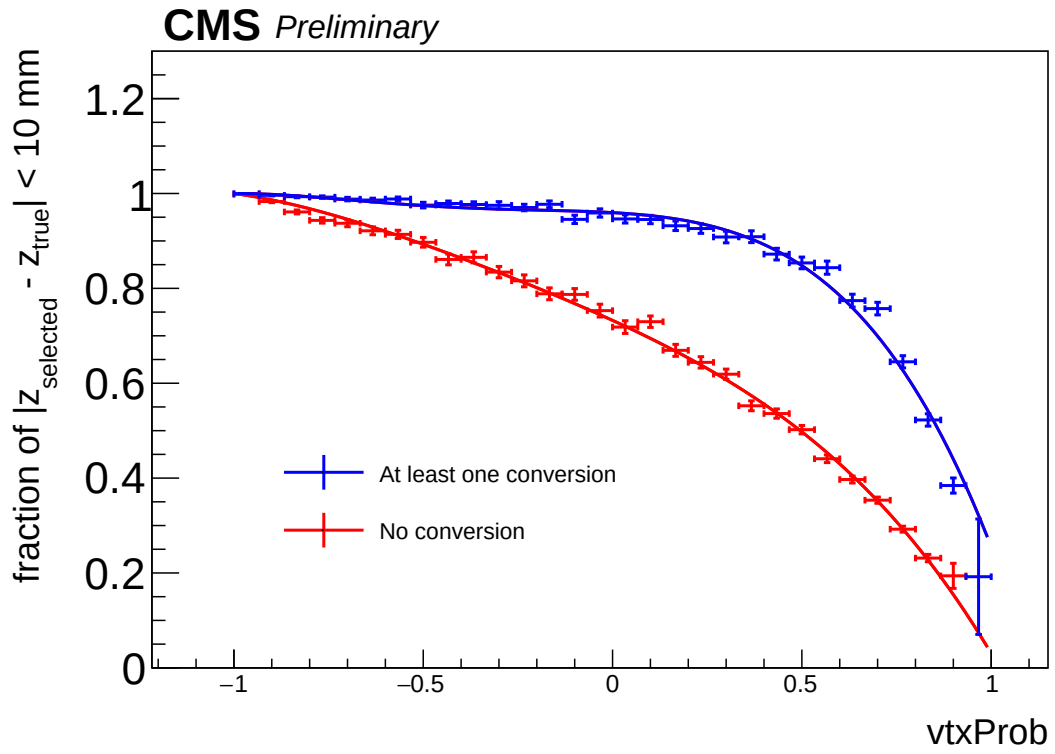


Figure 6.31: The vertex identification probability to choose the correct diphoton vertex as a function of the second BDT output ("MVApob") for converted photons (in blue) and for unconverted photons (in red).

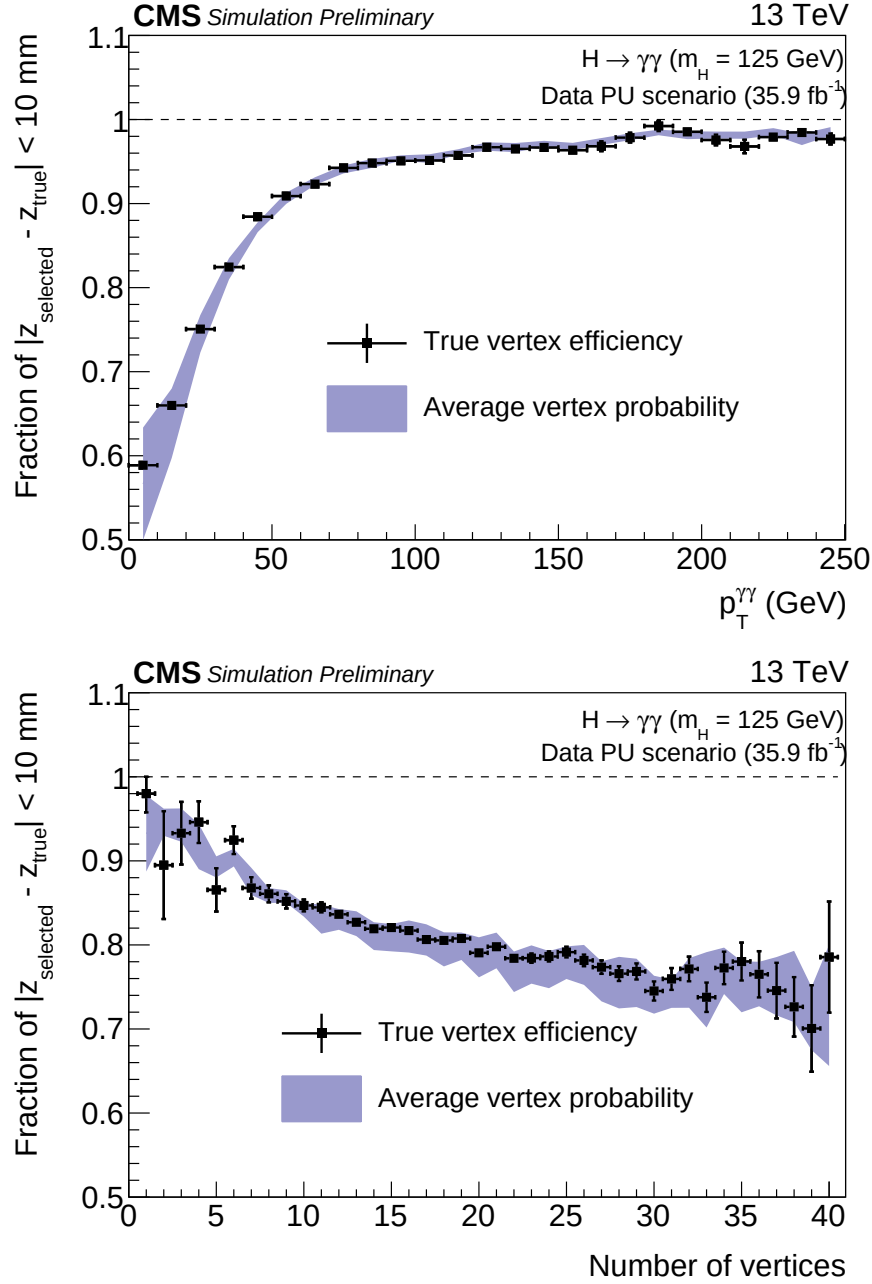


Figure 6.32: Comparison of the true vertex identification efficiency and the average estimated vertex probability as a function of the reconstructed diphoton p_T (top) and number of vertices (bottom) in simulated $H \rightarrow \gamma\gamma$ events with $m_H = 125$ GeV. The average vertex probability is superimposed with its uncertainty derived from $Z \rightarrow \mu\mu$ validation. The distributions are reweighted to match the final dataset PU and from beam spot size differences between data and simulation.

samples and event selection criteria are identical to that described in section 6.1.4. $Z \rightarrow \mu\mu$ events are selected and the tracks associated to the two muons removed from the input of the BDT calculation, in order to mimic a diphoton system. The vertex that is tagged by the dimuon system is used as a reference to study the performance of the BDT. If the z coordinate of the vertex selected by the $H \rightarrow \gamma\gamma$ vertex identification algorithm lies within 1 cm of the z coordinate of the dimuon system tagged vertex, then the vertex is taken to be correctly identified.

The normalized distribution of the per event probability of correct diphoton vertex is shown in Fig. 6.33 separately for the vertices correctly selected and for misassigned vertices in $Z \rightarrow \mu\mu$ events in data and simulation. The distributions are in fair agreement, and the agreement is slightly improved by the beam spot reweighting of simulated events. The efficiencies as a function of the output MVA value for data and simulation and their ratio are shown in Fig. 6.34. Most of the events are in the first bin of this efficiency, as most of the events have a correctly chosen vertex. This ratio is used to compute an error (blue band) on the vertex probability to choose the correct vertex, which is shown in Fig. 6.32. The efficiencies for data and simulation are in good agreement after beam spot reweighting in simulated events.

6.2.4 Validation with $\gamma + \text{jet}$ events for converted photons

The performance of the BDT for the calculation of the per event probability of correct diphoton vertex choice is validated using $\gamma + \text{jet}$ events with converted photons. The datasets, MC simulated samples and event selection criteria are identical to that described in Section 6.1.5. The $\gamma + \text{jet}$ events with converted photons are selected and the tracks associated with the jets are excluded from the input of the BDT calculation, in order to mimic a diphoton system. The vertex that is tagged by the jet is used as a reference to study the performance of the BDT. If the z coordinate of the vertex selected by the $H \rightarrow \gamma\gamma$ vertex identification algorithm lies within 1 cm of the z coordinate of the jet-tagged vertex, then the vertex is taken to be correctly identified.

The normalized distribution of the BDT score for the calculation of the per-event probability of the correct diphoton vertex is shown in Fig. 6.35 (top) separately for the vertices that are correctly selected and for misassigned vertices (if the z coordinate of the selected vertex does not lie within 1 cm of the z coordinate of the jet-tagged vertex) in $\gamma + \text{jet}$ events in data and simulation. The correct and misassigned vertices are well separated. The distributions are also shown after reweighting for the beamspot size difference between data and simulation.

The normalized distribution of the per event probability of correct diphoton vertex is shown in Fig. 6.35 (bottom) separately for the vertices correctly selected and for misassigned vertices in $\gamma + \text{jet}$ events in data and simulation. The probability to select good vertex for converted photons is higher than for unconverted ones.

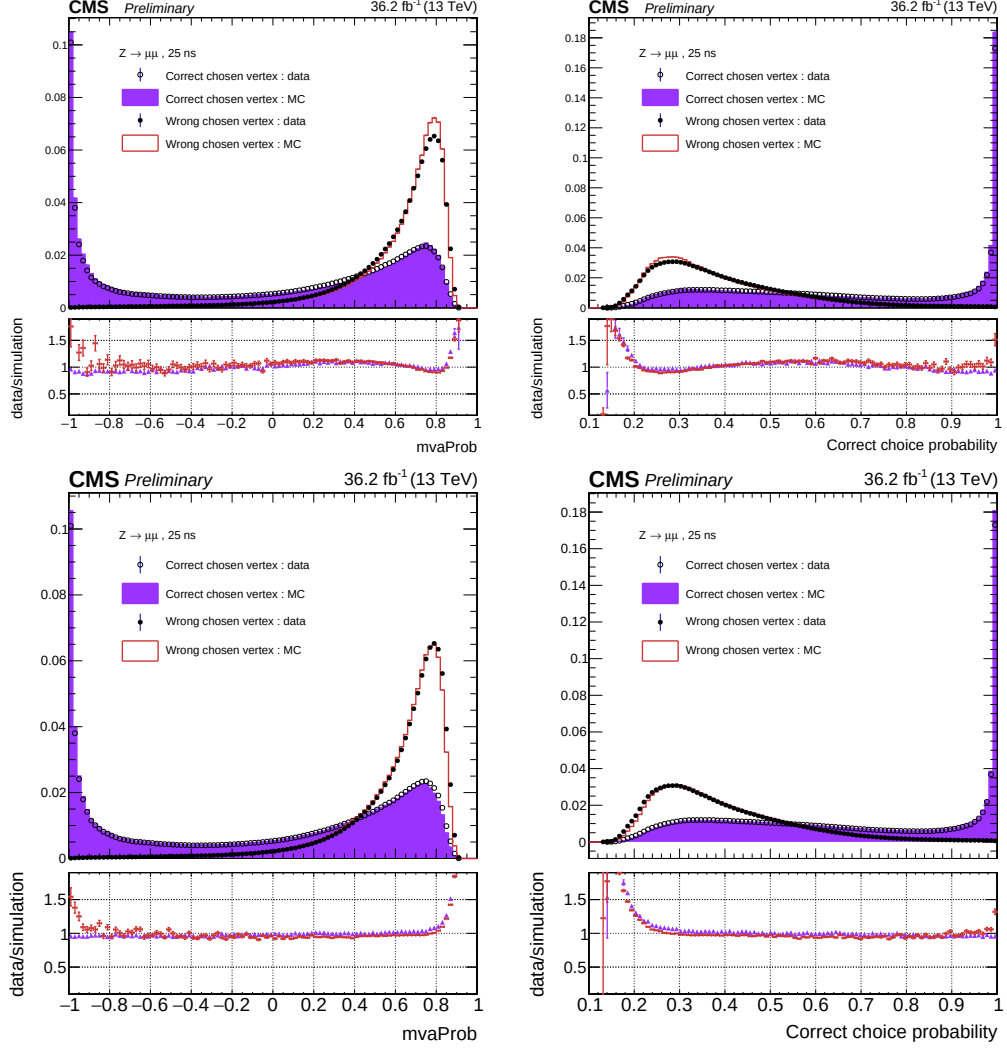


Figure 6.33: Normalized distributions of the second BDT score "MVAprob" (left) and of the per event probability of correct diphoton vertex (right) for correctly selected vertices in data (open black circles) and simulation (purple histogram) and for misassigned vertices in data (closed black circles) and simulation (open red histogram) in $Z \rightarrow \mu\mu$ events. In the two bottom plots, simulated events are reweighted to match the Δz distribution in data ("beam spot reweighting").

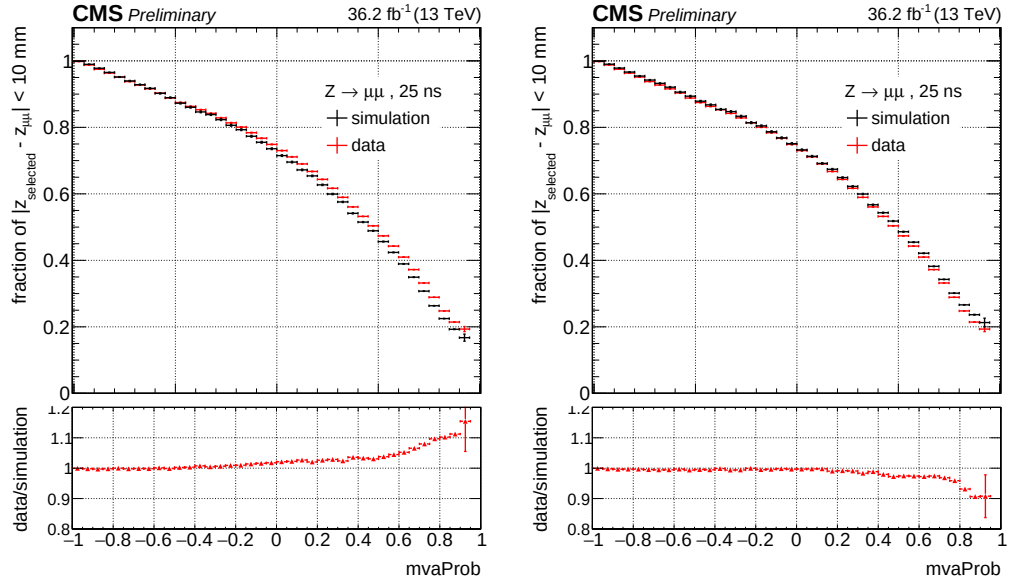


Figure 6.34: Efficiency as a function of the second BDT score "MVAprob" to find the vertex within 1 cm of the true one using $Z \rightarrow \mu\mu$ events for data and simulation. On the right, simulated events are reweighted to match the Δz distribution in data ("beam spot reweighting"), whereas on the left they are not.

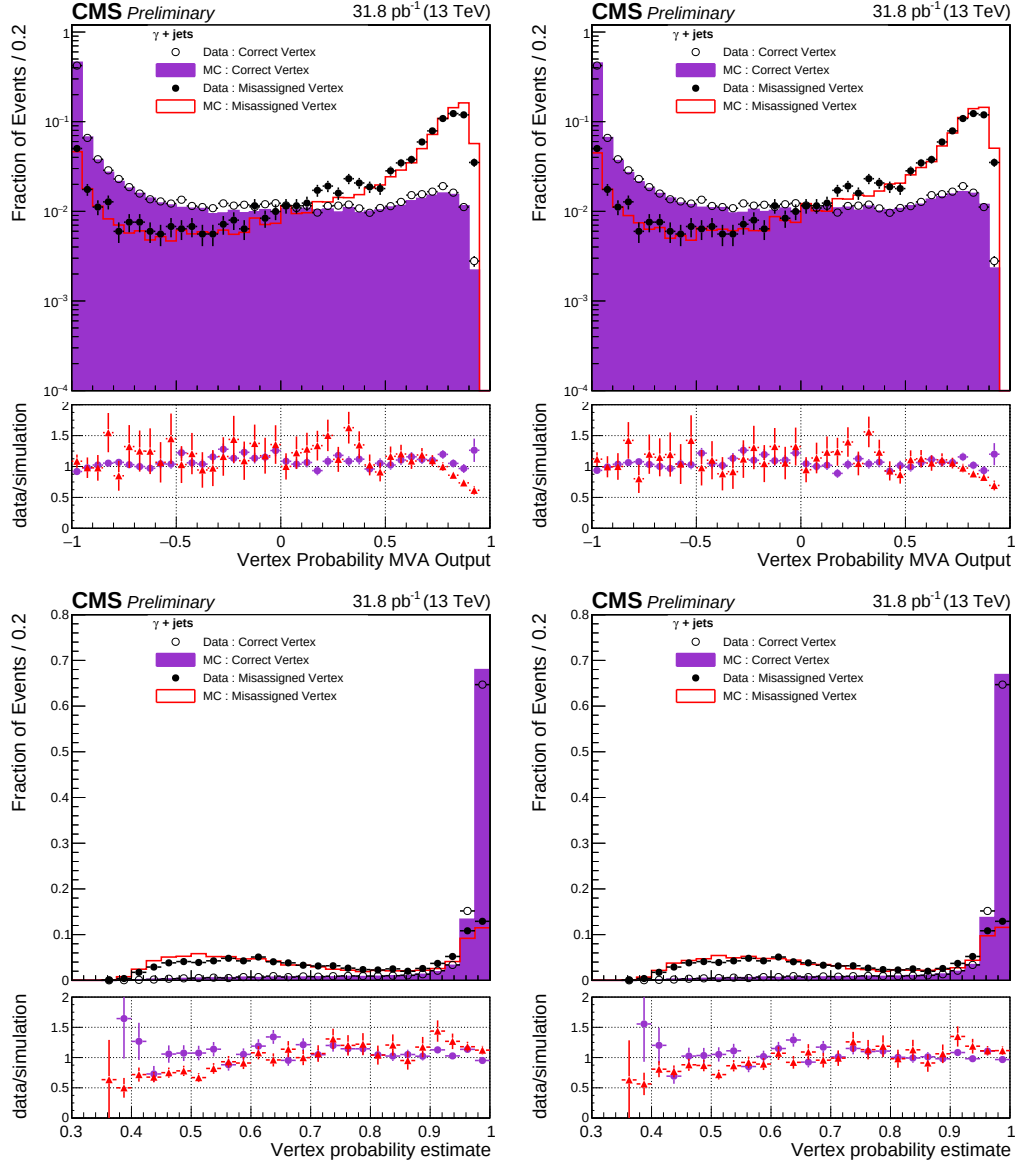


Figure 6.35: Normalized distributions of the BDT score "MVAprob" (top) and of the per event probability of correct diphoton vertex (bottom) for correctly selected vertices in data (open black circles) and simulation (purple histogram) and for misassigned vertices in data (closed black circles) and simulation (open red histogram) in $\gamma + \text{jet}$ events with converted photons. The plots on the left are not reweighted for the beam spot size differences between data and simulation, while the plots on the right are reweighted.

6.3 Corrections and systematic uncertainties

6.3.1 Corrections to simulation

To account for the observed differences between simulation and data, scale factors are applied to the Higgs boson simulation based on the observed differences in $Z \rightarrow \mu\mu$ events. Those scale factors are applied as a function of the Higgs boson p_T first. The remaining differences between data and MC as a function of the number of vertices are also corrected.

Fig. 6.36 shows the data/simulation ratio versus p_T (left) and the number of vertices (right) for 3 different scenarios:

- using $Z \rightarrow \mu\mu$ events with the nominal simulation reweighted for beam spot width differences between data and simulation (in black in all plots);
- using $Z \rightarrow \mu\mu$ events with a simulation with a realistic beam spot width, but with 50 times less statistics (in red in top plots), MC fluctuations are too large for this simulation to be used;
- using $\gamma + \text{jet}$ events for all photons (converted and non converted) with the nominal simulation reweighted for beam spot width differences between data and simulation (in blue in middle plots). The trends are similar to what is seen in Z , but statistical fluctuations are larger.

In the bottom plots of fig. 6.36, the red point are the data/MC ratio obtained using $Z \rightarrow \mu\mu$ events with the nominal simulation reweighted for beam spot width differences between data and simulation, after applying the corrections versus p_T . By construction, the left plot becomes flat and the right one, in red, shows the final correction to be used versus the number of vertices.

To take into account those differences between data and simulation, in the final analysis, simulation events are weighted with c_R for right vertices events, where c_R is defined as:

$$c_R = \frac{\epsilon(\text{data})}{\epsilon(\text{MC})} \quad (6.9)$$

where ϵ is the right vertex efficiency determined from $Z \rightarrow \mu\mu$ events. The weights used for wrong vertices events c_W are defined so that they preserve the total number of events:

$$c_W = \frac{1 - \epsilon_H c_R}{1 - \epsilon_H} \quad (6.10)$$

where ϵ_H is the right vertex efficiency from the Higgs boson simulation, using all production modes with their respective cross-sections and reweighted to the data PU.

The final corrections for event with the right vertex and the wrong vertex are shown in Fig. 6.37.

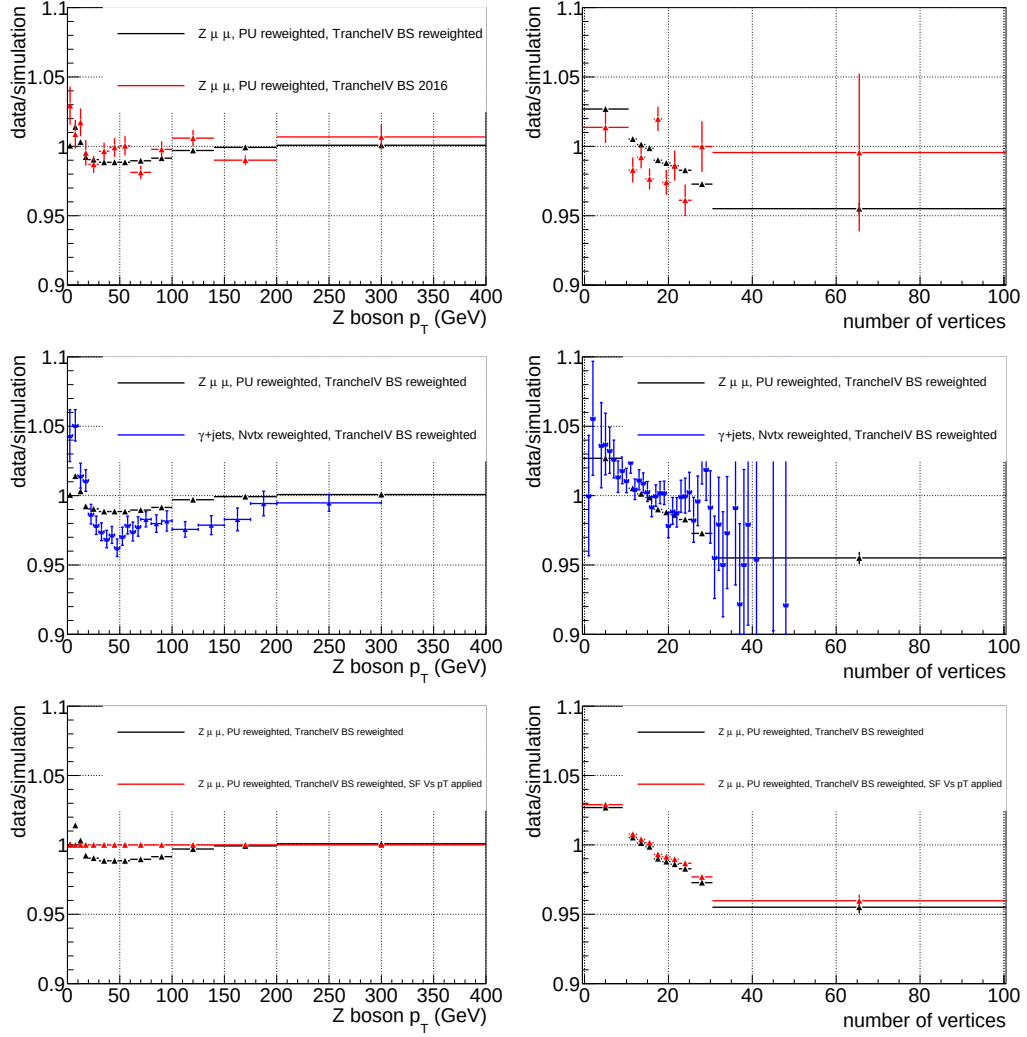


Figure 6.36: Data/simulation ratios for the vertex efficiency versus p_T (left) and the number of vertices (right) for $Z \rightarrow \mu\mu$ events with the nominal simulation reweighted to the data beam spot (black) and with a simulation with more realistic beam spot (red), they are shown on top. Data/simulation ratios for the vertex efficiency versus p_T (left) and the number of vertices (right) for $\gamma + \text{jet}$ events for all photons (converted and non converted) with the nominal simulation reweighted to the data beam spot are also shown (blue), they are shown in the middle. Data/simulation ratios for the vertex efficiency versus p_T (left) and the number of vertices (right) for $Z \rightarrow \mu\mu$ events with the nominal simulation reweighted to the data beam spot (black) and also with corrections as function of p_T applied (red), they are shown on the bottom.

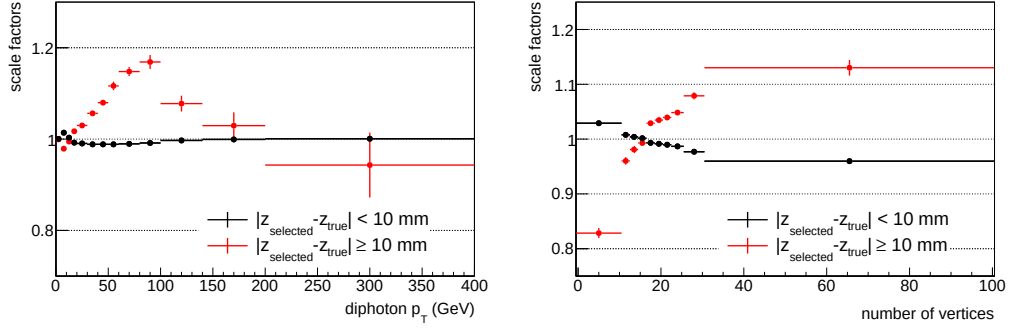


Figure 6.37: Final scale factors applied to right (red) and wrong (black) vertex events versus the Higgs boson p_T (left) and the number of vertices (left).

6.3.2 Systematic uncertainties

Applying those weights to the Higgs boson simulation changes the total vertex efficiency by 1% from 81% to 80%. This change is taken as a systematic uncertainty in the final analysis, together with other sources (uncertainties from the underlying event and from the Higgs boson p_T). The uncertainties are also estimated from:

- possible mismodelling of the underlying event (UE) in simulation,
- possible mismodelling of the Higgs boson p_T in simulation.

The uncertainty from UE is computed by comparing vertex efficiencies from Higgs simulation where UE tune [190] is varied. These efficiencies are shown in Fig. 6.38. The efficiency changes by +2% -1%.

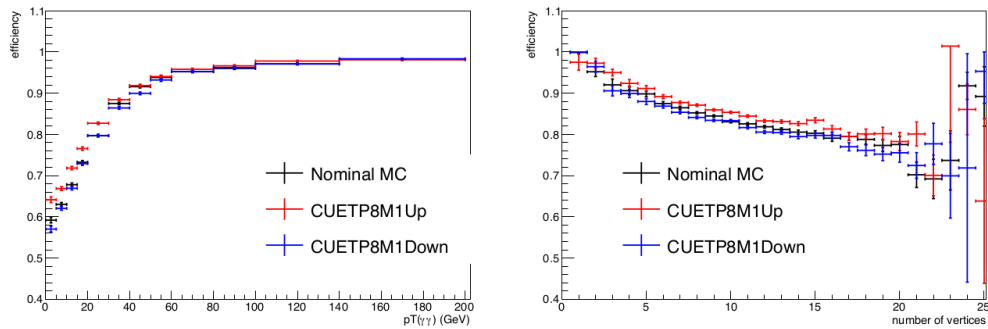


Figure 6.38: The efficiency to identify correct vertex for simulation with nominal underlying event - in black, with up variation - in red, with down variation - in blue.

The uncertainty coming from the Higgs boson p_T descriptions is computed using simulations with QCD renormalisation and factorisation scales which are twice larger or smaller with respect to the nominal one. The efficiencies to find the correct vertex estimated on simulations with different QCD scales along with their ratio to the nominal simulation are shown in Fig. 6.39.

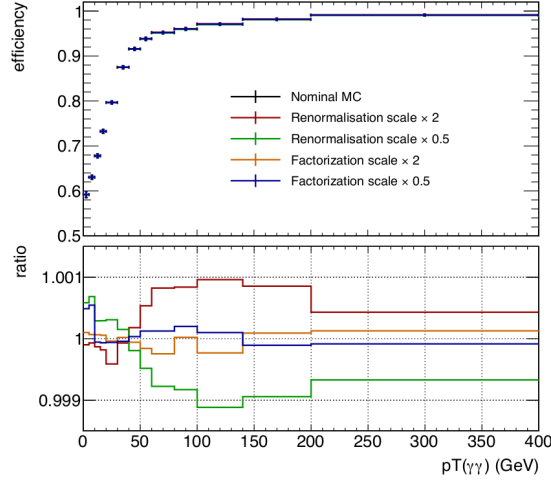


Figure 6.39: The efficiencies to find the correct vertex estimated on simulations with different QCD scales with their ratio to simulation with nominal scale. The simulation with renormalization scale twice more(less) the nominal one is shown in red(green), the simulation with the factorisation scale with two times more(less) with respect to nominal one is depicted in orange(blue).

The uncertainty from possible different Higgs boson p_T descriptions was estimated to range from $+0.5\%$ to -0.2% , as listed in Table 6.3.

Scale changed	Efficiency ratio
Renormalisation x 2	1.005 ± 0.002
Renormalisation x 0.5	0.998 ± 0.002
Factorisation x 2	1.000 ± 0.002
Factorisation x 0.5	1.004 ± 0.002

Table 6.3: Efficiency ratio with respect to the nominal MC simulation for different renormalization and factorisation scales.

The total systematic uncertainty coming from the vertex finding efficiency was conservatively computed to be 2%, summing in quadratures the different possible sources. Its impact on final analysis is negligible.

6.4 Conclusion

In this chapter the study for vertex identification algorithm for the $H \rightarrow \gamma\gamma$ analysis at 13 TeV was presented. The vertex identification efficiency is 80% with an average pileup of 23. The total systematic uncertainty is 2% and its impact on final analysis is negligible.

Chapter 7

Photon identification in CMS

Introduction

In proton-proton collisions the most important source of background to prompt photons, originating from hard interactions, are photons coming from neutral mesons decays.

Light neutral mesons, e.g. π^0 and η , are mainly produced in jets fragmentation. The background for prompt photons tends to be dominated by π^0 mesons that take an important fraction of the total jet p_T and are thus relatively isolated from jet activity in the detector. The photons from the decay of neutral pions are collimated, i.e with small opening angle, and are reconstructed as a single photon. These photons are called "fake" in this chapter. In the ECAL barrel, the minimum separation between two photons from the decay of a π^0 with $p_T = 15$ GeV is about the same as the ECAL crystal size.

The goal of photon identification is to reduce this background, exploiting the differences between prompt and fake photons in shower shapes in the ECAL and isolations against additional jet energy coming from fragmentation.

Two photon identification algorithms are used in CMS: one is based on an approach using selection requirements applied to a set of individual variables, i.e. cut-based approach, and the second one relies on a multivariate technique. The latter one is the subject of this chapter. The cut-based approach can be found in CMS public document [164].

This chapter is split into two parts: the general photon identification algorithm globally used by the CMS collaboration is described in section 7.1 and the photon identification algorithm adapted specifically for $H \rightarrow \gamma\gamma$ analysis in section 7.2. The first one is my service task contribution to the "Egamma" (EGM) CMS working group, and the second one is one of my contributions to the $H \rightarrow \gamma\gamma$ working group. For both photon identifications, the $H \rightarrow \gamma\gamma$ framework is used, and as they are very similar, I describe the methodology in details in the first part. In the second part I mention the differences and show specific results for the $H \rightarrow \gamma\gamma$ analysis.

The discriminating variables are discussed for the general algorithm in section 7.1.1

and the one for the $H \rightarrow \gamma\gamma$ analysis in section 7.2.1. Different photon preselections are used. They are described in sections 7.1.2 and in section 5.5 for $H \rightarrow \gamma\gamma$ photon identification. The results for both algorithms are shown in sections 7.1.3 and 7.2.2. Although I did not derive the corrections for simulation myself, they are shown for completeness in section 7.1.4 for the general algorithm. The validation procedure of the $H \rightarrow \gamma\gamma$ algorithm, using $Z \rightarrow e^+e^-$ events in data and simulations, is discussed in section 7.2.3.

7.1 Photon identification in CMS

The photon identification algorithm uses two types of variables which are distinctly different for prompt and fake photons: the electromagnetic shower shape variables and the isolation variables. This information is used to build an optimal discriminating variable.

7.1.1 Discriminating variables

As has been already mentioned above, the most photon-like jets result from fragmentation into neutral meson (π^0 or η) decaying into two collimated photons. Even if they cannot be well distinguished, such objects have wider shower shapes in the calorimeter on average than a single photon, which is demonstrated in Fig 7.1. This is true in particular along the η axis of the cluster, because the discrimination power along the ϕ axis is partially washed out by the magnetic field, which expands the electromagnetic cluster along the ϕ direction for converted photons. In addition, because π^0 or η are the results of jet fragmentation, there are additional charged and neutral particles present in the event. They tend to be produced close to the photon. This gives us a set of discriminating variables named "isolation". It consists of the sum of particular type of energies (electromagnetic, charged from charged particles, neutral from neutral particles) in a cone around the reconstructed photon.

Shower shape variables

The shower shapes variables used for photon identification are :

- R_9 - the energy sum of the 3×3 crystals centered on the most energetic crystal ("seed") in the "supercluster" divided by the raw energy of the supercluster. The raw energy is the initial sum of reconstructed energy deposits. A "supercluster" is formed by the photon reconstruction algorithm (as described in section 5.2) from clusters of the energy deposits in the ECAL.
- $S_4 = E_{2 \times 2} / E_{5 \times 5}$ - the ratio of the maximum energy in a 2×2 crystal grid and the energy in a 5×5 crystal grid centered on the seed crystal.

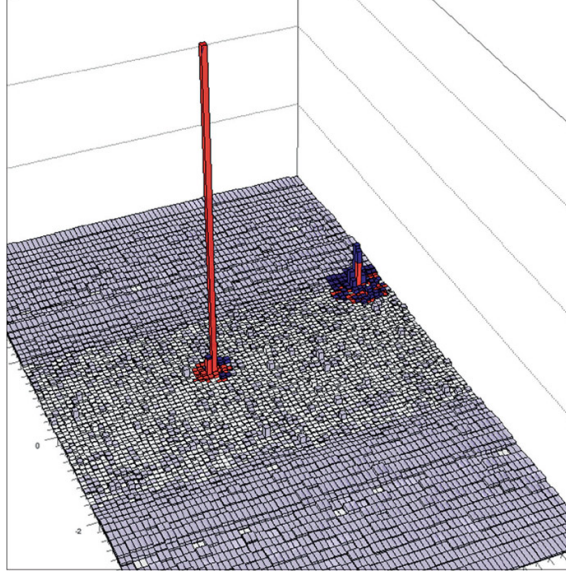


Figure 7.1: Energy deposit in the ECAL from simulation, showing a prompt photon (a narrow peak) and a non-prompt photon reconstructed from neutral meson decay products (a wider signature).

- $\sigma_{i\eta i\eta}$ - the energy weighted standard deviation of single crystal η (in crystal index) within the 5×5 crystals centered on the seed crystal:

$$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_{5 \times 5} (\eta^i - \bar{\eta}^i)^2 w_i}{\sum w_i}} \quad (7.1)$$

where $\bar{\eta} = \frac{\sum \eta_i w_i}{\sum w_i}$, $w_i = \max(0, 4.7 + \log \frac{E_i}{E_{5 \times 5}})$ is a per-crystal weight, and the sum is running over all crystals with energy deposit greater than 0.

- $cov_{i\eta i\Phi}$ - the covariance of the single crystal η and ϕ (in crystal index) values within the 5×5 crystals centered on the seed crystal:

$$\sigma_{i\eta i\phi} = \sqrt{\frac{\sum_{5 \times 5} (\eta^i - \bar{\eta}^i)(\phi_i - \bar{\phi}_i) w_i}{\sum w_i}} \quad (7.2)$$

- $SC \ \eta - width \ (\sigma_\eta)$ - the energy-weighted standard deviation of single crystal η in detector coordinate within the supercluster. The weight per-crystal is the ratio of the single crystal energy to the supercluster energy.
- $SC \ \Phi - width \ (\sigma_\phi)$ - the energy-weighted standard deviation of single crystal Φ in detector coordinate within the supercluster. The weight per-crystal is the ratio of

the single crystal energy to the supercluster energy.

- Preshower σ_{RR} (only for endcap) - the standard deviation of the shower spread in preshower x and y planes.

Isolation variables

The photon isolation is measured exploiting the information provided by the Particle Flow event reconstruction. The Particle Flow algorithm combines the information from the tracker, the calorimeters, and the muon detectors, and it aims to reconstruct the four-momenta of all particles in the event, classifying them as charged and neutral hadrons, photons, electrons and muons as described in section 3.7. The photon isolation variables are obtained by summing the transverse momenta of charged hadrons (charged isolation), photons (photon isolation), and neutral hadrons (neutral isolation), inside an isolation region of radius ΔR in the (η, ϕ) plane around the photon direction. It is computed as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ where $\Delta\eta$ and $\Delta\phi$ are the distances in η and ϕ respectively, between the photon direction and the direction of selected objects in the cone. When calculating the photon isolation, the Particle Flow photons falling in a pseudorapidity slice of size $\Delta\eta = 0.015$ are excluded from the sum. Similarly, when constructing the charged isolation, summing the transverse momenta of charged hadrons, a region of $\Delta R = 0.02$ is excluded to remove the conversions.

In the LHC Run II, the pileup is higher compared to the one in the Run I. PileUp Per Particle Identification (PUPPI) is one of the techniques developed in CMS to fight against pileup [191]. It takes Particle Flow objects (charged/neutral hadrons, photon and charged leptons) as input and it assigns a weight for each of them. This weight reflects the probability for the Particle Flow candidate to originate from the primary vertex of interest in the event. The performance of the isolation computed with PUPPI weights and its impact on the photon identification performance was checked.

The following isolation variables are used in photon identification algorithm:

- Charged isolation computed with respect to the primary hard vertex. Charged hadrons are reliably associated with the reconstructed primary vertices, however, the association of photons with a primary vertex is often less than certain, and an incorrect choice of the vertex gives a random isolation sum consistent with an isolated photon. The photon is used in the training, if it originates from the hard primary vertex (this was checked at generator level, so the chosen vertex is the "true" hard vertex) and the charged isolation is computed with respect to it. In the previous iteration of the photon identification, the vertex associated to the photon was not checked at generated level.
- Charged isolation computed with respect to the "worst" vertex. The "worst" vertex is the one which yields the largest isolation sum. This variable is useful because a

prompt photon is generally isolated with respect to any vertex and the fake photon is not.

- Photon isolation computed with respect to the primary hard vertex. The vertex is the same as in the charged isolation computation.

From a technical point of view, the isolation sums are computed with a CMS framework named "The Common Isolation Toolkit" (CITK). It is a flexible and common interface for calculating isolation quantities. It includes both the isolation computed with and without the PUPPI weights.

Additional variables

Additional variables are used to strengthen the discrimination between signal and background by accounting for the dependence of the shower-shape and isolation variables on the pileup present in the event:

- ρ - the median energy per unit area,
- $SC \ \eta$ - pseudorapidity of the supercluster corresponding to the reconstructed photon. It is computed from the pseudorapidity of the vector joining the point (0,0,0) to the reconstructed supercluster position,
- $SC \ E_{RAW}$ - the sum of crystal energy in the supercluster corresponding to the reconstructed photon,
- $ES \ E/SC \ E_{RAW}$ (only for endcap) - the ratio of energy deposited in the preshower and the sum of crystal energy in the supercluster.

7.1.2 Method

The photon identification is based on a multivariate analysis, employing a boosted decision tree (BDT) method with the "Gradient Boost" from the TMVA framework [173].

The BDT is trained using γ +jet 13 TeV simulated samples. The prompt photons, used as signal in the BDT training, and the non-prompt photons, used as background, are taken from a "double electromagnetic-enriched" γ + jet simulated samples, shown in section 5.1. To reproduce the total phase space, γ + jet process is generated in three different p_T and mass ranges, which are listed in section 5.1. Then the three sub-samples are weighted according to their cross-sections. Prompt photons and non-prompt photons are required to pass the preselection:

- $p_T > 18 \text{ GeV}$,

- ECAL barrel-endcap region is excluded by asking $|\eta_{SC}| < 1.442$ for the barrel and $1.566 < |\eta_{SC}| < 2.5$ for the endcap.

In addition, the prompt photons are required to match a generated photon which originates from a hard parton, while the non-prompt photons are all the remaining photons. To make the BDT training as independent as possible from the photon kinematics, the p_T and supercluster η distributions of the prompt photons (signal) are reweighted to match the ones of the non-prompt photons (background), named "2D reweighting". Half of the sample is used for training while the other half is used to assess the performance of the photon identification. Two trainings are performed separately for the barrel and the endcap.

7.1.3 Results

The p_T and η_{SC} distributions for fake photons and prompt ones before and after 2D reweighting are shown in Fig. 7.2 for barrel and endcap. The distributions of shower shape input variables for fake and prompt photons (before and after 2D reweighting) are shown in Fig. 7.3 and 7.4. As expected, fake photons have wider electromagnetic shower profile than prompt photons. The additional input variables distributions are shown in Fig. 7.5.

The isolation study using PUPPI weights. The charged photon isolation is depicted in Fig. 7.6 and photon isolation in Fig. 7.7, with and without the PUPPI weights, as well as the background efficiency versus signal efficiency ("ROC") curves, showing their performance. One can see that adding the PUPPI weights enhances the background rejection power in the photon isolation. There is no improvement for the charged isolation. This is expected as charged isolation is computed taking only charged hadrons from the primary vertex and the PUPPI weights are computed using the same charged hadrons.

Two photon identification trainings were performed using isolation with and without PUPPI weights. Their performances are shown in Fig. 7.8. One can see that after combining photon isolation with all input variables, the effect of the improvement coming from PUPPI weights decreases in the barrel and vanishes in the endcap.

Finally, the isolation computed without the PUPPI weights was chosen, given the small performance gain with respect to technical complications.

Detector issues impact on photon identification. Several detector issues were present in 2016, affecting the dataset, and resulting in data-simulation discrepancies for the input variables. The main issues were:

- ECAL pedestal¹ drift with time, impacting the energy reconstruction.

¹Pedestal is a response of a device when the signal is absent.

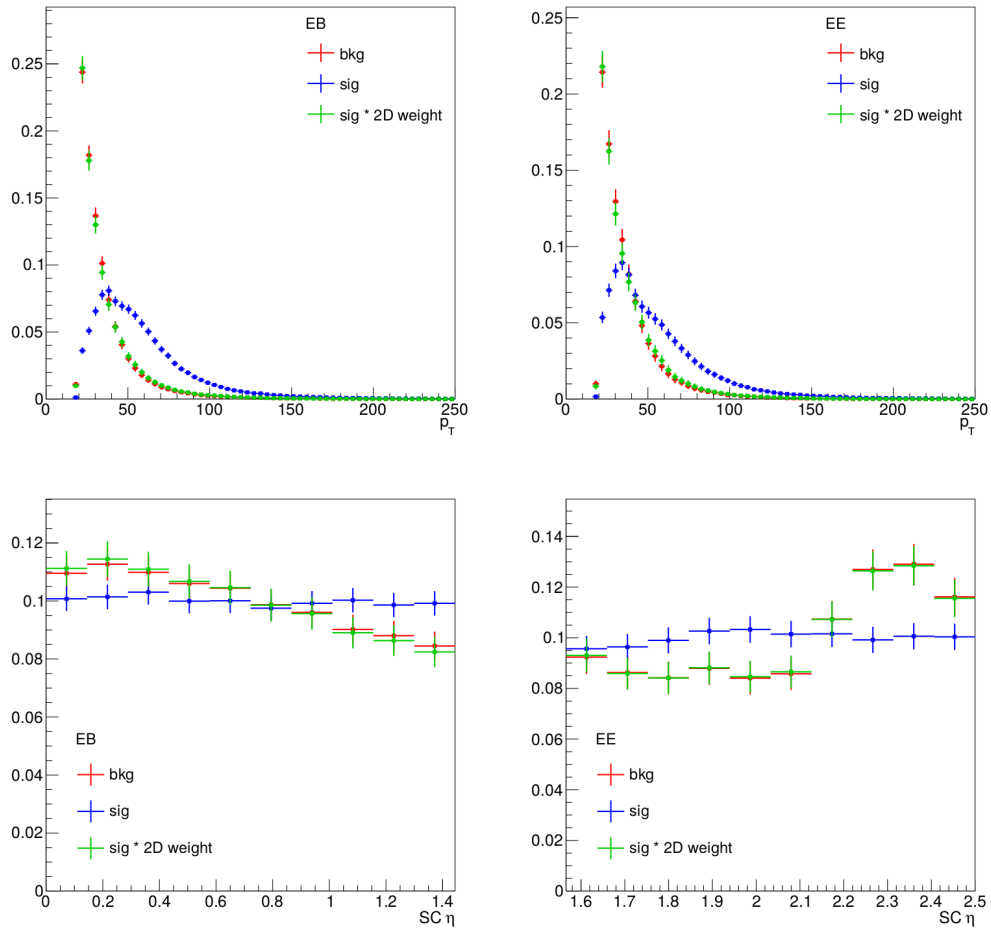


Figure 7.2: p_T and η_{SC} distributions for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green.

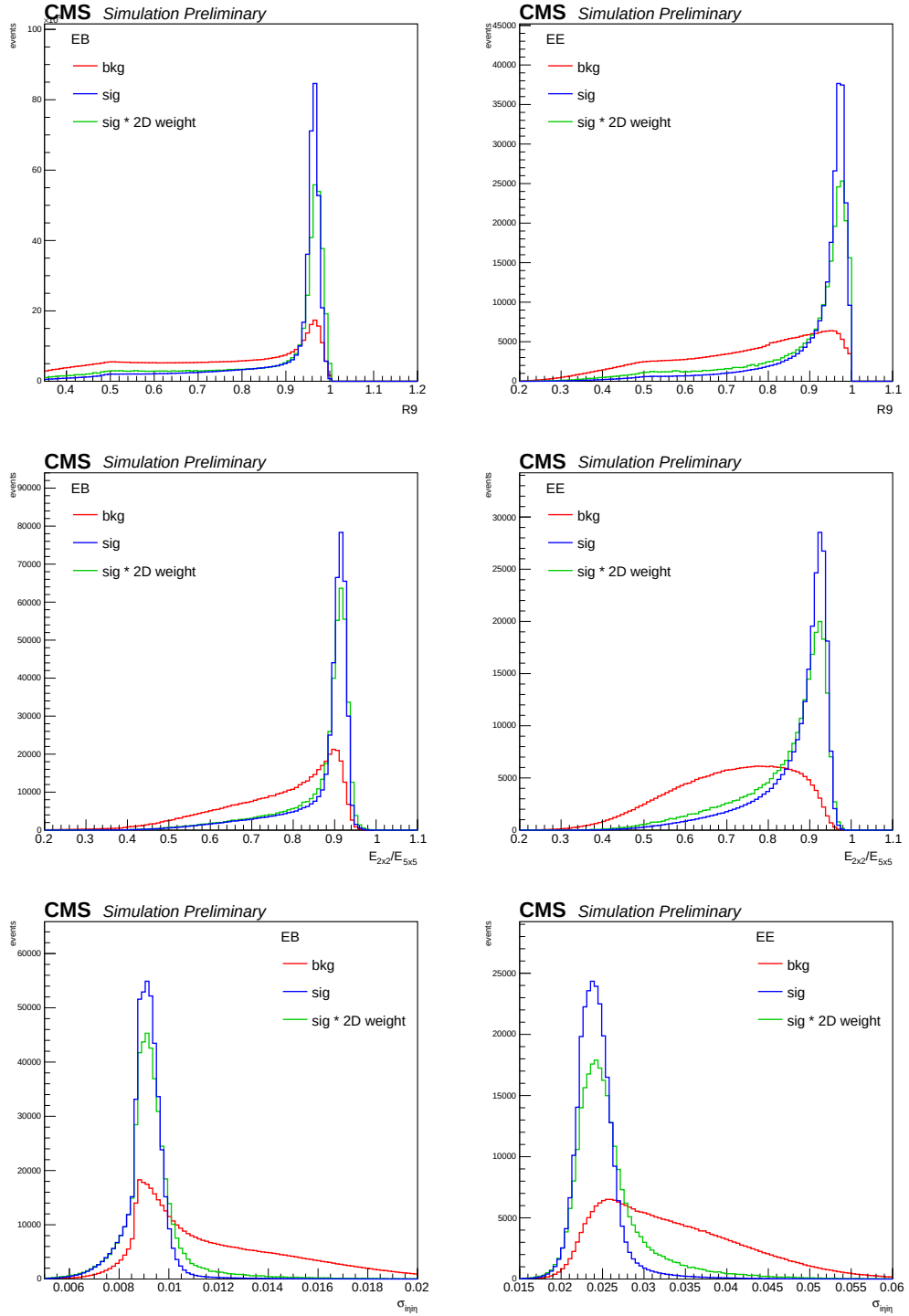


Figure 7.3: Shower shape input variables (R_9 (top), $S_4 = E_{2 \times 2} / E_{5 \times 5}$ (middle), $\sigma_{ii \eta}$ (bottom)) for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights - in green. Fake photons distributions are normalized to prompt photons ones.

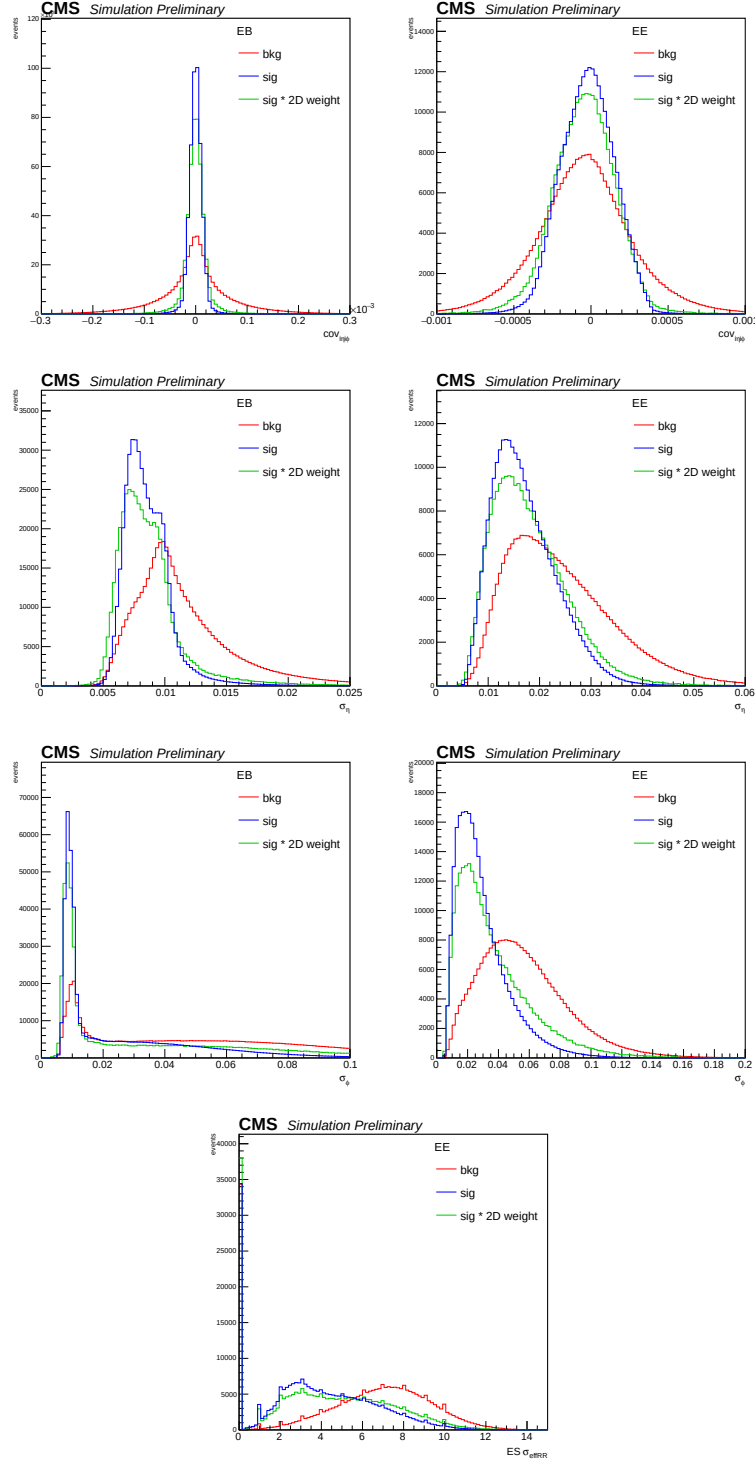


Figure 7.4: Shower shape input variables ($cov_{\eta\Phi}$ (top), $SC \eta$ -width and $SC \Phi$ -width (middle), Preshower σ_{RR} (bottom) for endcap only) for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green. Fake photons distributions are normalized to prompt photons ones.

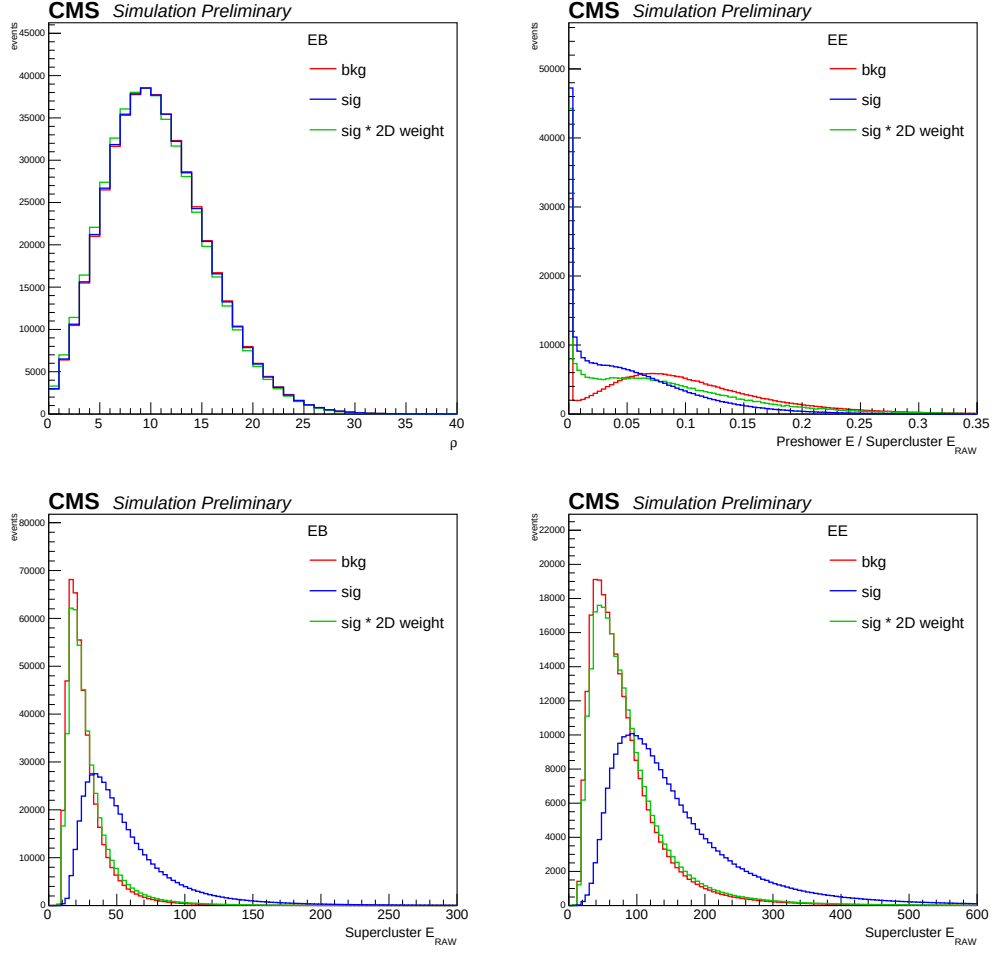


Figure 7.5: Additional input variables (ρ and $ES E/SC E_{RAW}$ (top), $SC E_{RAW}$ (bottom)) for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green. Fake photons distributions are normalized to prompt photons ones.

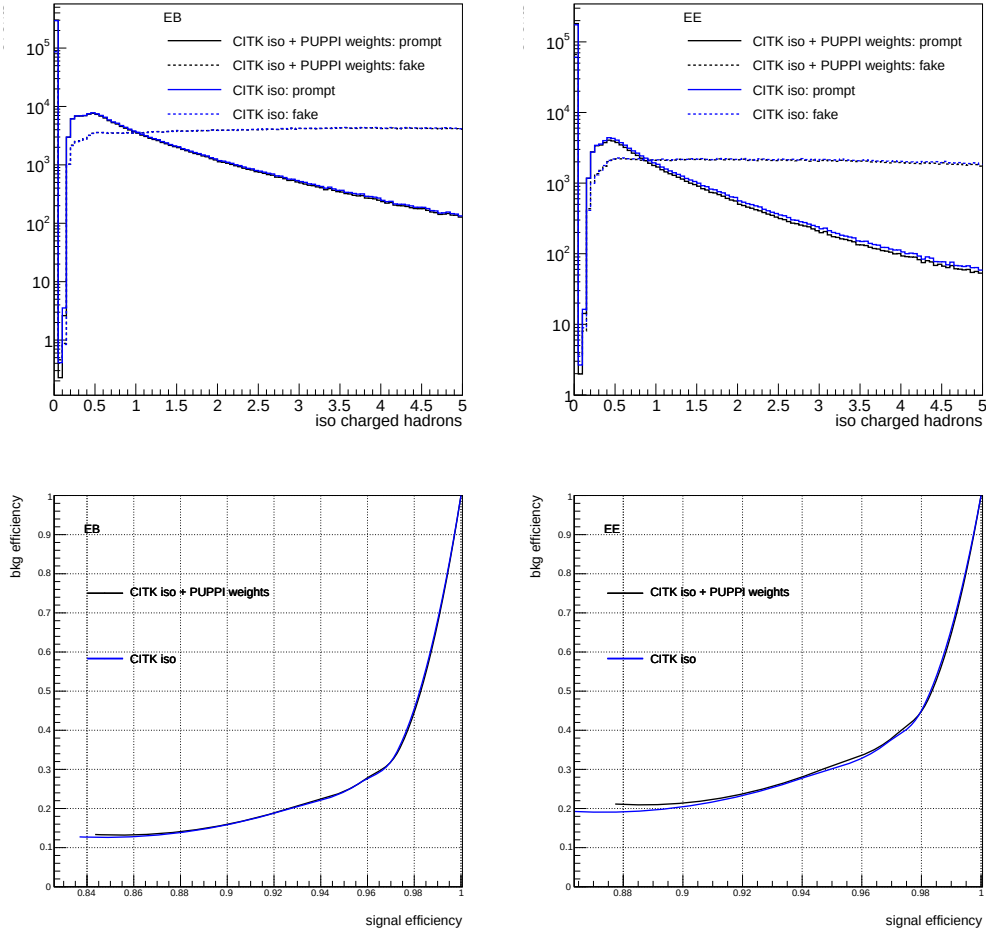


Figure 7.6: Top: charged isolation for the barrel (left) and for the endcap (right). Fake photons distributions are shown in dashed lines, prompt photons ones in solid lines. Fake photons distributions are normalized to prompt photons ones. The isolation computed without PUPPI weights is shown in blue, and with PUPPI weights in black. Bottom: ROC curve for charged isolation for the barrel (left) and for the endcap (right), computed without PUPPI weights in blue, and with PUPPI weights in black. As the PUPPI algorithm uses charged hadrons coming from primary vertex, there is no gain for charged isolation computation.

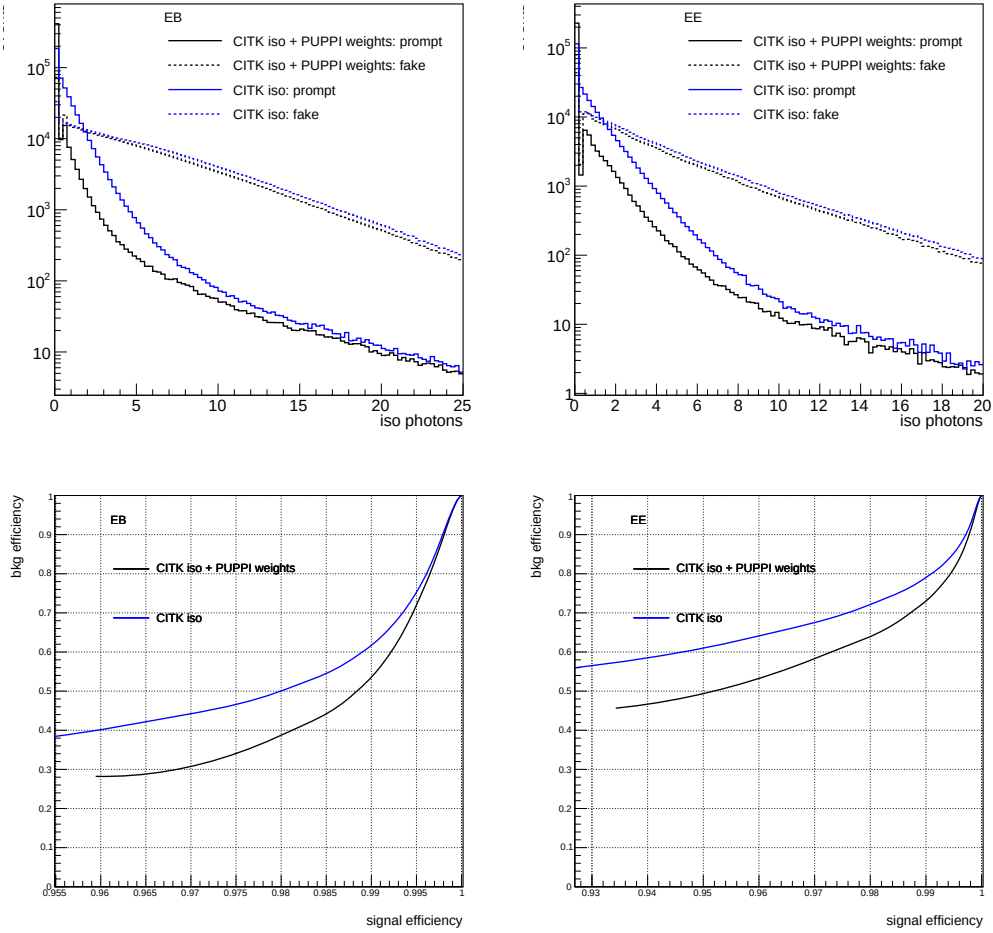


Figure 7.7: Top: photon isolation for the barrel (left) and for the endcap (right). Fake photons distributions are shown in dashed lines, prompt photons ones in solid lines. Fake photons distributions are normalized to prompt photons ones. the isolation computed with the CITK framework is shown in blue, without PUPPI weights, and with PUPPI weights in black. Bottom : ROC curve for charged isolation for the barrel (left) and for the endcap (right), computed without PUPPI weights in blue, and with PUPPI weights in black. Including PUPPI weights improves the background rejection by 30% in the barrel and by 16% in the endcap for signal efficiency $\sim 96\%$.

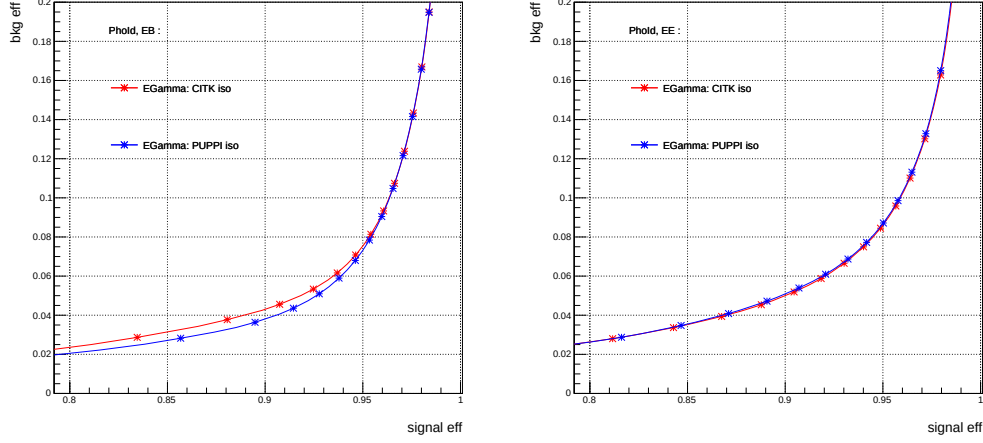


Figure 7.8: Photon identification performance with PUPPI weights included is shown in blue, without weights in red. The isolation with PUPPI weights gives the overall improvement in the barrel of $\sim 10\%$ in background rejection at 90% of signal efficiency, which vanishes at 95% and higher of signal efficiency. In the endcap no improvement was observed.

- ECAL crystals transparency loss in the endcaps, resulting in higher single channel noise with respect to simulation.

To fully resolve the disagreement, a new data reconstruction is necessary. It will use a better description of the detector conditions.

It was found that the most affected input variable in photon identification is the photon isolation in the endcap. The discrepancy in photon isolation leads to time-dependent corrections for the simulation. The corrections are computed as a ratio of photon identification efficiency in data and simulation. To solve this issue, it was decided to decrease the impact of the photon isolation variable in the algorithm, while keeping the best possible performance. Three different modifications for the photon identification training (in the endcap only) were tested:

1. No photon isolation in the training. The performance of this option is presented in Fig. 7.9 (in violet) and its ratio to the nominal photon identification (with photon isolation) is shown in the bottom part of the figure (in violet). One can see that the background rejection decreases by 12% - 5% for the same signal efficiency.
2. No photon isolation in the training and additional requirement. For practical reasons we define the corrected photon isolation as $\max(PhoIso - \rho \times EA - f(p_T), 0)$, where EA is the area of the isolation region weighted by a factor that takes into account the dependence of the pileup transverse energy density on pseudorapidity, and $f(p_T)$ equals $0.0034 \times p_T$ for the endcap. The effective areas have been determined in $\gamma + \text{jet}$

events. To make the algorithm blind to highly electromagnetic-isolated objects, which are poorly modelled, the final variable $\max(\text{PhoIso} - \rho \times EA - f(p_T), 2.5)$ is used. For this option we require corrected photon isolation to be equal to 2.5. It removes $\sim 28\%$ of background and $\sim 4\%$ of signal. In this case the photon isolation variable itself have no impact in the training, and its performance is shown in blue in Fig. 7.9. One can see that it improves the background rejection by 5%-7% with respect to the first option.

3. Corrected photon isolation in the training. We add the corrected photon isolation as in the previous option: $\max(\text{PhoIso} - \rho \times EA - f(p_T), 2.5)$ to the training. No additional cuts are required. The result is shown in green in Fig. 7.9, and the background rejection is similar to the one from the previous option.

Finally, option 3 was chosen, as it keeps better performance with respect to option 1 and it allows the choice of the working point with signal efficiency higher than 96%. In this way, decreasing the impact of the photon isolation in photon identification, the time-dependence in corrections for the simulation was mitigated.

The final photon identification training was included and validated in the official EGM framework. Its performance was compared to the previous iteration and shown in Fig. 7.10. Both trainings are evaluated on the same events with the pileup conditions ~ 17.5 . Background efficiency decreases by $\sim 50\%$ -10% in the barrel and by $\sim 30\%$ -10% in the endcap for the same signal efficiency. The reasons of the improvement are:

- the algorithm is trained requiring photons to originate from the true hard vertex, and this vertex is used for the isolation variables computation,
- the new photon identification algorithm is adapted for higher pileup than its previous version (~ 17.5 vs ~ 11) and it is trained on simulations which better reproduce the 2016 data taking conditions.

7.1.4 Corrections for simulation

One well established approach to measure the particle identification efficiency is "Tag and Probe" method [168] and it is used within the EGM group to compute the corrections for the simulation corresponding to photon identification. The "Tag and Probe" method uses the Z boson resonance to select an unbiased set of particles like electrons or muons. $Z \rightarrow e^+e^-$ events are selected with a trigger requiring at least 1 electron. One tag electron is required to match the trigger level electron and to pass a tight selection requirement. The other lepton from the Z boson decay, named probe lepton, is selected with very loose requirements. This allows to select a very pure sample of unbiased electrons. At the same time the invariant mass of the two leptons is required to be compatible with the

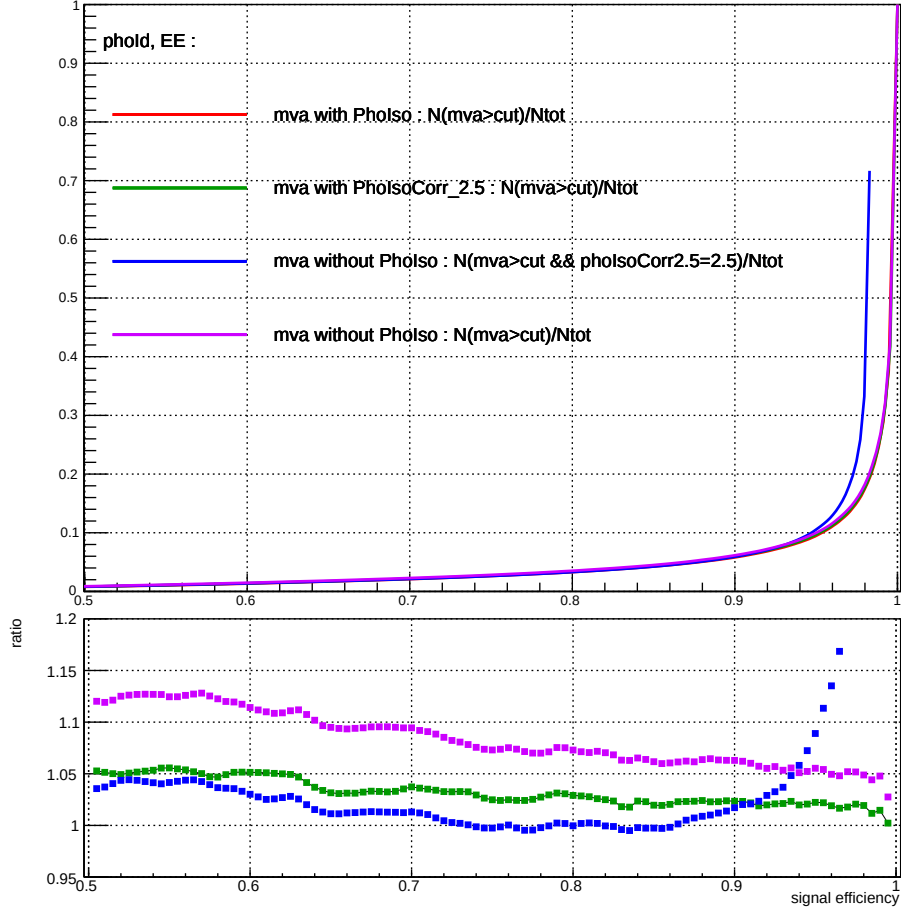


Figure 7.9: The ROC curves for the photon identification algorithm with different photon isolation configurations and their ratio with respect to the algorithm with the photon isolation included. The method without the photon isolation is shown in violet and its ratio also, the one with the corrected photon isolation in green with its ratio in blue. Removing the photon isolation from the method decreases background rejection by 12%-5% depending on signal efficiency, correcting the photon isolation improves this number by 7%-5%, and correcting the photon isolation and cutting at 2.5 gives similar performance.

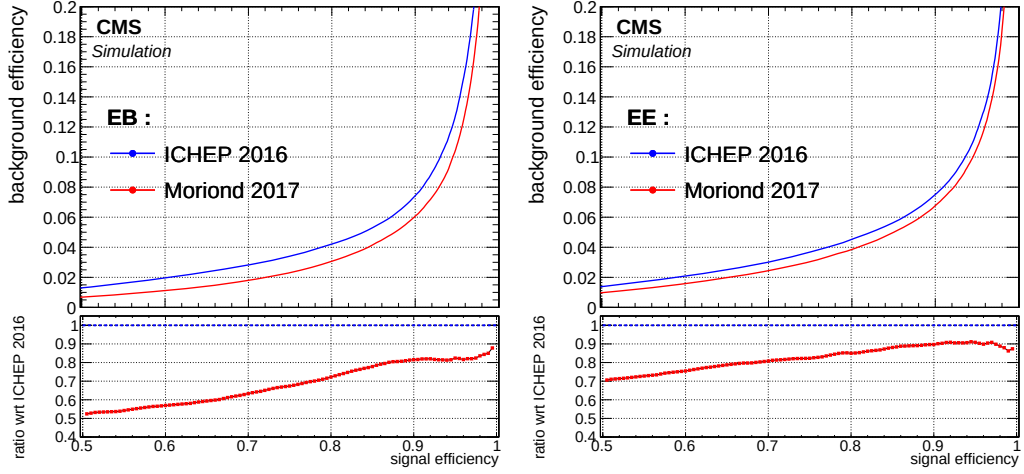


Figure 7.10: The ROC curves (background efficiency as function of signal efficiency) for the photon identification training used for ICHEP'16 (in blue), corresponding to the previous version of the optimisation, and the one used for Moriond'17, described in this chapter (in red). Left plot refers to the barrel, the right one to the endcap. The background rejection improves by $\sim 50\%$ - 10% in the barrel and by $\sim 30\%$ - 10% in the endcap with the new algorithm.

Z boson mass to further improve the purity of the sample. The identification criteria is tested on the probe lepton in both data and simulation events. This allows to correct the simulation in order to reproduce the efficiency of the selection criteria observed in data. The efficiency itself is measured by counting the number of "probe" particles that pass the desired selection criteria $\frac{P_{pass}}{P_{all}}$, where P_{pass} is the number of probes passing the selection criteria and P_{all} is the total number of probes counted using the Z boson resonance.

The efficiency observed in data as a function of p_T and η is shown in Fig. 7.11 for photon identification at a working point that is designed to have 90% of the signal efficiency ("Tight working point"). Final corrections for the simulation for the photon identification tight working point with their uncertainties are shown in Fig. 7.12. Since the photon identification efficiencies measured in data are different than those predicted by the simulation, the corrections have to be applied to simulated events so that this difference is taken into account. The corrections go up to 10% due to detector issues described above, as they affect not only photon isolation, but also shower shape variables. New reconstruction of data with problems fixed ² is going to improve the corrections.

²The improvements with new data reconstruction in alignment and calibrations include the update in pixel conditions for radiation damage corrections, the updated SiPixel alignment, the fix stability of the ECAL scale ("pedestals") and also HCAL corrections for radiation damage.

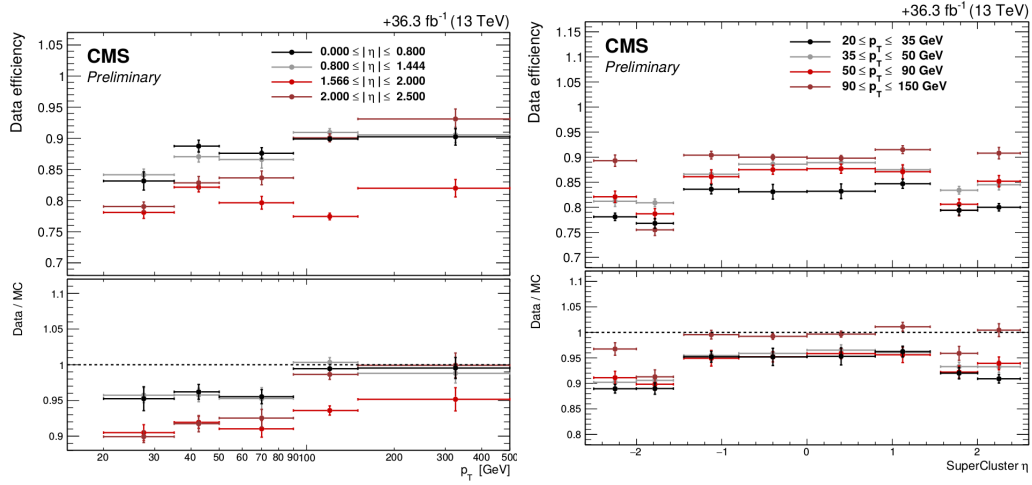


Figure 7.11: Data efficiency as a function of p_T (left) for different η ranges and η (right) for different p_T ranges. The data to simulation ratio is shown below. The efficiencies of the photon identification MVA tight working point are computed by the EGM group using the full 2016 dataset.

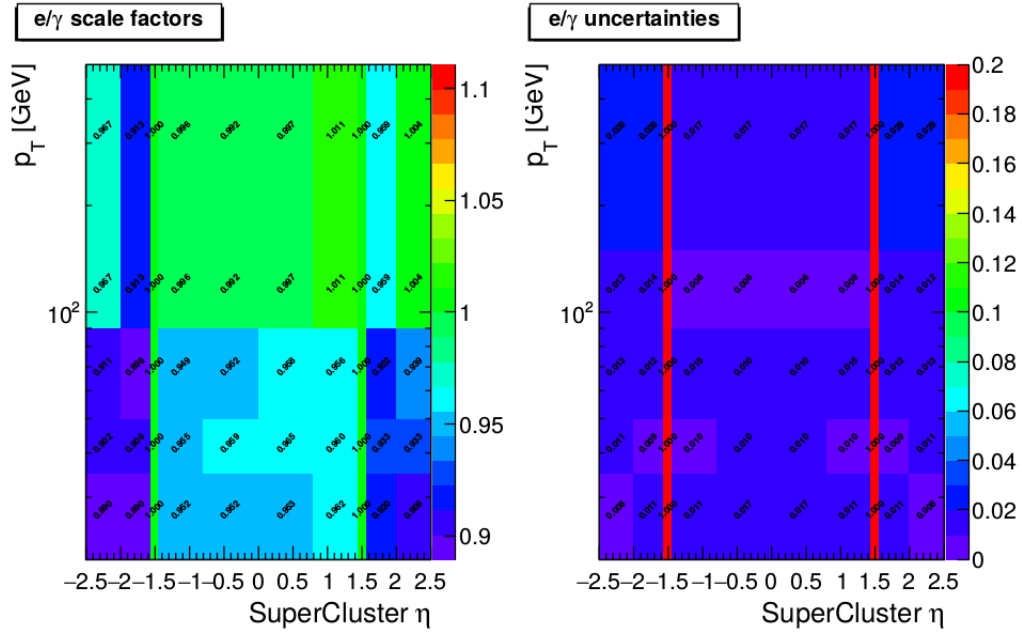


Figure 7.12: Corrections for simulation (left) and their uncertainties (right) of the photon identification MVA tight working point are computed by the EGM group using the full 2016 dataset.

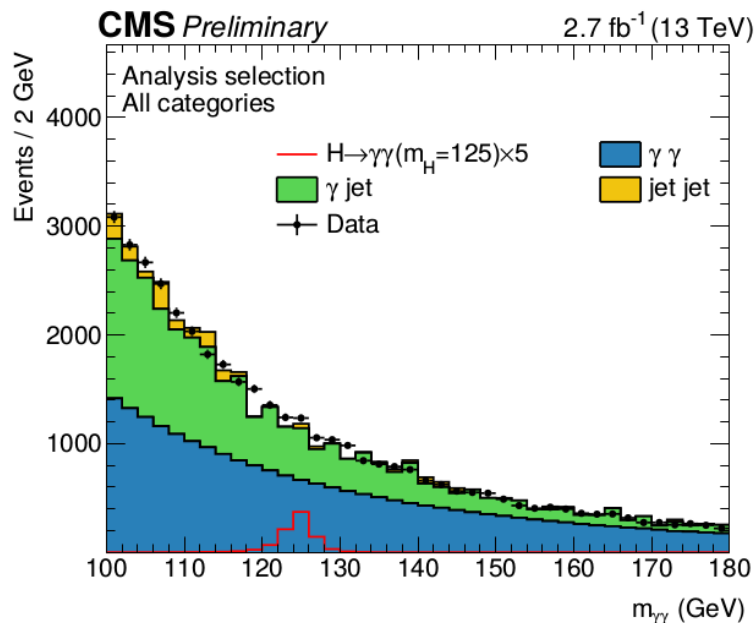


Figure 7.13: The diphoton invariant mass distribution for events passing the $H \rightarrow \gamma\gamma$ analysis selections. The background from simulation is presented in filled histogram and data in back points. The Higgs boson signal (scaled by 5) is shown in red. The data distribution is produced with the 2015 dataset of 2.7 fb^{-1} .

7.2 Photon identification in the $H \rightarrow \gamma\gamma$ analysis

The challenge in the $H \rightarrow \gamma\gamma$ channel is to identify a small peak in the diphoton mass distribution over a background which is at least an order of magnitude higher. The diphoton events include potential Higgs boson events, but mostly background ones. This can be seen from Fig. 7.13. There are "reducible" and "irreducible" background components, where the first one consist of dijet and $\gamma + \text{jet}$ events in which jets are misidentified as photons, and the second one of diphoton events in which the diphoton originates from the hard vertex. The goal of the photon identification in the $H \rightarrow \gamma\gamma$ analysis is to distinguish prompt photons from non-prompt photons, with the latter coming primarily from high momentum neutral mesons decaying to two photons, if both photons are included in the same supercluster and mimic a single photon. In order to do so, the Boosted Decision Tree (BDT) classifier (the same as in the previous section), is applied after the $H \rightarrow \gamma\gamma$ preselections described in section 5.5. The BDT output value obtained for a single photon is used as a photon identification input variable for the diphoton event classification, as mentioned in section 5.9. The BDT type of the classifier and input variables are the same as in the study for the EGM group described above.

7.2.1 Method

The method is exactly the same as described in section 7.1.2. The differences with respect to general the EGM photon identification are:

- the photon preselection is defined to match the $H \rightarrow \gamma\gamma$ trigger and described in details in section 5.5.
- While for EGM photon identification isolation variables are computed with respect to the true hard vertex, in the $H \rightarrow \gamma\gamma$ photon identification the $H \rightarrow \gamma\gamma$ vertex is used. It is selected by the multivariate algorithm described in chapter 6.

The difference with respect to the previous iteration (shown at ICHEP'16 conference) of the $H \rightarrow \gamma\gamma$ photon identification is:

- adding the E_{ES}/E_{RAW} variable in the endcap. It improves the background rejection by $\sim 8\%$ at signal efficiency 90% as shown in Fig. 7.14.

The input variables for the $H \rightarrow \gamma\gamma$ photon identification algorithm are shown in Fig. 7.15 - 7.18 for fake photons and prompt photons³ before and after 2D reweighting.

7.2.2 Results

The output of the $H \rightarrow \gamma\gamma$ photon identification is shown in Fig. 7.19 both for ICHEP'16 and LHCP'17 and it is used to estimate the performance of the algorithm. Both of them are constructed using the same events from γ +jet simulation, reweighted to match data pileup conditions (~ 23). Figure 7.20 shows the background efficiency versus the signal efficiency for the new photon identification (in green) and for the photon identification used for the ICHEP'16 analysis (in blue). Fig. 7.20 shows that the new photon identification performs slightly better than the one for ICHEP'16. A small improvement is coming from the fact that the new algorithm is adapted to higher pileup conditions. In the barrel all variables and BDT method parameters are the same. In the endcap new variable E_{ES}/E_{RAW} gives an improvement which compensates the loss from reducing the impact of photon isolation.

The photon identification output has been checked for the signal samples and for the $H \rightarrow \gamma\gamma$ backgrounds: γ + jet, dijet, and diphoton+jets. Fig. 7.21 shows the photon identification BDT score of the lower-scoring photon in diphoton pairs with an invariant mass, $m_{\gamma\gamma}$, in the range $100 < m_{\gamma\gamma} < 180$ GeV, for events passing the preselection, in data and for simulated background events. The distribution of the sum of all the simulated background events is normalized to data keeping the relative ratio of the single background components. The sum of all the backgrounds is consistent with data, even if some discrepancies are visible in the high-score region. These discrepancies are taken into account in the treatment of systematics uncertainties, presented in section 5.13.4.

³The definitions of prompt and fake photons is the same as in EGM photon identification.

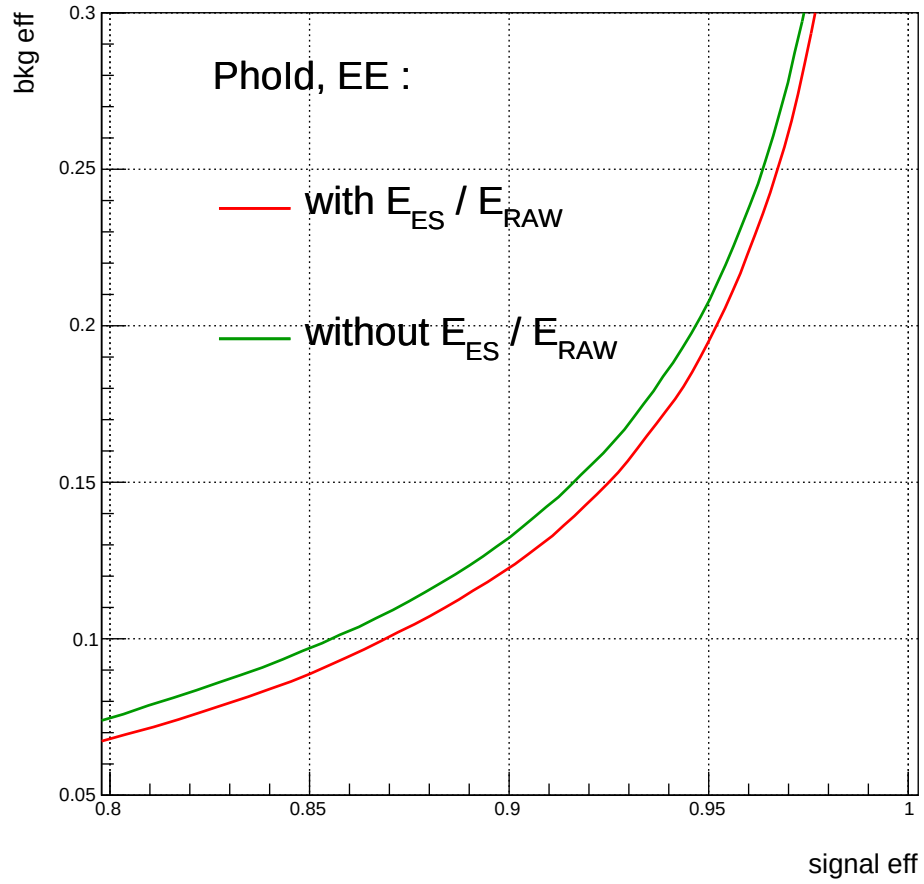


Figure 7.14: Performance of the MVA in the endcap trained with (in red) and without (in green) the E_{ES}/E_{RAW} variable. The improvement in background rejection is about 8% at signal efficiency of 90% and 95%.

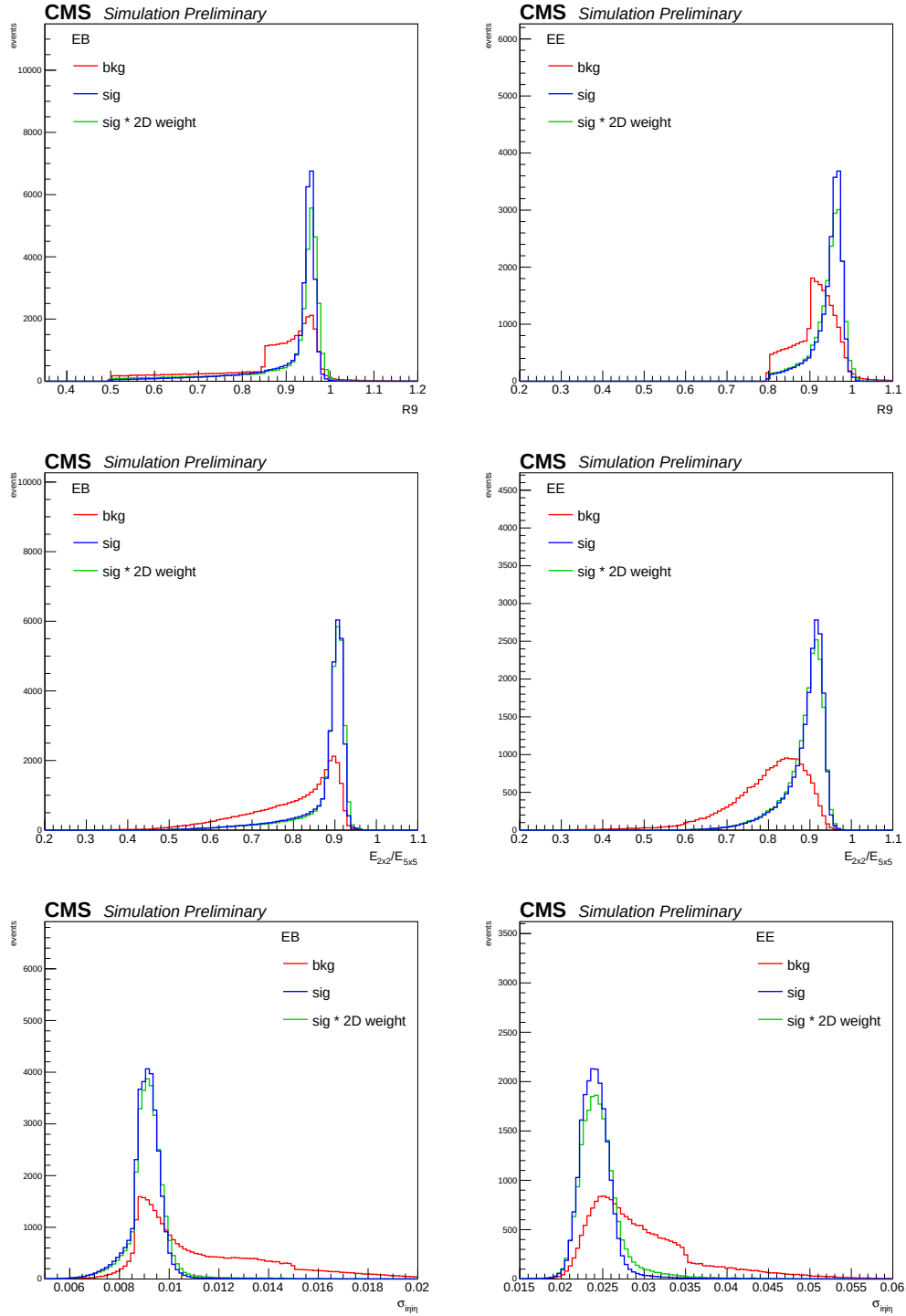


Figure 7.15: The shower shape input variables (R_9 (top), $S_4 = E_{2 \times 2}/E_{5 \times 5}$ (middle), $\sigma_{\eta\eta}$ (bottom)) for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights - in green. Fake photons distributions are normalized to the prompt photons ones.

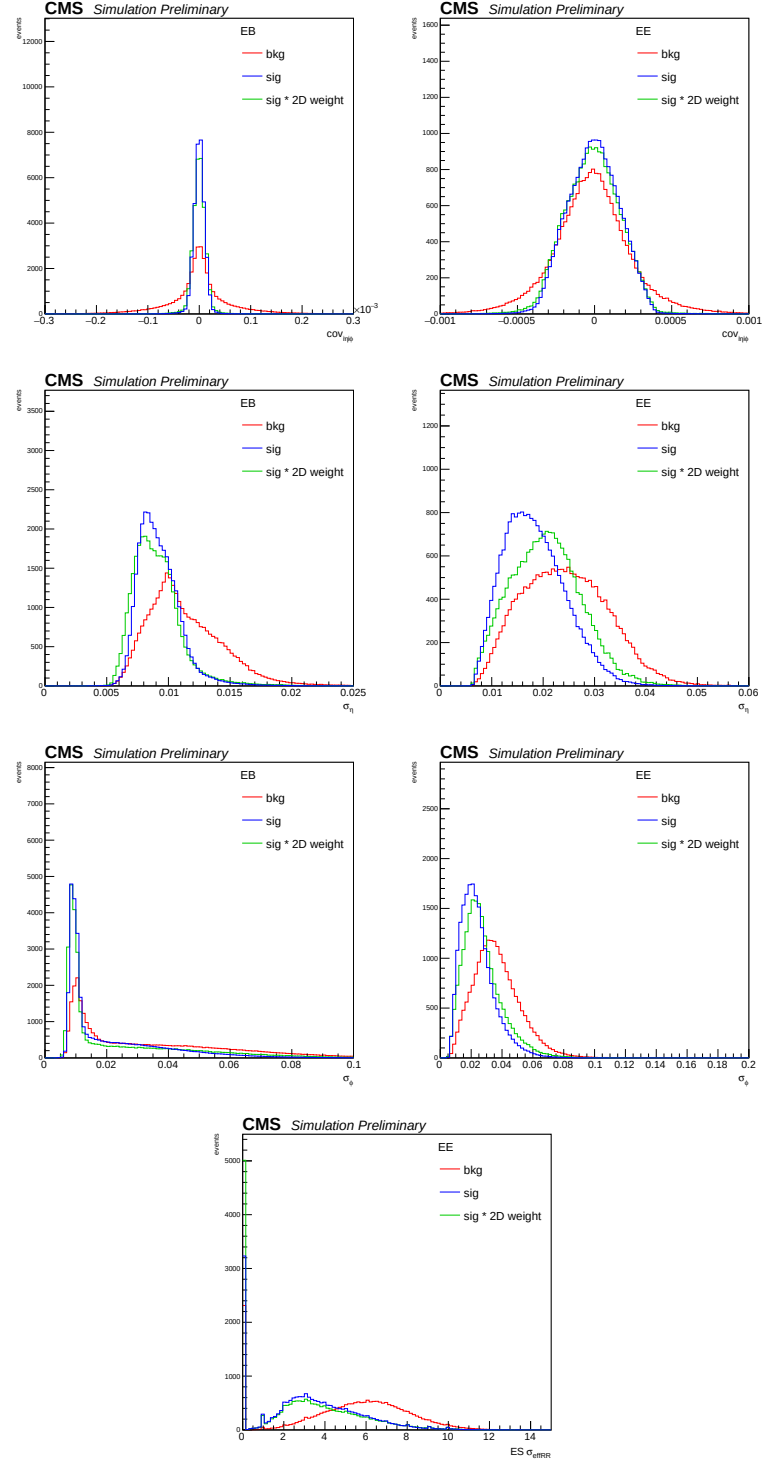


Figure 7.16: The shower shape input variables ($cov_{\eta i \Phi}$ (top), $SC \eta - width$ and $SC \Phi - width$ (middle), Preshower σ_{RR} (bottom) for endcap only) for the barrel (left) and for the endcap (right). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green. Fake photons distributions are normalized to the prompt photons ones.

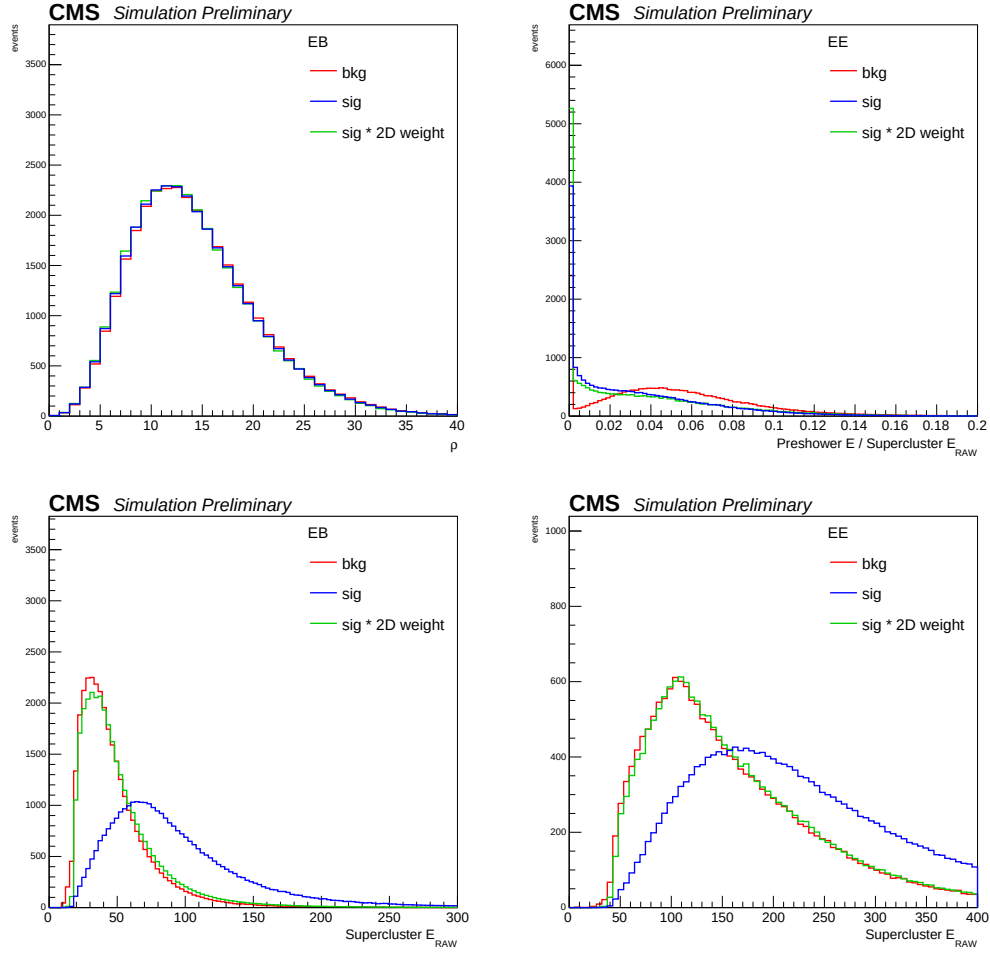


Figure 7.17: Additional input variables (ρ (top, left), $ES\ E/SC\ E_{RAW}$ (top, right), $SC\ E_{RAW}$ (bottom, for the barrel on the left, and for the endcap on the right)). Fake photons distributions are shown in red, prompt photons ones in blue, and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green. Fake photons distributions are normalized to the prompt photons ones.

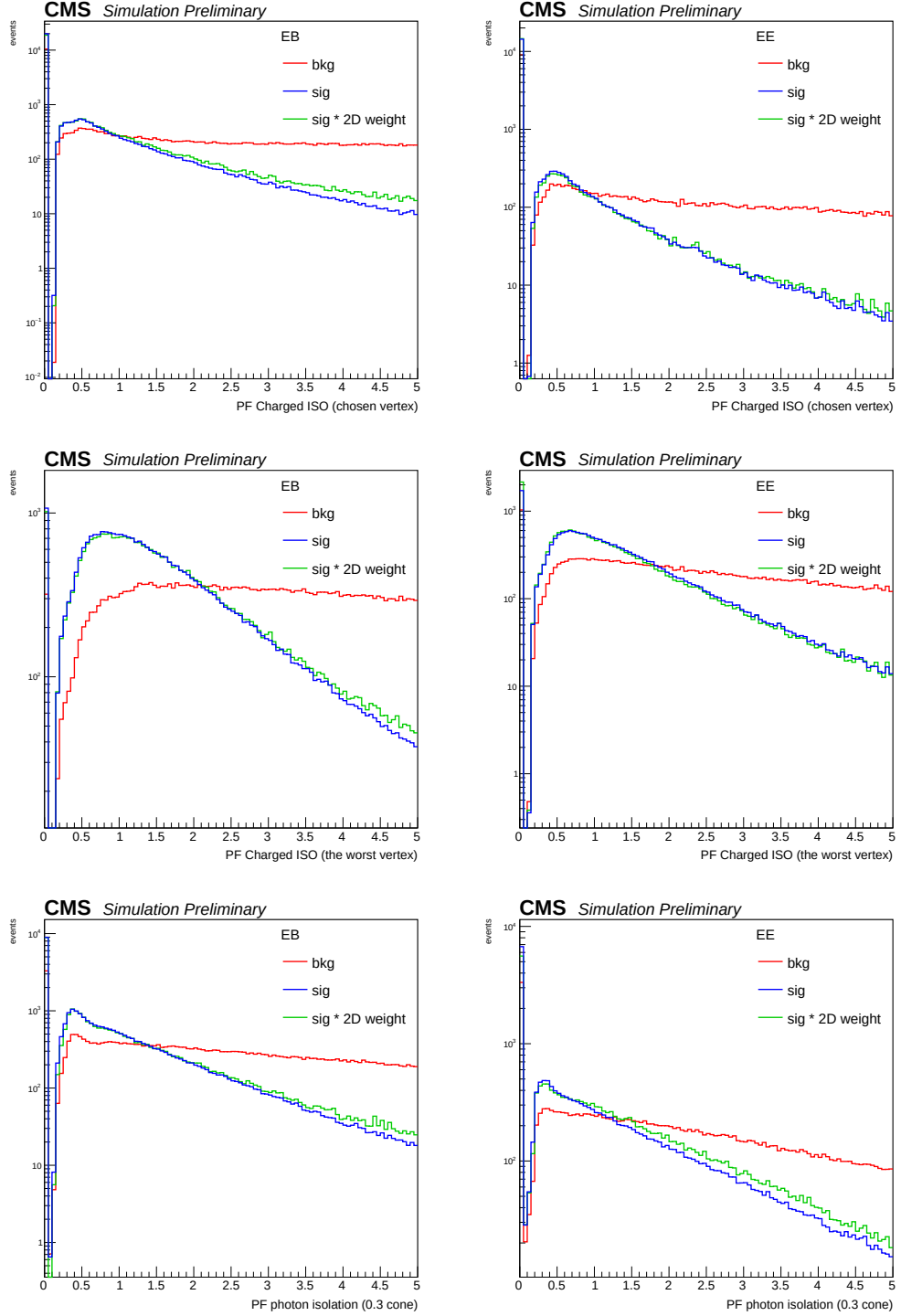


Figure 7.18: Isolation input variables (charged isolation (top), charged isolation with the largest sum (middle), photon isolation (bottom)) for the barrel (left) and for the endcap (right), fake photons distributions are shown in red, prompt photons ones in blue and prompt photons ones reweighted with 2D $p_T - \eta_{SC}$ weights in green. Fake photons distributions are normalized to the prompt photons ones.

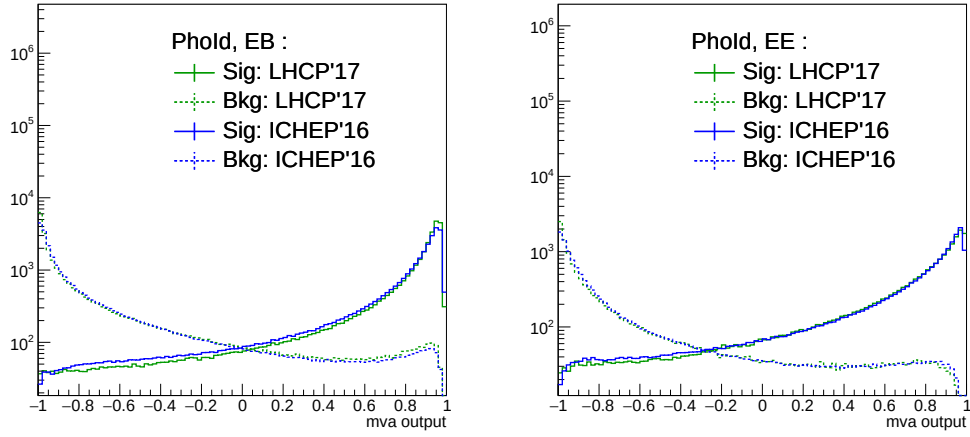


Figure 7.19: Photon identification output for both the new LHCP'17 training (in green) and the ICHEP'16 training (in blue). The left plot refers to the barrel, the right plot to the endcap.

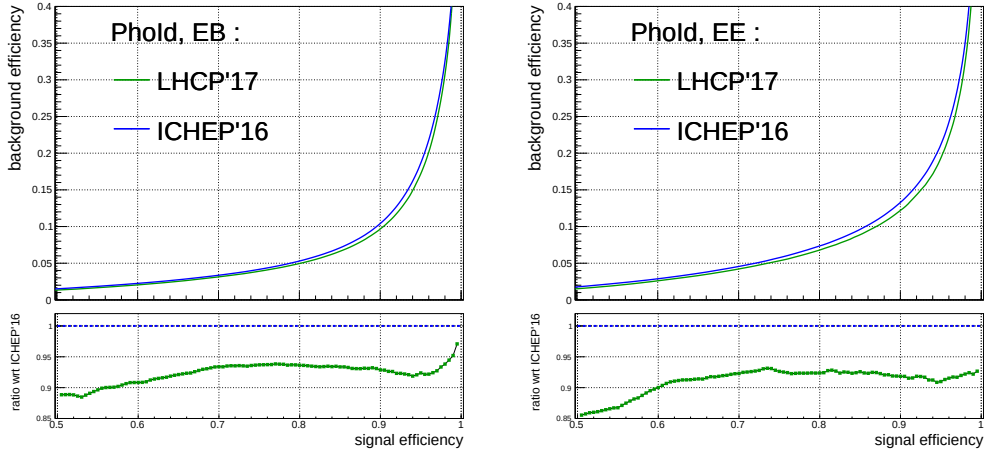


Figure 7.20: ROC curve of the background efficiency as a function of the signal efficiency of the training, both for the new LHCP'17 training (in green) and for the ICHEP'16 training (in blue). The left plot refers to the barrel, and the right plot to the endcap.

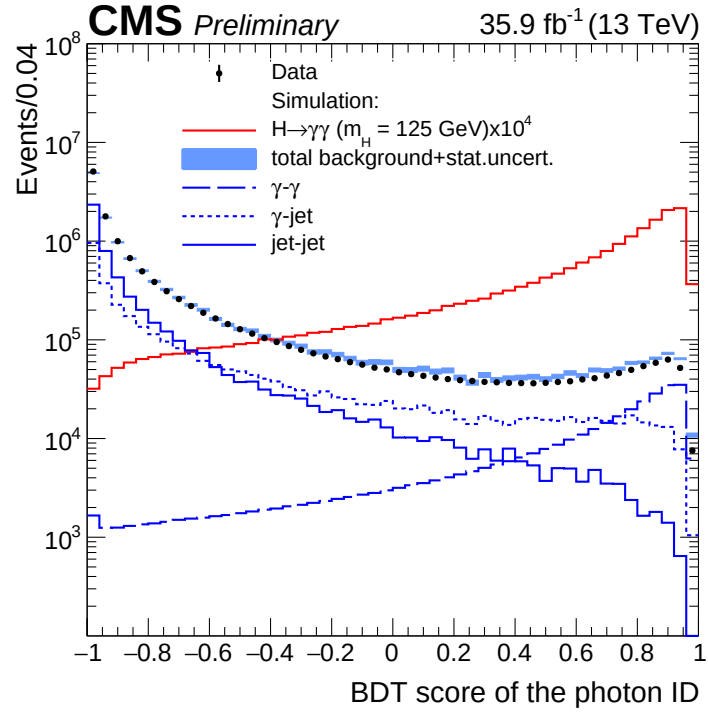


Figure 7.21: Photon identification BDT score of the lower-scoring photon of diphoton pairs with an invariant mass in the range $100 < m_{\gamma\gamma} < 180$ GeV, for events passing the preselection in the 13 TeV dataset (points), and for simulated background events (cyan histogram). Histograms are also shown for different components of the simulated background, in which there are either two, one, or zero prompt candidate photons. The distribution of the sum of all the simulated background events is scaled to data preserving the relative ratio of the single components, generated at leading order. The red histogram corresponds to simulated Higgs boson signal events.

Efficiencies relative to the photon preselection, described in section 5.5, are shown in Fig. 7.22 as functions of p_T , η and number of vertices in the event. A working point, with signal efficiency 95% for the barrel and 90% for the endcap, is used only to illustrate the performance of the photon identification. In the analysis the selection at -0.9 is applied on the output as an additional preselection, it guarantees 99% efficiency on the signal photons and rejects $\sim 49\%$ of the background ones. The photon identification output values for each photon are used as an input for a multivariate event classifier, described in section 5.9.

7.2.3 Data-simulation comparison

The shower shape variables are sensitive to the accuracy of the simulation of the detector response. A validation on data is performed for all input variables and also for the output one, since it is necessary to know how well the variables are modelled in the simulation. It is used to compute the systematic uncertainty on photon identification. This validation is done using $Z \rightarrow e^+e^-$ events, where both electrons are reconstructed as photons. Even if the Z boson differs from Higgs boson, a comparison for $Z \rightarrow e^+e^-$ events in data and simulation gives a good estimate of the expected data-simulation agreement.

The $Z \rightarrow e^+e^-$ events have been selected with an inverted electron veto cut applied as a part of the preselection cuts. In addition the leading and sub-leading "photons" are required to have $p_T > 30$ and 25 GeV respectively. Each "diphoton" pair is required to have invariant mass between 70 and 110 GeV. Single electron HLT trigger requiring the electron to pass tight identification criteria (listed in section 5.11.1) and $p_T > 27$ GeV has been applied both for data and simulation.

Due to the detector issues described above several variables were affected, mostly R_9 , S_4 , SC η -width, $\sigma_{in\eta}$ and photon isolation (both in barrel and endcap). They are shown in Fig. 7.23 and 7.24.

To improve the data/simulation agreement in shower shape variables, they were corrected using a histogram remapping with a sample of probes from the Z boson events. The photon isolation was corrected using "stochastic isolation correction" method. This method adds randomly some energy into isolation cone to reproduce the energy obtained in data. The variables after corrections application are shown in Fig. 7.25 and 7.26, and they show better agreement between data and simulation. The final photon identification training is done with the corrections applied to simulation.

A systematic uncertainty is introduced in the analysis to cover the differences between data and simulation. A shift on the photon identification output value in simulation of ± 0.03 is performed. In spite of that, a small discrepancy was observed in the low score tail. Since one of the main sources of systematic uncertainty on the diphoton MVA output comes from photon identification, the imperfect coverage of the photon identification tail is reflected in the diphoton MVA tail. In order to take this into account, it was decided to combine the ± 0.03 shift with a linear correction that expands uncertainty at low MVA

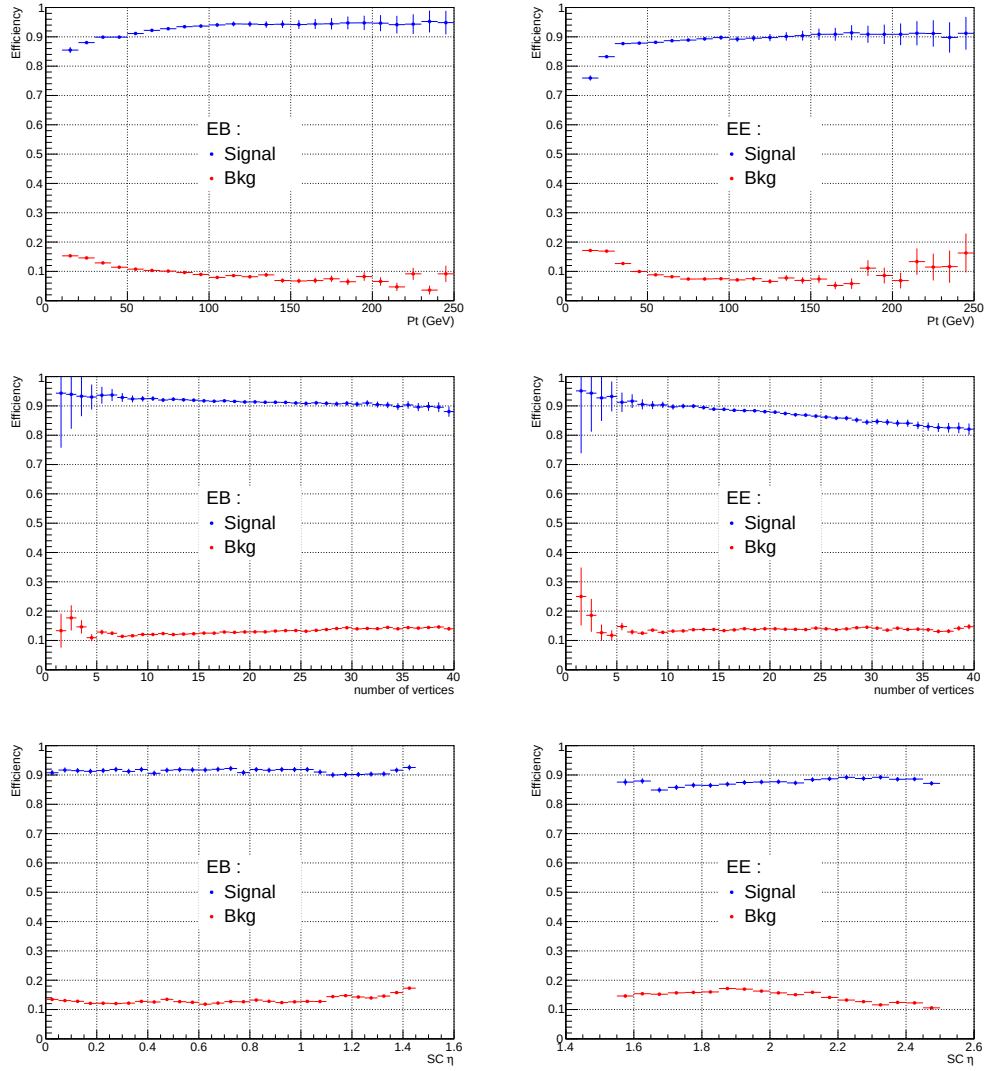


Figure 7.22: Signal and background efficiency versus p_T (top), supercluster η (middle) and number of vertices (bottom). The left plots refer to the barrel, and the right plots to the endcap.

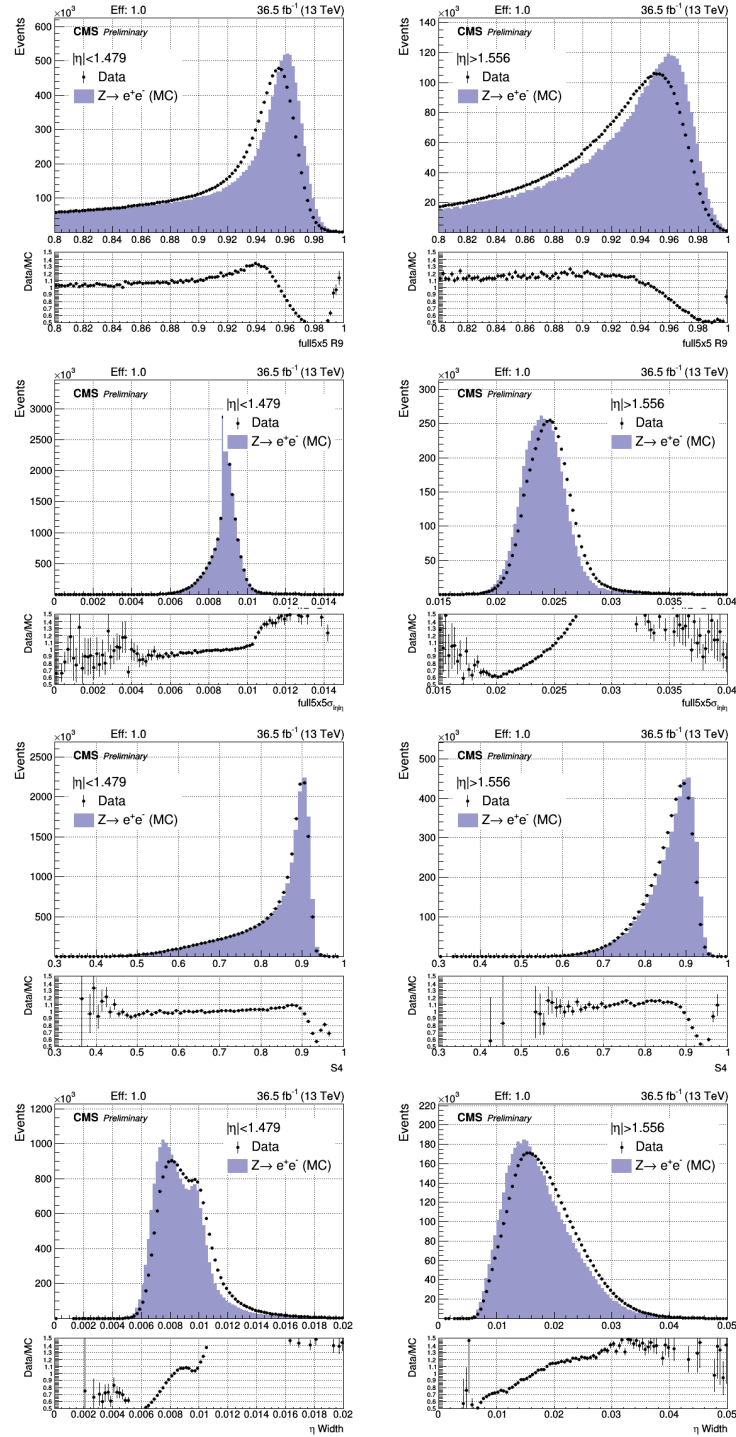


Figure 7.23: Data/MC comparison of the distributions of the shower-shape variables before applying the corrections: R_9 , $\sigma_{i\eta i\eta}$, S_4 and $SC \eta - width$ for both ECAL barrel (left) and endcap (right).

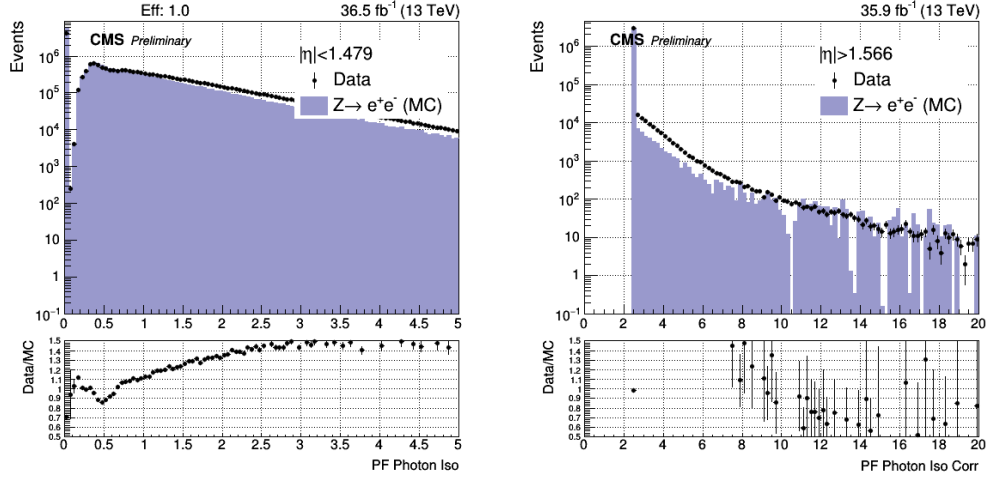


Figure 7.24: Data/MC comparison distribution for photon isolation for both ECAL barrel (left) and endcap (right) before applying the corrections.

scores. The final $H \rightarrow \gamma\gamma$ photon identification output is shown in Fig. 7.27 after all corrections with systematic band, which covers data/simulation difference.

7.3 Conclusion

In this chapter, the new photon identification algorithm of CMS is presented. The performance of the general EGM algorithm is improved by 30%-10% in the barrel and by 15%-10% in the endcaps with respect to its previous version. Several issues of the CMS ECAL manifested themselves during data taking in 2016. Most of them were mitigated both in the EGM and the $H \rightarrow \gamma\gamma$ photon identification algorithms. New data reconstruction is necessary to fully resolve these issues.

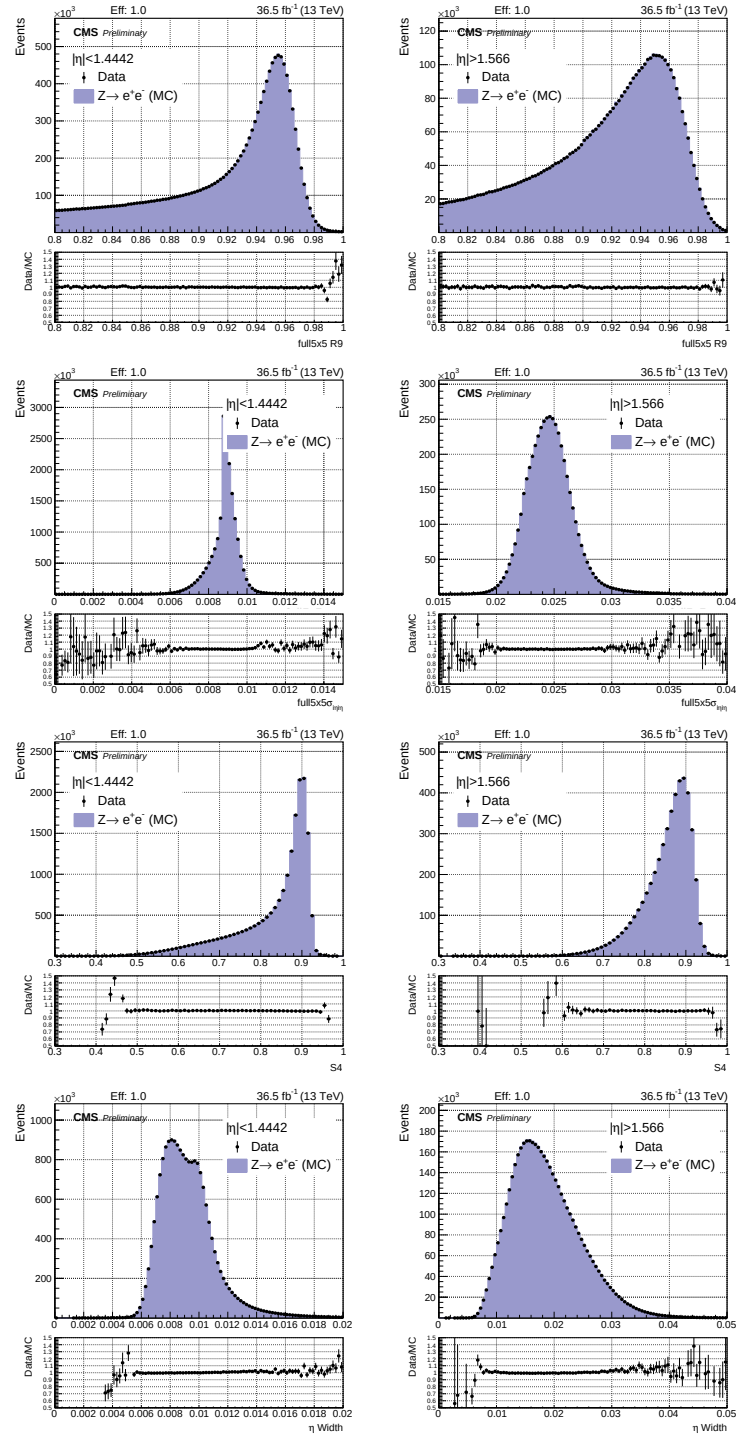


Figure 7.25: Data/MC comparison of the distributions of the shower-shape variables after applying the corrections: R_9 , $\sigma_{i\eta\eta}$, S_4 and SC η - width for both ECAL barrel (left) and endcap (right).

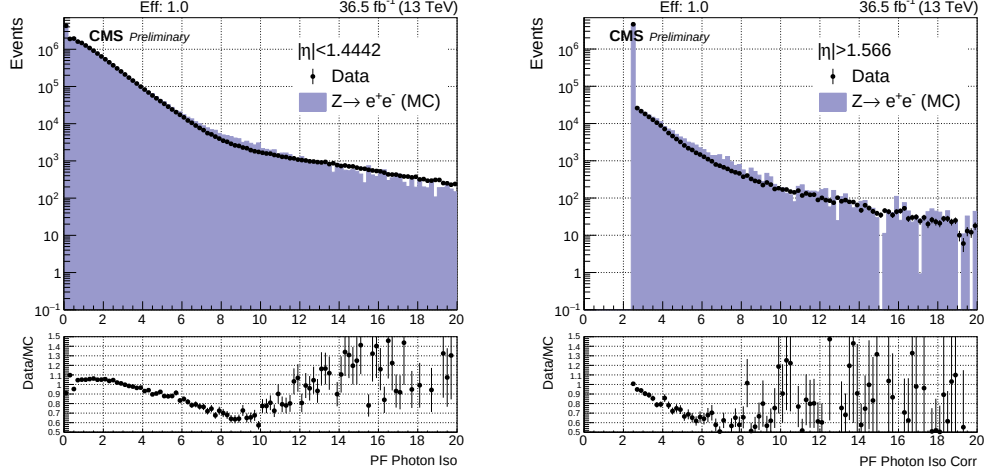


Figure 7.26: Data/MC comparison distribution for photon isolation for both ECAL barrel (left) and endcap (right) after applying the corrections.

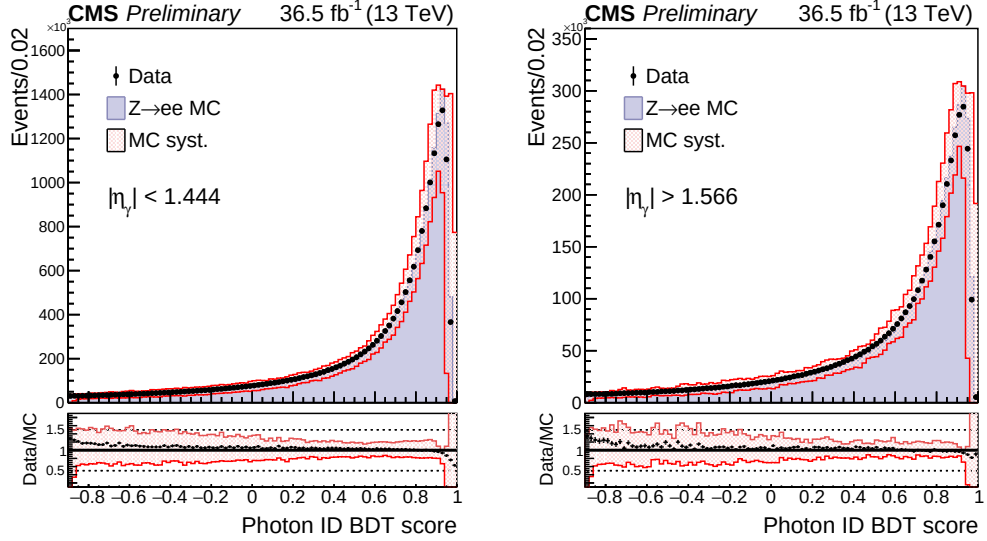


Figure 7.27: Photon identification BDT score for electrons from $Z \rightarrow e^+e^-$ reconstructed as photons in the ECAL barrel (left) and endcap (right). The distributions in data are compared to those in simulated Drell-Yan events. The shaded bands correspond to a shift of ± 0.03 applied to the score in simulated events. The corresponding ratios of data to simulation are shown in the bottom panels.

Chapter 8

The $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis

Introduction

In this chapter, I present the exclusive search for the Higgs boson production in association with a pair of top quark-antiquark ($t\bar{t}H$).

Since the discovery of a new boson at LHC in 2012, experimental studies have focused on determining the consistency of its properties with the expectations for the standard model (SM) Higgs boson. One striking feature of the SM Higgs boson is its strong coupling to the top quark compared to the other SM fermions. Based on the top quark large mass, the top-quark Yukawa coupling is expected to be of order one. The coupling between the top quark and the Higgs boson can be measured directly using the rate of $t\bar{t}H$ production. Several new physics scenarios [79–81] predict the existence of heavy top-quark partners, that would decay into a top quark and a Higgs boson. Observation of a significant deviation in the $t\bar{t}H$ production rate with respect to the SM prediction would be an indirect indication of unknown phenomena.

$H \rightarrow \gamma\gamma$ decay mode has a very clean signature in the detector (two isolated high energy photons) as well as an excellent mass resolution of $\sim 1\%$. The $t\bar{t}H$ with a subsequent decay $H \rightarrow \gamma\gamma$ search is difficult, due to $t\bar{t}H$ small cross-section of ~ 507 fb at a center of mass energy 13 TeV and $H \rightarrow \gamma\gamma$ branching ratio of $\sim 0.2\%$.

In the $t\bar{t}H, H \rightarrow \gamma\gamma$ search, events are classified based on objects arising from the top quarks decays. The top quark decays with $\sim 100\%$ probability to a W boson and a b quark, and the W decays either to a lepton and a neutrino, or a pair of quarks. Two categories of events are defined in order to increase the sensitivity, a hadronic category, where both W decay hadronically, and a leptonic category, where at least one of the W decays leptonically. The main background contributions to the $t\bar{t}H, H \rightarrow \gamma\gamma$ process comes from the production of top quarks and either real or misidentified photons in the final state, as well as the production of high- p_T photons in association with many jets, including heavy-flavour jets.

This analysis is part of the inclusive $H \rightarrow \gamma\gamma$ analysis. In this chapter the strategy employed for the ICHEP'16 conference is described in section 8.2 including its results. Several studies were performed after ICHEP'16 to improve the sensitivity, and they are shown in section 8.3 together with the improvements they bring with respect to the previous strategy.

8.1 Strategy for $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis

Photons originating from the Higgs boson are required to pass the preselection criteria described in section 5.5 and identification criteria shown in chapter 7. Events with a selected pair of photons are further required to pass additional criteria to be classified in $t\bar{t}H$ categories. Both leptonic and hadronic categories require the presence of jets and b-jets. For the leptonic category at least one lepton (electron or muon) is required and for the hadronic category no leptons in the event are specifically required.

The background in the $H \rightarrow \gamma\gamma$ analysis is estimated by fitting the data, so it is useful to test the background modelling in an independent control sample. The control sample is constructed using data events that have a pair of photons where one of the photons is required to pass the preselection and photon identification criteria, while for the other photon, the photon identification criteria is inverted and the preselection criteria are not applied. The kinematic properties of the control sample are different from the ones of the signal region, so a two dimensional reweighting is performed in the control sample based on the p_T and η of the photons to match those in the signal region. A control sample with similar kinematic properties as the data, yet statistically independent is thus obtained. The diphoton mass distributions after the reweighting are shown in Fig. 8.1, together with the distributions for the data sidebands (data events with $100 < m_{\gamma\gamma} < 115$ GeV and $135 < m_{\gamma\gamma} < 180$ GeV) with different jet selections.

8.2 $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis for ICHEP'16

The analysis presented in this section is public and can be found in the CMS document [160].

The object definitions are discussed in section 8.2.1 and the selection cuts for both categories in sections 8.2.2 and 8.2.3. In order to decrease the background, the diphoton MVA cut is optimized for each category. It is presented in section 8.2.4. The results are shown in section 8.2.5.

8.2.1 Object definition

Jets. Jets are reconstructed as described in section 5.10.2. A working point of PUJID algorithm output giving 99% jet efficiency is used. Selected jets are required to have $p_T >$

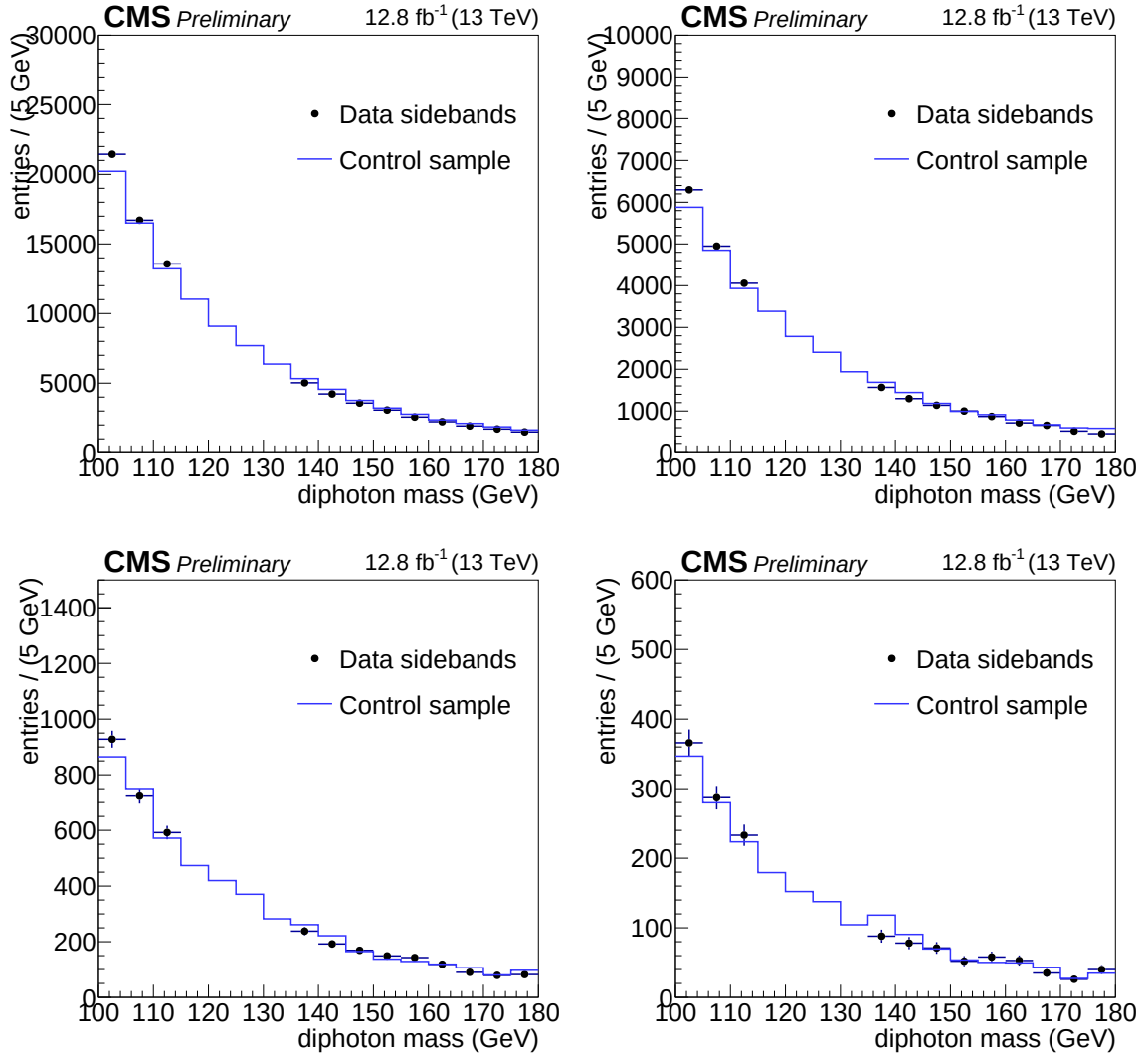


Figure 8.1: Distributions of the invariant diphoton mass from the control sample (blue) and the data sidebands (black) for different criteria: at least 2 jets (top left), at least 3 jets (top right), at least 3 jets and 1 b-jet (bottom left), at least 4 jets and 1 b-jet (bottom right). Control sample is reweighted by p_T and η of the photons, to match the ones in data sidebands. Control sample is normalized to match the integral of data in the full mass range $100 < m_{\gamma\gamma} < 180$ GeV.

25 GeV and $|\eta| < 2.4$.

b-Jets. The properties of the bottom hadrons can be used to identify the hadronic jets arising from b quarks. These hadrons have relatively large masses, long lifetimes and daughter particles with hard momentum spectra. Due to their long lifetime ($\sim 1.5 \times 10^{-12}$ s) bottom hadrons fly some millimetres in the tracker, creating a secondary vertex and displaced tracks, thus have a large and positive impact parameter.

Jets are identified as originating from b-quark (named b-tagged jet or b-jet) using the combined secondary vertex algorithm [192] that exploits the information on the impact parameters of the tracks and the reconstructed secondary vertices within the jets in a multivariate algorithm. The algorithm provides a continuous output ranging from 0 to 1, where high values indicate that the jet is more likely to originate from b quark, while low values indicate that the jet is more consistent with light-flavor quarks or gluons. The efficiency to tag b-jets and the rate of misidentification of non-b-quark jets depend on the chosen working point. For the working point used in $t\bar{t}H$ channels, the b-tagging efficiency is $\sim 69\%$ for jets originating from b-quarks and the probability of mistagging for jets originating from light quarks or gluons is 1% [193].

Electrons and muons. Electrons and muons are defined in the same way as described in section 5.11.1. Muons are required to have $p_T > 20$ GeV and $|\eta| < 2.4$. Selected electrons are required to have $p_T > 20$ GeV and pseudorapidity $|\eta| < 1.4442$ or $1.566 < |\eta| < 2.5$.

8.2.2 Leptonic channel

The leptonic category is optimised for semi-leptonic and leptonic decays of $t\bar{t}$ decays in $t\bar{t}H$ events, i.e. $t\bar{t} \rightarrow b\ell\nu_\ell\bar{b}q\bar{q}'$ and $t\bar{t} \rightarrow b\ell\nu_\ell\bar{b}\ell'\nu_{\ell'}$, where ℓ denotes either a muon or an electron. The requirement of having a high p_T isolated lepton in the final state reduces the background significantly.

The event selection for $t\bar{t}H$ leptonic category requires :

- leading (sub-leading) photon $p_T > m_{\gamma\gamma}/2$ ($m_{\gamma\gamma}/4$),
- at least one selected lepton ℓ with $p_T > 20$ GeV,
- the lepton should have $\Delta R(\ell, \gamma) > 0.4$,
- specific to the electron channel : the electron-photon mass $m_{e\gamma}$ is such that $|m_{e\gamma} - m_Z| > 10$ GeV, where m_Z refers to the mass of the Z boson,
- at least 2 selected jets with $p_T > 25$ GeV, $|\eta| < 2.4$, the angular distance between photon and jet $\Delta R(jet, \gamma) > 0.4$ and the angular distance between jet and lepton $\Delta R(jet, \ell) > 0.4$,

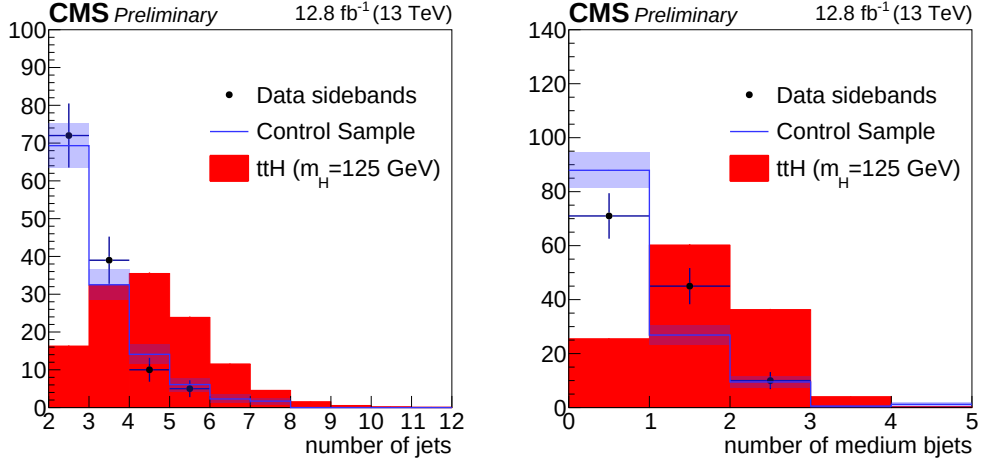


Figure 8.2: Jet multiplicity (left) and the b-jet multiplicity (right), defined with the medium working point of the b-jet discriminator, for events with at least two jets and at least one lepton. Data from the signal region sidebands (black markers) and data from the control sample (blue histogram) are compared to simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region sidebands.

- at least one of the jets has to be b-tagged with $p_T > 25$ GeV.

The distributions of the number of jets and b-jets prior to jet and b-jet selections are shown in Fig. 8.2 for both $t\bar{t}H$ signal and background events.

8.2.3 Hadronic channel

The hadronic category is designed for the events with two W bosons decaying into jets in $t\bar{t}H$ events, i.e. $t\bar{t} \rightarrow b\bar{q}q'\bar{b}q\bar{q}'$. For the $t\bar{t}H$ hadronic category, events passing the full analysis preselection are required to pass the following selections:

- leading (sub-leading) photon $p_T > m_{\gamma\gamma}/2$ ($m_{\gamma\gamma}/4$),
- no leptons (defined according to the leptonic category),
- at least 5 jets with $p_T > 25$ GeV,
- at least one of the jets in the event has to be b-tagged with $p_T > 25$ GeV.

The distributions of the number of jets and b-jets prior to jet and b-jet selections are shown in Fig. 8.3 for both $t\bar{t}H$ signal and background events.

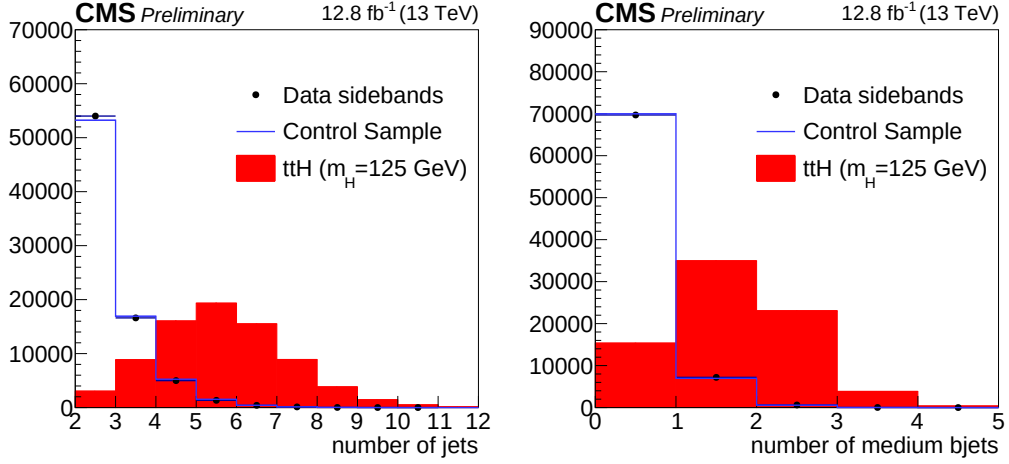


Figure 8.3: Jet multiplicity (left) and the b-jet multiplicity (right), defined with the medium working point of the b-jet discriminator, for events with at least two jets and no leptons. Data from the signal region sidebands (black markers) and data from the control sample (blue histogram) are compared to simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region sidebands.

8.2.4 Diphoton MVA cut optimization

In order to improve the background rejection, a diphoton MVA (defined in section 5.9) cut has been optimized in both hadronic and leptonic categories. The simulated sample was used for the signal, while a control sample has been used for the background. A comparison of the distributions of diphoton MVA between signal, control sample, obtained by inverting the $t\bar{t}H$ b-tag requirement and data sidebands for both hadronic and leptonic categories is shown in Fig. 8.4.

Hadronic channel. The efficiency of the different diphoton MVA cuts on the $t\bar{t}H$ signal and on the control sample is shown in Fig. 8.5. The selection has been optimized maximizing the statistical significance on the signal strength of the $t\bar{t}H$ process (noted further $\mu_{t\bar{t}H}$). The significance is $s = \sqrt{-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)}$, where λ is a statistical test defined in Eq. A.3 in appendix A. The best significance was found to be with diphoton MVA > 0.5 .

Leptonic channel. The leptonic category has a much lower number of events both in the data sidebands and in the control sample. An optimization similar to the one performed for the hadronic category is difficult due to the large statistical fluctuations in the control sample. Since the background in the leptonic category is expected to be smaller than in the hadronic one, we choose to loosen the selection to -0.4 which has a 96% efficiency on signal.

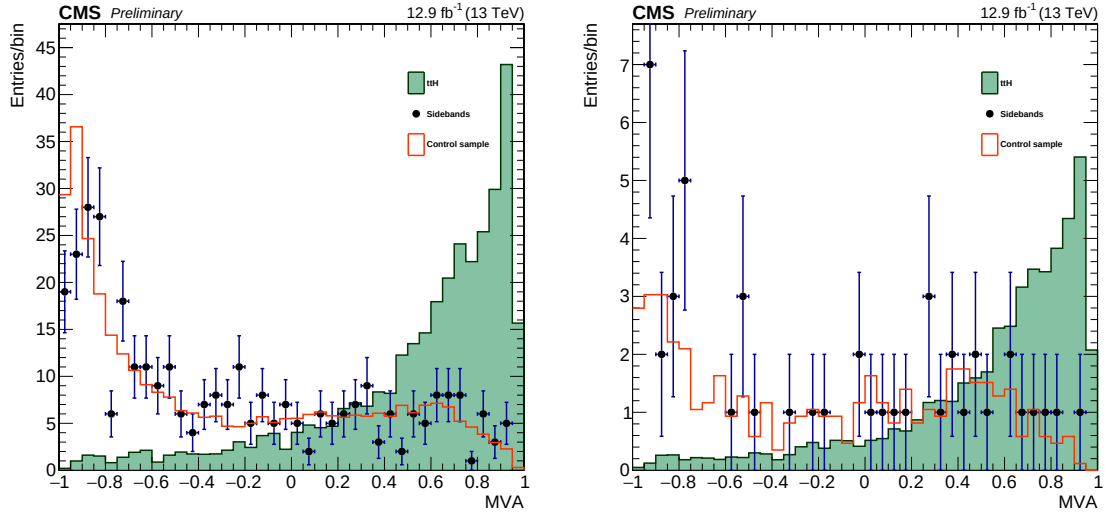


Figure 8.4: Distribution of diphoton MVA for signal, data sidebands and control sample in the hadronic (left) and leptonic (right) categories for the 12.9 fb⁻¹ dataset. The histograms are normalized to the area of the sidebands to compare the shape of the distributions.

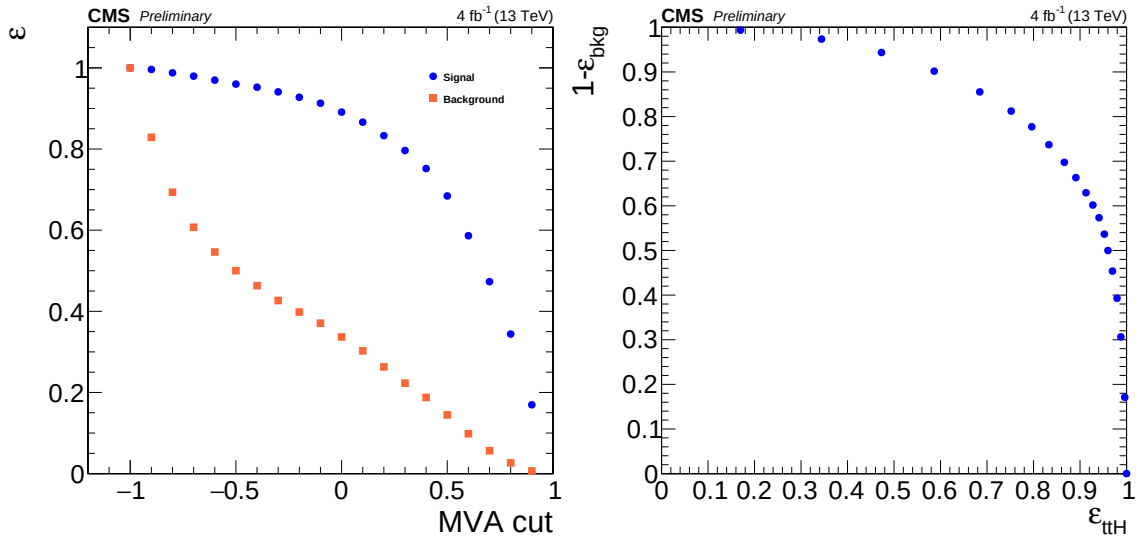


Figure 8.5: Efficiency on signal and background as a function of the cut on the diphoton MVA output (left). Performance ROC curve (right).

8.2.5 Results

Table 8.1 shows the expected number of signal events for the $t\bar{t}H$ leptonic and hadronic categories. The total number is broken by percentage contribution of each Higgs boson production mode. The mass resolution measured both by the width of narrowest interval containing 68% of the invariant mass distribution, σ_{eff} , and by the full width at half maximum of the distribution divided by 2.35, σ_{HM} . The background estimate per GeV in every category is obtained from a fit to the data.

Event Categories	SM 125GeV Higgs boson expected signal								Bkg (GeV ⁻¹) at $m_{\gamma\gamma} = 125$ GeV
	Total	ggH	VBF	WH	ZH	$t\bar{t}H$	σ_{eff}	σ_{HM}	
$t\bar{t}H$ hadronic	2.42	16.78 %	1.28 %	2.52 %	2.39 %	77.02 %	1.39	1.21	1.12
$t\bar{t}H$ leptonic	1.12	1.09 %	0.08 %	2.43 %	1.06 %	95.34 %	1.61	1.35	0.42

Table 8.1: The expected number of signal events per category. The σ_{eff} and σ_{HM} are also provided as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events per GeV for $m_{\gamma\gamma} = 125$ GeV is shown.

Results are extracted performing a simultaneous binned maximum likelihood fit to the diphoton invariant mass distribution in both $t\bar{t}H$ categories over the range $100 < m_{\gamma\gamma} < 180$ GeV and by minimizing $\lambda(\mu_{t\bar{t}H})$, which is defined in Eq. A.3 in appendix A.

The systematic uncertainties are included in the analysis as nuisance parameters and treated according to the frequentist paradigm. A description of the general methodology can be found in references [120, 194].

The final unblinded invariant mass distributions with the background-only and signal plus background fits are shown both for $t\bar{t}H$ hadronic and leptonic channels in Fig. 8.6.

Per-channel compatibility and signal strength for each of the analysis categories are shown at the top in Fig. 8.7 and the signal strength per production mode is depicted at the bottom.

During this iteration of the analysis, VH categories were not included, so the VH signal strength was set to one (i.e as expected in SM).

In addition a two-dimensional scan of the signal strength $\mu_{ggH}, \mu_{t\bar{t}H}$ for the fermionic production modes and μ_{VH}, μ_{VBF} for the vector boson production modes with the mass profiled in the fit, is performed. Fig. 8.8 shows the 68% and 95% confidence level contours, the best-fit values are $\mu_{ggH}, \mu_{t\bar{t}H} = 0.80 \pm_{-0.18}^{+0.14}$ and $\mu_{VH}, \mu_{VBF} = 1.59 \pm_{-0.45}^{+0.73}$.

All results, obtained with ICHEP'16 dataset (12.9 fb^{-1}) are in agreement with the SM within the uncertainties.

8.3 The $t\bar{t}H, H \rightarrow \gamma\gamma$ search new strategy in 2017

The analysis presented in this section is public and can be found in the CMS document [195].

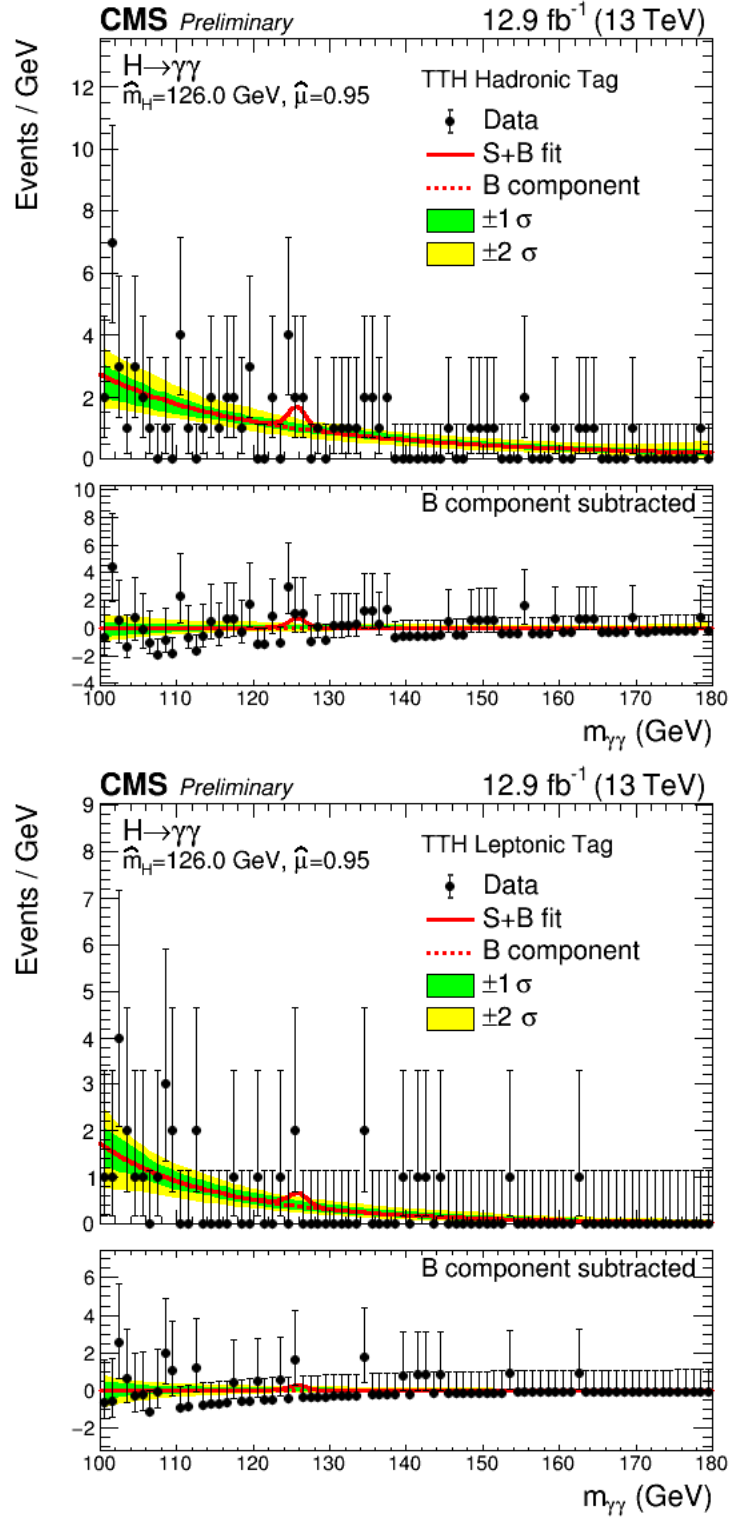


Figure 8.6: The unblinded distribution of $m_{\gamma\gamma}$ in data in each $t\bar{t}H$ hadronic (top) and leptonic (bottom) category with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

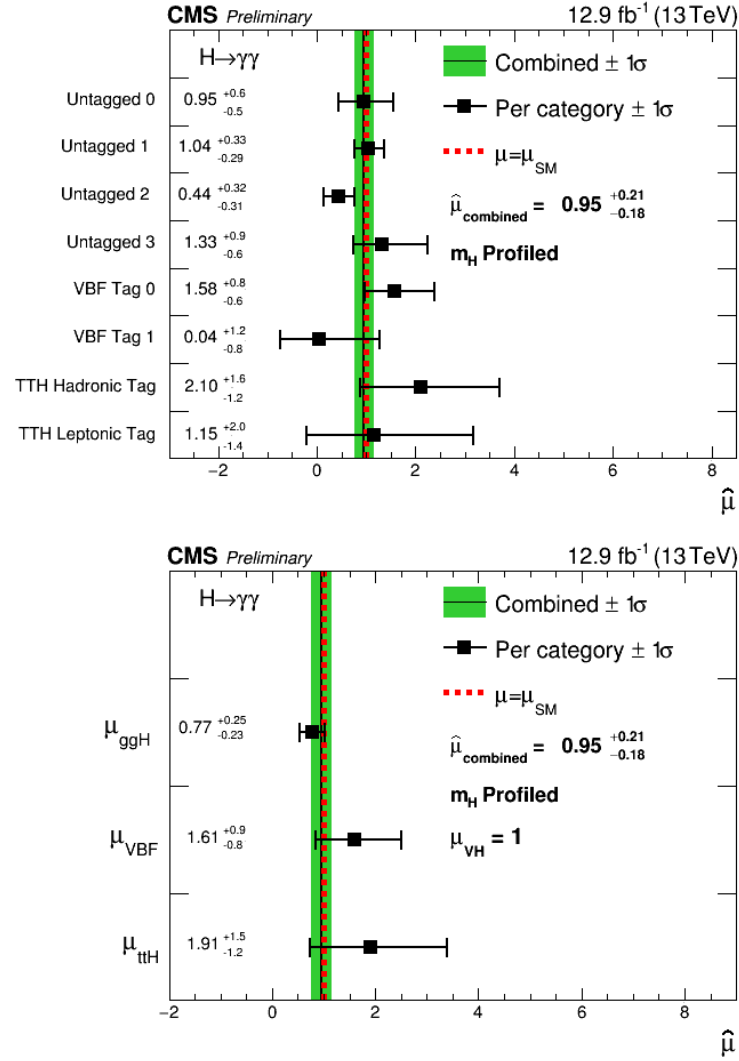


Figure 8.7: Signal strength (black) measured for each analysis category (top) and for each process (bottom) with profiled m_H , compared to the overall signal strength (green band) and to the expected value in the standard model (red dashed line).

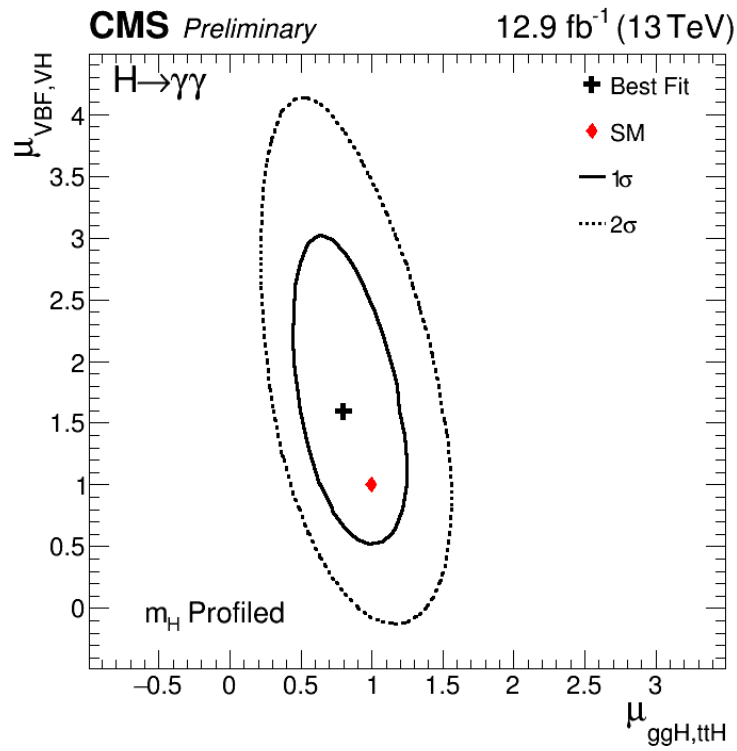


Figure 8.8: The two-dimensional best-fit (black cross) of the signal strength for fermionic and bosonic production modes compared to the standard model expectation (red diamond). The mass of the Higgs boson is profiled in the fit, the solid line represents 1σ (68%) confidence level and dashed line 2σ (95%).

The general strategy is similar to the one described in previous section. In the leptonic category cut optimization is done and described in section 8.3.2. In the hadronic category, the approach has evolved to a multivariate discriminator and shown in section 8.3.3.

8.3.1 Object definition

Jets and electrons are defined in the same way as in section 8.2.1. The b-jet discriminator has been changed with regard to ICHEP'16. We use two different working points denoted as loose and medium. For the medium working point the b-tagging efficiency is $\sim 62\%$ ($\sim 12\%$) for jets originating from a b (c) quark and the mistagging probability for light jets is $\sim 1\%$. For the loose working point the b-tagging efficiency is $\sim 80\%$ ($\sim 36\%$) for jets originating from a b (c) quark and the mistagging probability for light jets is $\sim 10\%$. The b-jet and c-jet efficiencies as function of jet p_T for three different working points are shown in Fig. 8.9.

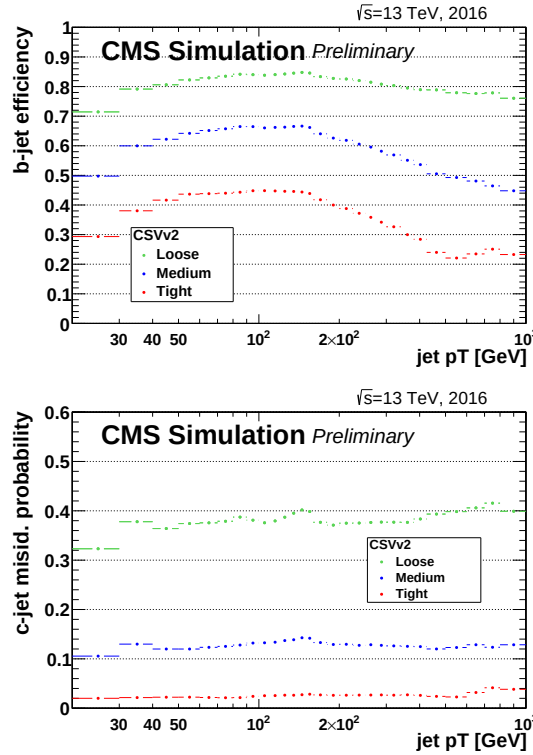


Figure 8.9: b-jet and c-jet efficiency as a function of the jet transverse momentum. The efficiencies are obtained on simulated $t\bar{t}$ events using jets within tracker acceptance with $p_T > 30$ GeV. The last bin includes the overflow.

The muon selection is very similar to the one in section 8.2.1, except for the isolation computation that is described in section 8.3.2.2.

8.3.2 Leptonic channel

The event selection for $t\bar{t}H$ leptonic category is very similar to the one described in section 8.2.2, and the differences are described below.

8.3.2.1 The $|m_{e,\gamma} - m_Z|$ cut optimization

Using simulated $t\bar{t}H$ samples for the signal and $Z\gamma$ for the background it was found that a window between the photon-electron mass and the Z boson mass of 10 GeV ($|m_{e,\gamma} - m_Z| > 10$ GeV) was too wide as the Z boson background is already very small after the $t\bar{t}H$ selection without any cut on $|m_{e,\gamma} - m_Z|$. This is illustrated in Fig. 8.10, where events with electrons passing photon identification and preselection cuts are shown in red, and events with electrons passing the $t\bar{t}H$ selection without any cut on $|m_{e,\gamma} - m_Z|$ (fraction of passing events is 1.5%) are shown in green.

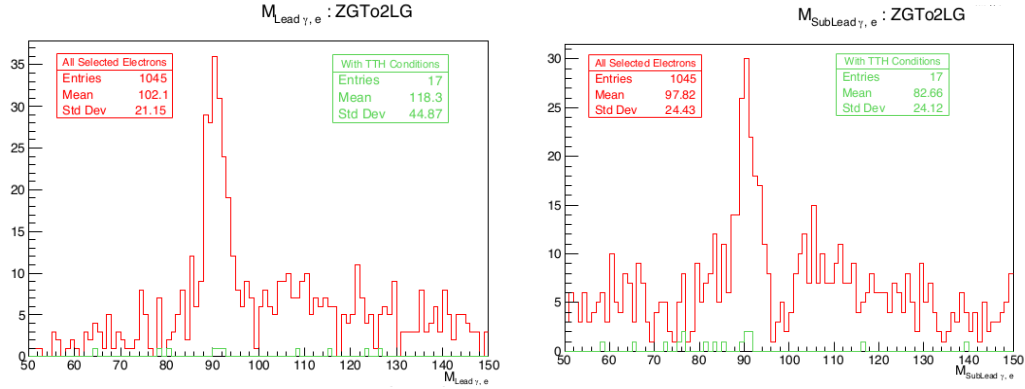


Figure 8.10: $m_{e,\gamma}$ distribution for leading (left) and subleading (right) photon. The electrons before passing photon identification and preselection cuts are shown in red, and events with electrons passing the $t\bar{t}H$ selection without any cut on $|m_{e,\gamma} - m_Z|$ are shown in green. Only 1.5% of events pass $t\bar{t}H$ selection without any cut on $|m_{e,\gamma} - m_Z|$.

The window size was released to $|m_{e,\gamma} - m_Z| > 5$ GeV, which led to a 23% signal efficiency increase in the electron channel translating into a 9% signal efficiency increase in the $t\bar{t}H$ leptonic channel. The background level was checked using data sidebands and was found the same. It was also checked that no loss in hadronic channel was introduced.

8.3.2.2 Muon isolation

While for ICHEP'16 muon isolation was computed in a $R = 0.4$ fixed cone size (I), a p_T dependent cone size is used in 2017, and called mini-isolation (I_{mini}). The size R of the

cone depends on the lepton p_T as:

$$R = \frac{10}{\min(\max(p_T(\mu), 50), 200)} \quad (8.1)$$

The isolation sum is corrected for contributions originating from pileup interactions through the area-based estimate of the pileup energy deposited in the cone. The reason for using mini-isolation is that in busy environments like $t\bar{t}H$ events, with multiple jets and photons, the fixed cone size isolation is not efficient enough. Fig. 8.11 shows that the mini-isolation (in green, pink marker shows the chosen working point $I_{\text{mini}} \leq 0.06$) performs better than the fixed cone isolation (in blue, red marker shows the chosen working point $I \leq 0.25$) in the $t\bar{t}H$ environment. For $I_{\text{mini}} \leq 0.06$, the signal efficiency is improved by 8% while keeping the same background level.

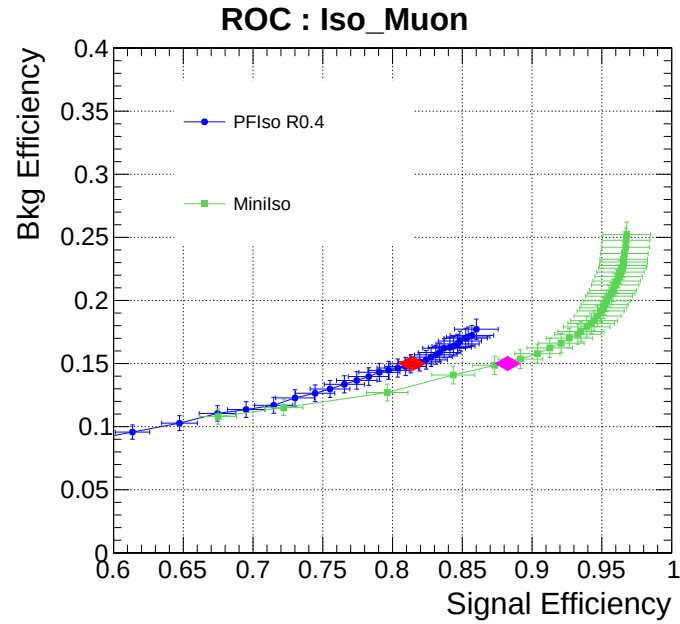


Figure 8.11: Signal efficiency versus the background efficiency for signal muons (matched to generated level muons) and non-prompt muons in the $t\bar{t}H$ events from the simulated signal sample for fixed cone size isolation in blue with red marker showing the chosen working point, and for mini-isolation in green with pink marker indicating chosen working point. $I_{\text{mini}} \leq 0.06$ has the same background efficiency as $I \leq 0.25$, but 8% higher signal efficiency.

For the next iteration of the analysis electron mini-isolation is going to be included as well.

8.3.2.3 Final cuts

Events are required to pass the following requirements to be selected for the $t\bar{t}H$ leptonic category:

- leading (sub-leading) photon $p_T > m_{\gamma\gamma}/2$ ($m_{\gamma\gamma}/4$),
- at least one selected lepton ℓ with $p_T > 20$ GeV,
- the angular distance between lepton and photon should be $\Delta R(\ell, \gamma) > 0.35$,
- specific to the electron channel : the electron-photon mass $m_{e\gamma}$ is such that $|m_{e,\gamma} - m_Z| > 5$ GeV, where m_Z refers to the mass of the Z boson,
- at least 2 selected jets with $p_T > 25$ GeV, $|\eta| < 2.4$, $\Delta R(jet, \gamma) > 0.4$ and $\Delta R(jet, \ell) > 0.4$,
- at least one of the jets with $p_T > 25$ GeV has to be b-tagged with the medium working point.

The distributions of the number of jets and b-jets prior to jet and b-jet selections are shown in Fig. 8.12 for both $t\bar{t}H$ signal and background events.

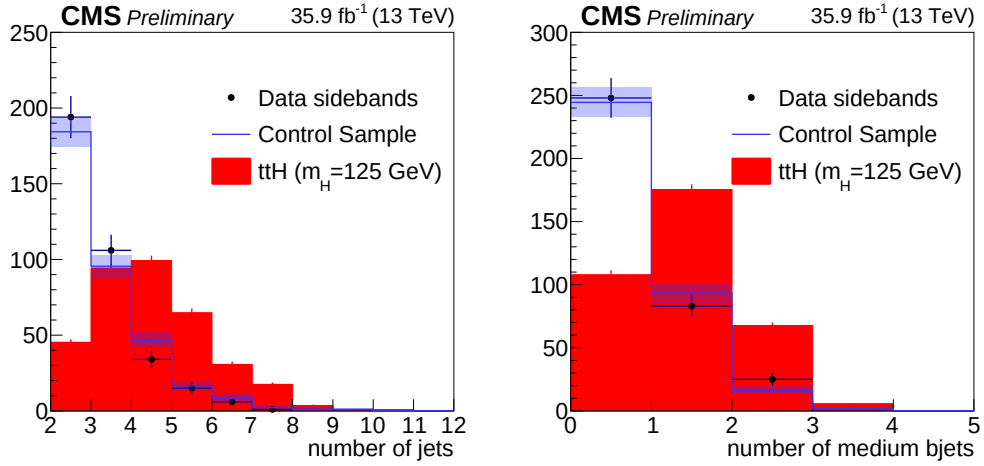


Figure 8.12: Jet multiplicity(left) and the b-jet multiplicity(right), defined with the medium working point of the b-discriminator, for events with at least two jets and at least one lepton. Data from the signal region sidebands(black markers) and data from the control sample(blue histogram) are compared to the simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region sidebands.

8.3.2.4 Expected improvement

To estimate the improvement, we use the figure of merit $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$, where λ is defined in the appendix A.

The signal yields are estimated on simulation using all Higgs boson production modes. The signal mass distributions are modelled with double gaussian, as shown at the top in Fig. 8.13. The background is modelled using data sidebands diphoton mass distribution fit with an exponential, as shown at the bottom in Fig. 8.13.

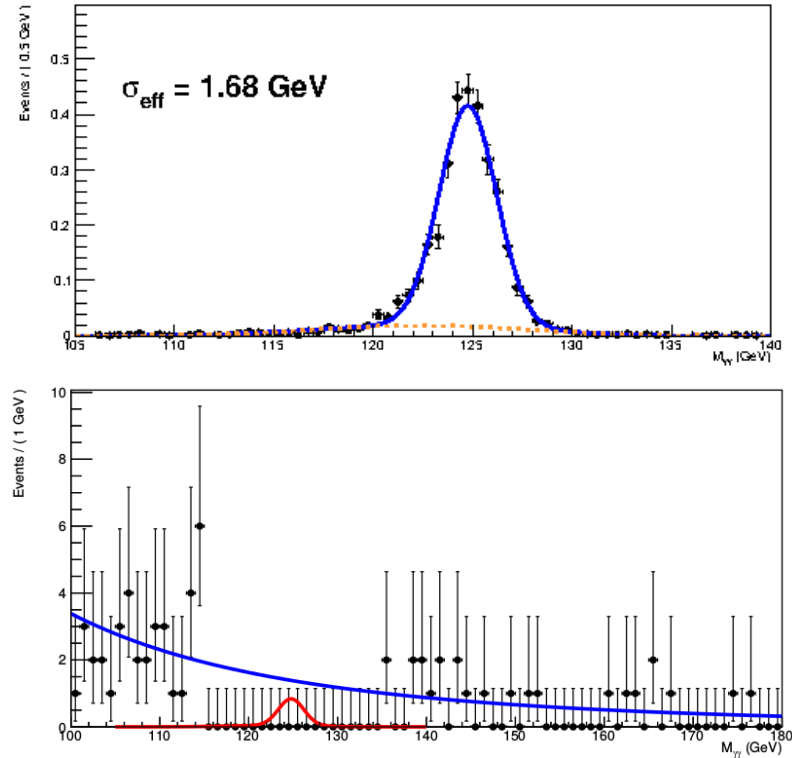


Figure 8.13: Invariant diphoton mass distribution of signal (top) and background (bottom) for optimized $t\bar{t}H$ leptonic selections. The full 2016 dataset is used, corresponding to 35.9 fb^{-1} .

The profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H})$ scan as function of $\mu_{t\bar{t}H}$ is shown in Fig. 8.14 for the ICHEP'16 strategy, the strategy with mini-isolation for muons, and the strategy with mini-isolation for muons and optimized $|m_{e,\gamma} - m_Z|$ cut. The narrower the profile likelihood is, the better precision on $\mu_{t\bar{t}H}$. The value of $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ goes from about 0.9 to about 1.05, which means $\sim 7\%$ improvement on sensitivity to $\mu_{t\bar{t}H}$. It is computed as $\sqrt{0.9/1.05}$, as the significance is $s = \sqrt{-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)}$. Most of the improvement comes from using the muon mini-isolation.

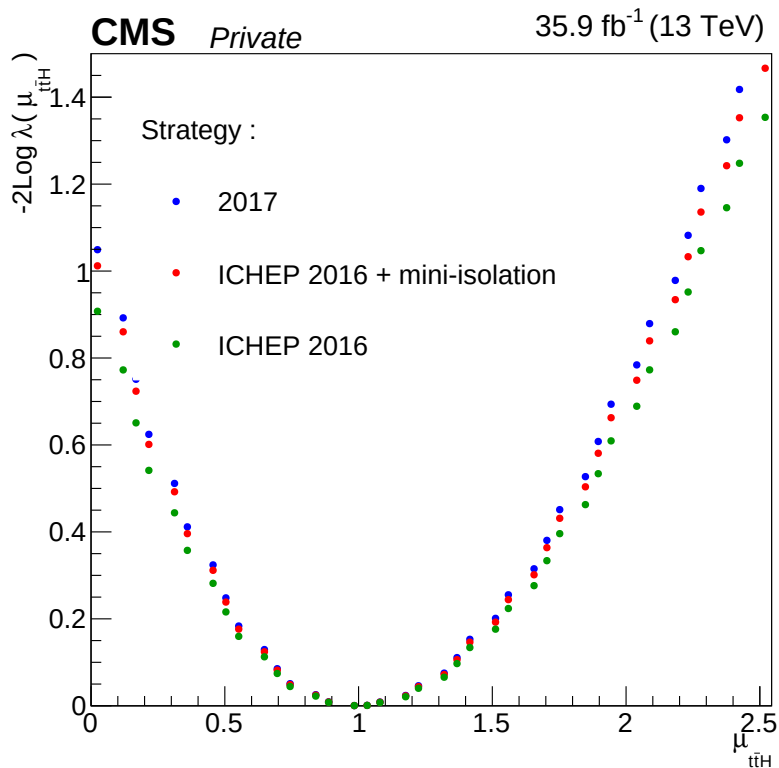


Figure 8.14: The profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H})$ scan as a function of $\mu_{t\bar{t}H}$ for the ICHEP'16 $t\bar{t}H$ leptonic strategy (in green), adding mini-isolation (in red) and optimized $|m_{e,\gamma} - m_Z|$ cut in (in blue). Each strategy has a cut on diphoton MVA > -0.4 . The main improvement comes from mini-isolation. Only the statistical uncertainty is taken into account.

8.3.3 Hadronic channel

The strategy to define the $t\bar{t}H$ hadronic category was changed with regard to the ICHEP'16 (described in section 8.2.3). Instead of cutting on the number of jets and b-jets, a multivariate discriminator is used to distinguish signal from background.

8.3.3.1 Multivariate discriminator $t\bar{t}H$ MVA

A multivariate discriminator, noted $t\bar{t}H$ MVA, is defined using the following input variables:

- the number of jets with $p_T > 25$ GeV (nJets),
- the maximum value of the b-jet discriminator (maxBTag),

- the second maximum value of the b-jet discriminator (secondMaxBTag),
- the leading jet p_T (leadJetPt).

After a cut on the diphoton MVA the main background for this channel comes from diphoton+jets production (this sample is described in section 5.1). In consequence, the discriminator was trained on $t\bar{t}H$ simulation for the signal and diphoton+jets simulation for the background. A specific preselection is applied before the training, where at least 3 jets and 1 loose b-jet are required. The b-tag discriminator values were corrected in the simulation to properly reproduce the b-tag discriminator in $t\bar{t}$ events in data [193]. The corrections for simulation are applied before the training.

The control sample is constructed using data events that have a pair of photons where one of the photons is required to pass the preselection and the photon identification criteria, while for the other photon, the photon identification criteria is inverted and the preselection criteria are not applied. The kinematic properties of the control sample are different from the ones of the signal region. We apply 2D weights of photons p_T and η , then the p_T of the leading jet is also reweighted. A control sample with similar kinematic properties as data, yet statistically independent is thus obtained. Since the diphoton MVA cut cannot be explicitly applied on the control sample with one of the photon identification inverted, it is mimicked by a normalization to data sidebands for different cuts on diphoton MVA. The diphoton mass distribution without any cut on diphoton MVA is shown in Fig. 8.15, for the control sample, data sidebands and the $t\bar{t}H$ simulation.

The distributions of the number of jets, the leading jet p_T , maximum b-tag value and second maximum b-tag value, for events with two photons with transverse momentum greater than, respectively, $m_{\gamma\gamma}/3$ and $m_{\gamma\gamma}/4$, with at least three jets, one 1 b-loose jet and no cut on diphoton MVA are shown in Fig. 8.16 for the $t\bar{t}H$ signal simulation, $t\bar{t}H$ background from data sidebands and control sample. One can see that the agreement is good between data sidebands and control sample. Also there is a good discrimination of the $t\bar{t}H$ signal from background in the chosen input variables.

The distribution of the diphoton MVA variable is shown in Fig. 8.17, for the signal simulation and data sidebands after the preselection cuts. It has a good signal-background separation. The distributions of the output variable, $t\bar{t}H$ MVA, are shown in Fig. 8.18, for signal simulation, data sidebands and diphoton+jets simulation for the preselection cuts without any cut on diphoton MVA (left) and with diphoton MVA > 0.5 (right). The $t\bar{t}H$ MVA distribution with the control sample for different cuts on diphoton MVA is shown in Fig. 8.19. The signal peaks at high values of $t\bar{t}H$ MVA, while background peaks at low values giving a good separation.

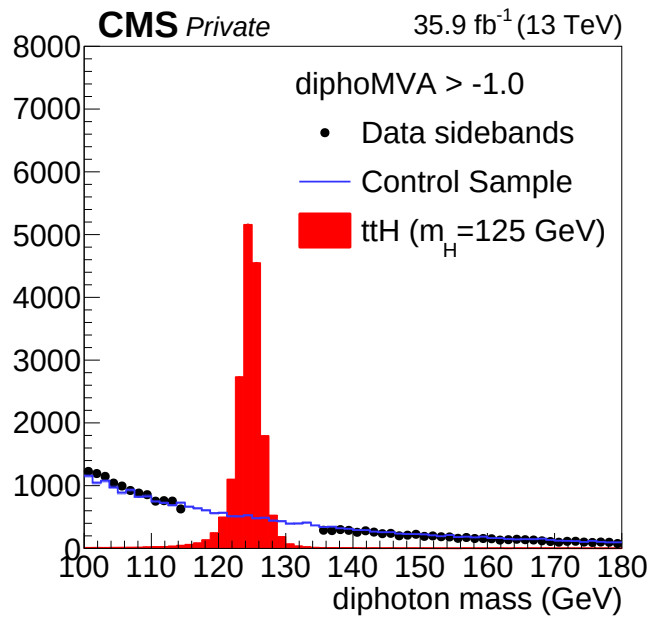


Figure 8.15: Diphoton mass distributions without diphoton MVA cut for events with at least three jets and one 1 b loose jet. Data from the signal region sidebands (black markers) and data from the control sample (blue histogram) are compared to simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region data sidebands.

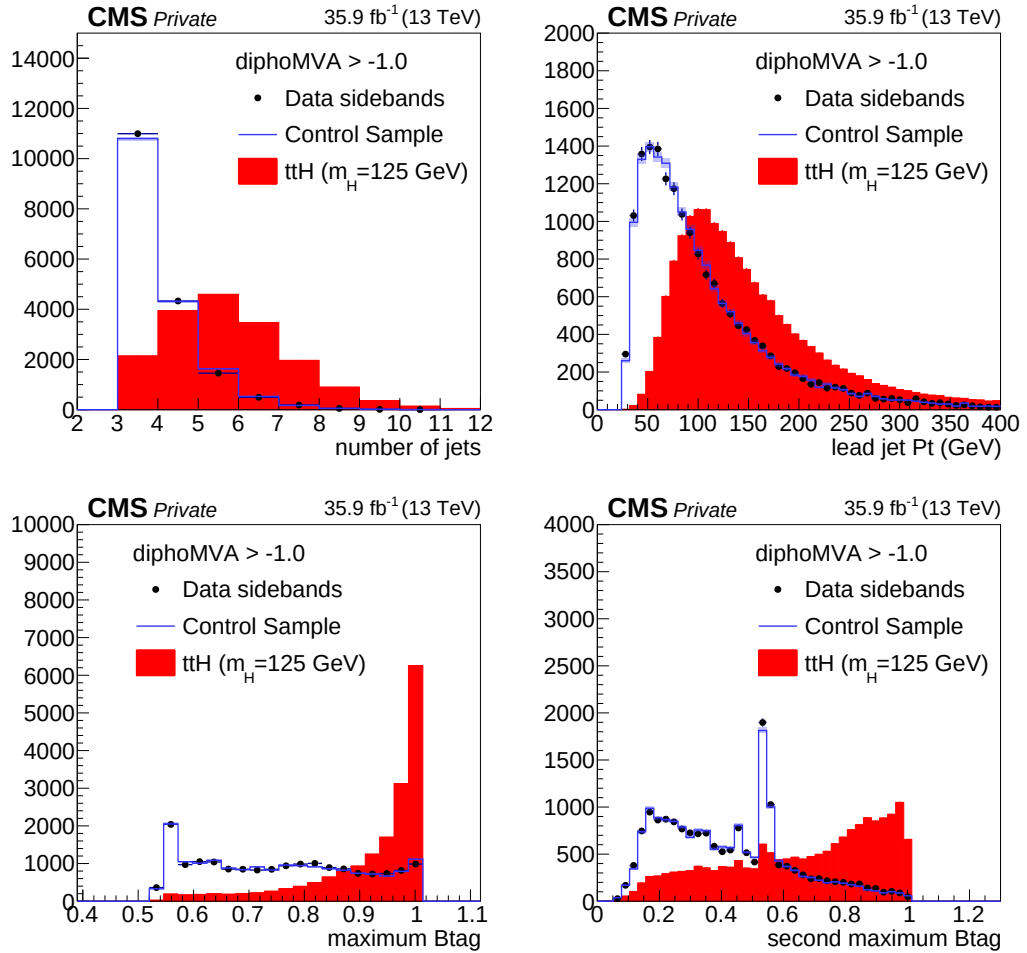


Figure 8.16: Number of jets distributions (top, left), the leading jet p_T (top, right), the maximum b-tag value (bottom, left) and second maximum b-tag value (bottom, right) without diphoton MVA cut for events with at least three jets and one 1 b loose jet. Data from the signal region sidebands (black markers) and data from the control sample (blue histogram) are compared to simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region data sidebands.

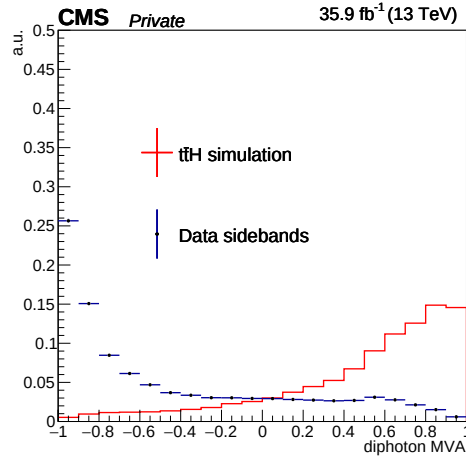


Figure 8.17: The diphoton MVA distribution for $t\bar{t}H$ signal simulation (red) and data sidebands (black) for the preselection cuts. All contributions are normalized to one.

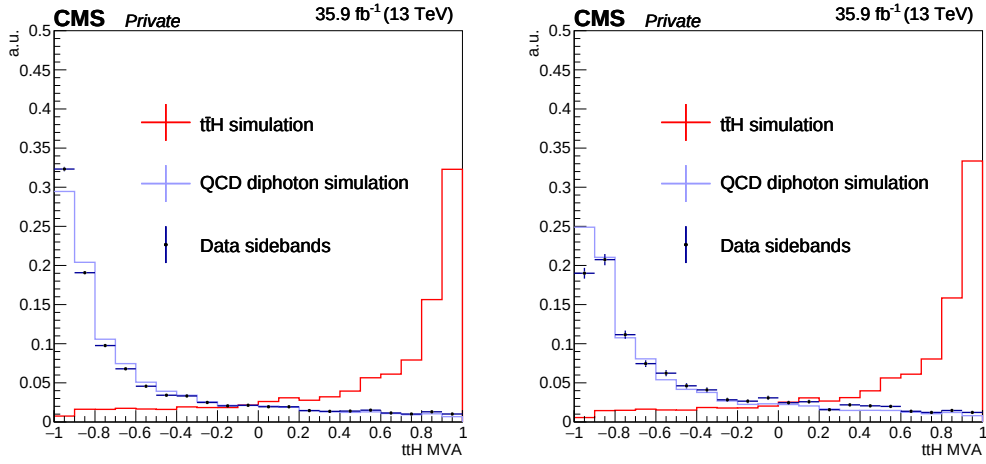


Figure 8.18: $t\bar{t}H$ MVA distributions for $t\bar{t}H$ signal simulation (red), data sidebands (black) and diphoton+jets simulation (blue) for the preselection cuts without any diphoton MVA cut (left) and with diphoton MVA > 0.5 (right). All contributions are normalized to one.

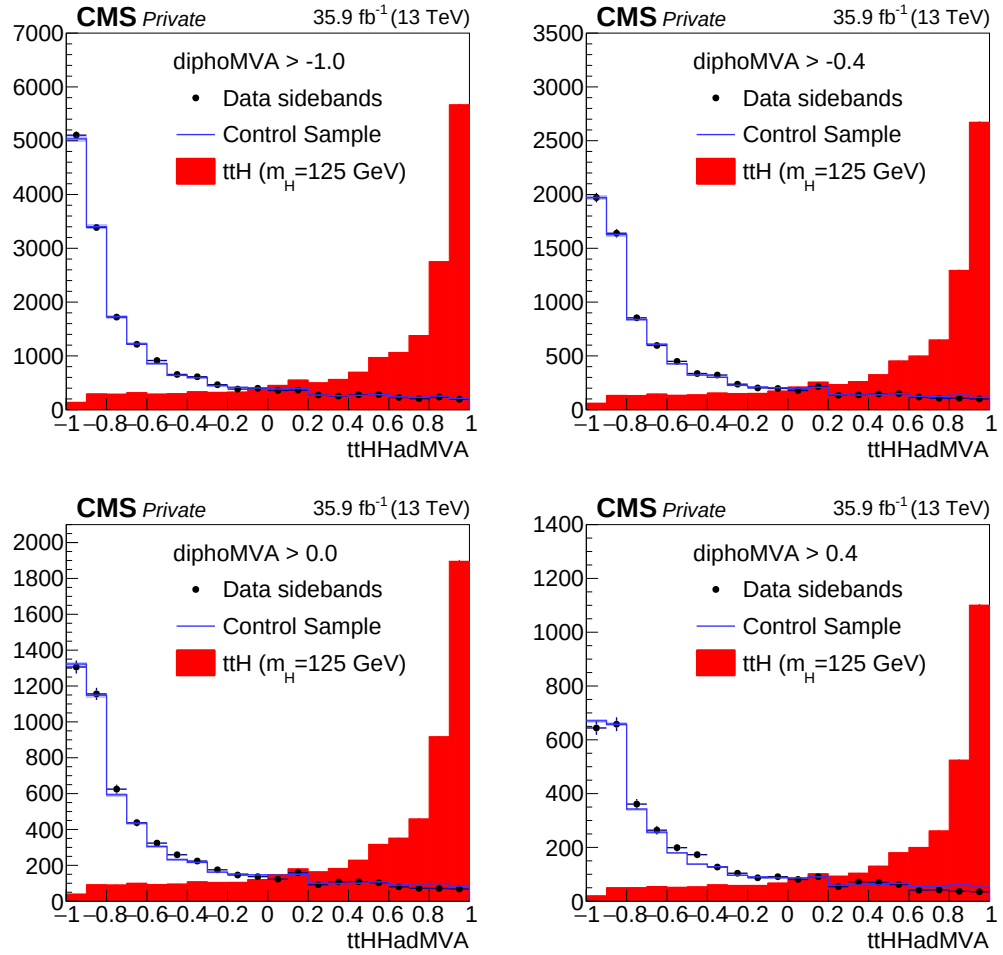


Figure 8.19: $t\bar{t}H$ MVA distribution for events with at least 3 jets, 1 b loose jet, no leptons and without diphoton MVA cut (top, left), with diphoton MVA > -0.4 (top, right), with diphoton MVA > 0.0 (bottom, left) and diphoton MVA > 0.4 (bottom, right). Data from the signal region sidebands (black markers) and data from the control sample (blue histogram) are compared to simulated $t\bar{t}H$ (red histogram) events. All contributions are normalized to the integral of the signal region sidebands.

8.3.3.2 Optimization

Yields, other modes contamination and simplified significance. In this section, the signal and the background expected yields are derived to calculate the simplified significance and the contamination from other Higgs boson production modes. This study was done using diphoton MVA > 0.5 , which was the optimal cut for ICHEP'16. The obtained numbers are compared to the ones of the cut based analysis to estimate the improvement.

The signal yields for all Higgs boson production modes are estimated from the simulation, including all final analysis corrections (shown in section 8.4).

For the background estimate three different estimators are used:

- data sidebands,
- diphoton+jets simulation at leading order, listed in section 5.1,
- a control sample with one of the photon identification requirements inverted and reweighted as described in section 8.3.3.1.

The background diphoton mass distributions are modelled with an exponential or a power law fit used to compute the number of background events per GeV, $n_{\text{bkg/GeV}}$ at $m_H = 125$ GeV. Both diphoton+jets simulation and the control sample are scaled so that the total background yield matches the integral over the full mass range in data sidebands, when $t\bar{t}H$ MVA ≥ -1 and diphoton MVA > 0.5 . In this way, by construction, simplified significance is the same for $t\bar{t}H$ MVA ≥ -1 and diphoton MVA > 0.5 for three types of background estimates.

For signal, the effective mass resolutions σ_{eff} are derived for every cut and the simplified significance $s_{t\bar{t}H}$ is calculated in a $2 \sigma_{\text{eff}}$ window as:

$$s_{t\bar{t}H} = \frac{0.945 N_{t\bar{t}H}}{\sqrt{2\sigma_{\text{eff}} n_{\text{bkg/GeV}} + 0.945 N_{\text{sig}}}}, \quad (8.2)$$

where $N_{t\bar{t}H}$ is the total number of selected $t\bar{t}H$ events and N_{sig} is the total number of Higgs boson events produced not only in $t\bar{t}H$, but also ggH, VH and VBF modes.

Fig. 8.20 shows the expected signal yields (top) and the other modes contamination (bottom) as a function of the cut on $t\bar{t}H$ MVA. On the left they are shown for the full $t\bar{t}H$ MVA range from -1 to 1, and on the top right for the 0.5-1 range. The line represents the numbers obtained with the cut-based analysis. As one can see the purity of $t\bar{t}H$ selection improves comparing to the cut-based analysis for $t\bar{t}H$ MVA ≥ 0.6 , i.e. the ggH contamination is lower.

Fig. 8.21 shows the number of background events per GeV at 125 GeV estimated from data sidebands and the simplified significances obtained for the three different samples aforementioned. We have less background with respect to cut-based analysis for

$t\bar{t}H$ MVA > 0.65 . The simplified $t\bar{t}H$ significance is better than that of the cut-based analysis for $t\bar{t}H$ MVA > 0.3 for all background estimates.

Simplified significances from data sidebands and control sample are very similar. The one obtained using diphoton+jets simulation is slightly more optimistic. First, diphoton+jets is the main reducible background, but not the only one i.e. $\gamma + \text{jet}$ processes are not taken into account. Also some irreducible background processes were not included in this computation such as: $t\bar{t}\gamma\gamma$, $t\bar{t}\gamma + \text{jet}$ and $t\bar{t} + \text{jet}$. Finally, the diphoton+jets process we use has LO cross-section instead of NLO, as the latter does not have enough statistics.

As a conclusion, the simplified significance obtained with the MVA approach is improved by 20% compared to the one with the cut-based approach.

1D optimisation of $t\bar{t}H$ MVA cut with profile likelihood. In this section we use a different and more robust estimation of the significance with profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$. For each cut on $t\bar{t}H$ MVA, the signal mass distributions for all Higgs boson production modes are modelled using a double gaussian, as shown in Fig 8.22. The background is modelled with an exponential law using data sidebands and diphoton+jets simulation, as shown in Fig. 8.23.

A scan of profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H})$ versus $\mu_{t\bar{t}H}$ is performed. Its values at $\mu_{t\bar{t}H} = 0$ are shown on Fig. 8.24, where one can see that the best significance is achieved with a cut on $t\bar{t}H$ MVA at 0.75. The difference between significance computed with background estimate from data sidebands and simulation is explained by the different shapes of the $t\bar{t}H$ MVA. All numbers are computed with a cut diphoton MVA > 0.5 . The dashed line shows the cut-based strategy profile likelihood value. The conclusion obtained with the simplified significance is confirmed, the significance is improved by $\sim 20\%$ with MVA approach.

One can see that the sensitivity on $\mu_{t\bar{t}H}$ improves with the MVA approach.

2D optimisation of the cuts on $t\bar{t}H$ MVA and diphoton MVA. A two-dimensional scan of the $t\bar{t}H$ MVA cut and the diphoton MVA cut with the background estimated from data sidebands, simulation and control sample was performed. The background estimation from simulation here includes not only diphoton+jets, but also top-pair production with real or fake photons to improve the background description. The number of events in data sidebands (corresponding to 35.9 fb^{-1}) and normalization factors (to match the number of events in data sidebands) for simulation (N_{MC}) and control sample (N_{CS}) for different diphoton MVA cuts and no cut on $t\bar{t}H$ MVA are listed in Table 8.2.

The values of $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ are shown on Fig. 8.25 on the top (bottom) with background estimation from data sidebands (simulation). The optimal cuts are found to be $t\bar{t}H$ MVA ≥ 0.75 and diphoton MVA ≥ 0.4 . Tightening diphoton MVA ≥ 0.5 gives similar significance, but smaller signal efficiency, is therefore discarded. The optimization with the simulation for the background estimation (Fig. 8.25 bottom plot) is fully consistent with the one done on data sidebands (Fig. 8.25 top plot).

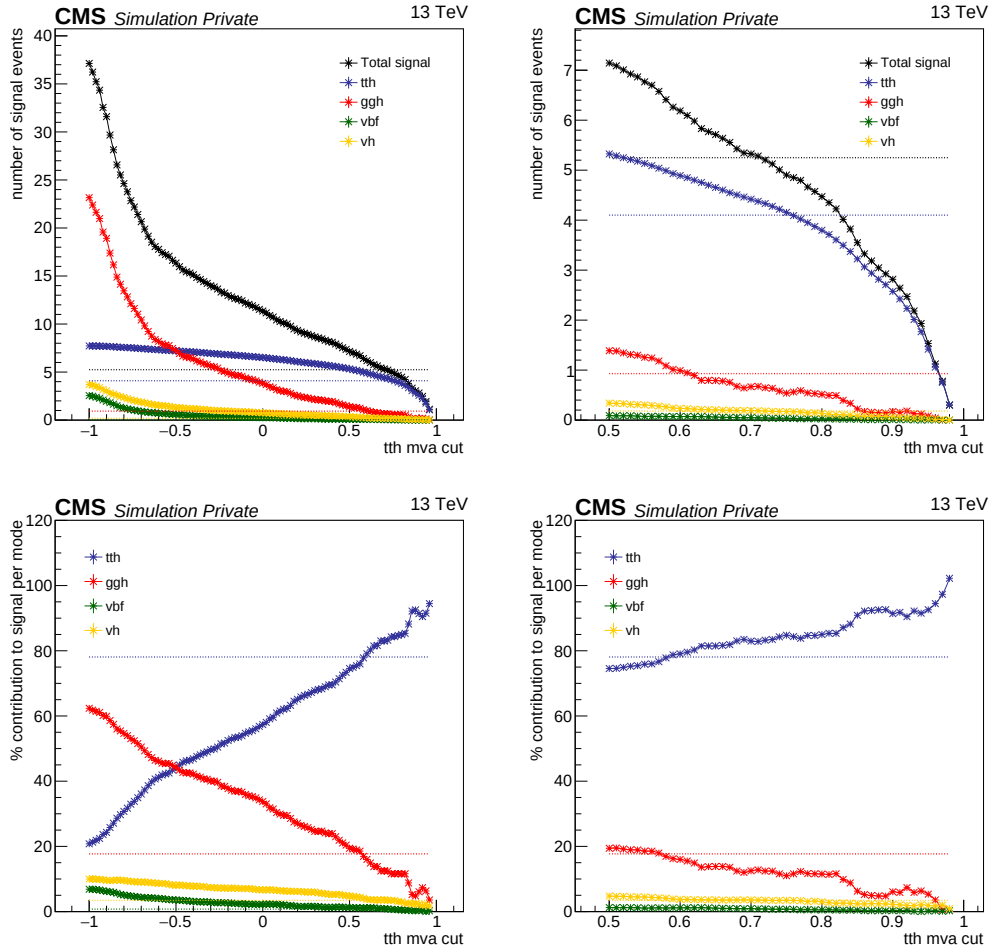


Figure 8.20: Expected number of $t\bar{t}H$ signal events (top) and expected contamination from other modes (bottom) versus the cut on $t\bar{t}H$ MVA. On the left plots, the full range of $t\bar{t}H$ MVA is shown, and on the right plots, $t\bar{t}H$ MVA > 0.5 and $t\bar{t}H$ MVA < 1 . The dashed lines show the cut-based strategy numbers. All numbers are computed with diphoton MVA > 0.5 .

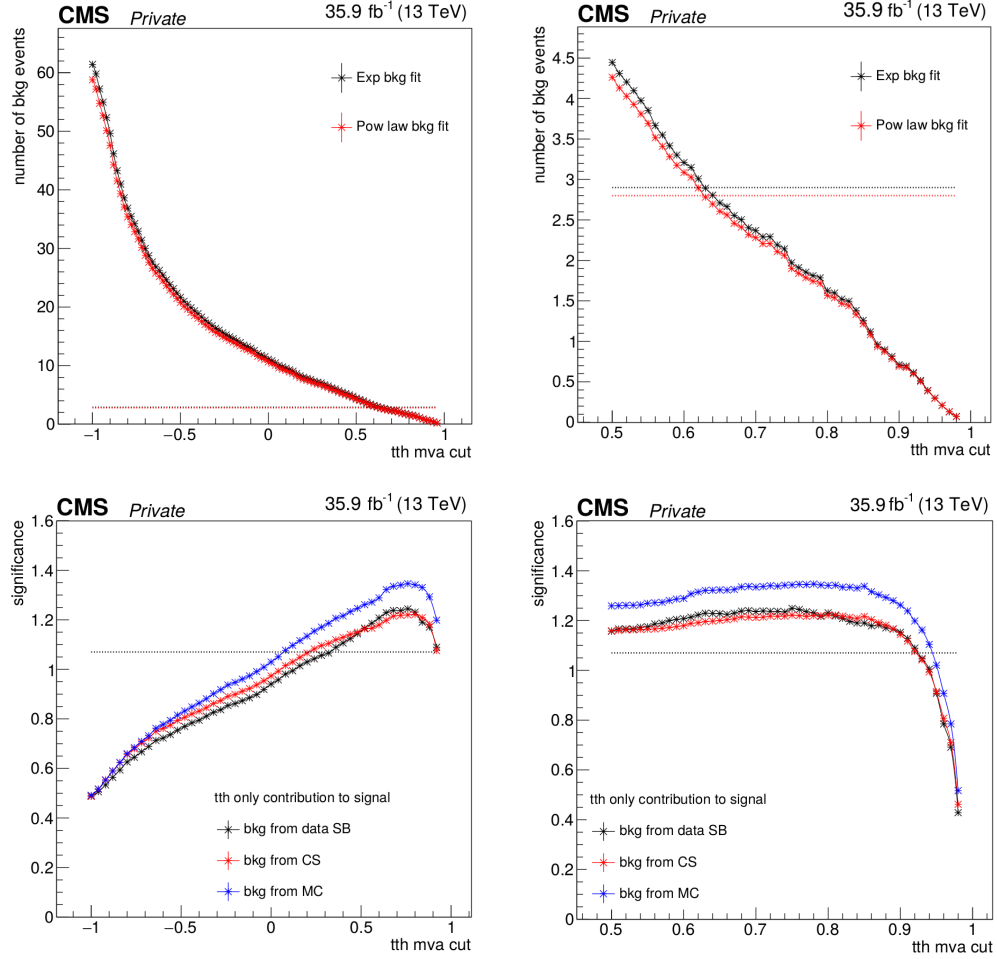


Figure 8.21: Expected number of background events per GeV at $m_H = 125$ GeV (top) estimated from data sidebands and simplified significances (bottom) for the 3 different background estimates described in the text. On the left plots, the full range of $t\bar{t}H$ MVA is shown, and on the right plots, $t\bar{t}H$ MVA > 0.5 and $t\bar{t}H$ MVA < 1 . The dashed lines show the cut-based strategy numbers. All numbers are computed with diphoton MVA > 0.5 . Simplified significance from data sidebands and control sample are very similar, the one from simulation is slightly more optimistic, because simulation does not describe perfectly the background. In all three cases the tendency and shape are the same. MVA approach gives 20% improvement of the significance.

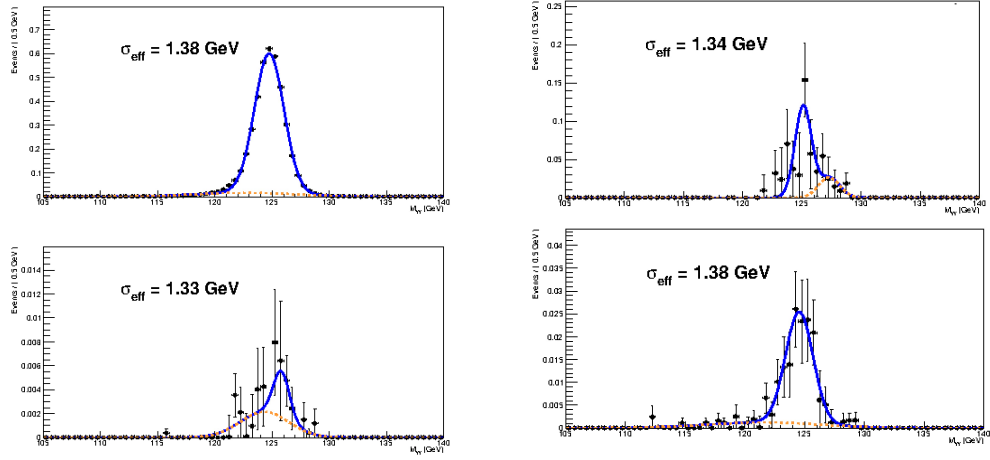


Figure 8.22: Signal models for $t\bar{t}H$ (top left), ggH (top right), VBF (bottom left), VH (bottom right) contributions, done with diphoton MVA > 0.5 and $t\bar{t}H$ MVA > 0.75 . Red and blue curves correspond to double gaussian.

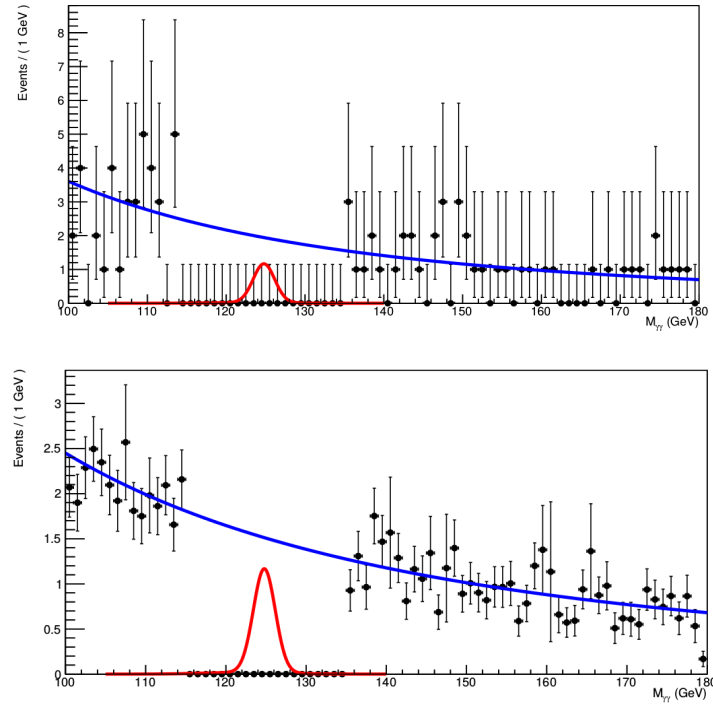


Figure 8.23: Background models on data sidebands (top) and on diphoton+jets simulation normalized to data sidebands (bottom), done with diphoton MVA > 0.5 and $t\bar{t}H$ MVA > 0.75 . Blue curve corresponds to background model and red one to signal model.

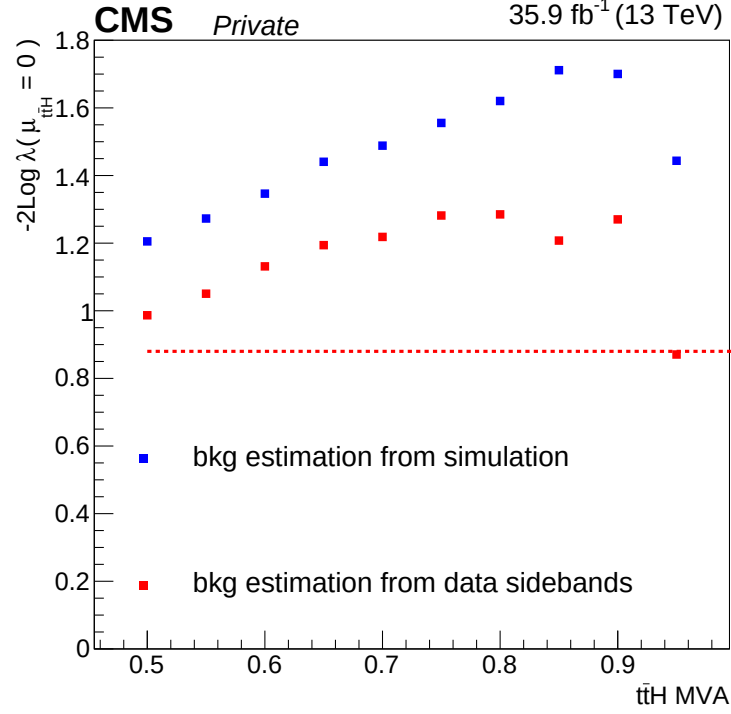


Figure 8.24: $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ versus the $t\bar{t}H$ MVA cut for background estimated from sidebands (red) and diphoton+jets simulation (blue), computed with diphoton MVA > 0.5 . The red dashed line corresponds to $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ obtained with the cut-based approach.

The optimization using the control sample for the background estimate is shown in Fig. 8.26. Since the diphoton MVA cut cannot be explicitly applied on the control sample with one of the photon identification inverted, it is mimicked by normalizing to data sidebands for different cuts on diphoton MVA. From Fig. 8.26 one can see that the optimal values are very close to the one obtained with data sidebands and simulation for the background estimate: $t\bar{t}H$ MVA ≥ 0.8 and diphoton MVA ≥ 0.4 .

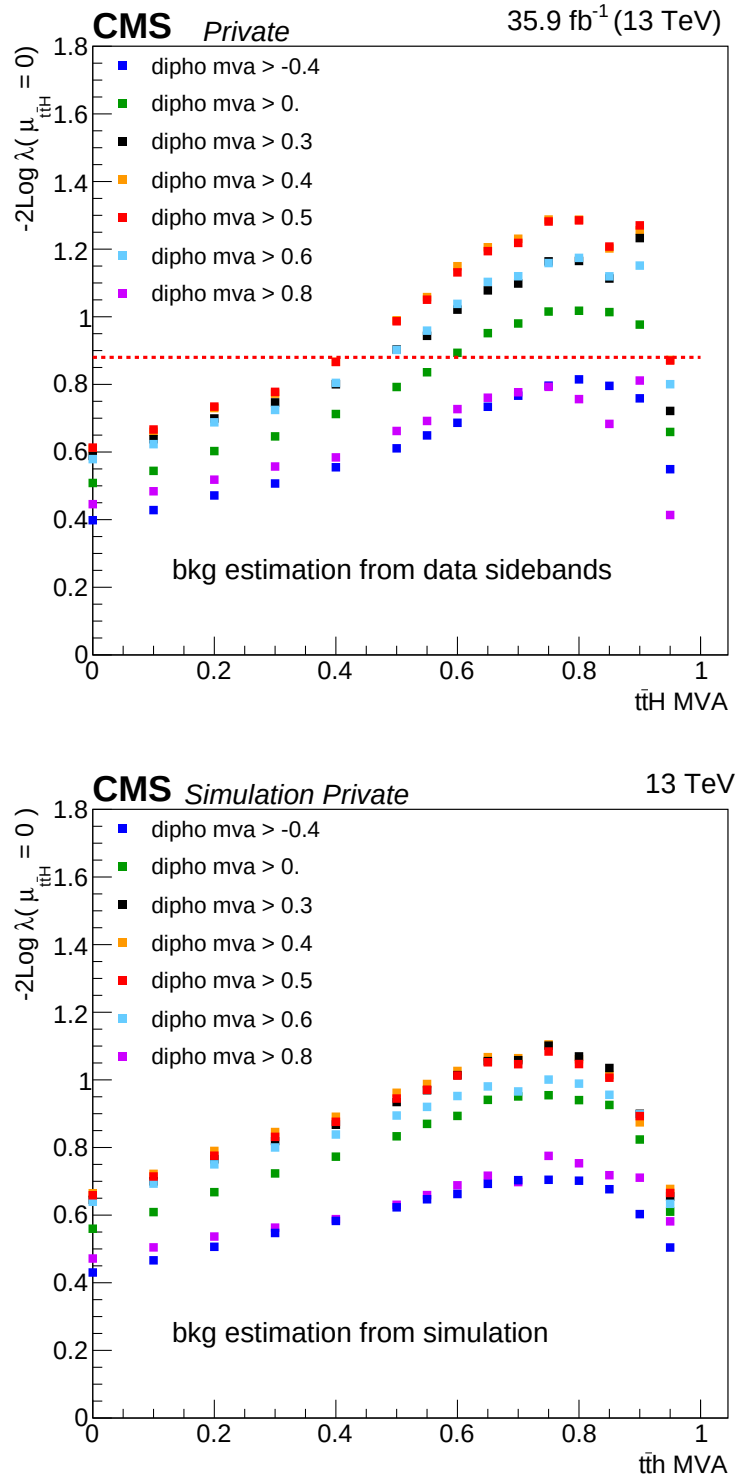


Figure 8.25: Top: Profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ versus the $t\bar{t}H$ MVA cut for the background estimated from sidebands for different values of the diphoton MVA cut. Bottom: Profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ versus the $t\bar{t}H$ MVA cut for the background estimated from simulation normalized to data sidebands for different values of the diphoton MVA cut. The optimization with simulation for the background estimation is fully consistent with the one done on data sidebands.

dipho MVA cut	Data sidebands	N_{CS}	N_{MC}
-0.4	14 451	0.356	2.63
0	9 525	0.228	2.1
0.3	6150	0.142	1.9
0.4	5096	0.116	1.9
0.5	4025	0.089	1.91
0.6	2790	0.058	1.91
0.8	841	0.0136	1.91

Table 8.2: The number of events in data sidebands (corresponding to 35.9 fb^{-1}) without any cut on $t\bar{t}H$ MVA and with different cuts on diphoton MVA. The normalization factors applied to the simulation and control sample to match the number of events in data sidebands are listed as well.

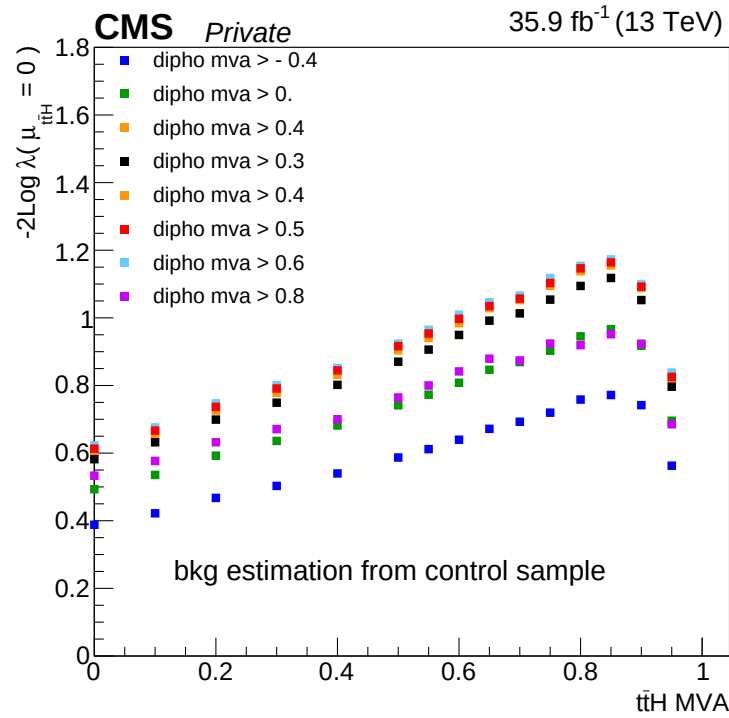


Figure 8.26: Profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ versus the $t\bar{t}H$ MVA cut for the background estimated from control sample reweighted and normalized to data sidebands for different values of the diphoton MVA cut. The shapes are the same for different "cuts" on diphoton MVA.

Final selection. After the optimization presented in the previous paragraphs, we found that the best cuts to define $t\bar{t}H$ hadronic category are:

- leading (sub-leading) photon $p_T > m_{\gamma\gamma}/3$ ($m_{\gamma\gamma}/4$)¹,
- at least 3 jets,
- at least one of the jets has to be loosely b-tagged with $p_T > 25$ GeV (loose working point),
- no leptons (defined according to the leptonic category),
- $t\bar{t}H$ MVA ≥ 0.75 ,
- diphoton MVA ≥ 0.4 .

8.3.3.3 Expected improvement

In order to have a fair comparison, the diphoton MVA cut for the cut-based strategy was re-optimized as described in section 8.2.4. The optimal value was found to be 0.6. The comparison between the re-optimised cut-based strategy, with diphoton MVA > 0.6 , and the optimised MVA strategy, with $t\bar{t}H$ MVA ≥ 0.75 and diphoton MVA ≥ 0.4 is shown in Fig. 8.27. The value of profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ increases from about 0.9 to about 1.3, which corresponds to a 20% improvement on the sensitivity to $\mu_{t\bar{t}H}$.

For the cut-based approach, the ggH contamination is found to be 18%, while for MVA approach it is lower and equal to 11%.

8.4 Systematic uncertainties

Possible sources of systematic uncertainties which are specific to $t\bar{t}H$ categories are listed below.

Jet energy scale and resolution. The systematic error on the jet energy scale is estimated by shifting the scale by $\pm 1\sigma$, where σ is the full jet energy scale uncertainty. The systematic error on jet energy resolution is estimated by varying the resolution according to the level of disagreement between the resolution measured in data and in simulation.

Lepton identification efficiency. For both electrons and muons, the uncertainty on the identification efficiency is computed by varying the correction for simulation (computed as a ratio of efficiency measured in data and simulation) by its uncertainty. The resulting difference in the signal efficiency is taken as systematic uncertainty ($\sim 1\%$).

b-Tagging efficiency, shape. For the leptonic channel, uncertainty on b-tagging efficiency is evaluated by varying the measured b-tagging correction within its uncertainty. For

¹Introducing $t\bar{t}H$ MVA let us to relax this cut to $p_T > m_{\gamma\gamma}/3$, with the previous cut-based strategy it was $p_T > m_{\gamma\gamma}/2$.

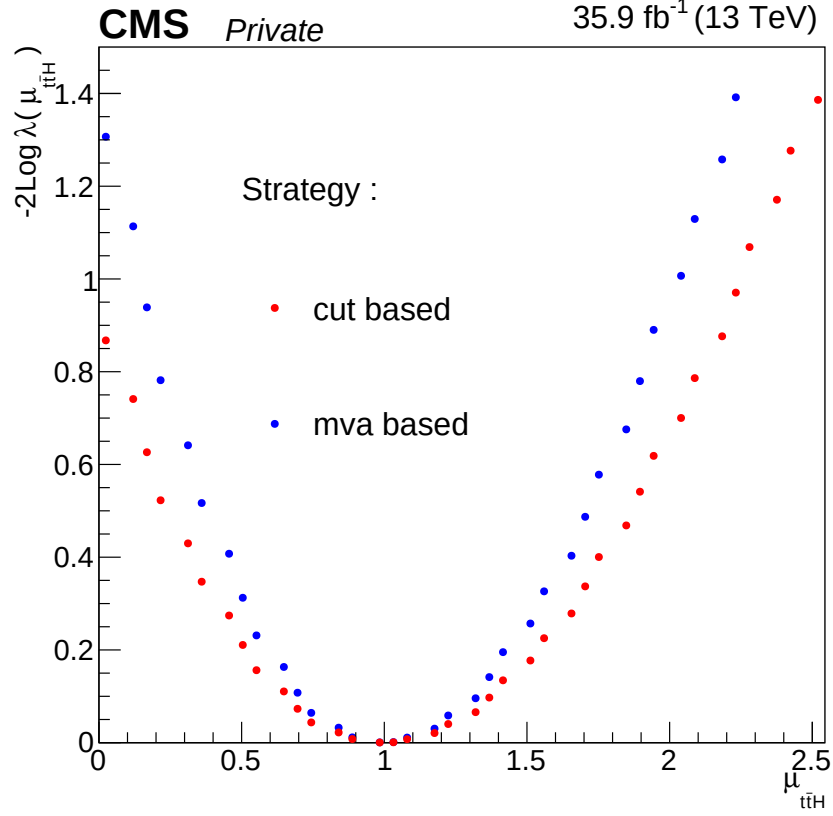


Figure 8.27: The value of profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ for the cut-based and the $t\bar{t}H$ MVA strategies, each strategy with their optimized diphotonMVA cut. The value of profile likelihood $-2\text{Log}\lambda(\mu_{t\bar{t}H} = 0)$ increases from about 0.9 to about 1.3, which corresponds to a 20% improvement on the sensitivity to $\mu_{t\bar{t}H}$. Only the statistical uncertainty is taken into account.

the hadronic channel, the correction for modifying the shape of the b-tagging discriminant in the simulated samples is varied by one standard deviation of the uncertainty. Different sources of uncertainties are taken into account, including the statistical uncertainties for the correction estimation for heavy flavour jets and light flavour jets, the contamination of light flavour jets into the correction calculation of heavy flavour jets, the contamination of heavy flavour jets into the correction calculation of light flavour jets and the effect of the jet energy scale uncertainty. The simulated signal yield changes by about 5% in the hadronic channel. The effect in the leptonic channel is less than 2%.

ggH contamination. The theoretical predictions for ggH contamination are not reliable when the Higgs boson is produced in association with a large number of jets $N_{jets} \geq 5$ radiated from the gluon-gluon initial state, the additional radiation being produced mostly

via parton showering. The systematic uncertainty on this contamination have been estimated taking into account several contributions:

- uncertainty due to the limited size of the simulated sample (10%)
- uncertainty from the parton shower modelling. This uncertainty is estimated using the observed difference in the jet multiplicity between the aMC@NLO predictions and data in $t\bar{t}$ +jets events (which are dominated by gluon fusion production $gg \rightarrow t\bar{t}$), with fully leptonic $t\bar{t}$ decays. This uncertainty is about 15% for the leptonic category selections and 35% for the hadronic category [184].
- uncertainty on the gluon splitting modelling. The difference between data and simulation of $\sigma(ttbb)/\sigma(ttjj)$ is measured. The fraction of events from ggH with real b-jets are then scaled by this value. A measurement of the scale factor is available in [185] and found to be 0.7. It gives an uncertainty on the ggH yield in $t\bar{t}H$ categories of 50%.

8.5 Conclusion

The $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis was shown. The results with the ICHEP'16 dataset (12.9 fb^{-1}) are consistent with SM expectations within the uncertainties. The analysis approach was modified to achieve better performance. In the hadronic channel the ggH contamination reduces by 7% and the sensitivity increases by 20%. In the leptonic channel the sensitivity increases by 7%. This new strategy was adopted for the $H \rightarrow \gamma\gamma$ analysis on the full 2016 dataset 35.9 fb^{-1} . The results of this improved analysis are shown in the next chapter.

Chapter 9

The $H \rightarrow \gamma\gamma$ and $t\bar{t}H$, $H \rightarrow \gamma\gamma$ analysis final results

9.1 Final results with the full 2016 dataset of 35.9 fb^{-1}

The results with the full 2016 dataset of 35.6 fb^{-1} were presented in LHCP'17 conference and can be found in the CMS public document [195].

The unblinded diphoton invariant mass distributions with the background-only and signal plus background fits are shown both for the $t\bar{t}H$ hadronic and leptonic categories in Fig. 9.1, for the VBF categories in Fig. 9.2, for VH categories in Fig. 9.3, and for the untagged categories in Fig. 9.4. The unblinded diphoton invariant mass distributions for all categories are summed are shown in Fig. 9.5.

The expected uncertainties on the measurement of the signal strength ($\mu = \sigma/\sigma_{SM}$) for the observation of a SM Higgs boson are obtained by generating an Asimov data set from the best-fit background model and injecting a signal of strength $\mu = 1$ at $m_H = 125 \text{ GeV}$.

A likelihood scan of the signal strength is performed, profiling all other nuisances. The results can be found in Fig. 9.6. In this scan, the mass of the Higgs boson was profiled in the same way as other nuisances in the fit. The best-fit signal strength measured for all categories combined is

$$\mu = 1.16^{+0.15}_{-0.14} = 1.16^{+0.11}_{-0.10}(\text{stat.})^{+0.09}_{-0.08}(\text{syst.})^{+0.06}_{-0.05}(\text{theo.}) \quad (9.1)$$

The expected and observed $t\bar{t}H$ signal strength (for leptonic and hadronic categories together and separately) are listed in Table 9.1 with statistical only and statistical plus systematic uncertainties. The uncertainty on $\mu_{t\bar{t}H}$ is dominated by statistical component. The main systematic uncertainties components are theoretical uncertainties of QCD scale yields, ggH contribution to $t\bar{t}H$ categories, and PDF and α_s modelling as shown in Table 5.6. The main experimental systematic uncertainties are coming from photon energy

scales and smearing, and photon identification. They are going to be improved with new data reconstruction for the $H \rightarrow \gamma\gamma$ publication.

Process	$\mu_{t\bar{t}H \text{ exp}} \pm (\text{stat.})$	$\mu_{t\bar{t}H \text{ exp}} \pm (\text{stat.} + \text{syst.})$	$\mu_{t\bar{t}H \text{ obs}} \pm (\text{stat.})$	$\mu_{t\bar{t}H \text{ obs}} \pm (\text{stat.} + \text{syst.})$
$t\bar{t}H$	$1^{+0.80}_{-0.69}$	$1^{+0.82}_{-0.70}$	$2.22^{+0.85}_{-0.77}$	$2.22^{+0.92}_{-0.79}$
$t\bar{t}H \text{ lept}$	$1^{+1.19}_{-0.96}$	$1^{+1.20}_{-0.96}$	$2.9^{+1.33}_{-1.13}$	$2.9^{+1.40}_{-1.15}$
$t\bar{t}H \text{ had}$	$1^{+0.92}_{-0.78}$	$1^{+0.96}_{-0.78}$	$1.5^{+0.93}_{-0.80}$	$1.5^{+1.00}_{-0.81}$

Table 9.1: The expected and observed $t\bar{t}H$ signal strength with statistical only and statistical plus systematic uncertainties.

The signal strength for each Higgs boson production mode is shown in Fig. 9.7.

A two-dimensional likelihood scan of the signal strength $\mu_{ggH}, \mu_{t\bar{t}H}$ for fermionic production modes and μ_{VH}, μ_{VBF} for vector boson production modes with the Higgs boson mass profiled in the fit, is performed. Fig. 9.8 shows the 68% and 95% confidence level contours, the best-fit values are

$$\mu_{ggH}, \mu_{t\bar{t}H} = 1.19^{+0.20}_{-0.18} \quad (9.2)$$

$$\mu_{VH}, \mu_{VBF} = 1.01^{+0.57}_{-0.51} \quad (9.3)$$

Fig. 9.9 shows the two-dimensional likelihood scans in the κ framework [96]. This framework was developed in order to take into account the currently best available SM predictions for Higgs cross sections, which include higher-order QCD and EW corrections [196, 197], while at the same time introducing possible deviations from the SM values of the couplings, the predicted SM Higgs cross sections and partial decay widths are dressed with scale factors κ_i . The scale factors κ_i are defined in such a way that the cross sections σ_{ii} or the partial decay widths Γ_{ii} associated with the SM particle i scale with the factor κ_i^2 when compared to the corresponding SM prediction. For example, $\kappa_\gamma^2 = \frac{\Gamma_{\gamma\gamma}}{\Gamma_{\gamma\gamma}^{\text{SM}}}$ and $\kappa_g^2 = \frac{\sigma_{ggH}}{\sigma_{ggH}^{\text{SM}}}$.

The first scan shows κ_f vs κ_V , which reflects the coupling modifiers of the Higgs boson to fermions and vector bosons respectively (on the top). The second scan shows κ_γ vs κ_g scan measures the Higgs boson couplings modifiers to photons and gluons (on the bottom). The κ parameters other than those varied are fixed to 1 in each case. Fig. 9.9 shows the 1 sigma and 2 sigma contours for each scan.

Expected and observed significance per production mode are listed in Table 9.2. The results obtained with the full 2016 dataset (35.9 fb^{-1}) are in agreement with SM within

the uncertainties. The signal strength is split per production mode and a 3.3 (1.5) σ excess from background-only hypothesis is observed (expected) in the $t\bar{t}H$ production mode.

Process	Expected significance	Observed significance
ggH	5.9σ	6.6σ
VBF	1.9σ	1.1σ
VH	1.2σ	2.4σ
$t\bar{t}H$	1.5σ	3.3σ

Table 9.2: The expected and observed significances per production mode.

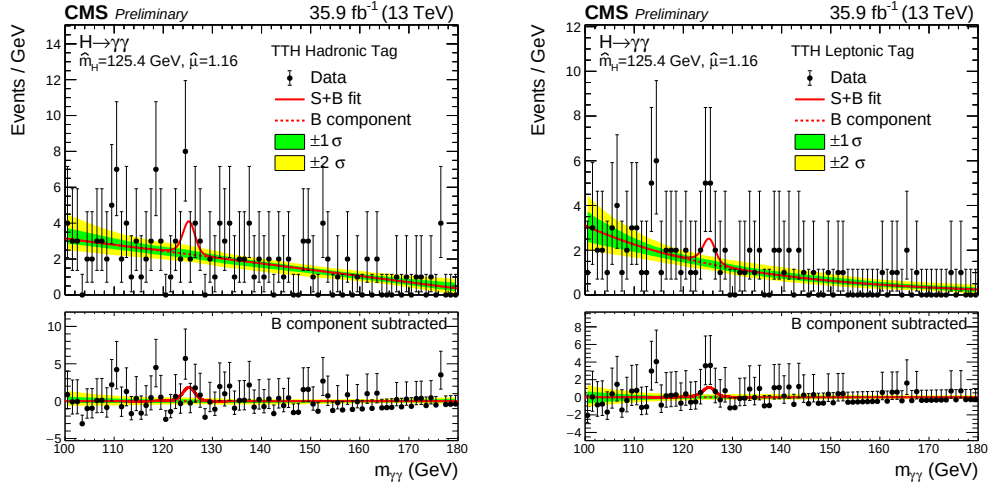


Figure 9.1: The unblinded distribution of $m_{\gamma\gamma}$ in data in each $t\bar{t}H$ hadronic (left) and leptonic (right) category with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

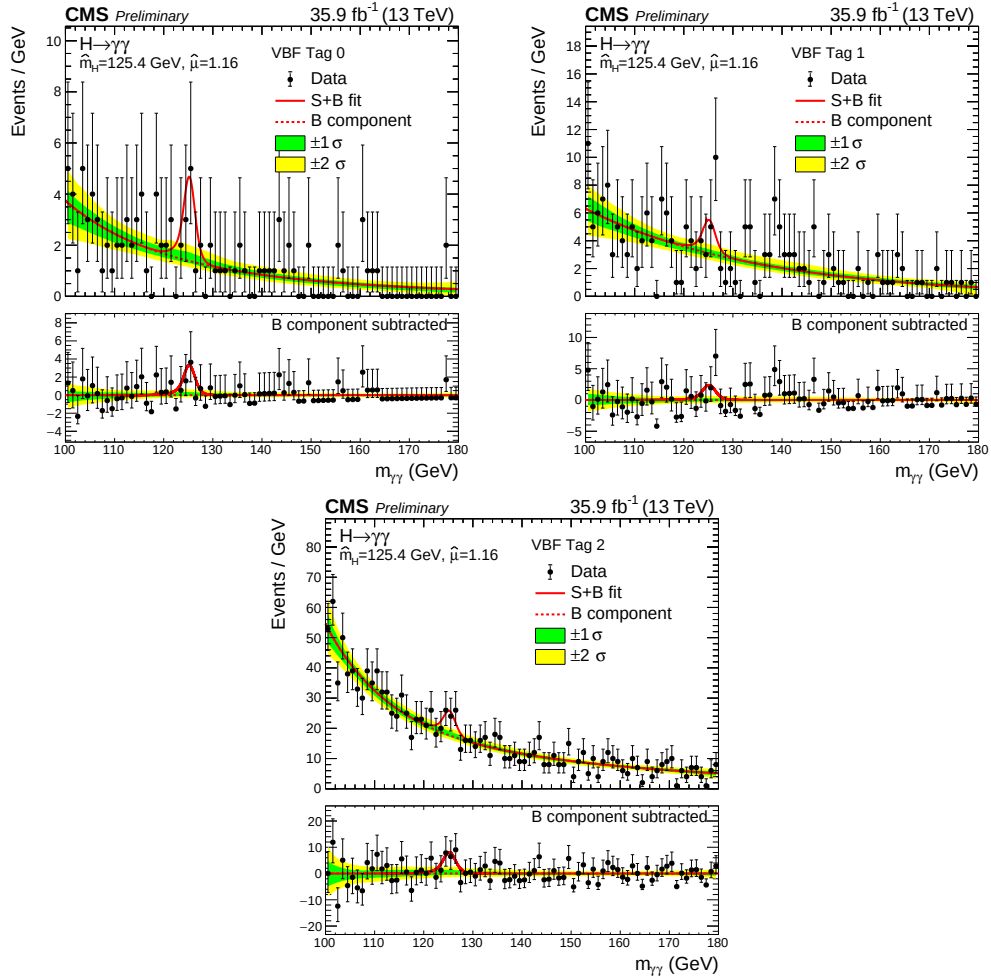


Figure 9.2: The unblinded distribution of $m_{\gamma\gamma}$ in data in each VBF category with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

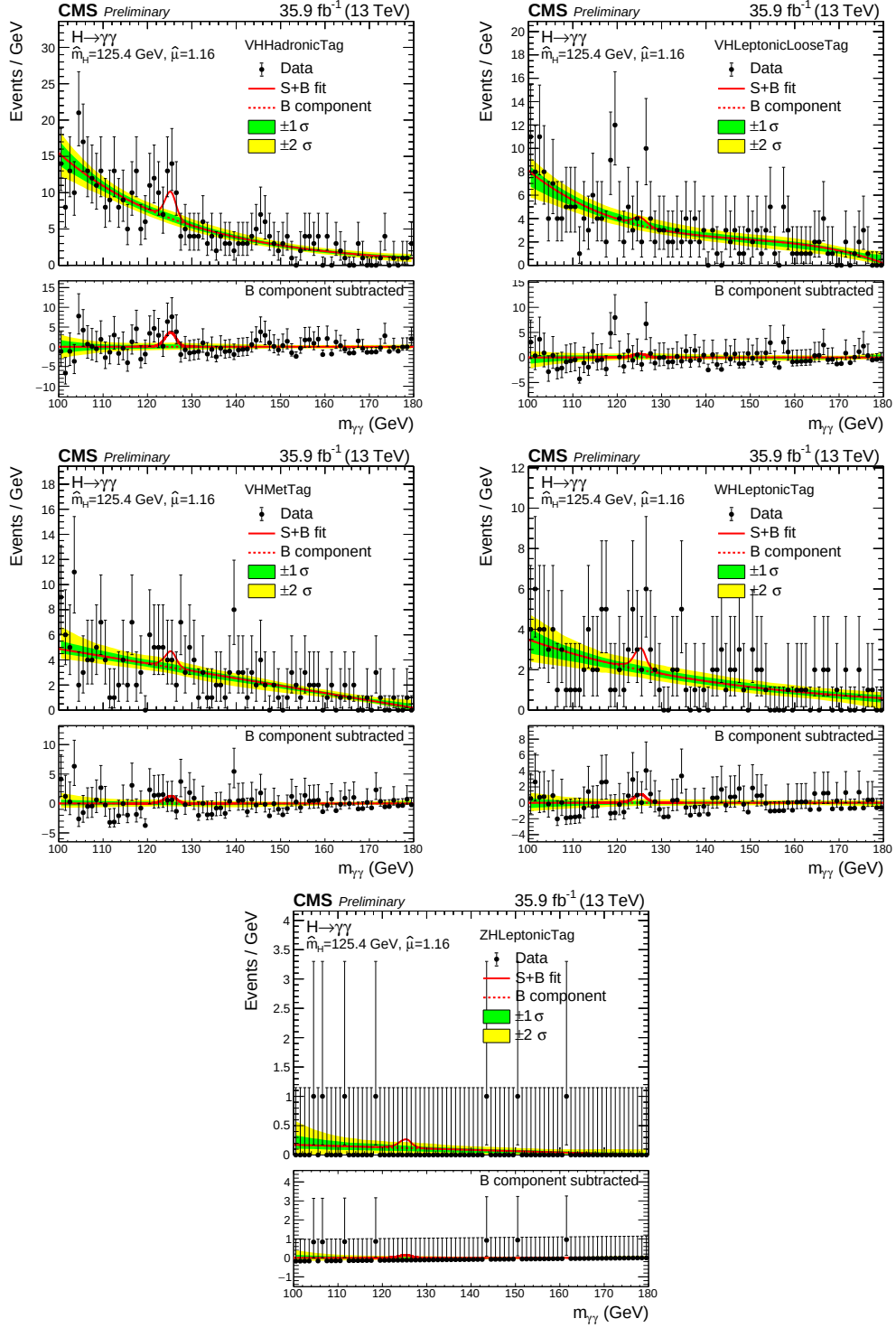


Figure 9.3: The unblinded distribution of $m_{\gamma\gamma}$ in data in each VH category with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

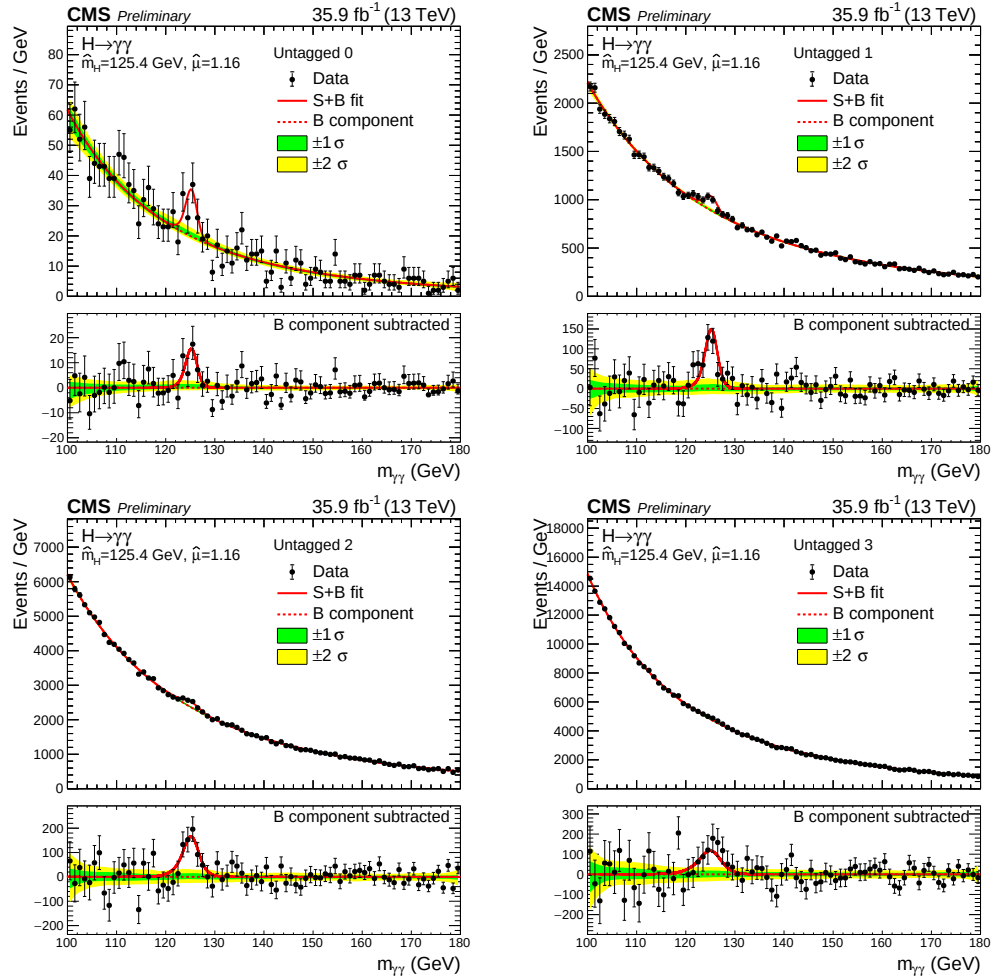


Figure 9.4: The unblinded distribution of $m_{\gamma\gamma}$ in data in each untagged category with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

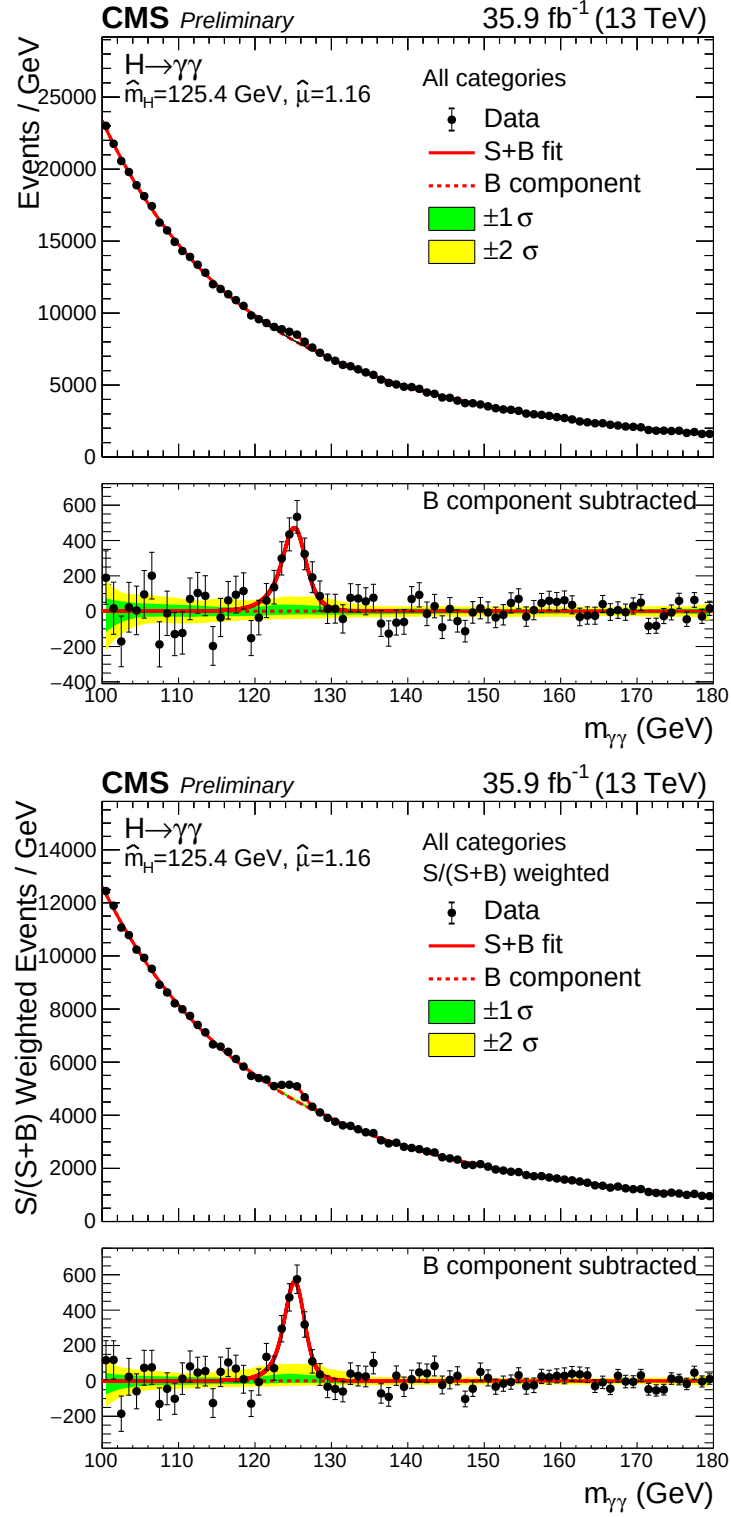


Figure 9.5: The unblinded distribution of $m_{\gamma\gamma}$ in data in all categories summed together (top) and all categories summed and weighted by their sensitivity (bottom) with the final background-only and signal plus background fits. The green and yellow bands give a measure of the statistical uncertainty in each 1 GeV bin. The corresponding signal model for each category is also shown.

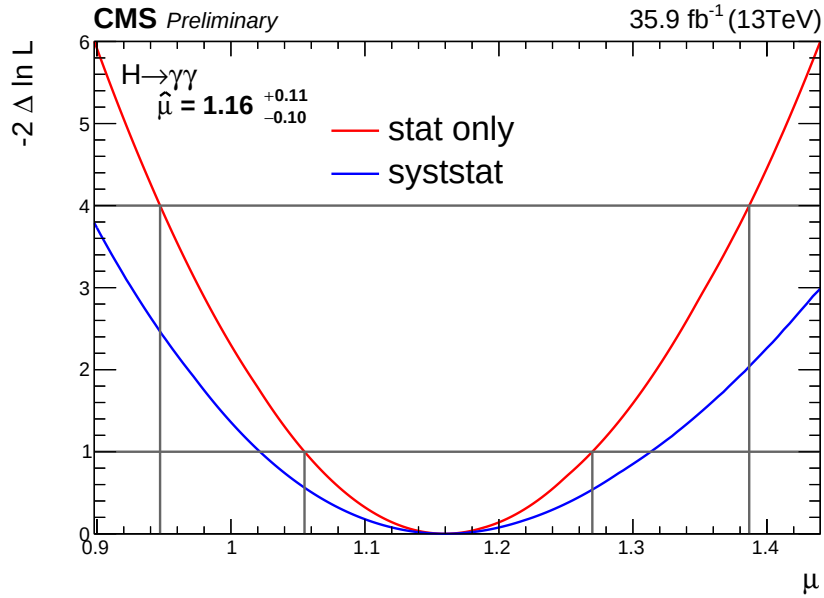


Figure 9.6: The observed measurement of the signal-strength μ with m_H profiled. The scans with both statistical uncertainties only and full systematics are shown in red and blue respectively.

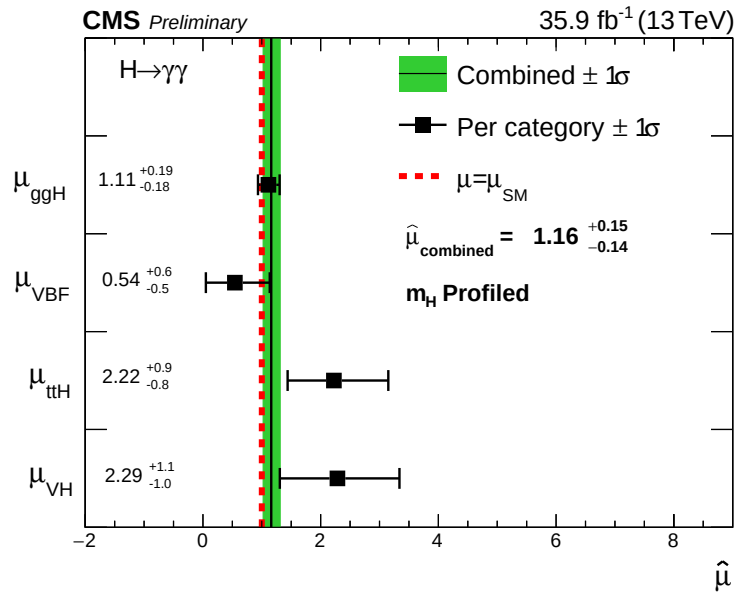


Figure 9.7: The signal strength (black) measured for each analysis category (top) and for each process (bottom) with profiled m_H , compared to the overall signal strength (green band) and to the expected value in the standard model (red dashed line).

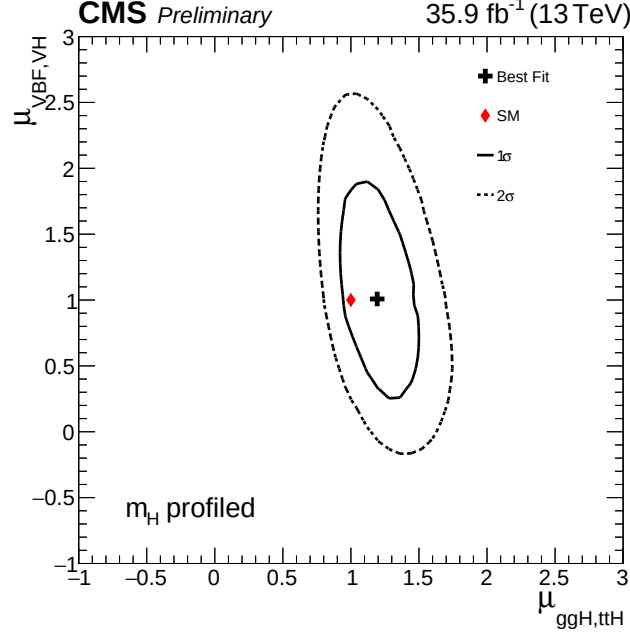


Figure 9.8: The two-dimensional best-fit (black cross) of the signal strength for the fermionic and bosonic production modes compared to the standard model expectation (red diamond). The mass of the Higgs boson is profiled in the fit, the solid line represents 1σ (68%) confidence level and dashed line 2σ (95%).

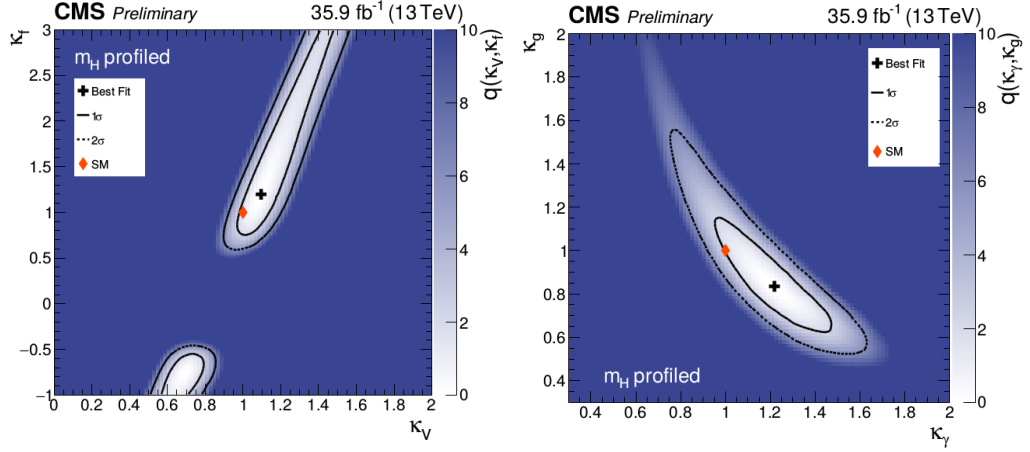


Figure 9.9: Top: The observed two-dimensional scan of the fermionic ($\mu_{\text{ggH},\text{ttH}}$) and vector boson ($\mu_{\text{VBF},\text{VH}}$) signal strengths relative to the SM prediction. The one and two sigma uncertainty contours are shown in the solid and dashed lines respectively. Bottom: The observed two-dimensional scan of the κ_g vs κ_γ . The one and two sigma uncertainty contours are shown in the solid and dashed lines respectively.

9.2 Conclusion

In this chapter, the properties of the standard model Higgs boson decaying into two photons are shown. The best fit signal strength obtained profiling m_H is found to be $\mu = 1.16^{+0.15}_{-0.14} = 1.16^{+0.11}_{-0.10}(\text{stat.})^{+0.09}_{-0.08}(\text{syst.})^{+0.06}_{-0.05}(\text{theo.})$. The best-fit values for the signal strength modifiers associated with the ggH and $t\bar{t}H$ production mechanisms, and with the VBF and VH production processes are measured; the best fit values for each modifier are $\mu_{\text{ggH}}, \mu_{t\bar{t}H} = 1.19^{+0.20}_{-0.18}$ and $\mu_{\text{VH}}, \mu_{\text{VBF}} = 1.01^{+0.57}_{-0.51}$. When $t\bar{t}H$ process is considered separately, the best-fit value is $\mu_{t\bar{t}H} = 2.22^{+0.9}_{-0.8}$, corresponding to a 3.3σ excess with respect to the absence of $t\bar{t}H$ production mode, and a 1.6σ excess with respect to the SM $t\bar{t}H$ prediction. The uncertainty on $\mu_{t\bar{t}H}$ is dominated by statistical component. To further improve the sensitivity in $t\bar{t}H$ leptonic channel, electron mini-isolation will be used. The studies of jets and photons kinematic variables are ongoing to include them into $t\bar{t}H$ MVA, therefore improve the sensitivity in $t\bar{t}H$ hadronic channel.

Conclusion

The discovery of the Higgs boson with a mass around 125 GeV has been announced by the ATLAS and the CMS collaborations on 4 July 2012. After the discovery, the analysis have moved to the measurement of the Higgs boson's properties. In this thesis, its properties in the $H \rightarrow \gamma\gamma$ decay channel are presented.

The central role of the electromagnetic calorimeter (ECAL) in the $H \rightarrow \gamma\gamma$ analysis has been described, paying special attention to the corrections for time-dependent response changes done by the laser monitoring system. During the LHC Phase II, ECAL will continue playing a crucial role in many physics analysis, including the $H \rightarrow \gamma\gamma$ one. The main concern for the current monitoring system is its radiation tolerance, which is most important for the light distribution system, and the PN diodes. A possible upgrade for the laser monitoring system is presented along with a study of its transparency measurement precision. It seems that fibres ageing effect occurs due to radiation. It has to be modelled and corrected. If after the correction the precision of the transparency measurement is higher than 1%, the current laser monitoring system has to be upgraded.

The $H \rightarrow \gamma\gamma$ analysis is described, with a particular focus on my personal contributions: vertex identification, photon identification and $t\bar{t}H$ analysis study. In the 2016 data, the vertex identification efficiency is 81%. The efficiency of the CMS general photon identification algorithm done with EGamma group was improved by 30%-10% in the barrel and by 15%-10% in the endcaps with respect to its previous version.

The $t\bar{t}H$, $H \rightarrow \gamma\gamma$ analysis was improved to achieve better performance. In the hadronic channel ggH contamination reduced by 7% and the sensitivity increased by 20%, while in the leptonic channel the sensitivity increased by 7%.

The best fit signal strength in $H \rightarrow \gamma\gamma$ channel was found to be $\mu = 1.16^{+0.15}_{-0.14} = 1.16^{+0.11}_{-0.10}(\text{stat.})^{+0.09}_{-0.08}(\text{syst.})^{+0.06}_{-0.05}(\text{theo.})$. The best-fit values for the signal strength modifiers associated with the ggH and $t\bar{t}H$ production mechanisms, and with the VBF and VH production processes were measured as well. The best fit values for each modifier are $\mu_{\text{ggH}}, \mu_{t\bar{t}H} = 1.19^{+0.20}_{-0.18}$ and $\mu_{\text{VH}}, \mu_{\text{VBF}} = 1.01^{+0.57}_{-0.51}$. When $t\bar{t}H$ process is considered separately, the best-fit value is $\mu_{t\bar{t}H} = 2.22^{+0.9}_{-0.8}$, corresponding to a 3.3σ excess over background-only hypothesis, and a 1.6σ excess with respect to the SM $t\bar{t}H$ prediction.

While this thesis work is approaching to the end, further explorations of the properties

of the Higgs boson using 13 TeV data are ongoing, especially on its mass measurement using an improved data reconstruction. This measurement will help to shed light on the (meta)stability of the electroweak vacuum [63, 64]. Further measurements of $\mu_{t\bar{t}H}$ accumulating more data will reveal its compatibility with the SM. The $t\bar{t}H$ production also opens a new field of SM tests. For instance CP violation can be probed via the top Yukawa coupling. Disentangling the scalar and pseudoscalar components can be done using a combination of measurements of $t\bar{t}H$, tH , and $\bar{t}H$ production modes [198].

Appendix A

Statistical methodology

To quantify or reject the presence of a signal in data and to estimate parameters, a common statistical procedure is used at the LHC. This methodology was developed in early Run I by the CMS and ATLAS collaborations in the context of the LHC Higgs Combination Group [194], in order to coordinate searches for the SM Higgs boson [111].

Likelihood function. The likelihood function is a function, which quantifies how the assumption of the theory is compatible with observed data. The method of maximum likelihood is a technique for estimating the values of the parameters given a finite sample of data [146].

To extract a result a simultaneous binned maximum-likelihood fit to the diphoton invariant mass distributions in all the event classes is performed over the range $100 < m_{\gamma\gamma} < 180$ GeV. Binned fits are used to speed up the computation. The chosen bin size, 250 MeV, is sufficiently small compared to the mass resolution so that no information is lost. The signal model is derived from MC simulation as described in section 5.13.1. The background is evaluated on the diphoton mass distribution, by fitting it in data and without any reference to simulation as discussed in section 5.13.2.

$$\mathcal{L} = \mathcal{L}(\text{data} | s(p, m_{\gamma\gamma}) + f(m_{\gamma\gamma})) \quad (\text{A.1})$$

where p represents the parameters of the signal model, that are allowed to vary in the fit, $s(p, m_{\gamma\gamma})$ is the parametric signal model and $f(m_{\gamma\gamma})$ the background fit function.

It runs over all channels considered in the analysis. The $\mathcal{L}$ functions stand for products of Poisson probabilities for the number of events in every bin of their channel:

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \prod_{i=1}^M \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \quad (\text{A.2})$$

where b_i is the background contribution, s_i the signal contribution according to the

nominal value of the model, and μ is the signal strength parameter; therefore $\mu=0$ corresponds to the background only hypothesis, while $\mu=1$ reproduces the SM signal hypothesis. The likelihood depends on the strength parameter μ and a set of nuisance parameters of the signal model collectively denoted θ , that are allowed to vary in the fit.

For the $H \rightarrow \gamma\gamma$ analysis a discrete profiling method, has been developed [179] to treat the uncertainty associated with the choice of the function used to fit the background. The choice of the function used to fit the background, in any particular event class, is included as a discrete nuisance parameter in the likelihood function used to extract the result.

Quantifying an excess. The test statistic is used to estimate how signal-like or background-like the data are. It is based on the profile likelihood ratio:

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \quad (\text{A.3})$$

where $\hat{\hat{\theta}}$ is the value of θ maximizing $\mathcal{L}$ for a given μ ; in other words, $\hat{\hat{\theta}}$ is the conditional maximum-likelihood estimator of θ and consequently is a function of μ itself. The denominator, instead, is maximized in an unconstrained way, thus $\hat{\mu}$ and $\hat{\theta}$ correspond to the true global maximum likelihood. Hence, by definition, the profile likelihood ratio is comprised between 1, when the hypothesized μ coincides with $\hat{\mu}$, showing thus great compatibility between the data and the hypothesis, and 0 when instead the assumed μ is at odds with $\hat{\mu}$ denoting in this way a high degree of incompatibility between the data and the hypothesis.

When trying to quantify the statistical significance of an excess over the background-only expectation hypothesis ($\mu = 0$), the following test statistic therefore is used:

$$\lambda(\mu = 0) = \frac{\mathcal{L}(0, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \quad (\text{A.4})$$

and the significance of an excess over the background-only expectation can be defined as :

$$s = \sqrt{-2 \text{Log} \lambda(\mu = 0)} \quad (\text{A.5})$$

Bibliography

- [1] A. Pais and R. H. Stuewer. Inward bound: Of matter and forces in the physical world. *Physics Today*, 39(8):61–62, aug 1986.
- [2] W. C. Röntgen. Ueber eine neue Art von Strahlen. *Sitzber.Physik.Med.Ges*, 137:1, 1895.
- [3] J.J Thompson. Cathode rays. *Phil.Mag.*, 44:293, 1897.
- [4] J.J Thompson. Cathode rays. *Nature*, 55:453, 1897.
- [5] E.Rutherford. Uranium radiation and the electrical conduction produced by it. *Phil.Mag.*, 47:109, 1899.
- [6] P.Villard. Sur le rayonnement du Radium. *Comp.Ren.*, 130:1178, 1900.
- [7] E.Rutherford. The scattering of α – and β – particles by matter and the structure of the atom. *Phil.Mag.*, 21:669, 1911.
- [8] E.Rutherford. Collision of α – particles with light atoms IV. An anomalous effect in Nitrogen. *Phil.Mag.*, 37:581, 1919.
- [9] W. Bothe and H. Becker. Künstliche Erregung von Kern- γ -Strahlen. *Zeitschrift für Physik*, 66(5-6):289–306, may 1930.
- [10] J. Chadwick. Possible existence of a neutron. *Nature*, 129(3252):312–312, feb 1932.
- [11] J. Chadwick. The existence of a neutron. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 136(830):692–708, jun 1932.
- [12] M.Planck. Zur theorie des gesetzes der energieverteilung im normalspektrum. *Verhandl. Dtsch. phys. Ges.*, 2:237, 1900.
- [13] A. Einstein. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, 322(6):132–148, 1905.

- [14] A.Einstein. Zur elektrodynamik bewegter körper. *Annalen der Physik. Leipzig*, 17:891, 1905.
- [15] A.Einstein. Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig. *Annalen der Physik. Leipzig*, 18:639, 1905.
- [16] A.Einstein. Die Feldgleichungen der Gravitation. *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin*, pages 844–847, 1915.
- [17] W. Heisenberg. Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen. *Zeitschrift für Physik*, 33(1):879–893, dec 1925.
- [18] M. Born, W. Heisenberg, and P. Jordan. Zur quantenmechanik. II. *Zeitschrift für Physik*, 35(8-9):557–615, aug 1926.
- [19] E. Schrödinger. Quantisation as a problem of proper values. part I. *Annalen der Physik. Leipzig*, 79:361, 1926.
- [20] E. Schrödinger. Quantisation as a problem of proper values. part II. *Annalen der Physik. Leipzig*, 79:489, 1926.
- [21] E. Schrödinger. On the relation between the quantum. mechanics of Heisenberg, Born, and Jordan, and that of Schrödinger. *Annalen der Physik. Leipzig*, 79:734, 1926.
- [22] G.E. Uhlenbeck and S. Goudsmit. Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons. *Naturwissenschaften*, 47:953–954, 1925.
- [23] P. A. M. Dirac. Quantised singularities in the electromagnetic field. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 133(821):60–72, sep 1931.
- [24] C. D. Anderson. The positive electron. *Physical Review*, 43(6):491–494, mar 1933.
- [25] S. H. Neddermeyer and C. D. Anderson. Note on the nature of cosmic-ray particles. *Physical Review*, 51(10):884–886, may 1937.
- [26] E.G Street, J.C.; Stevenson. Penetrating corpuscular component of the cosmic radiation. *Phys. Rev.*, 51:1005, 1937.
- [27] G. P. S. Occhialini and C. F. Powell. Nuclear disintegrations produced by slow charged particles of small mass. *Nature*, 159(4032):186–190, feb 1947.
- [28] E. Fermi. Versuch einer Theorie der β -Strahlen. I. *Zeitschrift für Physik*, 88(3-4):161–177, mar 1934.

-
- [29] E. C. G. Sudarshan and R. E. Marshak. Chirality invariance and the universal Fermi interaction. *Physical Review*, 109(5):1860–1862, mar 1958.
 - [30] R. P. Feynman and M. Gell-Mann. Theory of the fermi interaction. *Physical Review*, 109(1):193–198, jan 1958.
 - [31] C. N. Yang and R. L. Mills. Conservation of isotopic spin and isotopic gauge invariance. *Physical Review*, 96(1):191–195, oct 1954.
 - [32] M. Gell-Mann. A schematic model of baryons and mesons. *Physics Letters*, 8(3):214–215, feb 1964.
 - [33] G. Zweig. An $su(3)$ model for strong interaction symmetry and its breaking. Version 2. *Developments in the Quark Theory of Hadrons*, 1:22–101, 1964.
 - [34] E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, G. Miller, L. W. Mo, R. E. Taylor, M. Breidenbach, J. I. Friedman, G. C. Hartmann, and H. W. Kendall. High-energy inelastic e-p scattering at 6° and 10° . *Physical Review Letters*, 23(16):930–934, oct 1969.
 - [35] M. Breidenbach, J. I. Friedman, H. W. Kendall, E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, L. W. Mo, and R. E. Taylor. Observed behavior of highly inelastic electron-proton scattering. *Physical Review Letters*, 23(16):935–939, oct 1969.
 - [36] S. L. Glashow. Partial-symmetries of weak interactions. *Nuclear Physics*, 22(4):579–588, feb 1961.
 - [37] S. Weinberg. A model of leptons. *Physical Review Letters*, 19(21):1264–1266, nov 1967.
 - [38] A. Salam. Weak and electromagnetic interactions. *Conf.Proc.*, C680519:367–377, 1968.
 - [39] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters*, 13(9):321–323, aug 1964.
 - [40] P. W. Higgs. Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, 13(16):508–509, oct 1964.
 - [41] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. *Physical Review Letters*, 13(20):585–587, nov 1964.
 - [42] D. J. Gross and F. Wilczek. Ultraviolet behavior of non-abelian gauge theories. *Physical Review Letters*, 30(26):1343–1346, jun 1973.

- [43] H. D. Politzer. Reliable perturbative results for strong interactions? *Physical Review Letters*, 30(26):1346–1349, jun 1973.
- [44] G. 't Hooft and M. Veltman. Regularization and renormalization of gauge fields. *Nuclear Physics B*, 44(1):189–213, jul 1972.
- [45] G. 't Hooft and M. Veltman. Combinatorics of gauge fields. *Nuclear Physics B*, 50(1):318–353, jan 1972.
- [46] M. L. Perl et al. Evidence for anomalous lepton production in $e^+ - e^-$ annihilation. *Physical Review Letters*, 35(22):1489–1492, dec 1975.
- [47] M.L. Perl et al. Properties of the proposed τ charged lepton. *Physics Letters B*, 70(4):487–490, oct 1977.
- [48] R. Brandelik et al. Evidence for a spin-1 gluon in three-jet events. *Physics Letters B*, 97(3-4):453–458, dec 1980.
- [49] Ch. Berger et al. A study of multi-jet events in e^+e^- annihilation. *Physics Letters B*, 97(3-4):459–464, dec 1980.
- [50] G. Arnison et al. Experimental observation of isolated large transverse energy electrons with associated missing energy at $s = 540$ GeV. *Physics Letters B*, 122(1):103–116, feb 1983.
- [51] G. Arnison et al. Observation of the muonic decay of the charged intermediate vector boson. *Physics Letters B*, 134(6):469–476, jan 1984.
- [52] G. Arnison et al. Experimental observation of lepton pairs of invariant mass around 95 GeV/c² at the CERN SPS collider. *Physics Letters B*, 126(5):398–410, jul 1983.
- [53] G. Arnison et al. Observation of muonic Z^0 -decay at the $\bar{p}p$ collider. *Physics Letters B*, 147(1-3):241–248, nov 1984.
- [54] F. Abe et al. Observation of top quark production $\bar{p}p$ collisions with the collider detector at Fermilab. *Physical Review Letters*, 74(14):2626–2631, apr 1995.
- [55] S. Abachi et al. Observation of the Top Quark. *Physical Review Letters*, 74(14):2632–2637, apr 1995.
- [56] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016.
- [57] LHCb Collaboration. Observation of the resonant character of the Z(4430)-state. *Physical Review Letters*, 112(22), jun 2014.

- [58] LHCb Collaboration. Observation of $J/\psi p$ Resonances consistent with pentaquark states in $\lambda_b^0 \rightarrow J/\psi K^- p$ decays. *Physical Review Letters*, 115(7), aug 2015.
- [59] L. Álvarez-Gaumé and J. Ellis. Eyes on a prize particle. *Nature Physics*, 7(1):2–3, jan 2011.
- [60] J. Goldstone. Field theories with « superconductor » solutions. *Il Nuovo Cimento*, 19(1):154–164, jan 1961.
- [61] J., A. Salam, and S. Weinberg. Broken symmetries. *Physical Review*, 127(3):965–970, aug 1962.
- [62] Stephen P. Martin. A supersymmetry primer. In *Perspectives on Supersymmetry*, pages 1–98. World Scientific, jul 1998.
- [63] Giuseppe Degrandi, Stefano Di Vita, Joan Elias-Miró, José R. Espinosa, Gian F. Giudice, Gino Isidori, and Alessandro Strumia. Higgs mass and vacuum stability in the Standard Model at NNLO. *Journal of High Energy Physics*, 2012(8), aug 2012.
- [64] John Ellis. The Physics Landscape after the Higgs Discovery at the LHC, 2015.
- [65] M. Beyer. *CP Violation in Particle, Nuclear, and Astrophysics (Lecture Notes in Physics)*. Springer, 2010.
- [66] Ziro Maki, Masami Nakagawa, and Shoichi Sakata. Remarks on the unified model of elementary particles. *Progress of Theoretical Physics*, 28(5):870–880, nov 1962.
- [67] B. Pontecorvo. Inverse beta processes and nonconservation of lepton charge. *Sov. Phys. JETP*, 7:172–173, 1958. [Zh. Eksp. Teor. Fiz.34,247(1957)].
- [68] Richard P. Feynman, Fernando B. Morinigo, and William G. Wagner. *Feynman Lectures on Gravitation*. Addison-Wesley, 1995.
- [69] Plank collaboration. Planck2015 results. *Astronomy & Astrophysics*, 594:A13, sep 2016.
- [70] Yupeng Yang. Constraints on the basic parameters of dark matter using the planck data. *Physical Review D*, 91(8), apr 2015.
- [71] Gianfranco Bertone, Dan Hooper, and Joseph Silk. Particle dark matter: evidence, candidates and constraints. *Physics Reports*, 405(5-6):279–390, jan 2005.
- [72] Plank collaboration. Planck2015 results. *Astronomy & Astrophysics*, 594:A1, sep 2016.
- [73] R. Adam and al. Planck2015 results. *Astronomy & Astrophysics*, 594:A1, sep 2016.

- [74] P. J. E. Peebles and Bharat Ratra. The cosmological constant and dark energy. *Reviews of Modern Physics*, 75(2):559–606, apr 2003.
- [75] Giuliano Panico and Andrea Wulzer. *The Composite Nambu-Goldstone Higgs (Lecture Notes in Physics)*. Springer, 2015.
- [76] Roberto Contino, Christophe Grojean, Mauro Moretti, Fulvio Piccinini, and Riccardo Rattazzi. Strong double higgs production at the LHC. *Journal of High Energy Physics*, 2010(5), may 2010.
- [77] Gian Francesco Giudice, Christophe Grojean, Alex Pomarol, and Riccardo Rattazzi. The strongly-interacting light higgs. *Journal of High Energy Physics*, 2007(06):045–045, jun 2007.
- [78] Adam Falkowski. Effective Theory Approach to Physics Beyond the Standard Model, 2017.
- [79] Nima Arkani-Hamed, Andrew G. Cohen, and Howard Georgi. Electroweak symmetry breaking from dimensional deconstruction. *Physics Letters B*, 513(1-2):232–240, jul 2001.
- [80] Nima Arkani-Hamed, Andrew G Cohen, Emanuel Katz, and Ann E Nelson. The littlest Higgs. *Journal of High Energy Physics*, 2002(07):034–034, jul 2002.
- [81] Roberto Contino, Leandro Da Rold, and Alex Pomarol. Light custodians in natural composite Higgs models. *Physical Review D*, 75(5), mar 2007.
- [82] L. Evans and P. Bryant. LHC machine. *Journal of Instrumentation*, 3(08):S08001–S08001, aug 2008.
- [83] LHC Study Group. The Large Hadron Collider: Conceptual design. *CERN-AC-95-05-LHC*, 1995.
- [84] Valery Lebedev and Vladimir Shiltsev, editors. *Accelerator Physics at the Tevatron Collider*. Springer New York, 2014.
- [85] Stephen Holmes, Ronald S Moore, and Vladimir Shiltsev. Overview of the Tevatron collider complex: goals, operations and performance. *Journal of Instrumentation*, 6(08):T08001–T08001, aug 2011.
- [86] W.J. Stirling. Parton luminosity and cross section plots, Private communication, 2016.
- [87] Cinzia De Melis. The CERN accelerator complex. Complexe des accélérateurs du CERN. Jul 2016. General Photo.

-
- [88] Jean-Luc Caron. LHC Layout.. Schema general du LHC. AC Collection. Legacy of AC. Pictures from 1992 to 2002., Sep 1997.
 - [89] LHC Team. Lhc Report: end of 2016 proton-proton operation, 2016.
 - [90] A. Djouadi. The anatomy of electroweak symmetry breaking. *Physics Reports*, 457(1-4):1–216, feb 2008.
 - [91] Federico Demartin, Fabio Maltoni, Kentarou Mawatari, and Marco Zaro. Higgs production in association with a single top quark at the LHC. *The European Physical Journal C*, 75(6), jun 2015.
 - [92] K. A. Assamagan et al. The higgs working group: Summary report 2003, 2004.
 - [93] Robert V. Harlander and William B. Kilgore. Higgs boson production in bottom quark fusion at next-to-next-to-leading order. *Physical Review D*, 68(1), jul 2003.
 - [94] S. Dawson, C. B. Jackson, L. Reina, and D. Wackeroth. Higgs-boson production with one bottom-quark jet at hadron colliders. *Physical Review Letters*, 94(3), jan 2005.
 - [95] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 4. deciphering the nature of the Higgs sector. 2016.
 - [96] The LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 2013.
 - [97] G. Ridolfi. Search for the Higgs boson: theoretical perspectives, 2001.
 - [98] Kenneth G. Wilson. Renormalization group and critical phenomena. II. phase-space cell analysis of critical behavior. *Physical Review B*, 4(9):3184–3205, nov 1971.
 - [99] K Wilson. The renormalization group and the ϵ expansion. *Physics Reports*, 12(2):75–199, aug 1974.
 - [100] Roger Dashen and Herbert Neuberger. How to get an upper bound on the Higgs mass. *Physical Review Letters*, 50(24):1897–1900, jun 1983.
 - [101] Julius Kuti, Lee Lin, and Yue Shen. Upper bound on the Higgs-boson mass in the Standard Model. *Physical Review Letters*, 61(6):678–681, aug 1988.
 - [102] R.Sekhar Chivukula and Elizabeth H. Simmons. Custodial symmetry and the triviality bound on the Higgs mass. *Physics Letters B*, 388(4):788–792, nov 1996.
 - [103] A. D. Linde. Dynamical Symmetry Restoration and Constraints on Masses and Coupling Constants in Gauge Theories. *JETP Lett.*, 23:64–67, 1976. [Pisma Zh. Eksp. Teor. Fiz.23,73(1976)].

- [104] Steven Weinberg. Mass of the Higgs boson. *Physical Review Letters*, 36(6):294–296, feb 1976.
- [105] S. Coleman and E. Weinberg. Radiative corrections as the origin of spontaneous symmetry breaking. *Physical Review D*, 7(6):1888–1910, mar 1973.
- [106] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, and The LEP Working Group for Higgs Boson Searches. Search for the Standard Model Higgs boson at LEP. *Physics Letters B*, 565:61–75, jul 2003.
- [107] CDF and D0 Collaborations. Combination of Tevatron searches for the Standard Model Higgs boson in the W^+W^- decay mode. *Physical Review Letters*, 104(6), feb 2010.
- [108] ALEPH Collaboration, CDF Collaboration, D0 Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, SLD Collaboration, LEP Electroweak Working Group, Tevatron Electroweak Working Group, and SLD electroweak heavy flavour group. Precision electroweak measurements and constraints on the Standard Model. 2010.
- [109] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, SLD Collaboration, LEP Electroweak Working Group, and heavy flavour groups SLD electroweak. Precision electroweak measurements on the z resonance. *Physics Reports*, 427(5-6):257–454, may 2006.
- [110] J. F. de Trocóniz and F. J. Ynduráin. Hadronic contributions to the anomalous magnetic moment of the muon. *Physical Review D*, 71(7), apr 2005.
- [111] CMS Collaboration. Combined results of searches for the Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Physics Letters B*, 710(1):26–48, mar 2012.
- [112] ATLAS Collaboration. Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Physical Review D*, 86(3), aug 2012.
- [113] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, sep 2012.
- [114] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, sep 2012.
- [115] CMS Collaboration. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *Journal of High Energy Physics*, 2013(6), jun 2013.

- [116] CDF and D0 Collaborations. Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron. *Physical Review Letters*, 109(7), aug 2012.
- [117] CDF and D0 Collaborations. Higgs boson studies at the Tevatron. *Physical Review D*, 88(5), sep 2013.
- [118] CMS Collaboration. Measurement of the properties of a Higgs boson in the four-lepton final state. *Physical Review D*, 89(9), may 2014.
- [119] ATLAS Collaboration. Measurement of the Higgs boson mass from the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4l$ channels in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector. *Physical Review D*, 90(5), sep 2014.
- [120] CMS Collaboration. Observation of the diphoton decay of the Higgs boson and measurement of its properties. *The European Physical Journal C*, 74(10), oct 2014.
- [121] CMS Collaboration. Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the Standard Model predictions using proton collisions at 7 and 8 TeV. *The European Physical Journal C*, 75(5), may 2015.
- [122] CMS and ATLAS Collaborations. Combined measurement of the Higgs boson mass in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments. *Physical Review Letters*, 114(19), may 2015.
- [123] CMS Collaboration. Study of the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs. *Physical Review Letters*, 110(8), feb 2013.
- [124] ATLAS Collaboration. Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment. *The European Physical Journal C*, 76(1), jan 2016.
- [125] CMS and ATLAS Collaborations. 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. *Journal of High Energy Physics*, 2016(8), aug 2016.
- [126] CDF and D0 Collaborations. Combination of the top-quark mass measurements from the Tevatron collider. *Physical Review D*, 86(9), nov 2012.
- [127] ATLAS Collaboration. Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC. *Physics Letters B*, 726(1-3):88–119, oct 2013.

- [128] ATLAS Collaboration. Evidence for the spin-0 nature of the Higgs boson using ATLAS data. *Physics Letters B*, 726(1-3):120–144, oct 2013.
- [129] CMS Collaboration. Search for the associated production of the Higgs boson with a top-quark pair. *Journal of High Energy Physics*, 2014(9), sep 2014.
- [130] CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004–S08004, aug 2008.
- [131] CMS Collaboration. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.
- [132] W Adam, R Frühwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9–N20, jul 2005.
- [133] CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9(10):P10009–P10009, oct 2014.
- [134] CMS Collaboration. *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.
- [135] CMS Collaboration. *The CMS hadron calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.
- [136] Victor Daniel Elvira. Measurement of the Pion Energy Response and Resolution in the CMS HCAL Test Beam 2002 Experiment. Technical Report CMS-NOTE-2004-020, CERN, Geneva, Sep 2004.
- [137] CMS Collaboration. *The CMS muon project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.
- [138] CMS Collaboration. Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV. *Journal of Instrumentation*, 7(10):P10002–P10002, oct 2012.
- [139] CMS Collaboration. The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s} = 7$ TeV at the LHC. *Journal of Instrumentation*, 8(11):P11002–P11002, nov 2013.
- [140] CMS Collaboration. The CMS trigger system. *Journal of Instrumentation*, 12(01):P01020–P01020, jan 2017.
- [141] CMS Collaboration. Run 2 upgrades to the CMS level-1 calorimeter trigger. *Journal of Instrumentation*, 11(01):C01051–C01051, jan 2016.

-
- [142] CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector. Submitted to *JINST*, 2017.
 - [143] CMS Collaboration. Commissioning of the Particle-Flow Event Reconstruction with the first LHC collisions recorded in the CMS detector. Technical Report CMS-PAS-PFT-10-001, 2010.
 - [144] CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9(10):P10009–P10009, oct 2014.
 - [145] CMS Collaboration. Energy resolution of the barrel of the CMS electromagnetic calorimeter. *Journal of Instrumentation*, 2(04):P04004–P04004, apr 2007.
 - [146] Glen Cowan. *Statistical Data Analysis (Oxford Science Publications)*. Clarendon Press, 1998.
 - [147] Emanuele Di Marco. CMS electromagnetic calorimeter calibration and timing performance during LHC Run I and future prospects. Technical Report CMS-CR-2014-410, CERN, Geneva, Nov 2014.
 - [148] Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV. *Journal of Instrumentation*, 8(09):P09009–P09009, sep 2013.
 - [149] CMS collaboration. Laser monitoring system for the CMS lead tungstate crystal calorimeter. Technical Report CMS-NOTE-2007-028, CERN, Geneva, Nov 2007.
 - [150] CMS Collaboration. THE CMS ECAL Laser Monitoring System. In *Astroparticle, Particle and Space Physics, Detectors and Medical Physics Applications*. World Scientific, apr 2006.
 - [151] CMS ECAL energy inter-calibration precision with 2015 data. May 2016.
 - [152] Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV. *Journal of Instrumentation*, 8(09):P09009–P09009, sep 2013.
 - [153] Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. *Journal of Instrumentation*, 10(06):P06005–P06005, jun 2015.
 - [154] 2015 ECAL detector performance plots. Dec 2015.

- [155] M. Dejardin. Determination of the single pulse response of the CMS-ECAL electronics and application to the pulse shape reconstruction. Technical Report CMS detector note 2008-001, CERN, Geneva, Jan 2008.
- [156] ECAL Laser monitoring till end of 2016 and ECAL phi-symmetry. Feb 2017.
- [157] Hamamatsu. High speed photo-diodes S5973 series: 1GHz. Technical Report https://www.hamamatsu.com/resources/pdf/ssd/s5971_etc_kpin1025e.pdf.
- [158] Analog devices. ADG904 Wideband 2.5 GHz, 37 dB Isolation at 1 GHz, CMOS 1.65 V to 2.75 V, 4:1 Mux/SP4T. Technical Report <http://www.analog.com/media/en/technical-documentation/data-sheets/ADG904.pdf>.
- [159] First results on Higgs to gammagamma at 13 TeV. Technical Report CMS-PAS-HIG-15-005, CERN, Geneva, 2016.
- [160] Updated measurements of Higgs boson production in the diphoton decay channel at $\sqrt{s} = 13$ TeV in pp collisions at CMS. Technical Report CMS-PAS-HIG-16-020, CERN, Geneva, 2016.
- [161] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 2014(7), jul 2014.
- [162] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to PYTHIA 8.1. *Computer Physics Communications*, 178(11):852–867, jun 2008.
- [163] LHC Higgs Cross Section Working Group Collaboration. Handbook of LHC Higgs Cross Sections: 4. Deciphering the nature of the Higgs sector, 2016.
- [164] CMS Collaboration. Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. 2015.
- [165] Emilio Meschi, T Monteiro, Christopher Seez, and Pratibha Vikas. Electron Reconstruction in the CMS Electromagnetic Calorimeter. Technical Report CMS-NOTE-2001-034, CERN, Geneva, Jun 2001.
- [166] Pierre Billoir. Progressive track recognition with a Kalman-like fitting procedure. *Computer Physics Communications*, 57(1-3):390–394, dec 1989.
- [167] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois. Electron reconstruction in CMS. *The European Physical Journal C*, 49(4):1099–1116, jan 2007.

- [168] CMS Collaboration. Measuring Electron Efficiencies at CMS with Early Data. (CMS-PAS-EGM-07-001), 2008.
- [169] CMS Collaboration. The CMS trigger system. *J. Instrum.*, 12(CMS-TRG-12-001.01):P01020. 122 p, Sep 2016.
- [170] CMS Collaboration. CMS ECAL first results with 2016 data. (CMS-DP-2016-031), Jun 2016.
- [171] John Erthal Gaiser. *Charmonium Spectroscopy From Radiative Decays of the J/ψ and ψ'* . PhD thesis, SLAC, 1982.
- [172] Shervin Nourbakhsh. *Measurement of the mass of the Higgs Boson in the two photon decay channel with the CMS experiment*. PhD thesis, Minnesota U., 2013.
- [173] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr., M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, and A. Zemla. TMVA - Toolkit for Multivariate Data Analysis. 2007.
- [174] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2):1–19, feb 2011.
- [175] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, apr 2008.
- [176] CMS Collaboration. Pileup Jet Identification. Technical Report CMS-PAS-JME-13-005, CERN, Geneva, 2013.
- [177] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The Higgs boson machine learning challenge. In *Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42*, HEPML’14, pages 19–55. JMLR.org, 2014.
- [178] CMS Collaboration. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. 2016.
- [179] P. D. Dauncey, M. Kenzie, N. Wardle, and G. J. Davies. Handling uncertainties in background shapes: the discrete profiling method. 2014.
- [180] Federico Demartin, Stefano Forte, Elisa Mariani, Juan Rojo, and Alessandro Vicini. Impact of parton distribution function and α_s uncertainties on Higgs boson production in gluon fusion at hadron colliders. *Physical Review D*, 82(1), jul 2010.

- [181] Stefano Carrazza, Stefano Forte, Zahari Kassabov, José Ignacio Latorre, and Juan Rojo. An unbiased Hessian representation for Monte Carlo PDFs. *The European Physical Journal C*, 75(8), aug 2015.
- [182] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 3. Higgs properties: Report of the LHC Higgs cross section working group, 2013.
- [183] Jon Butterworth, Stefano Carrazza, Amanda Cooper-Sarkar, Albert De Roeck, Joel Feltesse, Stefano Forte, Jun Gao, Sasha Glazov, Joey Huston, Zahari Kassabov, Ronan McNulty, Andreas Morsch, Pavel Nadolsky, Voica Radescu, Juan Rojo, and Robert Thorne. PDF4LHC recommendations for LHC Run II. 2015.
- [184] Measurement of the differential cross section for $t\bar{t}$ production in the dilepton final state at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-TOP-16-011, CERN, Geneva, 2016.
- [185] CMS Collaboration. Measurement of the cross section ratio $t\bar{t}b\bar{b}/t\bar{t}b\bar{b}+jj$ using dilepton final states in pp collisions at 13 TeV. 2016.
- [186] CMS Luminosity Measurements for the 2016 Data Taking Period. Technical Report CMS-PAS-LUM-17-001, CERN, Geneva, 2017.
- [187] Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector. Technical Report CMS-PAS-JME-16-004, CERN, Geneva, 2016.
- [188] CMS Collaboration. Identification of b-quark jets with the CMS experiment. 2012.
- [189] Mia Tosi. Performance of tracking, b-tagging and jet/MET reconstruction at the CMS high level trigger. *Journal of Physics: Conference Series*, 664(8):082055, dec 2015.
- [190] CMS collaboration. Event generator tunes obtained from underlying event and multiparton scattering measurements. *The European Physical Journal C*, 76(3), mar 2016.
- [191] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup per particle identification. *Journal of High Energy Physics*, 2014(10), oct 2014.
- [192] C. Weiser. A combined secondary vertex based B-tagging algorithm in CMS. 2006.
- [193] Identification of b quark jets at the CMS Experiment in the LHC Run 2. Technical Report CMS-PAS-BTV-15-001, CERN, Geneva, 2016.
- [194] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011.

-
- [195] Measurements of properties of the Higgs boson in the diphoton decay channel with the full 2016 data set. Technical Report CMS-PAS-HIG-16-040, CERN, Geneva, 2017.
 - [196] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables, 2011.
 - [197] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 2. Differential Distributions, 2012.
 - [198] John Ellis, Dae Sung Hwang, Kazuki Sakurai, and Michihisa Takeuchi. Disentangling Higgs-top couplings in associated production. *Journal of High Energy Physics*, 2014(4), apr 2014.

Acknowledgements

I wish to thank people working in CMS collaboration and LHC; people who have built and maintained those amazing machines, people who provide and control data, and people who work on $H \rightarrow \gamma\gamma$ analysis.

This thesis work was performed under the supervision of **Julie**. Your careful planning let me discover different components of experimental physicist's work. I thank you for your guidance at each stage, your help to overcome obstacles and become a better physicist. You have shown me how to take responsibility both in the working group and collaboration. Our discussions and your careful comments shaped my manuscript.

Gautier, I thank you for the enthusiasm, encouragement and availability even though you had many responsibilities.

Fabrice, I am grateful for your availability, brilliant ideas for our analysis, and your guidance in Egamma group. I thank you for your detailed comments, they have improved this manuscript. Also I am grateful for your very kind comments during the defense preparations.

Saranya, thank you for the team work on the analysis, and your patience to reply on my questions.

Marc D. and Jean-Louis, I received my first and very exciting experience in detector work with you! Thank you for teaching me, answering my questions with kindness and patience.

Marc B., I am grateful for our weekly group meetings, and your suggestions for our analysis. I appreciate the fact that you read my manuscript, while you were very busy. Your great comments improved a lot the style and the content of this document. Thank you for participating in all rehearsals I have done.

Nathalie, my "marraine", you always kept track of my work and were available when I was in need of sincere conversation. I am extremely grateful for reading my manuscript during your vacations, and your thoughtful comments!

I am grateful to **members of the jury** for their comments and questions, which improved my manuscript.

I wouldn't be able to arrive to the end (event to the beginning of the thesis) without **Ludwik**. You came to my university in Dnipro, and shown me the way to the world of

particle physics. I cannot thank enough for your guidance and friendship during these years.

CMS Saclay group, you became for my second family. I want to thank **Carlotta**, who helped me to understand a lot about analysis, technicalities, and collaboration itself. You've become my friend, who guided me also in the world of mountain hiking. I will never forget our trip from Gstaad to l'Etivaz (thanks **Benny** for choosing it...)! **Martina**, I am happy that we shared this hike, as well as other adventures! I thank you for the support, advice and endless conversations. I thank both of you for the "Italian spirit" we had in the group. **Guilia**, my friend and "co-bureau", I am grateful for your kindness, support and laughs we had in the office! **Clement**, thank you for telling us the incredible stories from your life, and also for your "Friday lunch" e-mails. Guilia and Clement, I wish you good luck, no... I wish you strength (both physical and moral), and belief in yourself during the year you are facing. I thank **Özgür** for sharing your contacts, and knowledge about HGCAL project, to which I am planning to contribute.

I wish to thank **all students of SPP**, who made those three years extremely bright: Héliou, Julien, Clotilde, Liam, Eloi (mille merci pour ton courage parler français avec moi au tout début de ma thèse!), Laura, Mafalda, Matthias, Marta, Francesco and Ciro. It was fantastic!

Olga merci pour ton soutien, pour nos conversations francs et intéressants! Tu me manque pendant mon dernier année, mais je suis heureuse que notre amitié continue.

Nata and Artem I am grateful for your presence during my thesis defense. It was very important and touching that you arrived despite the fact of 3 deadlines! I love you and grateful for your friendship during ... 12 years... what? Anyway, next year it is your turn, and there is no way I miss it. **Stas and Olya** thank you for coming for my defense, it was so nice! You have changed my life in enormous way as well, you know it. I will be always grateful for that.

My mom, I thank you for being available and supportive during all my life. I am infinitely grateful for the education you gave me, it was always our priority no matter what happened.

Je remercie à **Pierre** qui fait ma vie magnifique. Merci pour ton soutien, optimisme, curiosité, l'ouverture d'esprit et ta gentillesse.

Titre : Recherche du boson de Higgs se désintégrant en deux photons et produit en association avec une paire de quarks top dans l'expérience CMS au LHC

Mots-clés : *boson de Higgs, expérience CMS, collisionneur LHC, modèle standard, interaction faible*

Résumé : Dans cette thèse, la mesure des propriétés du boson de Higgs dans le canal de désintégration en deux photons avec l'expérience CMS au Grand Collisionneur de Hadrons (LHC) est présentée. L'objectif de ce travail est l'étude du mode de production associé à une paire de quark top ($t\bar{t}H$). Ce mode représente le seul accès direct au couplage de Yukawa du quark top, un paramètre fondamental du Modèle Standard. Le mode de production $t\bar{t}H$ est un processus très rare. Il est de l'ordre de deux ordres de grandeur plus petits que la production principale du boson de Higgs par fusion de gluons. À 13 TeV, le mode de production $t\bar{t}H$ est environ 4 fois plus grand qu'à 8 TeV. Cette thèse reprend les études réalisées à 8 TeV, où l'échantillon de données ne suffisait pas pour établir une observation de ce mode de production. Bien que le canal en deux photons ne soit pas celui qui présente le plus grand rapport de branchement (seulement 0,2%), il est très prometteur en raison de son excellente résolution en masse (1%). De plus, sa signature est très propre dans le détecteur. Le canal de désintégration en deux photons est particulièrement intéressant puisqu'il s'agit du seul canal permettant l'étude de tous les modes de production : la fusion de gluons, la fusion de bosons vecteurs, les productions associées avec des bosons W ou Z ou avec une paire de quarks top. Le document commence par une introduction théorique du Modèle Standard et la physique du boson de Higgs au LHC, suivie d'une description du détecteur CMS. Pour obtenir une excellente résolution de masse dans le canal de désintégration en deux photons, le calorimètre électromagnétique doit être calibré. Le système de monitoring de la transparence des cristaux du calorimètre électromagnétique de CMS par le système laser joue un rôle important dans la chaîne d'étalonnage et est décrit en détail. Sur le long terme, le système de monitoring laser devra être amélioré car le niveau de rayonnement influence son électronique. Je présente mon travail sur l'amélioration possible du système de monitoring laser, ainsi que l'étude de sa précision possible. L'analyse inclusive $H \rightarrow \gamma\gamma$ a eu plusieurs itérations pour les conférences en 2016 et 2017. La stratégie pour 2017 est décrite dans ce document. Une classification des événements sert à maximiser la signification du signal et à étudier les modes spécifiques de production du boson de Higgs. Mes contributions à l'analyse $H \rightarrow \gamma\gamma$ consistent en l'identification du vertex primaire, l'identification du photon et l'étude du mode de production $t\bar{t}H$. L'analyse $t\bar{t}H, H \rightarrow \gamma\gamma$ est présentée pour deux itérations en 2016 et 2017, en mettant l'accent sur les améliorations dans l'analyse de 2017. Enfin, les résultats de l'analyse inclusive et $t\bar{t}H, H \rightarrow \gamma\gamma$, en utilisant l'ensemble complet de données 2016 correspondant à une luminosité intégrée de $35,9 \text{ fb}^{-1}$, sont présentés.

Title: Search for the Higgs boson decaying to two photons and produced in association with a pair of top quarks in the CMS experiment

Keywords: *Higgs boson, CMS experiment, LHC collider, standard model, weak interaction*

Abstract: In this thesis, the measurement of the Higgs boson properties in the diphoton decay channel with the CMS experiment at the Large Hadron Collider is presented. The focus of this work is the $t\bar{t}H$ production mode, as it is the only direct access to the top quark Yukawa coupling, a fundamental parameter of the Standard Model. $t\bar{t}H$ is a very rare process, two orders of magnitude smaller than the dominant Higgs boson production by gluon fusion. At 13 TeV, $t\bar{t}H$ production is about 4 times larger than at 8 TeV. This thesis takes over the studies performed at 8 TeV, where the statistics was not enough for an observation of $t\bar{t}H$. Despite a very small branching ratio (only about 0.2%), the two photons decay channel of the Higgs boson is very promising, because of its excellent mass resolution (about 1%). Moreover, its signature in the detector is very clear. The diphoton decay channel is also of particular interest as it is the only channel allowing the study of all production modes: gluon fusion, vector boson fusion, associated productions with a W or a Z bosons, or with a top quark pair. The document starts with a theoretical introduction about the Standard Model and Higgs boson physics at LHC, followed by a description of the CMS detector. To achieve an excellent mass resolution in the $H \rightarrow \gamma\gamma$ channel, the electromagnetic calorimeter has to be calibrated. The laser monitoring system plays an important role in the calibration chain and it is described in details. On the long term, the laser monitoring system will have to be upgraded as level of radiation influences its electronics. I present my work on the possible upgrade of the laser monitoring system, along with the study of its possible precision. $H \rightarrow \gamma\gamma$ inclusive analysis had several iterations for conferences in 2016 and 2017. The strategy for 2017 is described in this document. An event classification is used to maximize the signal significance and to study specific Higgs boson production modes. My contributions to the $H \rightarrow \gamma\gamma$ analysis are primary vertex identification, photon identification and the study of the $t\bar{t}H$ production mode. The $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis is shown for two iterations in 2016 and 2017, with the emphasis on improvements in 2017 analysis. Finally, the results of the inclusive and $t\bar{t}H, H \rightarrow \gamma\gamma$ analysis, using the full 2016 dataset corresponding to an integrated luminosity of 35.9 fb^{-1} , are shown.