Search for exotic decays of the Higgs boson using photons with the Compact Muon Solenoid experiment

by Tanvi Wamorkar

B.Sc. in Physics, University of Delhi
M.S. in Physics, Northeastern University

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

May 20, 2021

Dissertation Committee:

Professor Toyoko Orimoto, Chair
Professor Emanuela Barberis
Professor James Halverson
Professor Darien Wood

Acknowledgements

I have received incredible support from many people throughout the course of this PhD. It is difficult to express my gratitude in words, but I will try nonetheless.

First and foremost, I would like to thank my advisor, Professor Toyoko Orimoto. I have been extremely lucky to have an advisor who gave me the freedom to pursue my research interests and ensured that I had opportunities to explore diverse projects. I would like to thank Toyoko for always giving me a chance, for always considering my well-being and for giving me many amazing opportunities. I'll always look up to Toyoko as a great mentor.

I would also like to thank the rest of my thesis committee: Prof. Darien Wood, Prof. Emanuela Barberis and Prof. James Halverson. Their continuous helpful input and comments have greatly helped me improve my work throughout the course of my PhD. Additionally, I would also like to thank Prof. Louise Skinnari. Much of what I know about how to do a physics analysis has come from discussions at the weekly CMS Northeastern group meetings. I greatly appreciate the supportive and friendly work environment in the CMS Northeastern group and would like to thank Prof. Wood, Prof. Barberis, Prof. Orimoto and Prof. Skinnari for that.

I would like to thank Dr. Rafael Teixeira de Lima and Dr. Andrea Massironi who have now moved on from Northeastern but were instrumental in shaping the analysis presented in this thesis. They were always generous with their time and knowledge, and I have learnt a great deal from them. I would also like to thank the postdocs of our group: Dr. Badder Marzocchi, Dr. Yacine Haddad, Dr. David Morse and Dr. Ashley Parker. I have learnt a vast amount from their expertise. In particular I would like to thank Badder who has helped me with every part of the analysis presented in this thesis. I would like to thank my fellow PhD students and friends: Andrew Wisecarver, Chad Freer, Bingran Wang, Gustavo Salinas de Souza, Amin Aboubrahim, Abraham Tishelman-Charny, Vivan Nguyen, Gabriel Madigan, Jack Li, Yixao Zhang and Amrutha Krishna. I am incredibly lucky to be part of such an awesome group. Abe, Andrew, Chad, Gustavo, Amin and Bingran have been great friends throughout my PhD and I could not have survived graduate school without them. They have been there with me every step of the way and I consider them family. I will always be grateful for their friendship, support, encouragement and all the fun times. I wish them success and happiness in all their new ventures.

Lastly, I would like to thank my family: Aai, Baba, Kaka, Kaku, Mansi Tai, Tarush, Swayam Jiju, Ruchi and Anay. I am extremely fortunate to have such a loving, caring and supportive family. Their unwavering support and encouragement has been a constant in my life and this thesis would not have been possible without them. Aai, Baba: Thank you for giving me every opportunity I could have dreamed of and for always being proud of me.

Abstract of Dissertation

The search for answers to the fundamental questions of nature are being explored by the experiments at the Large Hadron Collider at CERN. The discovery of the Higgs Boson at the LHC marked the culmination of several decades of experimental search. However, the experimental investigation of the Higgs Boson is just ramping up. With its mass near 125 GeV, the Higgs Boson poses a naturalness question and hints at the possibility of fine tuning or the hierarchy problem in the Standard Model. The Higgs particle can thus be an excellent tool to probe the unknown. A search for the exotic decays of the 125 GeV Higgs Boson is presented in this thesis. The analysis selects event with at least four photons in the final state, and uses data sample of $\sqrt{s} = 13\,\text{TeV}$ proton-proton collisions collected by the CMS detector between 2016 and 2018, and corresponds to 132 fb^{-1}. No significant excess of events is observed over the background expected from the Standard Model, and upper limits are set on $\sigma(pp \to h) \times BR(h \to aa \to \gamma\gamma\gamma\gamma)$. The performance of the CMS Electromagnetic Calorimeter, which plays a crucial role in these results, is also presented in this document.

Table of Contents

List of Figures

Figure

List of Tables

Table

Chapter 1

Introduction

Particle physics is the study of the nature of the smallest, most fundamental particles that make up matter and their interactions. The Standard Model (SM) is the cornerstone of particle physics and is highly successful is describing the fundamental particles and interactions. Since 2010, when the Large Hadron Collider (LHC) began colliding protons, one of its primary goals was the finding the last missing piece of the SM: the Higgs boson. In 2012, the CMS (Compact Muon Solenoid) and ATLAS (A Toroidal LHC Apparatus) experiments discovered the Higgs boson and completed the understanding of the SM.

However, the SM does not provide explanations of mysteries such as Dark Matter, the weakness of the gravitational force, or the hierarchy problem. The SM is therefore, an incomplete theory. Various beyond the SM (BSM) models propose answers to these mysteries, and many of these models are centred around the Higgs boson. The small natural width of the Higgs makes it special because this enables the Higgs to easily mediate interactions with sectors of new physics. An important part of the LHC's Higgs program is to search for exotic decays of the Higgs boson which are not predicted by the SM. Such decays can provide a window into physics beyond the SM and are a rich phenomenological territory to explore.

Pseudoscalars are well motivated in many BSM scenarios: they may be the QCD axion solution to the strong CP problem, Nambu-Goldstone bosons of a broken symmetry or mediators to the dark sector. One of the pseudoscalar candidates could be the lighter

pseudoscalar of the NMSSM (Next to Minimal Super Symmetric Model) or a generic SM-single boson that couples to photons via a heavy vector like matter. The branching fraction of the pseudoscalar to photons can be very low in the first case, however the 4γ final state is basically SM background free. On the other hand, in some models the pseudoscalar may only be able to decay into photons, thus making the 4γ final state extremely important. The analysis presented in this thesis searches for exotic decays of the Higgs to light pseudoscalars with photons in the final state, and makes use of data sample of $\sqrt{s} = 13\,\text{TeV}$ proton-proton collisions collected by the CMS detector between 2016 and 2018, which corresponds to 132 fb^{-1}. The thesis is structured as follows.

Chapter 2 provides a description of the SM and motivates the use of Higgs as a probe to search for new physics.

Chapter 3 provides a description of the LHC design and operation, and of the CMS experiment. The various sub-detectors that constitute CMS and their roles within the experiment are discussed.

Chapter 4 details the CMS Electromagnetic Calorimeter. In particular, the details of energy reconstruction procedure by ECAL, and the calibration as well as alignment procedure of ECAL are described.

Chapter 5 describes the analysis which makes up the core part of my dissertation work. The analysis strategy, samples used in the analysis, details of the selections, and final results are discussed.

Finally, Chapter 6 concludes this thesis with a summary of the results, and discusses ideas for future improvements to this search.

Chapter 2

Theoretical Foundation

The Standard Model (SM) of particle physics provides a description of the dynamics of all known particles and interactions currently known. It is constructed using the principles of Quantum Field Theory, and is constrained to follow rules pertaining to several symmetries and allowed interactions. It has been incredibly successful in predicting the existence of particles before their discovery, and has passed many high-precision tests. However, it has several shortcomings and is possibly incomplete, especially at energies higher than those probed by the current particle accelerators. A wide range of models that describe physics beyond the SM (BSM) are possible extensions of the SM.

This chapter explores the fundamentals of the Standard Model, as well as the particles and interactions described by it. The phenomenology of the Higgs mechanism is also detailed in this chapter. The BSM models that use the Higgs boson as probe to search for new physics, and motivate the search performed in Chapter 5 are also discussed.

2.1 The Standard Model of Particle Physics

The SM of particle physics is a framework that describes our current understanding of the elementary particles and their interactions. Developed as a quantum field theory in the 1960s and 70s [29; 30; 31], the SM has been thoroughly examined and tested by multiple experiments over a large range of energy scales. It describes the interactions of elementary matter with three out of the four known forces: the electromagnetic force, the weak nuclear

force, and the strong nuclear force. This model has been exceptionally successful in predicting new particles and phenomena, which includes the prediction of the Higgs boson almost 50 years before its discovery in 2012, which completed the SM.

In the SM, each particle is represented as a field, and the theory is constructed by means of a set of symmetries which are used as a guide to write the most general renormalizable Lagrangian. The SM must follow a set of rules, the first of which is that the SM must be invariant under any choice of reference frame. That is why the SM is symmetric under translations, rotations, and boosts. The second rule that the SM must follow is to be gauge invariant. The SM is defined by the local $SU(3) \times SU(2) \times U(1)$ gauge symmetry, where each term roughly describes one of the three forces described by the Standard Model.

The particles of the SM can be divided into two categories: fermions, with half-integer spin, and bosons, with integer spin. Fermions can be further divided into two types of particles: quarks and leptons. Both, quarks and leptons, have half-integer spin, and obey Fermi-Dirac statistics. Gauge bosons, which are spin-1 particles, act as force carriers on the fermions. The only example of a spin-0 boson in the SM is the Higgs boson. The Higgs boson is associated with the mechanism that gives mass to other particles in the Standard Model. Figure 2.1 depicts all the known SM particles.

Sections 2.1.1 and 2.1.2 describe the particle and force content of the Standard Model.

2.1.1 Matter

All matter described in the SM is constituted of fermions, which are spin-half particles that can be further divided into two groups, quarks and leptons. Leptons undergo weak interactions, and quarks undergo both, weak and strong interactions.

2.1.1.1 Leptons

In figure 2.1, leptons can be found on the bottom left. They exist in three generations and have a flavor each: electron, muon, and tau. The massive leptons are arranged in order

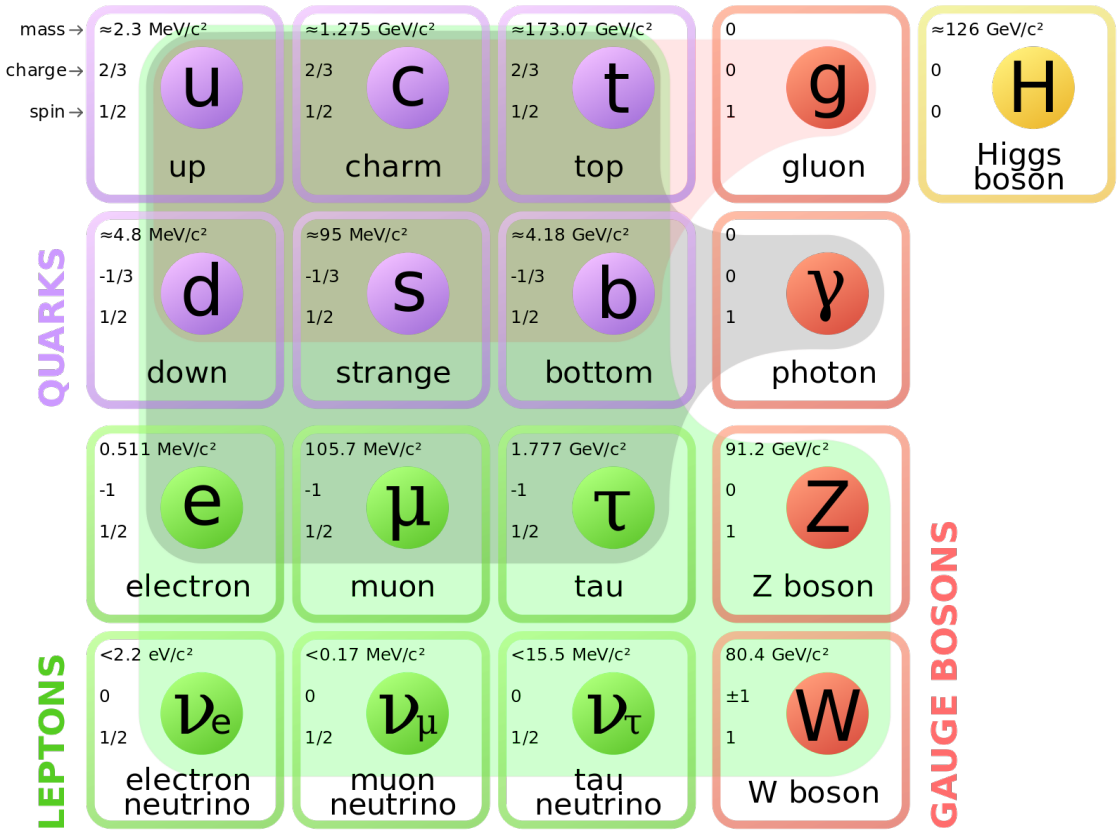


Figure 2.1: Standard Model particles [1].

of increasing mass, carry a negative electromagnetic charge, and have a positively charged anti-particle. The charged leptons interact under the electromagnetic and weak forces.

Additionally, there are three generations of neutrinos and they exist in the same flavors as the massive leptons. In case of neutrinos, the flavor eigenstates do not correspond to mass eigenstates, as a result of which neutrinos oscillate between flavor states as they propagate through space. Since they are not electromagnetically charged, neutrinos only interact via the weak force which makes it exceptionally difficult to detect them. It is still unknown if neutrinos have a separate anti-particle or are their own anti-particles.

Lepton number, L, defined as the number of leptons minus the number of anti-leptons in a state, is conserved by the SM. This rule is applicable to each lepton flavor. As a result, electron, the lightest charged lepton, is stable. The second and third generation charged leptons decay via the weak force to lighter leptons. An example of lepton number conservation is the decay of a muon to an electron and two neutrinos, via the W boson

$$\mu^+ \rightarrow e^+ + \nu_e^+ + \bar{\nu}_\mu \tag{2.1}$$

Processes such as $\mu^- \rightarrow e^- + \gamma$ are prohibited by lepton number conservation.

2.1.1.2 Quarks

In figure 2.1, quarks can be found on the top left. They are electromagnetically charged particles and interact weakly. They are distinguishable from leptons by their strong interactions. Like the leptons, quarks come in three generations and the generations are ordered by mass. The three generations of quarks come in pairs of up-type and down-type quarks, where the up-type quarks have electric charge $+\frac{2}{3}q_e$ and the down-type quarks have electric charge -$\frac{1}{3}q_e$, with q_e being the electron charge. There are six flavors of quarks: up, down, charm, strange, top and bottom. Additionally, quarks have a color charge, characterized by the colors red, green, and blue, and can interact via the strong force. Anti-quarks have a

color charge characterized by anti-red, anti-green, and anti-blue.

Individual quarks and fractional charges are never observed in nature due to color-confinement. Because of this requirement, stable particle states must be color-neutral. Quarks can either create two-particle composite state called *mesons*, which consist of one quark and one anti-quark with opposite color charges, or three-particle state called *baryons* which are bound states of quarks or anti-quarks with three different color states. Proton is the lightest color-neutral state and is made up of quarks *uud*. Highly unstable bound states of multiple quarks, such as the pentaquark discovered in 2015 [32] at the LHC, can also exist. The multi-quark bound states are collectively known as *hadrons*.

Similar to the leptons, the number of quarks in a state are conserved. The conservation of quarks is decribed in terms of the baryon number, B. Baryons have $B = 1$, anti-baryons have B = -1, and mesons have $B = 0$.

2.1.2 Gauge Bosons and Forces

In the Standard Model, forces are mediated by spin-1 bosons. Fermions, which are charged under the corresponding force's quantum numbers, interact via these force modifiers. For example, photon interacts with electromagnetically charged particles, and gluons only interact with color charged particles, quarks and themselves. Since all fermions are weakly charged, they interact with the mediators of the weak force, the W and Z bosons.

The core principle behind formulation of each of these forces is based on the requirement that the SM Lagrangian must be locally gauge invariant. This is achieved by addition of gauge fields to the Lagrangian, whose behavior under gauge transformations cancels out the gauge dependence of the free Lagrangian. This mechanism, however, only creates forces mediated by massless gauge bosons because addition of a mass term reintroduces gauge dependence. This problem is solved by the addition of the Higgs field, which provides mass terms for the weak gauge bosons as well as other particles and does not interfere with the gauge invariance of the Lagrangian.

The total symmetry group of the SM is $SU_C(3) \times SU_L(2) \times U_Y(1)$, where C stands for color, the charge of the strong force, L stands for left, because the weak force is left-handed, and Y stands for *hypercharge*, which is the charge of the unified electroweak force.

2.1.2.1 The Electromagnetic Force

The electromagnetic force is mediated by the photon (γ), which interacts with all electromagnetically charged particles, and is described by Quantum Electro-Dynamics (QED). Electromagnetism is the simplest example of a local gauge invariance generating a Lagrangian description of a force.

Dirac Lagrangians describe the free propagation of fermions and have the form

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.2}$$

The free Lagrangian for fermions needs to be invariant under a $U(1)$ local gauge transformation, $e^{iq\lambda(x)}$. This can be achieved by adding a term to the Lagrangian which cancels out the derivative terms arising from the x dependence of λ's:

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - \left(q\bar{\psi}\gamma^{\mu}\psi\right) A_{\mu} \tag{2.3}$$

where A_{μ} is a gauge field that transforms as follows:

$$A_{\mu} \rightarrow A_{\mu} + \partial_{\mu}\lambda \tag{2.4}$$

This vector field comes with an additional term,

$$\mathcal{L} = -\frac{1}{16\pi}F^{\mu v}F_{\mu v} + \frac{1}{8\pi}m_A^2 A^v A_v \tag{2.5}$$

In order to make the the field invariant under transformation, it is made massless. This

makes the final Lagrangian equal to

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - \frac{1}{16\pi}F^{\mu v}F_{\mu v} - \left(q\bar{\psi}\gamma^{\mu}\psi\right)A_{\mu} \tag{2.6}$$

When interpreted in a quantized manner, this Lagrangian descibes a field that interacts with particles with non-zero electromagnetic charge q via interactions with a masselss spin-1 boson, the photon. This charge is dependent on the energy scale of the interaction, μ, and the strength of the interaction is described by the electromagnetic coupling constant

$$\alpha_{EM}(\mu) = \frac{q(\mu)^2}{4\pi} \tag{2.7}$$

2.1.2.2 The Strong Force

Similar to the electromagnetic force, the strong force is also generated by a similar process of requiring local gauge invariance of the Lagrangian, except in this case it is required for an $SU(3)$ transformation. The theory of Quantum Chromodynamics (QCD) is used to describe the interactions of the strong force. It is given by the Lagrangian

$$\mathcal{L}_{\text{strong}} = -\frac{1}{4}G^{\alpha}_{\mu\nu}G^{\alpha\mu\nu} - \frac{1}{2}\bar{Q}_m \not{D} Q_m \tag{2.8}$$

here α represents the eight force carriers of the strong force, gluons with different color charges, and can take values that range from 1 to 8. The index m represents the three quark generations, and $G^{\alpha\mu\nu}$ is the field strength tensor, which is defined as

$$G^{\alpha}_{\mu\nu} = \partial_{\mu}G^{\alpha}_{v} - \partial_{\nu}G^{\alpha}_{\mu} + g_s f^{\alpha}_{\beta\gamma}G^{\beta}_{\mu}G^{\gamma}_{v} \tag{2.9}$$

where g_s is a function of the energy scale of the interaction μ, and is related to the strong coupling constant by

$$\alpha_s(\mu) = \frac{g_s(\mu)^2}{4\pi} \tag{2.10}$$

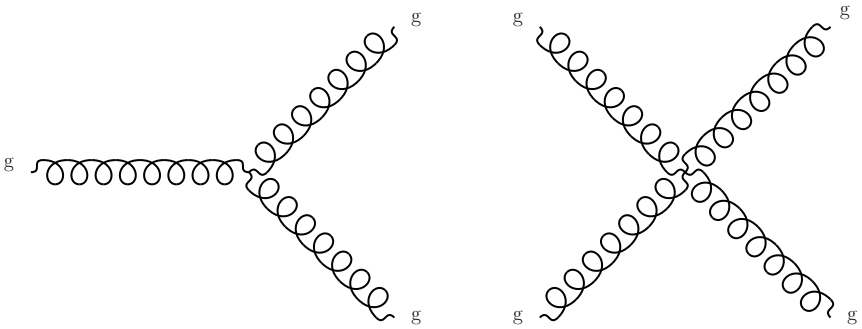


Figure 2.2: Feynman diagrams for gluon self interactions

The first term of the Lagrangian corresponds to the gluon self-coupling interactions. This involves the 2, 3 and 4 gluon field terms. The 2-field term represents gluon propagation, and the other terms correspond to gluon self-interaction terms. Because gluons are charged by the force they carry, unlike photons, they can undergo self-interactions. Figure 2.2 depicts the Feynman diagrams corresponding to gluon self-interactions.

In the second term of the Lagrangian, $\not{D}Q_m$ is the covariant derivative acting on the quark field. Since quarks are charged under the strong, electromagnetic, and weak forces, the covariant derivative includes relevant terms such that the Lagrangian corresponding to each of the three force's is gauge invariant. As a result, this term introduces four types of quark-boson interactions that are depicted in figure 2.3.

An assumption of quantum field theory is that particles can propagate without interaction. This assumption is correct as long as the coupling constants for each force are much less than one. However, this is not always true for the strong coupling constant, α, which changes as a function of energy of an interaction according to the equation

$$\mu_R^2 \frac{d\alpha_s}{d\mu_R^2} = \beta(\alpha_s) = -\left(b_0\alpha_s^2 + b_1\alpha_s^3 + b_2\alpha_s^4 + \ldots\right) \tag{2.11}$$

where μ_R^2 represents the renormalization scale, and b_n is the correction to the β function based on diagrams with n loops [2]. It should be noted that the overall negative sign dictates the energy dependence of α_s, which becomes very small at high energy scales and asymptoti-

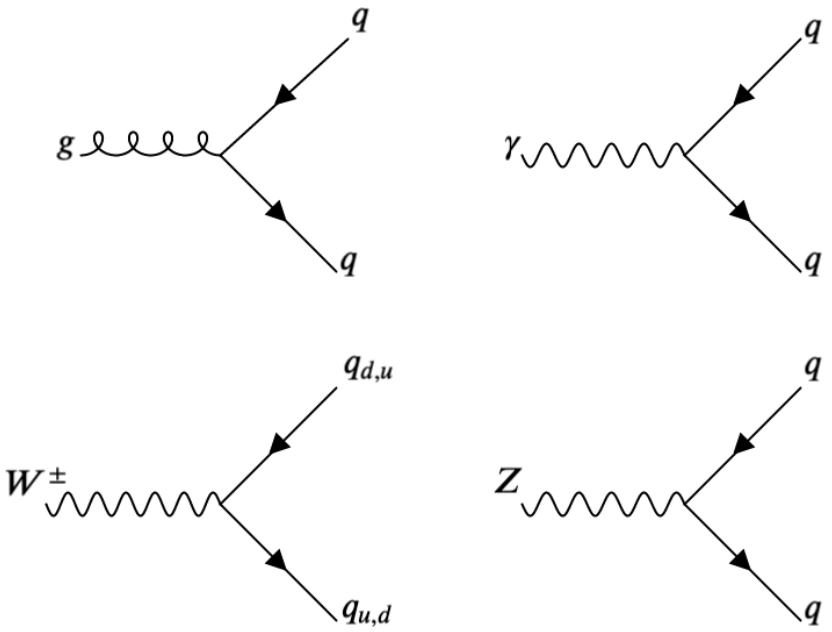


Figure 2.3: Feynman diagrams corresponding to quark couplings to various gauge bosons.

cally increases at low energies. This translates to the understanding that at small distance scales $\alpha_s \ll 1$, and α_s approaches 1 at large distance scales. Therefore, the perturbative approximation is no longer valid at large distances ($\sim 10^{-6}$m). The boundary between these regimes differentiates energies at which QCD can be considered perturbatively and those at which it cannot. This boundary is sometimes referred to as Λ_{QCD}. The LHC is capable of producing high energy quarks and probe distances of about 10^{-9}m, but the quarks lose energy by radiating gluons and eventually enter the regime below Λ_{QCD}. The transition between these regimes is an extremely complicated process and largely dictates the way that strongly charged particles appear in the CMS detector.

2.1.2.3 The Electroweak Force

If the W and Z bosons were massless, a similar gauge transformation mechanism applied on the $SU(2)$ symmetry would have been sufficient to describe the W and Z bosons of the SM. Since they do in fact have masses, a different mechanism is required to generate massive

force carriers. Before this mechanism was understood and they were discovered, the large masses of the W and Z bosons were proposed so as to unify the electromagnetic and weak forces into the electroweak force. The large masses of the W and Z bosons were vital in explaining the discrepancy of the strength of the two forces.

The symmetry group $SU(2)_L \times U(1)_Y$, where L refers to left-handed fields, and Y refers to the quantum number for hypercharge, is responsible for generating the unified electroweak force. Y is defined as $Y = 2(Q - T_3)$, where Q is the electromagnetic charge, and T_3 is the third component of the weak isospin T, the quantum number related to the weak interaction.

The gauge bosons arising from this unified theory include a triplet, W, with coupling g_W, and a single field B, with coupling $g\prime/2$. Since electroweak symmetry is broken, mixing between these two states occurs. As a result, the force carriers are produced: the first two states of the triplet W give rise to $W^{\pm}$, and mixing of the last state of W and B gives rise to the Z and the photon field A.

The electroweak Lagrangian can be written as

$$\mathcal{L}_{\text{electroweak}} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{fermions}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}} \tag{2.12}$$

where the first term can be written as

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4} W^{a\mu\nu} W^{a}_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B\mu\nu \tag{2.13}$$

here, the a indices represent the generators of $SU(2)$, and take values that range from 1 to 3. The $SU(2)$ generators can be written as

$$W^{a}_{\mu v} = \partial_\mu W^{a}_{v} - \partial_\nu W^{a}_{\mu} + g_W \epsilon_{abc} W^{b}_{\mu} W^{c}_{i} \tag{2.14}$$

Interactions between the three gauge bosons, as shown in figure 2.4, are generated by the gauge portion of the Lagrangian when written in the mass eigenstate basis.

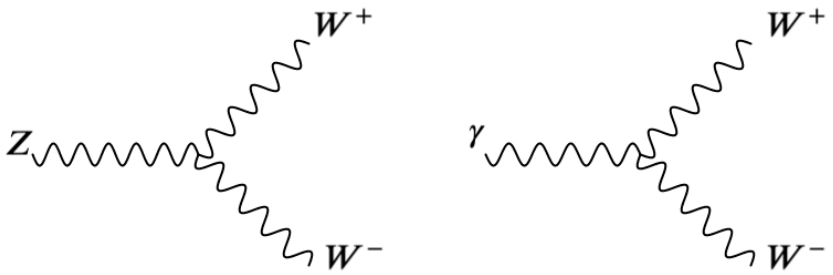


Figure 2.4: Feynman diagrams corresponding to trilinear gauge couplings in the SM.

The fermion part of the electroweak Lagrangian is written as

$$\mathcal{L}_{\text{fermion}} = -\frac{1}{2}\bar{L}_m \not{D} L_m - \frac{1}{2}\bar{Q}_m \not{D} Q_m - \frac{1}{2}\bar{U}_m \not{D} U_m - \frac{1}{2}\bar{D}_m \not{D} D_m - \frac{1}{2}\bar{E}_m \not{D} E_m \quad (2.15)$$

In this equation, L is the left-handed lepton doublet, Q is left-handed quark doublet, U is the right-handed singlet for up-type quarks, D is the right-handed singlet for down-type quarks, and E is the right-handed singlet for electrons, muons and taus. m index takes values from 1 to 3, and represents the three generations.

The weak force treats left and right handed fermionic fields differently, it only interacts with the left-handed fields. The first term in this Lagrangian produces weak interactions as shown in figure 2.5.

The remaining pieces of the electroweak Lagrangian are related to the Higgs field.

2.1.2.4 The Higgs Mechanism

The Higgs field is a scalar field with the Lagrangian

$$\mathcal{L}_{\text{Higgs}} = \frac{1}{2}\left(\partial_\mu \phi\right)^* \left(\partial^\mu \phi\right) + \frac{1}{2}\mu^2 \phi^* \phi - \frac{1}{4}\lambda^4 \left(\phi^* \phi\right)^2 \quad (2.16)$$

where ϕ is a complex scalar, $\phi = \phi_1 + i\phi_2$. A unique feature of this Lagrangian is that its ground state does not occur at $\phi = 0$. Due to its perturbative nature, quantum field theory is only valid when expanded around a ground state. For the Higgs Lagrangian, the ground

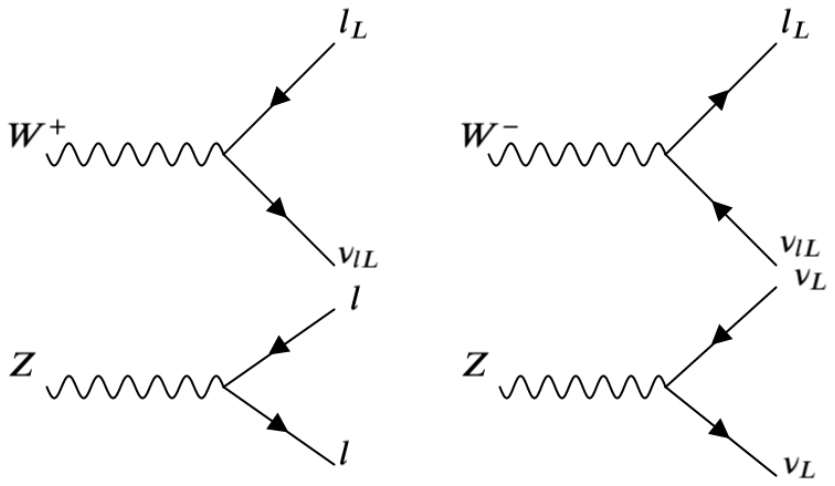


Figure 2.5: Feynman diagrams corresponding to weak couplings to leptons in the SM.

state must satisfy

$$\phi_1^2 + \phi_2^2 = \frac{\mu^2}{\lambda^4} \tag{2.17}$$

The original Lagrangian can then be writted in terms of a field, $v + H(x)$, centred around the ground state with energy equal to the vaccuum expectation value defined as

$$v = \frac{\mu}{\lambda^2} \tag{2.18}$$

This results in a Lagrangian with non-imaginary mass. As a result of the spontaneous symmetry breaking process, the $SO(2)$ rotational symmetry of the Lagrangian is lost, and only $U(1)$ rotational symmetry remains. Under a local $U(1)$ transformation, $\phi \rightarrow e^{l\theta(x)\psi}$, it is possible to make the Lagrangian invariant by adding a massless gauge field A^μ, and by using the covariant derivative. The non-zero ground state results in many cross terms, which leads to creation of terms for mass of one of the scalar bosons as well as for the gauge field. Finally, this leaves only one massless scalar boson which can be removed from the theory via local $U(1)$ transformations. This ultimately produces a theory with one massive scalar boson, the Higgs boson, and a massive gauge field, W.

Figure 2.6 depicts the Feynman diagrams of the Higgs interaction with weak gauge

bosons. Higgs self-interaction terms, which produce vertices describing 3 and 4 Higgs interactions, are also included in the Lagrangian, and are depicted in figure 2.7.

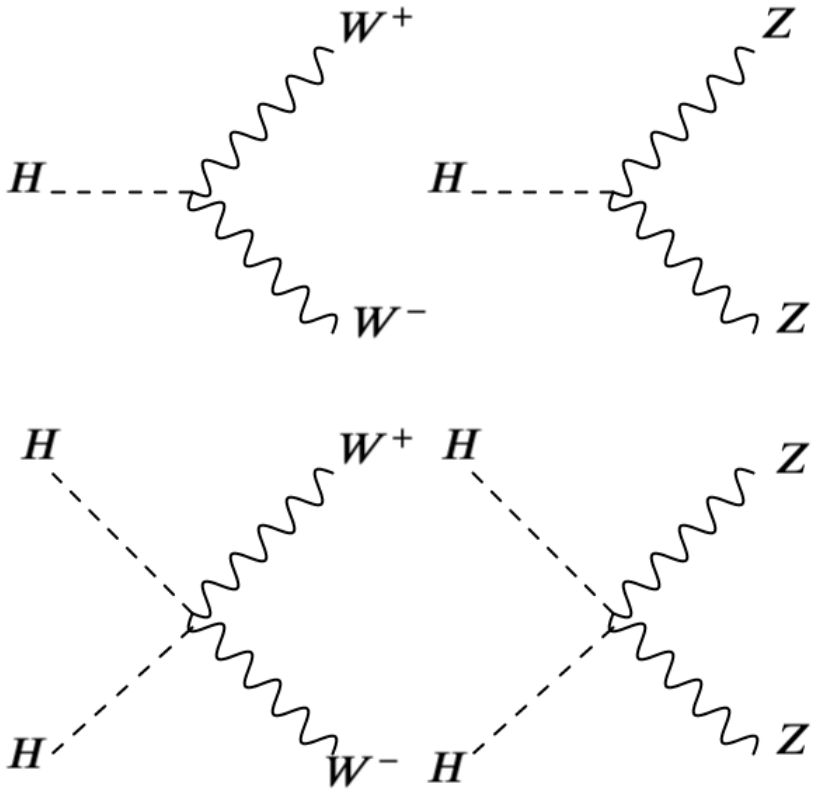


Figure 2.6: Feynman diagrams corresponding to Higgs couplings to the weak gauge bosons in the SM.

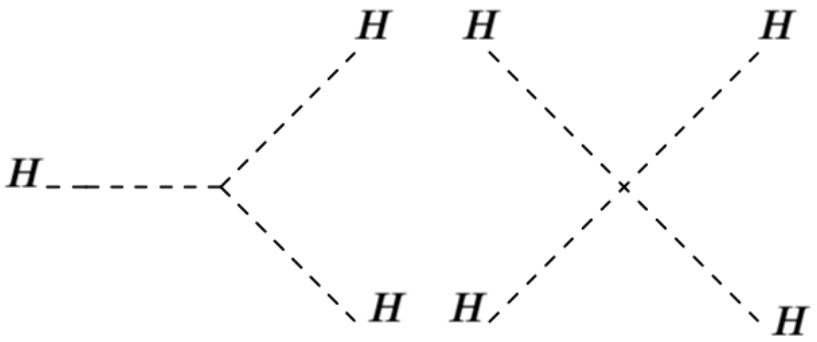


Figure 2.7: Feynman diagrams corresponding to Higgs self-interactions in the SM.

This leaves us with the remaining part of the Lagrangian, $\mathcal{L}_{\text{Yukawa}}$. This part describes

the interaction of the Higgs field with fermions, and is written as

$$\mathcal{L}_{Yukawa} = -\Gamma^{e}_{mn}\bar{L}_m\phi E_n - \Gamma^{u}_{mn}\bar{Q}_m\phi U_n - \Gamma^{d}_{mn}\bar{Q}_m\phi D_n + \text{ h.c.} \tag{2.19}$$

where *h.c.* corresponds to the hermitian conjugate term. The indices of Γ matrices correspond to the fermion generations. When diagonalized, Γ matrices are proportional to the fermion masses.

Feynman diagrams for couplings of the Higgs to quarks and leptons can be seen in figure 2.8.

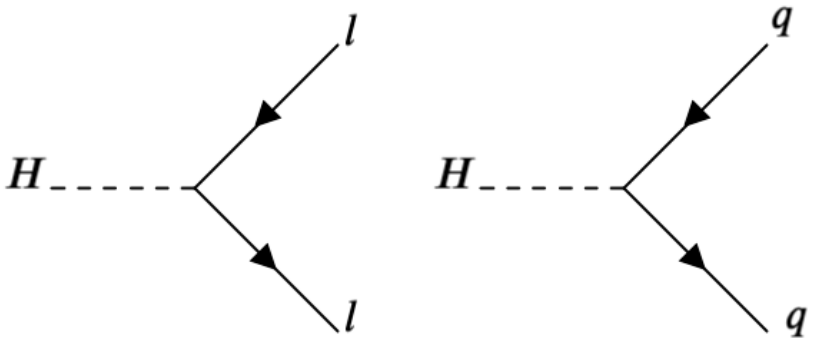


Figure 2.8: Feynman diagrams corresponding to Higgs couplings to fermions in the SM.

2.2 The Higgs Boson

The Higgs mechanism, explained in section 2.1.2.4, gives rise to the massive Higgs Boson, and is a particle of great interest since it generates the mass of all the other elementary particles. In 2012, the CMS [33] and ATLAS experiments [34] observed the Higgs boson, using approximately 5 fb^{-1} of 7 TeV and 8 TeV data. The observed Higgs boson was found to be compatible with the SM. This discovery led to the Nobel prize to be awarded to the leading theorists behind the electroweak symmetry breaking mechanism: Peter Higgs and Francois Englert.

2.2.1 Higgs Boson Phenomenology at the LHC

The Higgs boson mass dictates the cross section and branching fraction of the Higgs decays at colliders. At the LHC, the leading production mechanisms of the Higgs boson, in decreasing order of cross-section are: gluon-gluon fusion (ggF), vector boson fusion (VBF), vector boson-associated production (VH), and associated production with $t\bar{t}$and $b\bar{b}$. The Feynman diagrams corresponding to these production modes are shown in figure 2.9. The predicted Standard Model cross sections for these production modes are listed in table 2.1 and plotted with theoretical uncertainties in figure 2.10.

Production Mode	Cross section (pb)
ggH	48.71
VBF	3.78
WH	1.37
ZH	0.88
ttH	0.51
bbH	0.49
tH	0.07

Table 2.1: Cross section values, in units of pb, for the main Higgs boson production mechanisms at the LHC, for $\sqrt{s} = 13TeV$ and $m_H = 125$ GeV [28].

The SM Higgs couples with all massive vector gauge fields and charged fermions, and

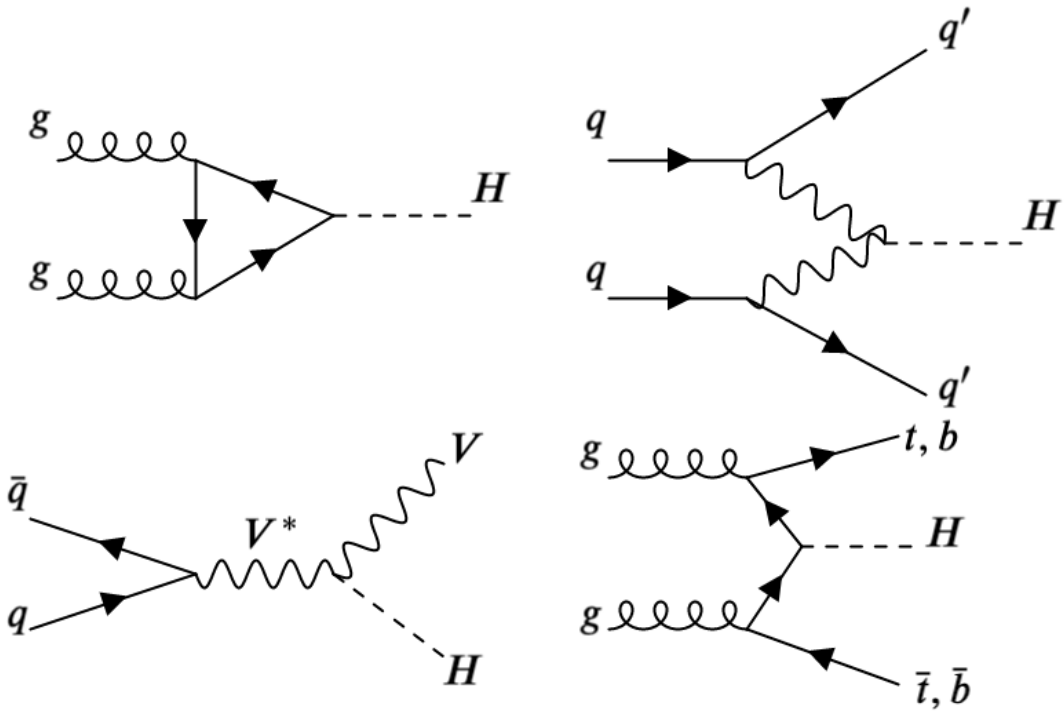


Figure 2.9: Feynman diagrams depicting the leading production modes of the Higgs Boson at the LHC. Top left corresponds to Higgs production through gluon-gluon fusion, top right corresponds to Higgs production through vector boson fusion, bottom left corresponds to Higgs production through vector boson associated production, and bottom right corresponds to Higgs production through associated production with heavy quarks ($t\bar{t}$and $b\bar{b}$).

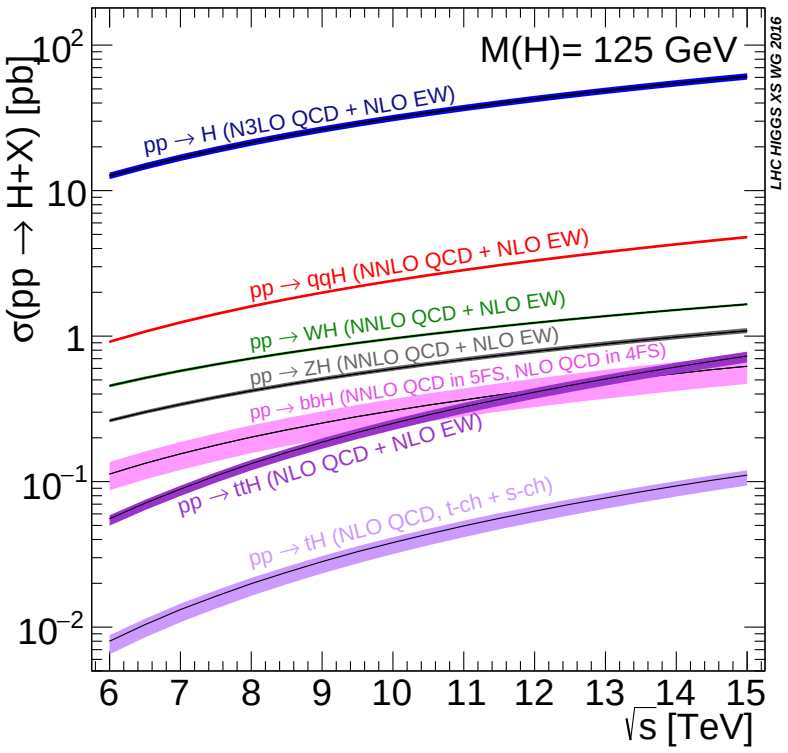


Figure 2.10: The Standard Model Higgs boson production cross section, for pp collisions, is shown as a function of center-of-mass energy, $\sqrt{s}$, for $m_H = 125$ GeV. The colored bands indicate theoretical uncertainties [2].

therefore can decay into all such particles. The coupling strength of the Higgs is directly proportional to the mass of the decay product. Additionally, the Higgs can also decay into massless particles, such as the gluon and photon, via loop diagrams. The branching fractions for the primary decay modes of the Higgs are listed in table 2.2 and plotted with theoretical uncertainties in figure 2.11.

As is evident, from table 2.2 and figure 2.11, the decay to $b\bar{b}$ has the largest branching fraction. However, at hadron colliders, the multijet events from QCD processes have a large cross section, therefore, making the $H \rightarrow b\bar{b}$ analysis very challenging. In spite of their low branching fractions, the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^*$ channels are highly sensitive. The sensitivity of these channels relies on the relatively low background and precise ability to reconstruct the mass of the Higgs boson from the decay products.

Decay channel	Branching Fraction
H $\rightarrow b\bar{b}$	58.72 %
H $\rightarrow W^{\pm}W^{\mp *}$	21.4 %
H $\rightarrow gg$	8.2 %
H $\rightarrow \tau^{+}\tau^{-}$	6.3 %
H $\rightarrow c\bar{c}$	2.8 %
H $\rightarrow ZZ^*$	2.6 %
H $\rightarrow \gamma\gamma$	0.23 %
H $\rightarrow Z\gamma$	0.153 %
H $\rightarrow \mu^{+}\mu^{-}$	0.022 %

Table 2.2: Branching fraction values of the main Higgs boson decay modes, for $m_H = 125$ GeV [28].

2.2.2 Current status of Higgs physics

Approximately 5 fb^{-1} of data was collected by the CMS and ATLAS collaborations during Run 1 of the LHC. This data was combined and used for experimental measurements of the Standard Model Higgs's production cross sections and decay modes. The summary results by CMS and ATLAS [3] are presented in terms of signal strength (μ), defined as the ratio of measured Higgs boson rate to its SM prediction, and can be seen in figure 2.12,

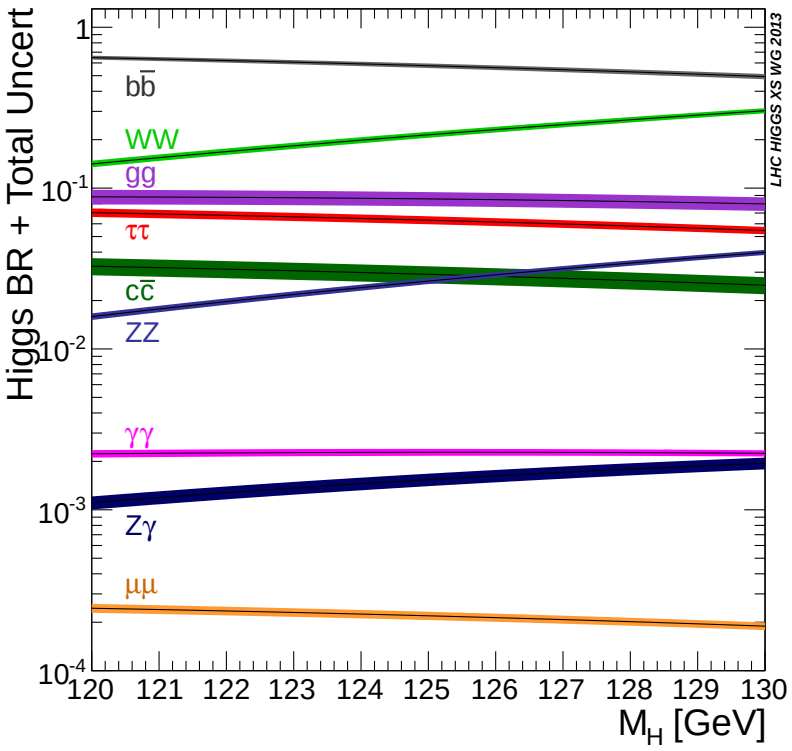


Figure 2.11: The branching ratio of the Standard Model Higgs boson are shown for values of m_H ranging from 120 to 130 GeV. The colored bands indicate theoretical uncertainties [2].

figure 2.13, and figure 2.14. There is generally good agreement between the experimental measurements and SM predictions, with all best fit values within two standard deviations of the SM predictions.

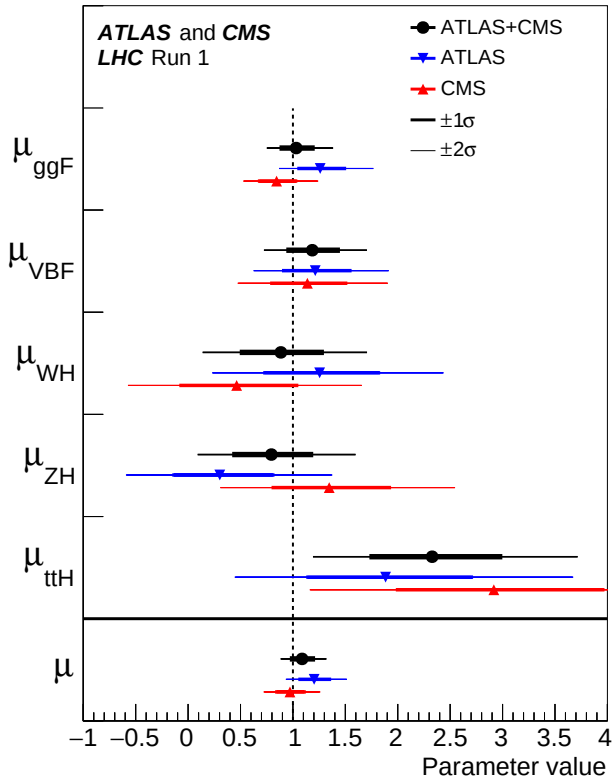


Figure 2.12: Best fit results for the production signal strengths for the combination of ATLAS and CMS data. The results from each experiment are also shown. The error bars indicate 1σ (thick lines) and 2σ (thin lines) intervals. The measurement of the global signal strength μ are also shown [3].

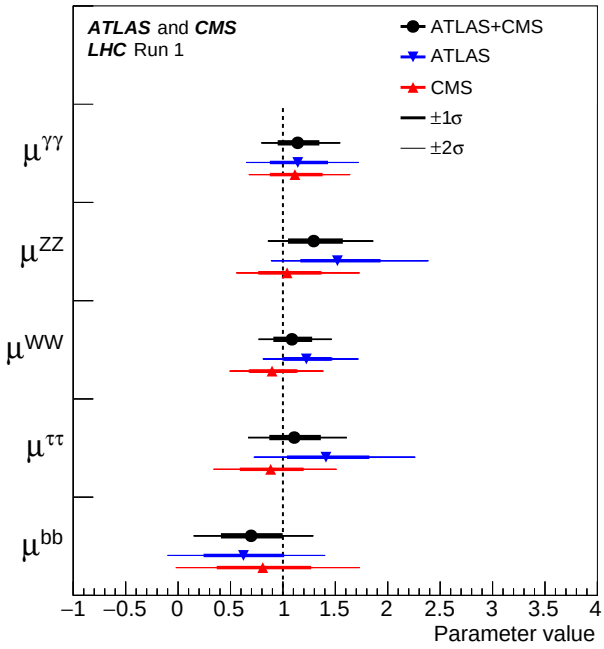


Figure 2.13: Best fit results for the decay signal strengths for the combination of ATLAS and CMS data, along with results from each experiment are shown. The error bars indicate 1σ (thick lines) and 2σ (thin lines) intervals [3].

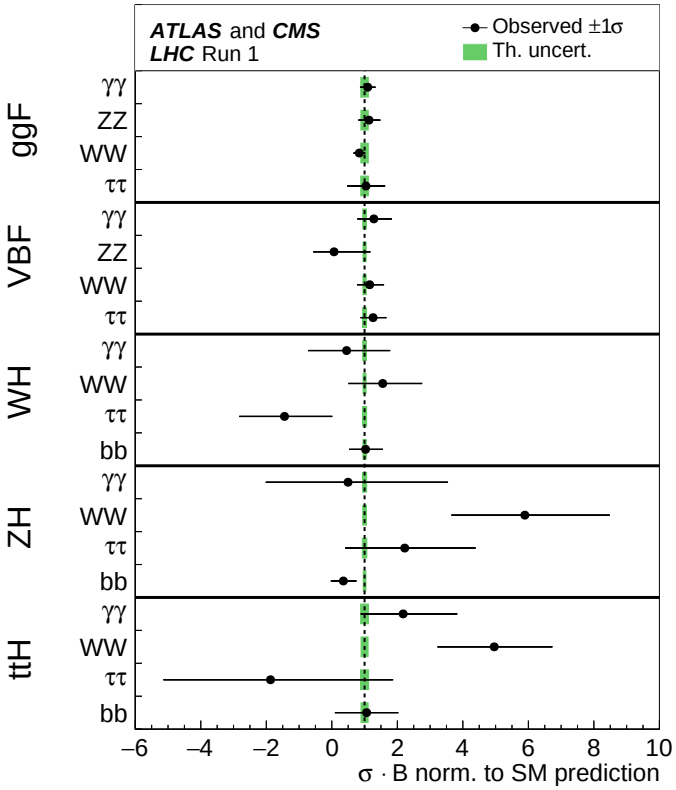


Figure 2.14: Best fit values of $\sigma_i.B^f$ for each specific channel i, as obtained from the generic parametrization with 23 parameters for the combination of the ATLAS and CMS measurements. The error bars indicate the 1σ intervals. The fit results are normalized to the SM predictions for the various parameters and the shaded bands indicate the theoretical uncertainties in these predictions. Only 20 parameters are shown because some are either not measured with a meaningful precision or not measured at all, and therefore fixed to their corresponding SM predictions. [3].

2.2.3 Problems in the Standard Model

The SM is a highly successful theory, and the experimental confirmation of the Higgs boson is its latest achievement. However, the SM has several shortcomings, the most glaring of which is the absence of gravity. The gravitational force is well understood at large scales by the General Theory of Relativity, but currently lacks any quantum description. The energy scale at which gravitational interactions become large is about 10^{19} GeV, and is known as the Planck scale. The Planck scale is about 16 orders of magnitude above the electroweak scale. Since the Higgs boson is a scalar, its mass is impacted by higher order corrections from loops that contain spin 0, 1/2 and 1 particles, which diverge as a function of the energy scale. If the SM was valid at all energy scales, these higher order corrections would cause the mass of the Higgs to be at the Planck scale. This is obviously in disagreement with the experimentally observed mass of the Higgs. One potential explanation for this observation is that the higher order corrections to the Higgs mass cancel out in a way such that we arrive at $m_H = 125$ GeV. This fine tuning is, however, not well motivated within the theory. This large discrepancy in the energy scales in the electroweak theory, Higgs mass, and electroweak vacuum versus the Planck scale, is known as the Hierarchy problem.

The absence of any explanation for the existence of Dark Matter (DM) is another shortcoming of the SM. First identified in 1933 [35] through the observation of galactic rotation curves, DM is named after its lack of electromagnetic interactions. By studying the speed of galaxy rotation, the existence of more mass than what could be accounted for by direct observations of the galaxy has been suggested. Additionally, other evidence for DM has been observed in measurements of the cosmic microwave background [36], and colliding clusters [37]. Particles that form DM have never been directly detected and evidence of DM remains to be seen at particle colliders. Beyond its density in the universe, and that it does not interact via the strong or the electromagnetic force, DM is not very well understood.

The Hierarchy problem and DM are just two questions out of a non-exhaustive list of

open questions that remain unanswered by the SM. Various models and extensions to the SM have been proposed to solve these problems.

2.2.4 Using the Higgs to search for new physics

Various models and extensions to the SM have been postulated to solve the problems discussed in section 2.2.3. Supersymmetry (SUSY) is perhaps the most popular amongst them. SUSY was proposed and developed in the 1970s [38; 39; 40]. It introduces a new symmetry that correlates bosons and fermions, and this symmetry is used to explain the higher order corrections. According to SUSY, for every fermion in nature, there must exist a boson superpartner with the same mass, and vice versa. However, till date there is no such experimental observation. This suggests that if SUSY exists, it must be a broken symmetry, and in order to explain the Hierarchy problem, the SUSY breaking scale must be such that the superpartner masses should not exceed $\mathcal{O}$ (1 TeV).

The Minimal Supersymmetric SM (MSSM) is the simplest supersymmetric extension of the SM. It adds the minimal amount of extra fields required to the current model: each SM fermion (boson) has one boson (fermion) superpartner. The MSSM model also introduces changes to the Higgs sector [41]: electroweak symmetry breaking is performed by two Higgs doublets, and this gives masses to both, up and down type quarks, and leptons. The extra Higgs doublets implies the existence of five new bosons: two neutral CP-even bosons, h, H (one of which is the 125 GeV Higgs boson discovered in 2012), one neutral CP-odd boson (pseudoscalar) a, and two charged bosons $\mathrm{H}^{\pm}$. The Next to Minimal Supersymmetric SM (NMSSM) is another extension of the SM and is very similar to MSSM [42]. The MSSM and NMSSM are both examples of models that add an extra scalar doublet to the Higgs sector of the SM. Such models are generally known as the Two-Higgs Doublet Models (2HDM) [43]. The extensions presented in the 2HDM models are found in various well motivated theories, such as SUSY [44] and axion models [45; 46]. The 2HDM models explain new phenomena related to the SM-like Higgs boson. For exotic production of a new resonance, for example

H, it must be heavier than the SM-like Higgs (h), which allows decays such as $H \to h + X$, where X can either be a Z, a, or another h. The SM-like Higgs boson can also decay exotically into a making decays such as $h \to aa$ possible, if a is lighter than $m_h/2 \approx 62.5$ GeV.

Addition of a complex scalar singlet to the 2HDM models, such that the singlet has no direct Yukawa couplings and acquires all of its couplings to SM fermions through its mixing with the Higgs doublets, allows for exotic Higgs decays of the form

$$h \to aa \to X\bar{X}Y\bar{Y} \tag{2.20}$$

where a is a pseudoscalar, and X,Y are SM fermions or gauge bosons. This setup is refered to as 2HDM+S. Depending on the value of Tanβ, where Tan$\beta = v_2/v_1$ (v_2, v_1 are the vacuum expectation values of the scalar doublets), various types of 2HDM+S models can be theorized:

- Type I (Figure 2.15): The branching ratio of a to SM particles is independent of Tanβ. The pseudoscalar couplings to all fermions are proportional to those of the SM Higgs, and the branching ratios are very similar to those of a simple scalar extension of the SM.

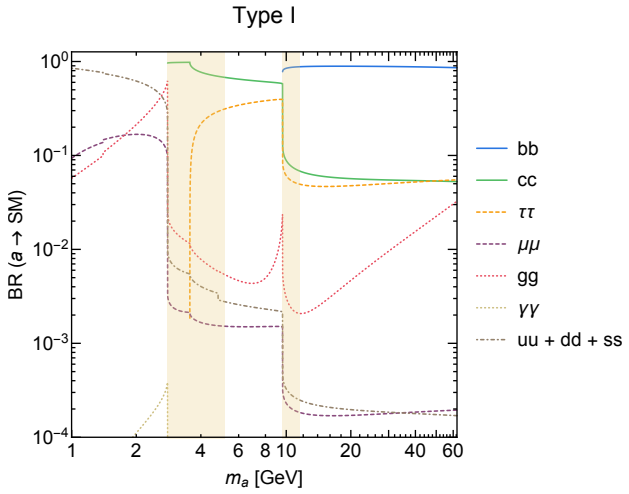
- Type II (Figure 2.16): Unlike Type I models, the branching ratios are Tanβ dependent, with decays to down-type fermions suppressed (enhanced) for down-type fermions for Tan$\beta < 1$ (Tan$\beta > 1$).

- Type III (Figure 2.17): The branching ratios are Tanβ dependent. For Tan$\beta > 1$, pseudoscalar decays to leptons are enanced over decays to quarks, and for Tan$\beta < 1$, decays to quarks are preferred over decays to leptons.

- Type IV (Figure 2.18): The branching ratios are Tanβ dependent.

The branching ratios of a to SM particles are only independent of Tanβ for Type I, and all types reduce to Type I for Tan$\beta = 1$. For Tan$\beta < 1$ and compared to NMSSM, the

pseudoscalar decays to up quarks and leptons are enhanced with respect to down quarks.

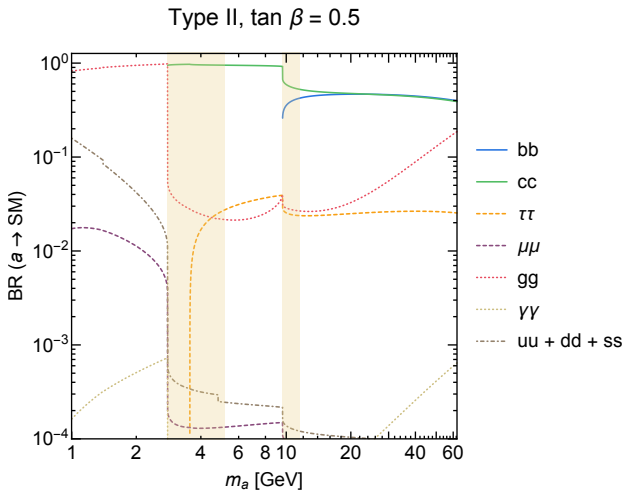


(a) Type I

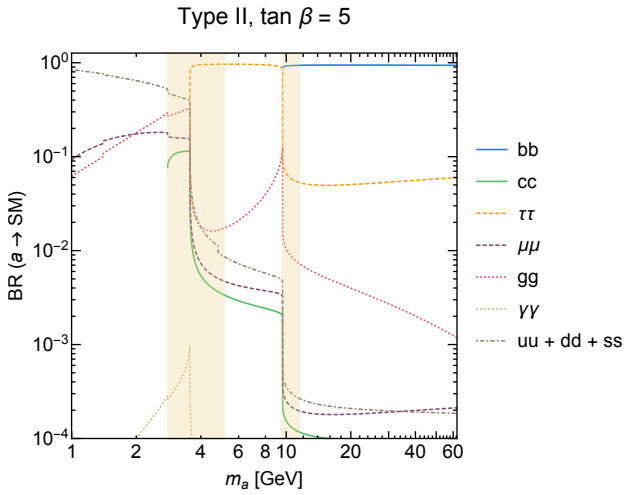
Figure 2.15: Branching ratios of a singlet-like pseudoscalar in the 2HDM+S Type I model [4].

These models are examples of how exotic physics can manifest itself in the Higgs sector. Exotic decays of the Higgs boson are a well motivated possibility even with the discovery of a Higgs particle consistent with the SM, and they may provide the only window into BSM physics at the LHC. The search for non-standard decays of the Higgs must be continued. Combined results from CMS and ATLAS have established an upper limit on branching ratio of Higgs to BSM physics, B_{BSM} equal to 0.34 at 95% CL, compared to an expected upper limit of 0.39. The negative log-likelihood scan is shown in figure 2.19. These results leave a large open phase space for Higgs decays to BSM physics. More details on these results can be found here [3].

The search for exotic decays of the Higgs Boson is being extensively pursued by both, the CMS and ATLAS experiments. The analysis explored in this thesis searches for decay of the 125 GeV Higgs boson to two light pseudoscalars, with each pseudoscalar subsequently decaying to a pair of photons. There are many theoretically well-motivated candidates for the pseudoscalar: it could be the light pseudoscalar of the NMSSM, a generic SM-singlet boson or any pseudoscalar that mixes with the CP-odd Higgses of the (N)MSSM. In most

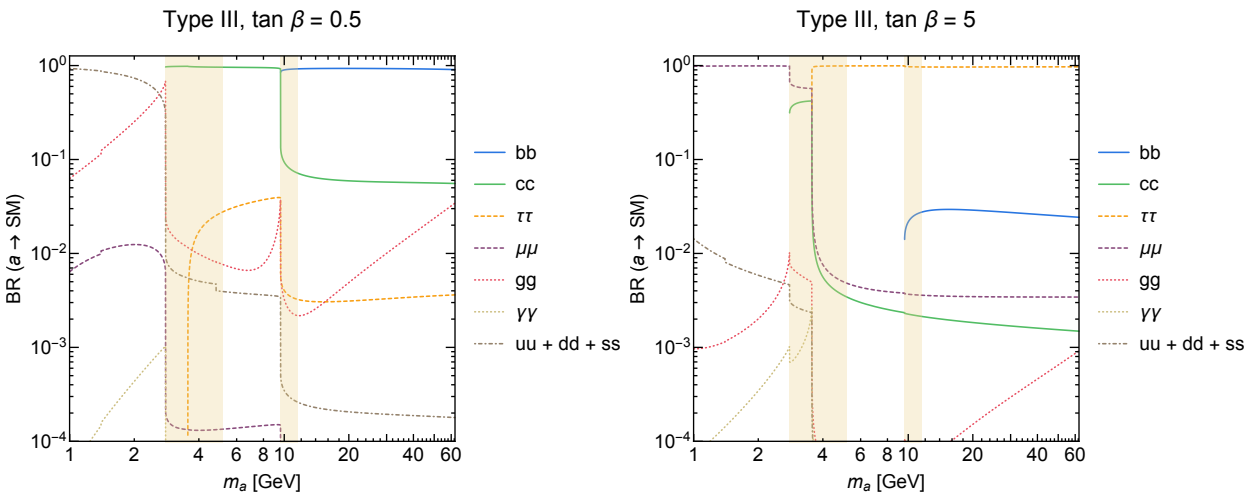


(a) Type II; Tanβ=0.5



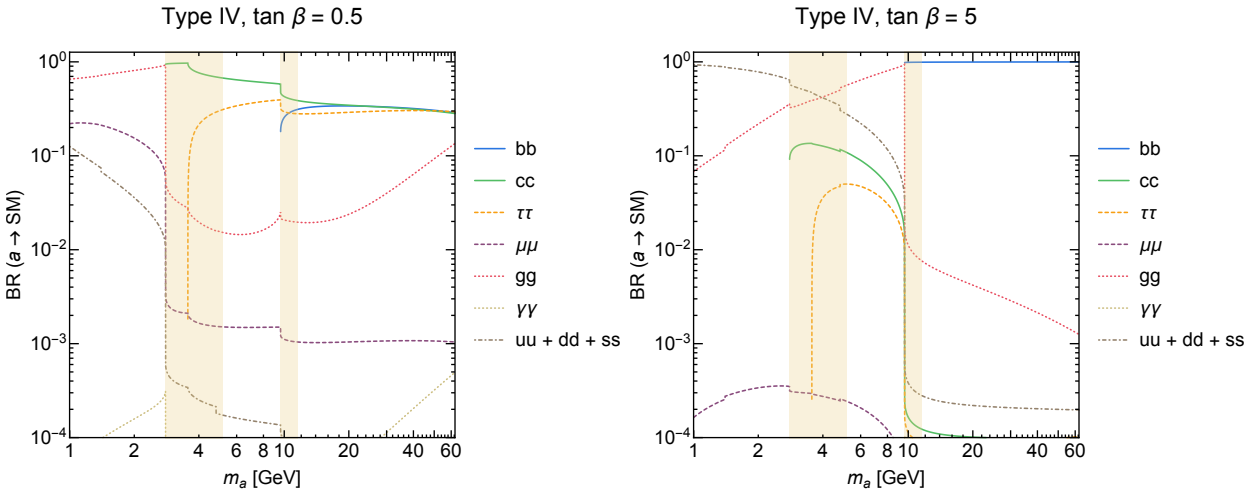
(b) Type II; Tanβ=5

Figure 2.16: Branching ratios of a singlet-like pseudoscalar in the 2HDM+S Type II model [4].



(a) Type III;Tanβ=0.5 (b) Type III;Tanβ=5.0

Figure 2.17: Branching ratios of a singlet-like pseudoscalar in the 2HDM+S Type III model [4].



(a) Type IV;Tanβ=0.5 (b) Type IV;Tanβ=5.0

Figure 2.18: Branching ratios of a singlet-like pseudoscalar in the 2HDM+S Type IV model [4].

models, the branching fraction of the pseudoscalar to photons is sub-dominant, however its clean experimental signature make it an interesting final state to explore. Additionally, in certain models where the pseudoscalar couples at the renormalizable level only to the Higgs, it may only be decay to photons. These reasons motivate the search described in Chapter 5.

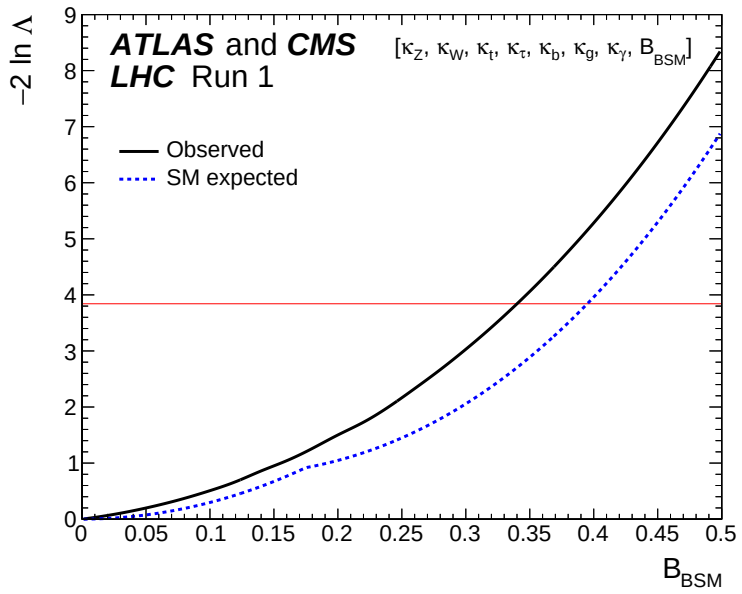


Figure 2.19: Observed (solid line) and expected (dashed line) negative log-likelihood scan of B_{BSM}, shown for the combination of ATLAS and CMS when allowing additional BSM contributions to the Higgs boson width. The parameters of interest from the list in the legend are varied in the minimisation procedure. The red horizontal line at 3.84 indicates the log-likelihood variation corresponding to the 95% CL upper limit. [3].

Chapter 3

The LHC and the CMS experiment

The analysis presented in this thesis uses pp collision data that was delivered by the Large Hadron Collider (LHC) and was recorded by the Compact Muon Solenoid (CMS) experiment. Section 3.1 provides an overview of the LHC and Section 3.2 describes the CMS experiment.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the largest particle accelerator and collider in the world [47], and it operates at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. It is a circular proton-proton collider with a circumference of 27 km, and is located in the tunnel that previously housed the Large Electron Positron (LEP) collider [48]. Unlike the previous large accelerators, LEP and Tevatron, which collided electrons-positrons and protons-antiprotons, both of the LHC beams contain protons. The use of protons is motivated by the fact that the LHC was built to test the Standard Model at unprecedented energies, provide data to probe rare processes, and finally to discover new particles. Additionally, the LHC was built as a pp and not as a $p\bar{p}$ machine since antiprotons are hard to produce and are unstable, which would limit the maximum achievable luminosity. The proton beams are collided at four points around the ring, which can be seen in figure 3.1.

Particle detectors, which detect and analyse the LHC collisions, surround the collision points. Two out of the four detectors, ATLAS [49] and CMS [10], are multipurpose in

Figure 3.1: The LHC collider ring overlaid on a map of Switzerland and France, with the four main LHC experiments are shown [5].

their conceptual design. They are complex detectors which are designed with the goal of discovering new processes, in addition to measuring SM ones. LHCb (Large Hadron Collider beauty) investigates properties related to the b quark [50]. ALICE (A Large Ion Collider Experiments) studies lead ion collisions at the LHC, with the goals of understanding the high energy densities that are produced as a result of collisions of massive particles [51].

3.1.1 The Large Hadron Collider Design

In order to achieve its primary goal of providing high luminosity proton-proton collisions at 13 TeV, the LHC must successfully accelerate a large number of protons and maintain them at a constant energy, in addition to organizing them into bunches which can be focused and collided at precise points. The complicated system of pre-accelarators that helps with these tasks is shown in figure 3.2.

The journey of producing a proton beams starts with an electromagnetic field that is used to strip off electrons from the hydrogen atoms to obtain protons. The protons are then

passed through a linear accelerator which uses radio frequency (RF) cavities, Linac 2. Here, alternating positive and negative electric fields are used to move the particles through the accelerator. Owing to the effect of the RF cavities, protons remain in separate bunches. This initiates the formation of the bunch structure which is used later for collisions. Through this journey, beams are kept focused with the help of quadrupole magnets. The beams attain an energy of 50 MeV by the time they arrive at the end of Linac 2. The proton beams are then injected into the Proton Synchroton Booster (PSB), the first circular accelerator in the pre-accelarator chain. Here, protons are accelerated to 1.4 GeV. Next, the proton beam is injected to the Proton Synchrotron (PS), which is a 600 m long circular accelerator consisting of RF cavities, and are accelerated to 25 GeV. Thereafter, the beam is sent to the Super Proton Synchrotron (SPS), the last accelerator before injection into the LHC, and the protons reach an energy of 450 GeV here.

The process of travelling through the various pre-accelerators gives the protons a bunch structure. The PS produces 72 bunches, separated by 25 ns in time, which are then injected into the SPS. Another important feature of this process is the gap in the proton bunch structure. In the absence of any gap, known as the injection kicker rise time, between the proton bunches, the changing magnetic field would direct particles out of the accelerator and produce extremely high amounts on unsafe radiation around the PS. Similarly, a gap between bunches is also required when protons travel from SPS to LHC. The entire injection process, which takes about 3 minutes, is repeated until the LHC is filled with over 2000 bunches [47].

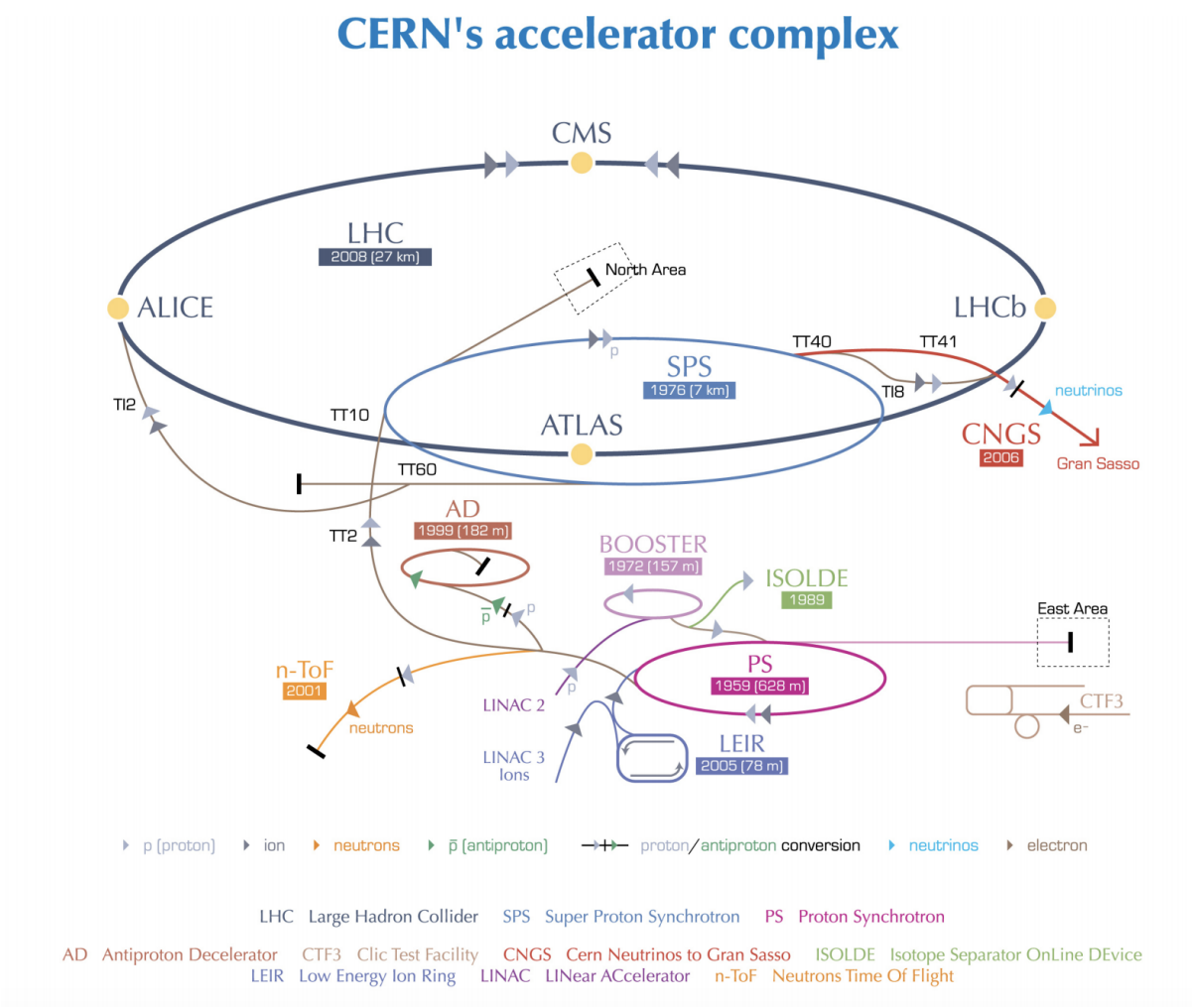


Figure 3.2: Sketch of the CERN accelerator complex. The four main LHC experiments, CMS, ATLAS, LHCb and ALICE, are also shown [6].

3.1.2 The Large Hadron Collider Operation

The LHC is a proton-proton collider, as opposed to a proton-antiproton collider, and therefore the two counter-rotating beams must be stored in separate rings, and be accelarated separately. Twin-bore superconducting magnets, such as the one shown in figure 3.3, are used to achieve this. They surround the two rings and accelerate them both. Quadrupole magnets are used at the four collision points to focus the beam, which cross at an interaction point at the center of each detector. There are over 6000 superconducting magnets in the LHC, and they are kept below their critical temperature of 1.9K with the help of liquid helium cooling.

The protons, when first injected into the LHC, must be accelerated with increasing RF frequencies. This process occurs over many turns through the machine. Simultaneously, the dipoles apply increasingly high magnetic field which applies force to bend the beam. Once protons reach a maximum energy, the quadrupole magnets help reduce the total transverse area of the beam and elongate the bunches. This process is called *squeezing*, and the shape produced as a result of this process determines the area in which the collisions occur, the beam spot, for the CMS detector. On ensuring that the beams have reached a stable energy and have been squeezed, the LHC is ready to provide collisions to the various experiments around the ring, and data taking can begin. Over a period of time, as collisions occur, the number of protons in the beam decrease, and when it reduces down to a value to require a new fill, or in case of any instability, the beam is dumped into a cavern filled with material to absorb the energy.

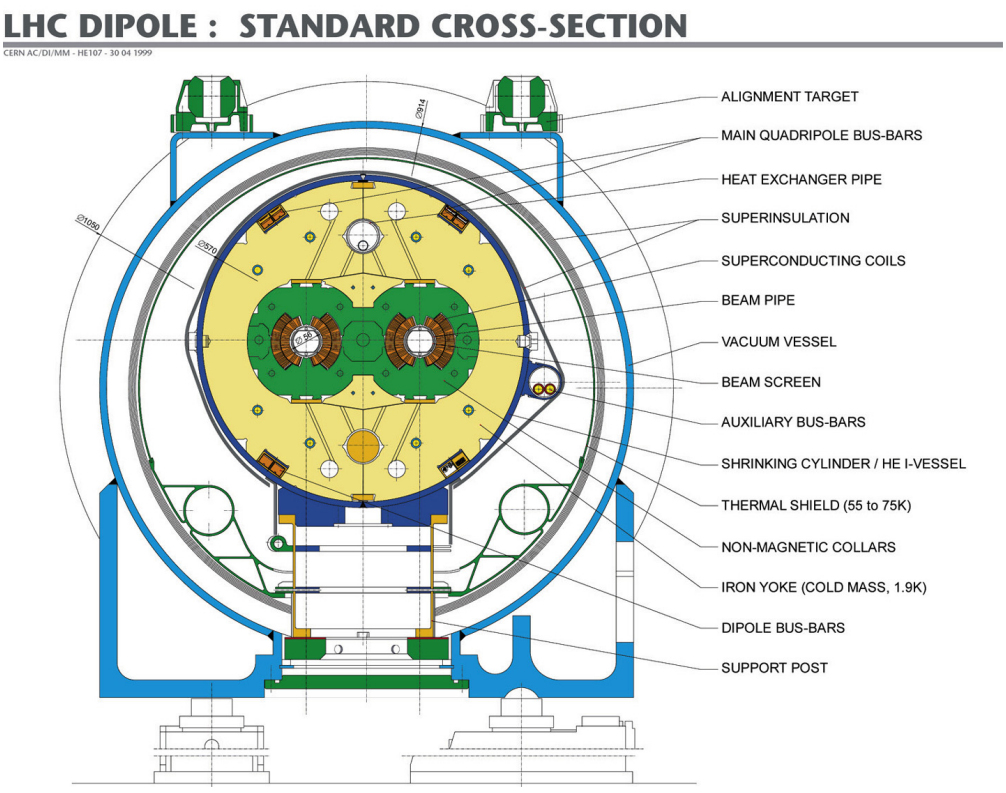


Figure 3.3: Cross section of an LHC dipole magnet [7].

3.1.3 Luminosity and Pile-Up in Proton-Proton Collisions

An important factor in the design of the LHC is the frequency of the collisions it can provide. The frequency should be high enough to produce various SM and BSM particles, which can be observed by the detectors. The number of events of a given type are given by

$$N_{\text{event}} = \int dt L \sigma_{\text{event}} \tag{3.1}$$

where L is the luminosity delivered by the LHC, and σ_{event} is the cross-section of the particular process. The cross-section value can vary over many orders of magnitudes depending on the type of process. Since processes that produce high mass particles are very rare, a large amount of luminosity is required to produce them. Another important factor to consider is that processes such as jet production occur very frequently, and must be produced at high enough rates such that they can be distinguished from other more common processes.

The instantaneous luminosity at the LHC is defined as

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta_*} F \tag{3.2}$$

where, N_b is the number of protons per bunch ($\sim 10^{11}$), n_b is the number of bunches in the beam ($\sim 10^3$), f_{rev} is the frequency with which the beam travels around the ring (1 kHz), γ_r is the relativitic gamma factor (~ 7000), ϵ_n is the normalized transverse beam emmittance ($\sim 4mm\,\mu$m), and β_* is the β function at the collision point which describes the transverse displacement of particles in the beam ($\sim 0.5m$). F is the reduction factor that arises due to the geometry of the beam crossings

$$F = \left(1 + \left(\frac{\theta_c \sigma_z}{2\sigma *}\right)^2\right)^{-1/2} \tag{3.3}$$

where θ_c is the crossing angle of the beams, σ_z is the RMS of the bunch length in the z

direction, and $\sigma*$ is the RMS of the bunch length in the transverse direction.

Figure 3.4 shows the integrated luminosity versus time. 163 fb^{-1} of pp collision data was delivered to the CMS detector between 2015 and 2018, out of which 150 fb^{-1} was recorded. The analysis described in this thesis makes use of 132 fb^{-1} of data collected between 2016 and 2018.

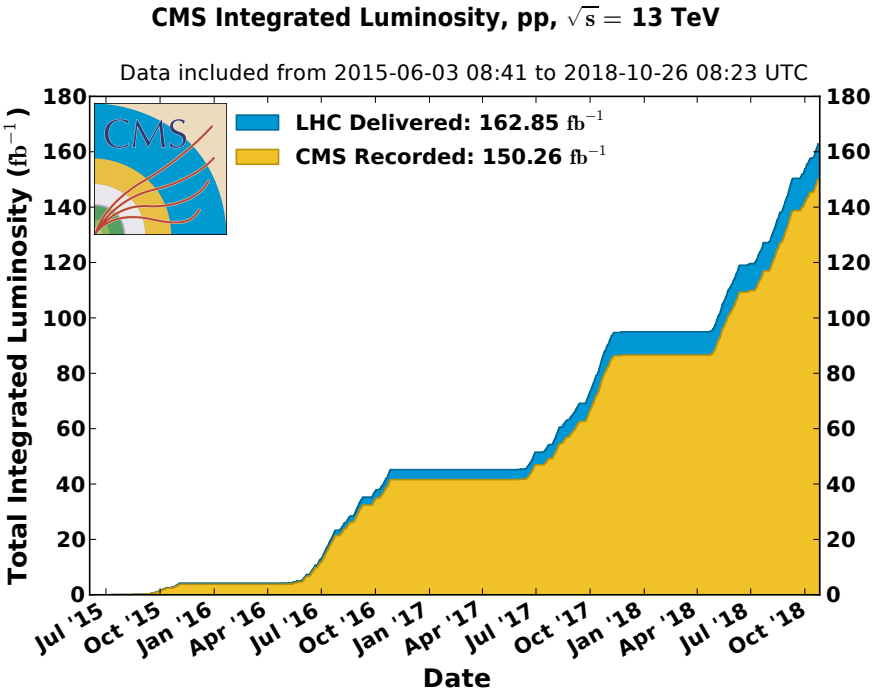


Figure 3.4: Total Integrated luminosity versus time delivered to the CMS (blue), total integrated luminosity recored by CMS (yellow) during stable beams for pp collisions at 13 TeV center-of-mass energy in 2015-2018 [8].

During Run 2 of the LHC, peak instantaneous luminosity of $\sim 10^{34} cm^2 s^{-1}$, which is roughly twice as large as the designed instantaneous luminosity, was observed in CMS. A result of this high instantaneous luminosity at the LHC is pile-up. Pile-up can be in-time, in which multiple collisions occur during the same bunch crossing, and out-of-time, in which detectors are read out over intervals longer than the time between bunch crossings. Two important activities associated with the amount of pile-up in a dataset are the number of primary vertices per event, NPV, and the average number of inelastic interactions per bunch

crossing, $\langle\mu\rangle$ which is defined as

$$\langle\mu\rangle = \frac{L \cdot \sigma_{\text{inelastic}}}{N_{\text{bunch}} \cdot f_{\text{LHC}}} \tag{3.4}$$

Figure 3.5 shows the mean $\langle\mu\rangle$, for data collected between 2015 and 2018. For the CMS experiment, the peak pileup for the full Run 2 dataset is approximately 34 interactions per bunch crossing.

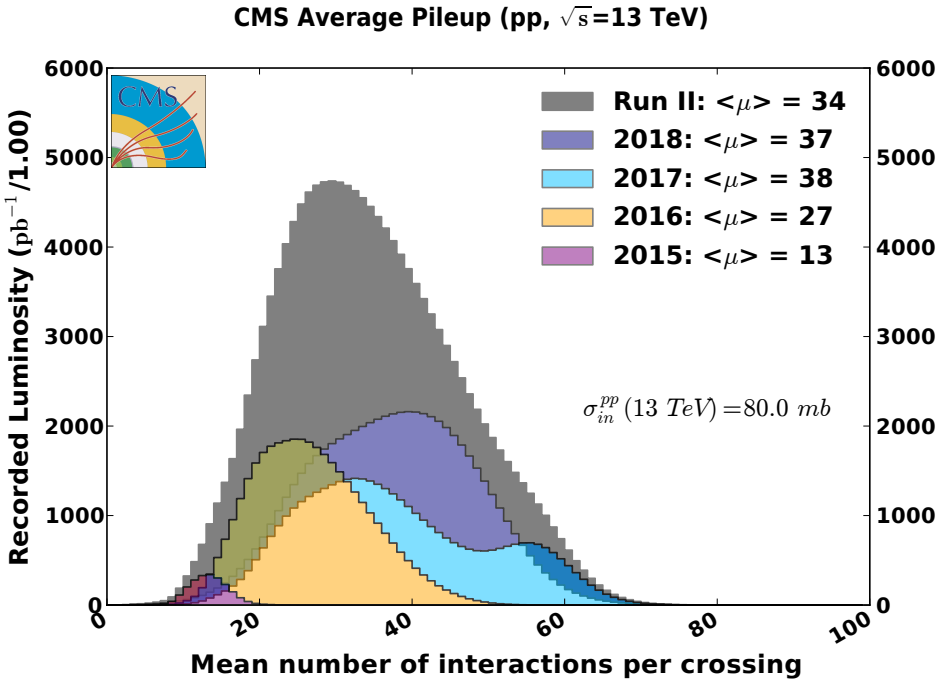


Figure 3.5: Mean number of interactions per bunch crossing for 13 TeV center-of-mass energy *pp* collision data collected between 2015 and 2018 by CMS. Data collected during different years are indicated by different colors [8].

3.2 The Compact Muon Solenoid Experiment

The Compact Muon Solenoid (CMS) is one of the two multipurpose detectors at the LHC. It is located at interaction point 5 (P5) in Cessy, France. CMS has a cylindrical shape, has a length of 21 m, a diameter of 15 m, total weight of 14 000 tonnes, and consists of a barrel and two endcap parts at either end of the barrel. Individual subdetectors, which specialize in providing different types of event information, are stacked together concentrically and are shown in figure 3.6. A primary feature of the CMS detector is the superconducting solenoid (Section 3.2.2) which produces a high magnetic field and enables CMS to achieve excellent momentum resolution. The muon tracker system (Section 3.2.5) is interspersed within the return yoke layers. The CMS calorimeters, lead-tungstate crystal electromagnetic calorimeter (ECAL) (Chapter 4), brass scintillator hadron calorimeter (HCAL) (Section 3.2.4), and the silicon tracker Section 3.2.3) are contained within the bore of the magnet coil. They will be described in more detail in the following sections.

3.2.1 Coordinate system

The CMS detector uses a right-handed coordinate system, as shown in figure 3.7. The collision point defines the origin of the CMS coordinate system, with the positive y-axis pointing vertically upward, the positive x-axis pointing radially inwards towards the center of the LHC, and the z-axis pointing along the beam direction (towards the Jura mountains).

The azimuthal angle, ϕ, is measured around the beam axis such that $\phi = 0$ along the positive x-axis. The polar angle, θ, is measured with respect to the beam axis. Some variables are widely used to characterize the kinematics of particles produced in LHC collisions, such as the transformation of the polar angle called the pseudorapidity (η)

$$\eta = -ln(tan\frac{\theta}{2}) \tag{3.5}$$

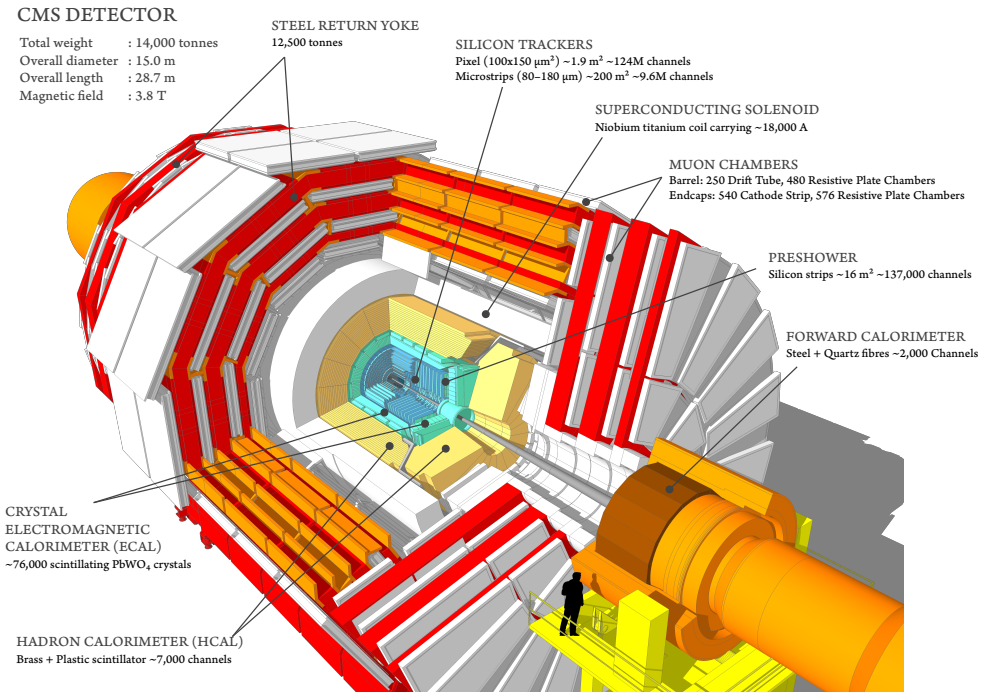


Figure 3.6: Sketch of the CMS experiment at the LHC. The different subdetectors are also highlighted [9].

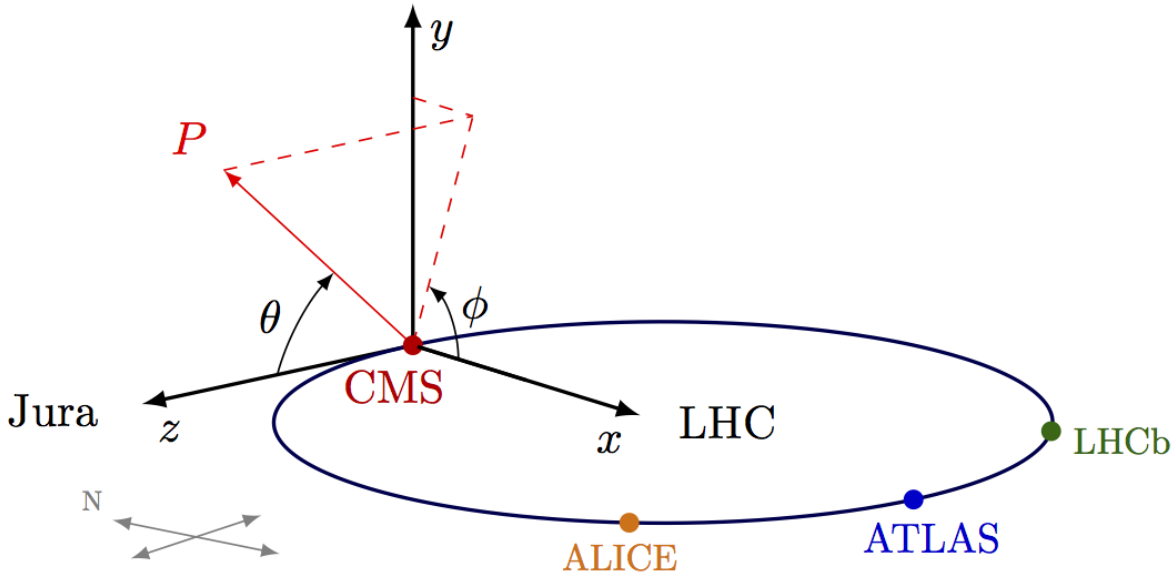


Figure 3.7: CMS coordinate system. The x, y, and z-axes are shown along with the geometric angles θ, η, and ϕ.

η is an often used quantity since differences in pseudorapidity are Lorentz-invariant.

Angular distance between two particles is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.6}$$

p_T and E_T are the momentum and energy measured transverse to the beam direction, respectively, and are computed from x and y components.

3.2.2 Magnet

The superconducting magnet, designed to reach a 3.8-T magnetic field, is one of the most prominent features of CMS. Weighing at about 220 tonnes, with a diameter of 6m, and length of 12.5m, it is the largest superconducting magnet ever built, and is located between the CMS calorimeters and muon detectors. Due to the number of ampere-turns required for generating a field of 3.8T, the winding of the the magnet is composed of four layers. The magnet's geometry is such that the the direction of the magnetic field is different at the CMS tracker and at the CMS muon system. This feature of the magnet is intentional since it provides an optimal measurement of the muon momentum by causing two bends in the path of the reconstructed muon, instead of only one. Figure 3.8 provides an artistic view of the superconducting magnet showing the five modules composing the cold mass inside the cryostat. The CMS magnet is also composed of a return yoke, which has an outer diameter of 14m and consists of three layers, and provides structural support to the entire experiment. Figure 3.9 shows a fish-eye view of the yoke with its clearly visible layers.

A distinctive feature of the CMS magnet is the high ratio between stored energy and cold mass (11.6 KJ/kg), which may cause large mechanical deformations of the magnet during energising. This creates unique engineering challenges pertaining to the magnet's ability to bear high amount of strain. Figure 3.10, which compares the energy over mass ratio as a function of the stored energy, shows that the CMS magnet can bear much larger strain and

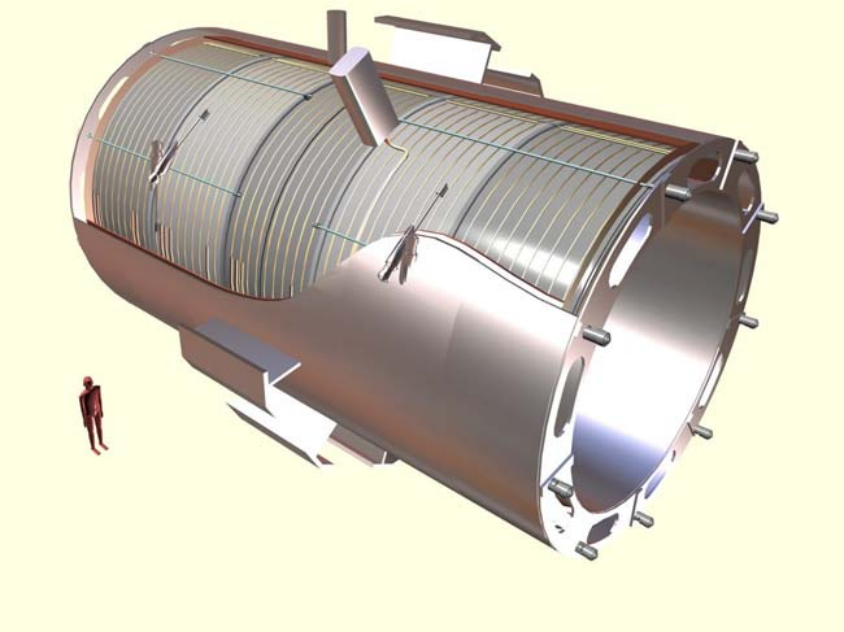


Figure 3.8: An artist's view of the CMS magnet. Five modules which compose the cold mass inside the cryostat, along with the magnet supporting system are shown [10].

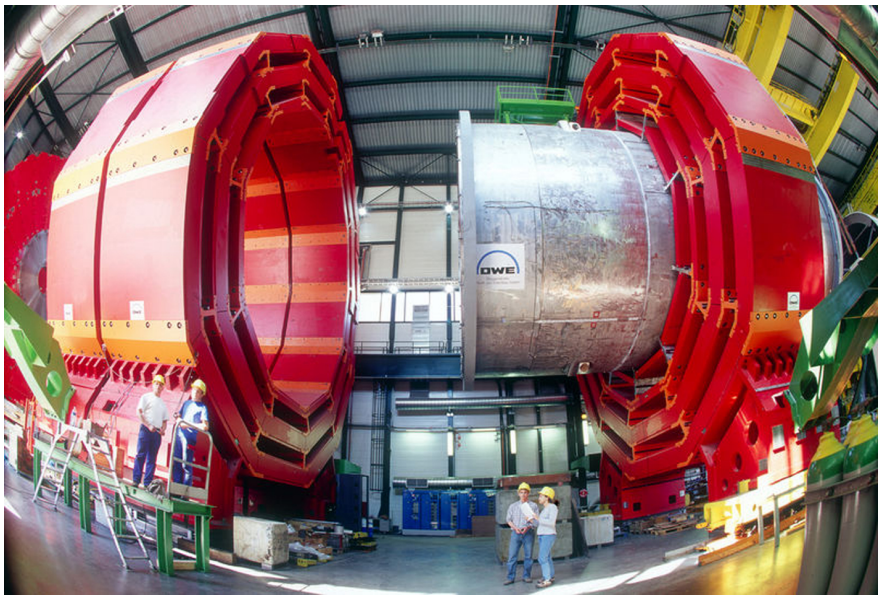


Figure 3.9: Fish-eye view of the CMS magnet yoke [11].

store greater energy than the other detector magnets.

To meet the requirements of providing accurate momentum and charge measurements of particles, bending of the charged particles in the magnetic field is essential. Additionally, as the p_T of the particles increases, its curvature decreases and it is thanks to the high magnetic strength provided by the CMS magnet, coupled with the excellent performance of the CMS tracker detector, which ensures that momentum of highly energetic particles can be accurately measured.

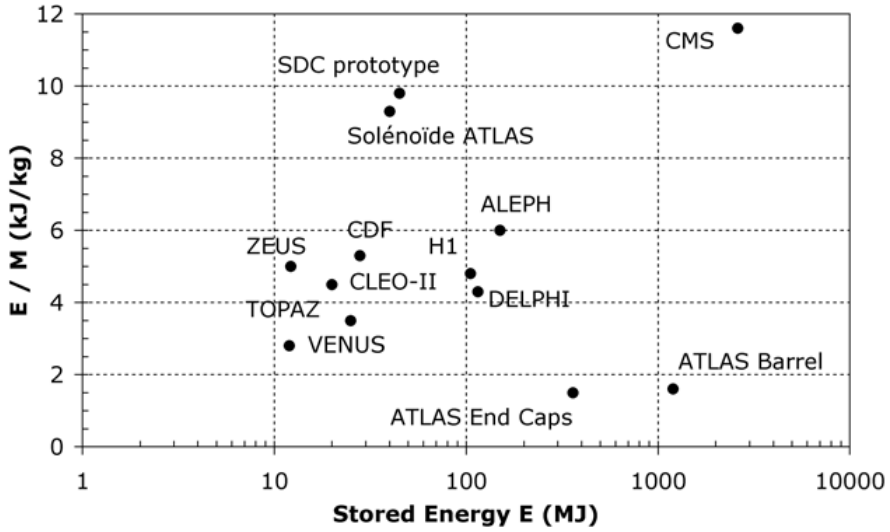


Figure 3.10: Energy over mass ratio (E/M) versus stored energy shown for several detector magnets [10].

3.2.3 Tracker

The tracker is the innermost layer of CMS, and its primary purpose is to provide accurate reconstruction of charged particle trajectories, such that the momenta of the particles can be precisely measured. Additionally, the tracker should be able to provide precise reconstruction of secondary vertices originating from e.g b quark decays. It utilizes the fact that charged particles, when moving through matter, lose energy by ionization. The particle trajectories, known as *tracks*, are built from different interactions of the incoming particle with the tracker. These interactions are referred to as *hits*. The primary goals of the CMS tracker have to be achieved while also ensuring that the particles are disturbed the least amount possible while passing through it. Therefore, in order to minimise multiple scattering, bremsstrahlung, photon conversion and nuclear interactions, the tracker should be made using as little material as possible. Being the innermost layer of CMS, the tracking sensors and readout electronics encounter harsh radiation conditions. This translates into the need for radiation hard materials. To provide precise vertexing, the tracker should also reach as close as possible to the beamline and at the same time it must also be ensured that hits from nearby particles are not assigned ambiguously. This requirement imposes a high detector segmentation of order 100 μm.

The CMS tracker is constructed entirely out of silicon. It has a length of 5.8 m and a diameter of 2.5m. A total of 75 million read-out channels result in a very fast response and provide a precision of 10 μm in position measurement, and transverse momentum resolution of 1-2 % for $p_T \approx 100$ GeV. A schematic diagram of the Phase-0 version of the tracking system is shown in figure 3.11.

In regions with radii below 10 cm, the track density is higher and the tracker is constituted of pixelated sensors which ensure an occupancy of less than 1%. The pixel detector is the closest detector to the interaction point and its major performance goals include providing precise impact parameter measurements which are necessary for efficient vertexing

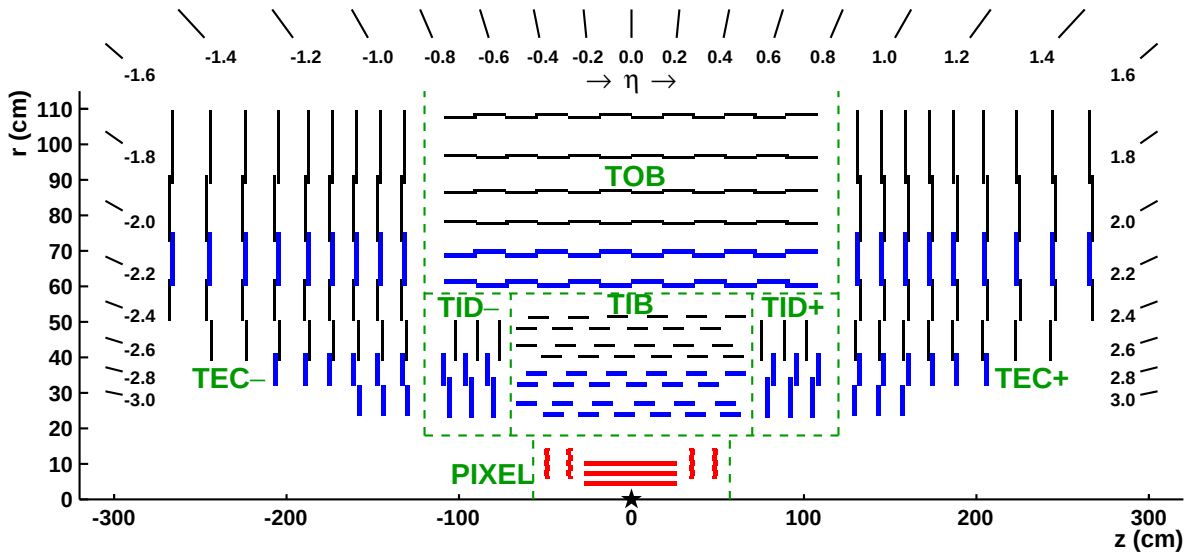


Figure 3.11: Schematic cross section of the Phase-0 CMS tracker in the r-z plane. Each line represents a detector module [12].

and providing seeds for track reconstruction. The pixel detector is composed of three barrel layers and two forward disks, located on each cylinder. They extend the acceptance of pixel detector to $|\eta| < 2.5$. To keep the temperature from rising too much above the operating temperature of $-10°$C, a cooling system is installed in the pixel detector. In 2017, the pixel detector was replaced with four layers in the barrel region and three disks in the forward region. This new design is referred to as the Phase-1 version and figure 3.12 displays its differences with respect to the Phase-0 version.

The strip tracker detector is placed after the pixels and it covers the regions with radii between 20 cm and 116 cm. Charged particles enter the strip detector after passing through the pixel layers. Strip detectors measure the position of particles in one direction. The strips have a typical cell size of 10cm $\times 80\mu$m and this ensures that the occupancy of a single strip is low. Moving further outwards, in the outer tracking region, the strips are longer which allows for a wider strip separation. The strip tracker detector is constituted of several subsystems: the tracker inner barrel (TIB), inner disks (TID), outer barrel (TOB), and tracker endcaps (TEC). These subsystems are highlighted in figure 3.11, and a photograph of the tracker

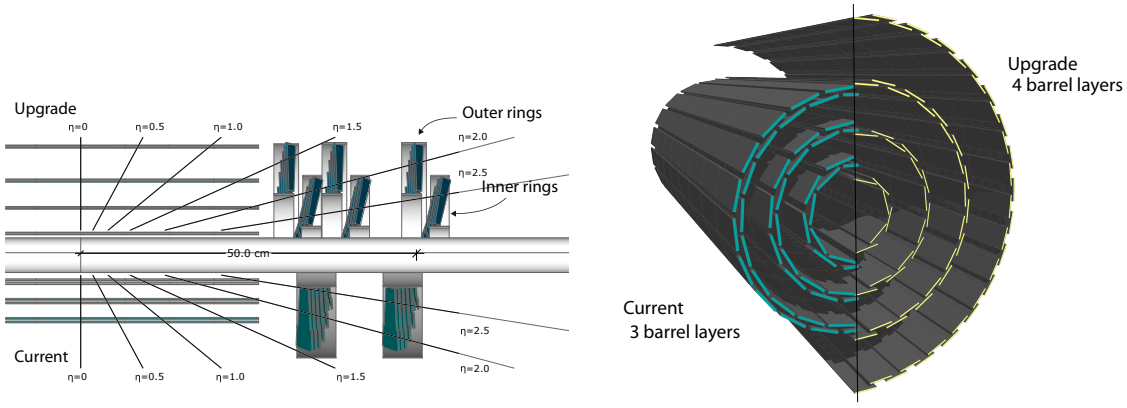


Figure 3.12: Comparison of the Phase-0 and Phase-1 pixel detectors. The left figure shows the layout of the various layers and disks in the Phase-0 and Phase-1 configurations, and the right figure shows the pixel barrel layers in a transverse-oblique view [13].

detector can be seen in figure 3.13.

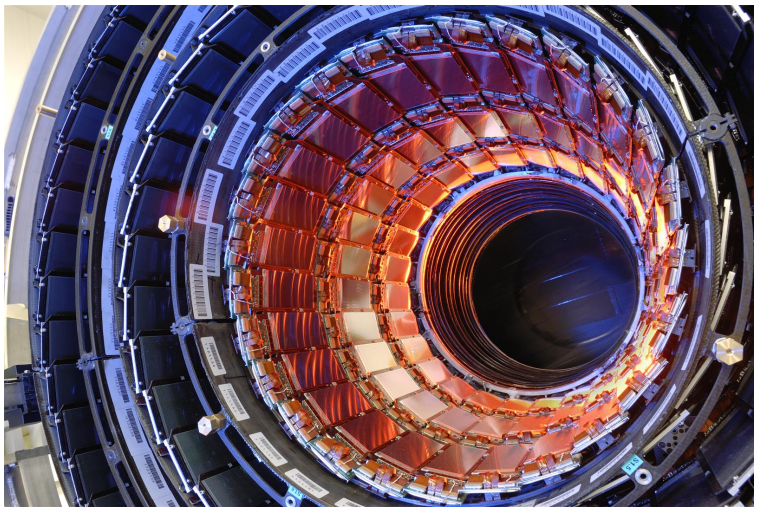


Figure 3.13: Photograph of the CMS strip tracker consisting of three layers of silicon modules [14].

The TEB/TID provide measurements of the particle trajectory in the r-ϕ direction, cover $r < 55$ cm and $|z| < 118$ cm, and deliver position measurement resolution of $\approx 13 - 38\mu$m. The TOB and TEC are made up of six and nine layers, respectively, and deliver position measurements in the r-ϕ direction with a resolution of $\approx 18 - 48\mu$m. In order to minimize the effect of radiation damage, the strip tracker is maintained at a temperature of -20°C.

The CMS tracker is the largest silicon tracker ever built with an active silicon area of

about 200 m^2. An overarching goal of the tracker is to maximize the measurement of the particle trajectory while keeping the material budget at a minimum. The tracker material budget, driven by the needs for cooling, structural support and cabling, grew beyond the original design. It reaches around $2X_0$ near the endcap transition region, where X_0 refers to radiation length which corresponds to the distance over which the energy of a charged particle is reduced by a factor of e.

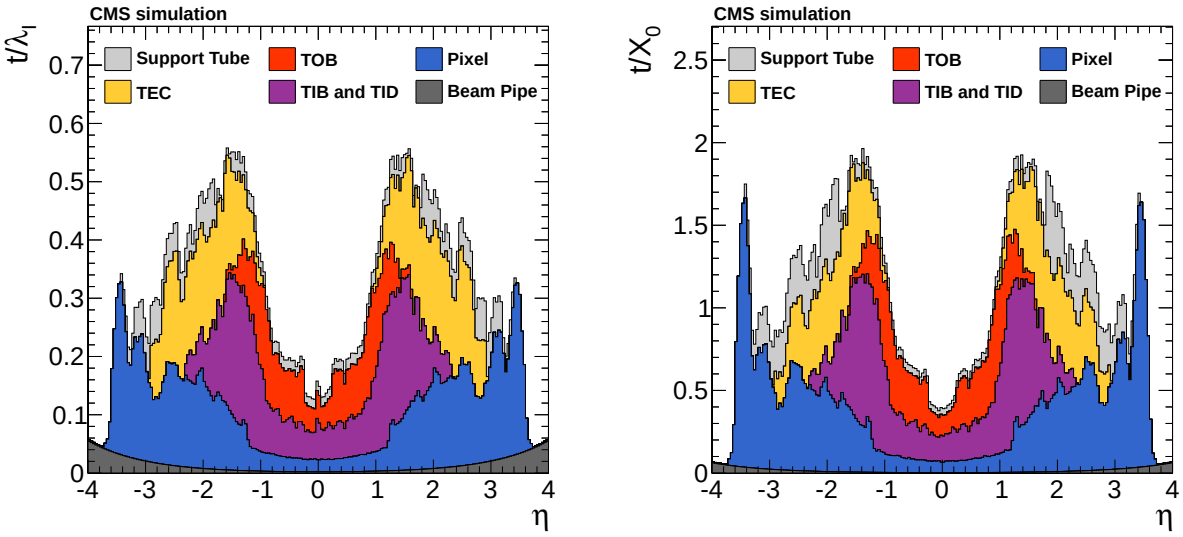


Figure 3.14: Total thickness, t, of the tracker material covered by a particle, is shown as a function of pseudorapidity η, expressed in units of radiation length, X_0 (left), and interaction length, λ_1 (right) [12].

3.2.4 The Calorimeter Detectors

Beyond the tracker, the next layers of CMS are the calorimeters. Calorimeters operate on the basic principle of making the particles interact with their active material such that they deposit most of their energy within the calorimeter. The size of the generated signal can then be used to infer the energy of the original particle. The position of the energy deposition of the particle helps in the particle's position measurement. The CMS calorimetry system is made up of an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL). Photons and electrons have a short radiation length and the electromagnetic showers created

by them are contained in the ECAL. The ECAL will be described in detail in Chapter 4, and the HCAL detector will be described in the following Section 3.2.4.1.

3.2.4.1 Hadronic Calorimeter

Quarks and gluons, due to the influence of strong force, undergo the process of hadronization. Hadronization is the process in which a colored particle produces multiple hadrons. The unstable hadrons further decay into stable particles, such as leptons and other stable hadrons. The hadronization process ultimately produces a conglomerate of particles, such as electrons (along with photons from bremsstrahlung), muons, pions, kaons, protons and neutrons. All these particles are combined to form an object known as *jet*. Precise measurement of the energy of hadrons is one of the principal goals of the hadronic calorimeter (HCAL). The HCAL is designed with the objective of detecting and absorbing hadrons, such that only muons and very weakly interacting particles such as neutrinos escape detection by the HCAL. It is precisely for this reason that the HCAL is designed to be as hermetic as possible; layers of the detector are staggered to make certain that there are no gaps that would allow hadrons to escape undetected.

Due to the compactness of CMS, HCAL had to be physically contained between the ECAL and the CMS solenoid. To ensure that any punch-through hadronic particles are also absorbed, a part of the HCAL is also located just behind the magnet coil. In total, the HCAL is divided into four parts: the barrel (HB), outer barrel (HO), endcap (HE) and forward (HF) section. These parts are highlighted in figure 3.15.

The HB and HE are sampling calorimeters, made of alternating layers of brass or steel which act as absorber materials, and fluorescent plastic scintillator tiles. The geometry of HB is as follows: it consists of 36 segments that cover a pseudorapidity region -1.4 $< \eta <$ 1.4. This translates to 2304 towers with a segmentation $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$. Figure 3.16 displays a photograph of the HB detector and numbering scheme for the HB wedges.

The HE consists of 14 towers in the η direction with a segmentation of 5° in ϕ and covers

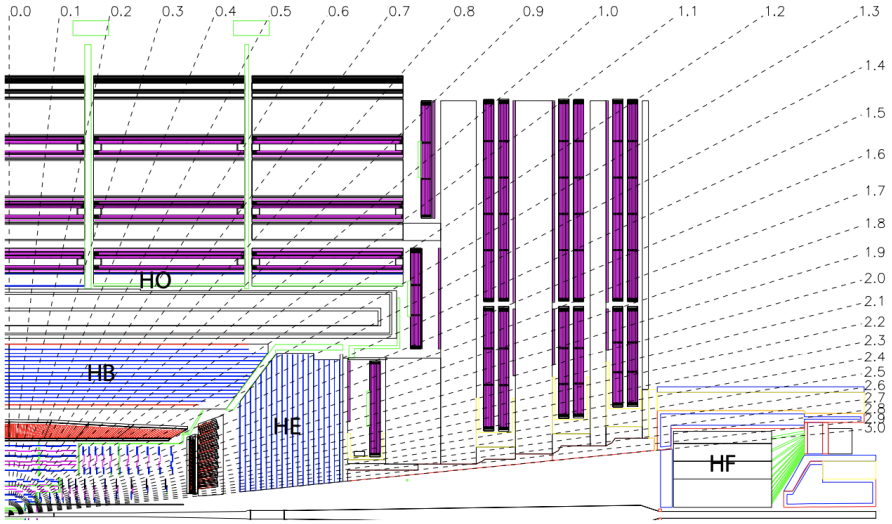


Figure 3.15: Layout of an HCAL quadrant, showing the barrel component (HB), the endcap (HE), the forward (HF) and the outer calorimeter (HO) [10].

a pseudorapidity region $1.3 < \eta < 3.0$. A layout of the HE calorimeter can be seen in figure 3.17. On hitting an absorber plate, the hadronic particle undergoes an interaction to produce several secondary particles. The secondary particles, by traversing through the absorber material, undergo interactions themselves and this process results in a cascade of particles. During this process, the blue-violet light emitted by the scintillators is absorbed by optical fibres that shift the wavelength of the light to the green region of the spectrum, which is finally sent by clear optical cables to readout boxes. As is typical for scintillators, the amount of light collected is proportional to the energy of the particle. The optical signals are converted into electronic signals by photodiodes and they are ultimately sent to the CMS data acquisition system.

The HO detector, placed just behind the magnet coil, acts as a tail-catcher and ensures that no hadronic particles escape detection. It consists of a layer of plastic scintillator tiles with a thickness of 10 mm and covers the pseudorapdity region -1.26 $< \eta <$ 1.26 . The scintillator tiles are grouped in 30° sectors in ϕ such that they match the segmentation of the muon chambers to which the HO is attached. Figure 3.18 shows the layout of the HO

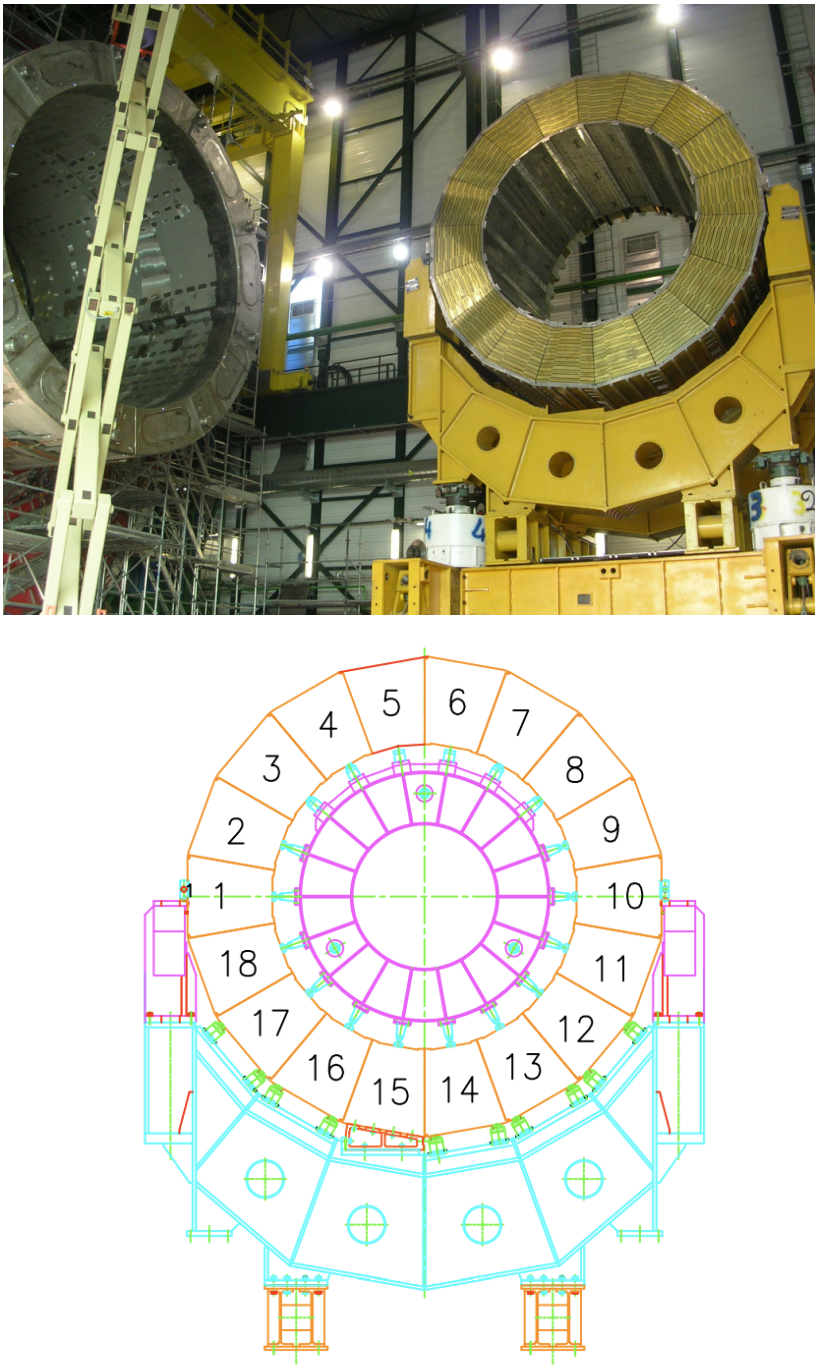


Figure 3.16: (Left) Photograph of HB minus, removed from the vacuum tank and rotated[15] and (right) numbering scheme for the HB wedges are shown. Wedge 1 is on the inside (+x direction) of the LHC ring [10].

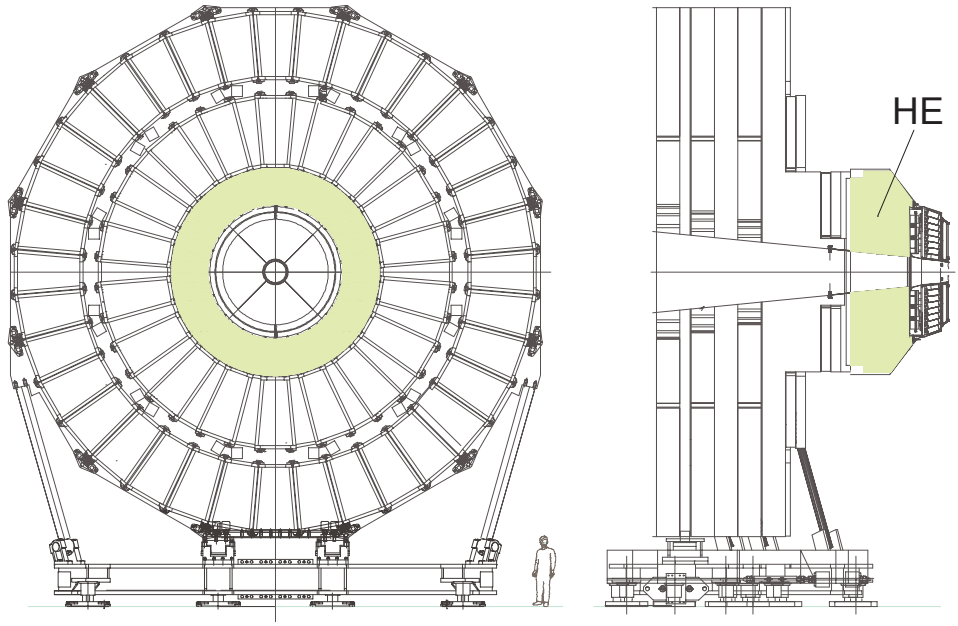


Figure 3.17: HE calorimeter mounted on the endcap iron yoke[10].

trays in CMS.

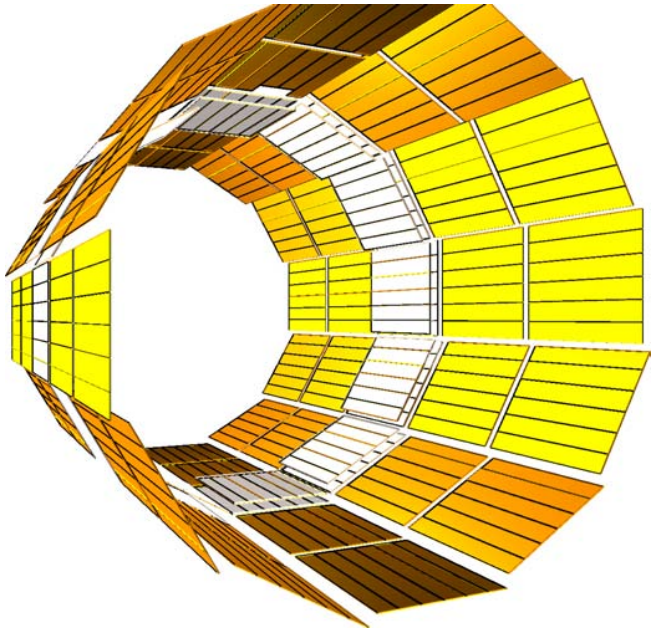


Figure 3.18: Layout of HO trays in the CMS detector [10].

The HF detectors provide coverage in the pseudorapidity range $3.0 < |\eta| < 5.0$, and are

positioned at a distance of 11.2 m from the interaction point, and are constituted of 900 towers and 1800 channels. They are heavily exposed to the collision and must therefore have high resistance to radiation. This requirement guides the design of the HF which leads to use of steel as absorber and quartz fibers as an active medium. On passing through the quartz fibers, particles generate Cherenkov light which is transported to photomultiplier tubes, where it is converted into electrical signals.

3.2.5 Muon Detectors

The muon system makes up the outermost region of the CMS detector. Because muons are about 200 times heavier than electrons, they undergo much less bending in the magnetic field than electrons with the same momentum. High energy muons are also much less sensitive to deflection in the materials that they traverse through, and do not suffer from high radiation losses. As a result, muons can easily pass through the CMS calorimeters and solenoid. These are the primary motivations behind adding a dedicated tracking system to CMS outside the solenoid, where muons are the only particles that would create a signal. Muons provide a very clean signature and are an important element of the physics program at CMS. For this exact reason, the decay of the Higgs boson to four muons is known as the 'golden channel', and it was one of the two final states in which the Higgs boson discovery was made [16]. Figure 3.19 shows the distribution of the four-lepton invariant mass in the said channel.

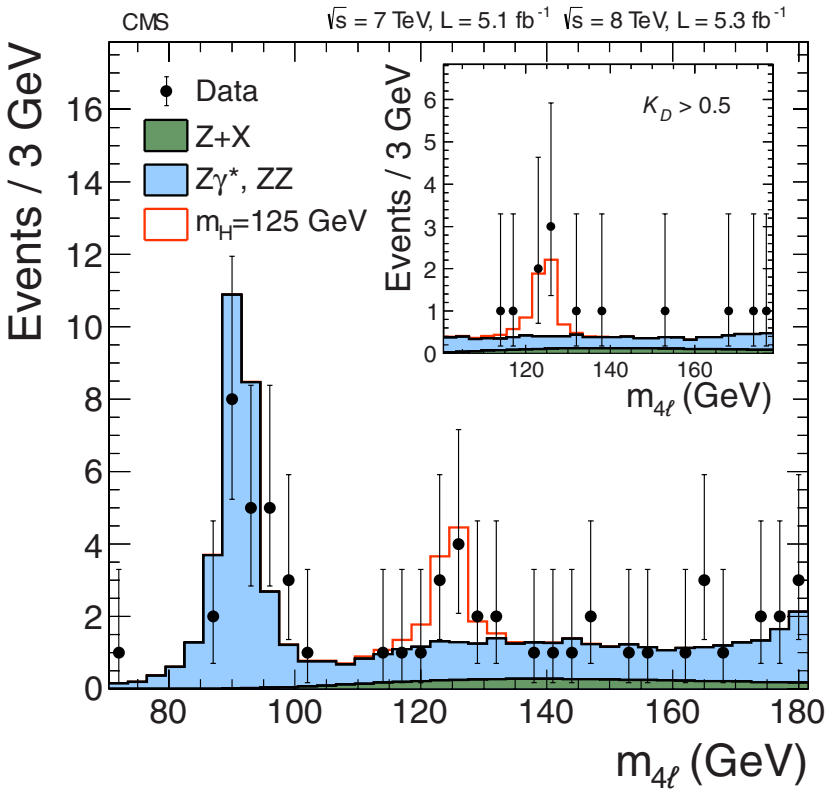


Figure 3.19: Distribution of the four-lepton invariant mass for the $ZZ \to 4l$ analysis. The black points represent the data, the filled histograms represent the background, and the open red histogram represents the signal expectation for the 125 GeV Higgs boson [16].

The CMS muon system can be divided into the following sections: the drift tubes (DT) and resistive plate chambers (RPC) in the barrel, and the cathode strip chambers (CSC) and RPCs in the endcap. RPCs provide very fast trigger information with rough position information, and DTs and CSCs provide highly accurate position information along with trigger information. DTs and RPCs are arranged in concentric cylinders in the barrel region, and the endcap disks contain different layers of CSCs and RPCs. These subsystems are highlighted in figure 3.20, and figure 3.21 displays a photograph of the muon chambers interspersed in the magnet return yoke.

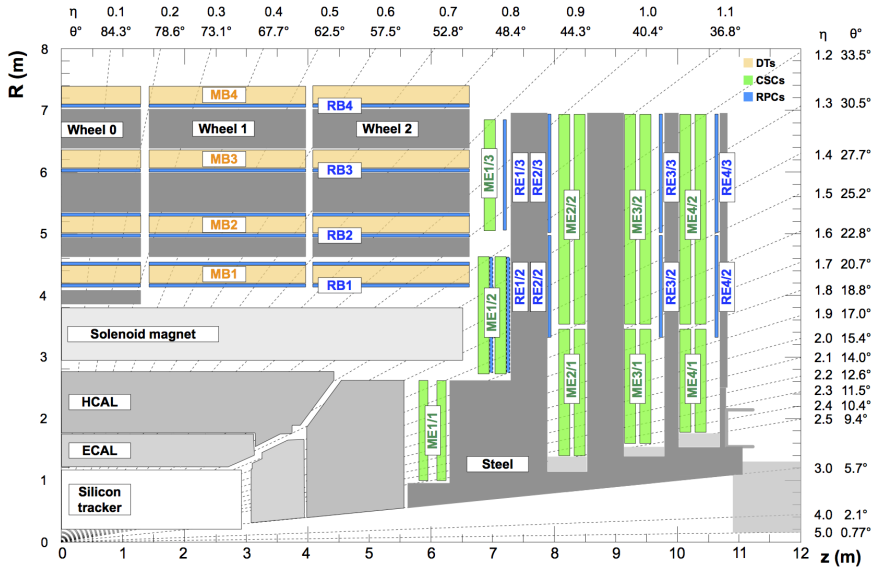


Figure 3.20: Layout of the CMS muon system with the various sub-detectors highlighted. The drift tubes (DTs) are labeled muon barrel (MB), the cathode strip chambers (CSCs) are labeled muon endcap (ME). The resistive plate chambers (RPCs), which are mounted in both the barrel and endcap regions are labeled RB and RE, respectively [17].

All of the muon subdetectors are gaseous detectors and operate on the common principle that as muons pass through the gas, they knock out electrons of the gas atoms which are then collected by wires or strips and provide a measurement of the electron energy.

Drift tubes comprise of a stretched wire within a gas volume. The tube is filled with a mixture of Ar and CO_2 gas. The basic unit of the drift tube is shown in figure 3.22. A charged particle, when passing through the gas, creates free electrons which then drift

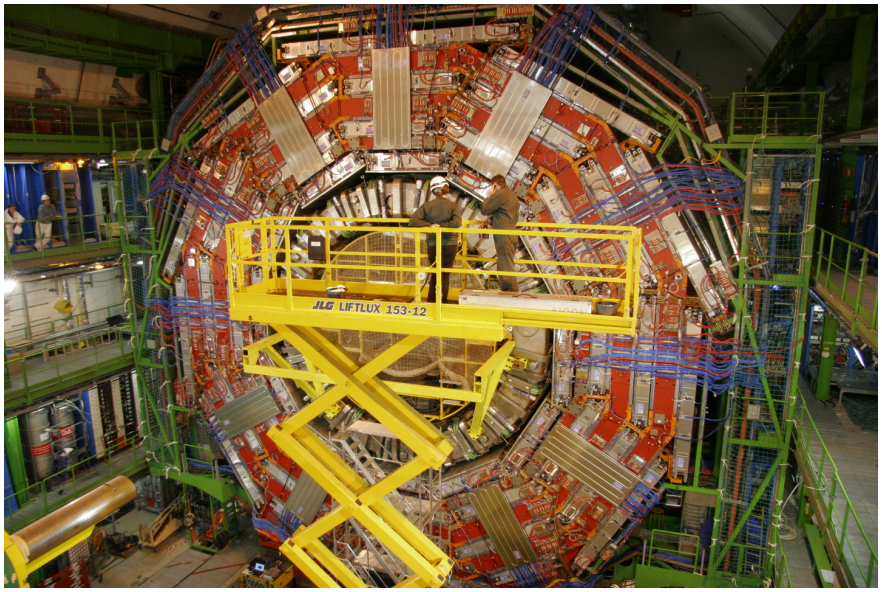


Figure 3.21: Photograph of the muon chambers [18].

towards the positively charged wire. To ensure that the position measurements are reliable, the spatial uniformity of the field along the path is important. The equipotential lines inside a DT cell are shown in figure 3.23, and it can be seen that the field lines are mostly parallel along the z axis. This ensures that electrons produced along the z-axis will have comparable drift times.

Each DT chamber consists of 12 aluminium layers and are, on an average, 2 m $\times$ 2.5 m in size. They are arranged in three layers of four, known as *superlayers*, each of which contain up to sixty 4 cm wide drift tubes. The superlayers are placed orthogonal to each other and therefore, the DTs can provide two coordinates for the muon position. The wires parallel to the beampipe measure the (r, ϕ) co-ordinate, and the wires perpendicular to the beampipe measure the z coordinate. The electron's drift speed and time of arrival at the wire helps determine how far the muon has travelled from the wire.

The layout of the CMS barrel muon DT chambers is shown in figure 3.24 for one of the five wheels. The chambers are identical in each wheel, with the exception of wheels -1 and +1, where the chambers in 2 sectors need to be shortened due to the presence of cryogenic chimneys for the magnet.

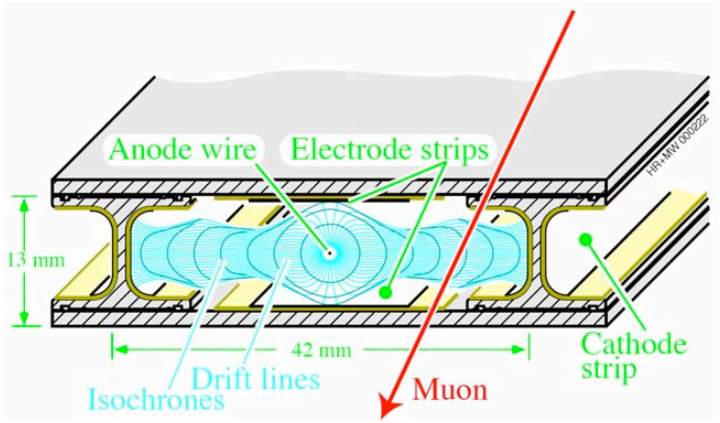


Figure 3.22: Diagram of drift tube cells [10].

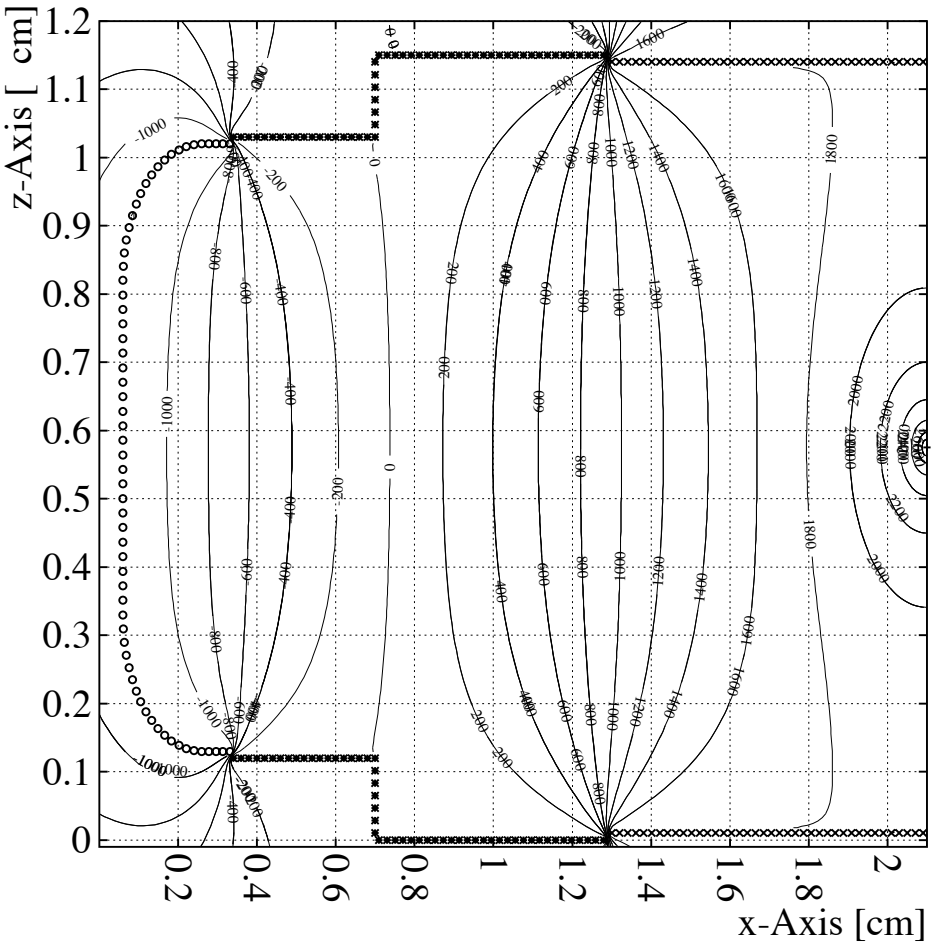


Figure 3.23: Equipotential lines in half of a drift cell are shown. The lines are labeled with the potential in volts [10].

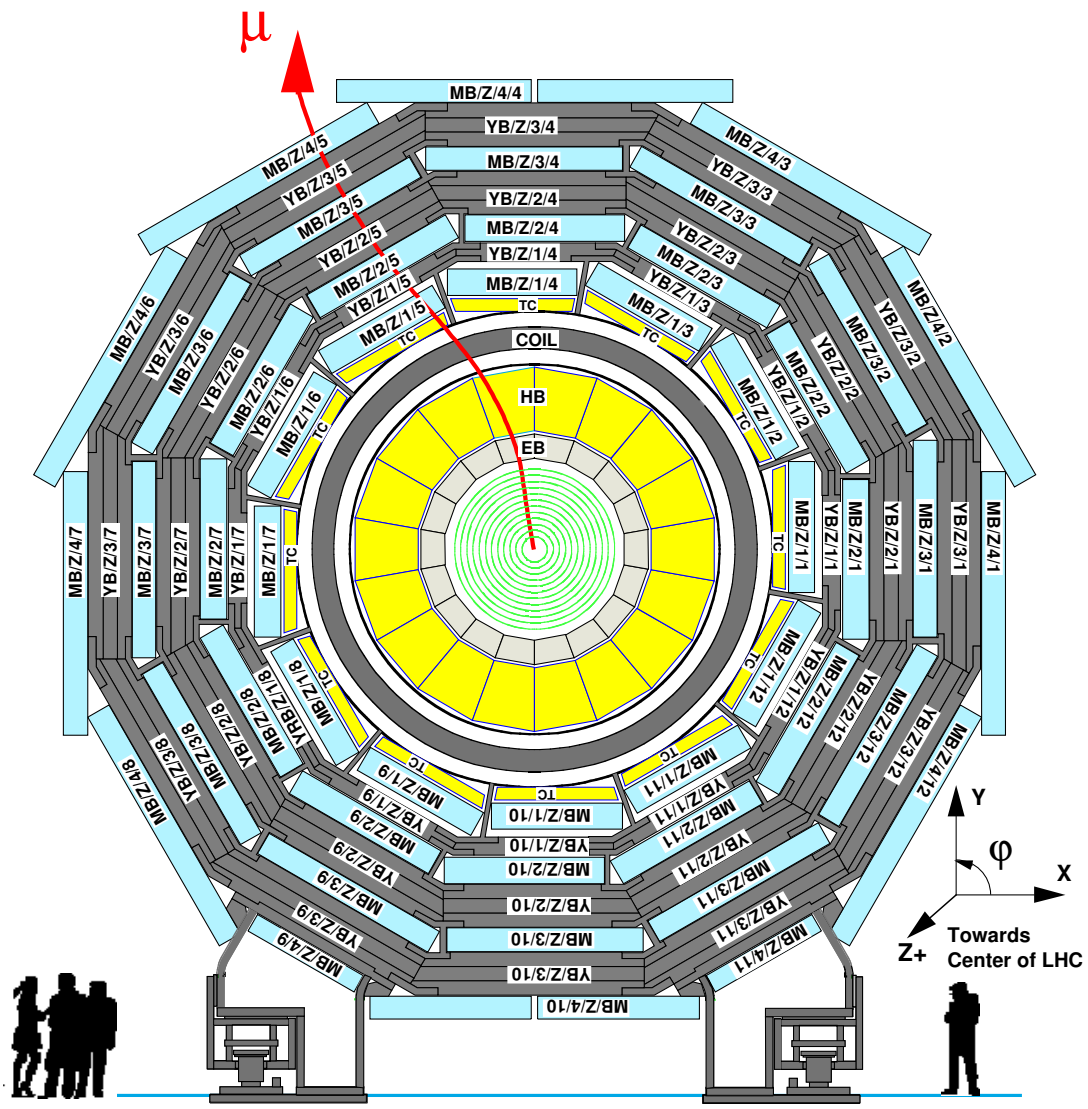


Figure 3.24: Layout of the CMS barrel muon DT chambers (in blue, labelled 'MB') in one of the five wheels [10].

Cathode strip chambers are placed in the endcaps, which has a significantly higher occupancy. Therefore, it's necessary for the cathode strip chambers to have much finer wire pitch and have the ability to measure coordinates of every avalanche, in addition to having high radiation resistance. The CSCs are multiwire proportional chambers, and consist of positively-charged anode wires crossed with negatively-charged copper cathode strips contained in a gas volume. A total of 468 trapezoidal shaped CSC chambers are present at both endcaps. Muons passing through them ionize the gas, which creates electrons and positive ions, that move towards the anode wires and copper cathode, respectively.

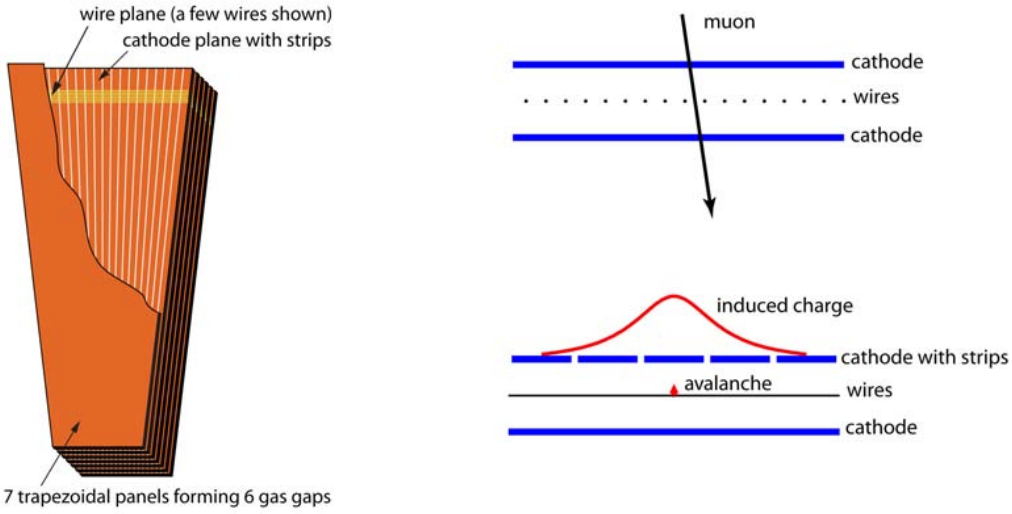
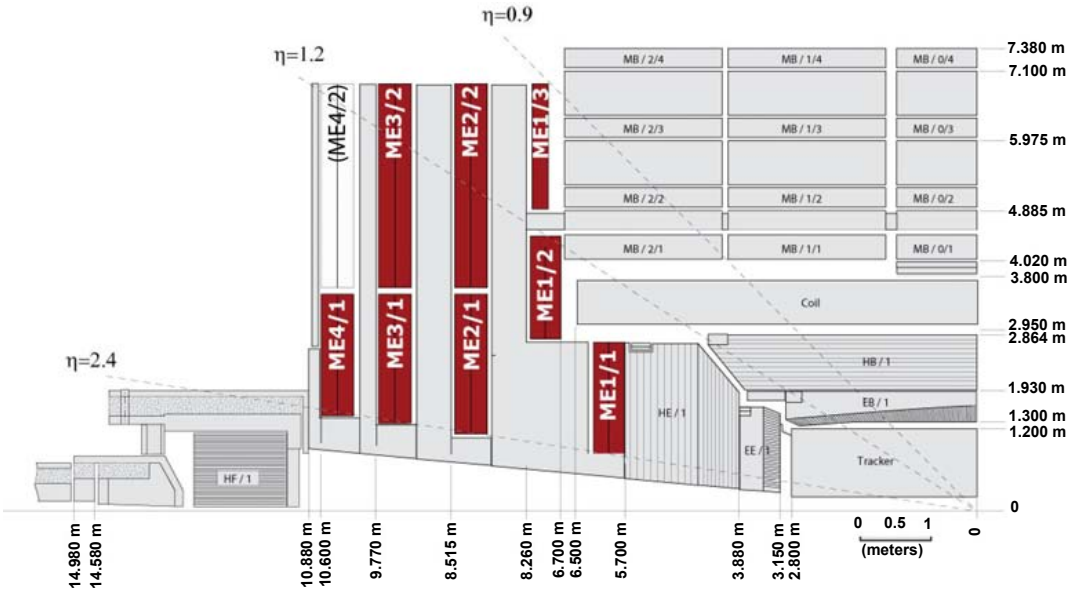


Figure 3.25: Layout of a CSC with seven trapezoidal panels are shown on the left. The anode wires and cathode strips, that constitute each CSC are also visible. The right is a schematic diagram depicting the principle of CSC operation [10].

The perpendicularly placed strips and wires provide two position coordinates; wires provide the radial coordinate r, and strips provide the ϕ coordinate. The closely spaced wires enable the CSCs to provide fast signals and allow CSCs to be used as an input to the trigger decision as well. Figure 3.25 shows the layout of a CSC, with the anode wires and cathode strips visible in the diagram. The figure also depicts the process of charge induction in a CSC. Figure 3.26 shows the arrangement of the muon endcap (ME) chambers, and a photograph of the ME2 station is also shown.



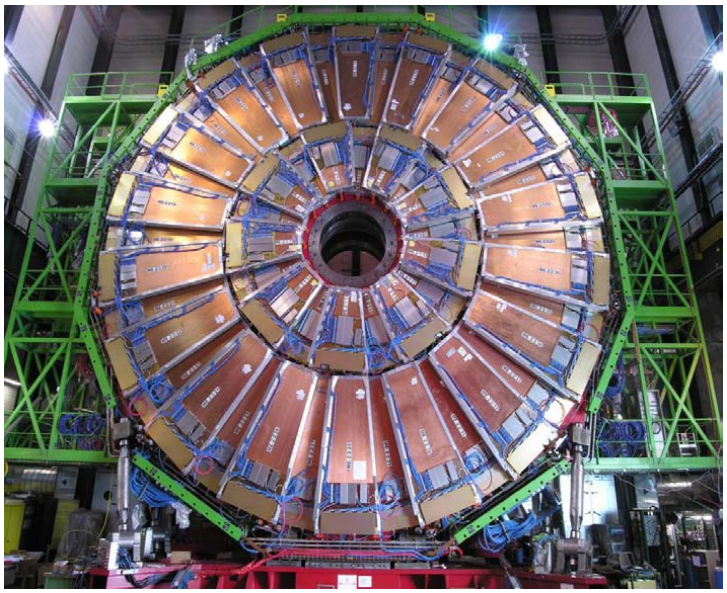


Figure 3.26: CSCs are arranged in groups that are labelled 'ME' (muon endcap) chambers. The placement of ME (1-4) is highlighted in the top figure. The bottom photo shows the ME2 station. Various chambers, belonging to the same station, overlap to provide continuous coverage in ϕ[10].

Resistive plate chambers are the final part of the CMS muon system. They are made of two parallel plates, a positively charged anode and a negatively-charged cathode, which are separated by a gas volume, the *gap*. Similar to the other muon sub-detectors, muons passing through the RPCs produced free electrons out of the gas atoms. The electrons then give rise to an avalanche which makes its way to the anode plate, and is detected by the metallic readout strips.

The RPC stations consist of 2 mm gaps, each under a potential of 9 kV. The barrel and endcap parts, consist of six and four layers of RPC stations, respectively. Figure 3.27 shows the layout of the RPCs with respect to the DTs and CSC chambers.

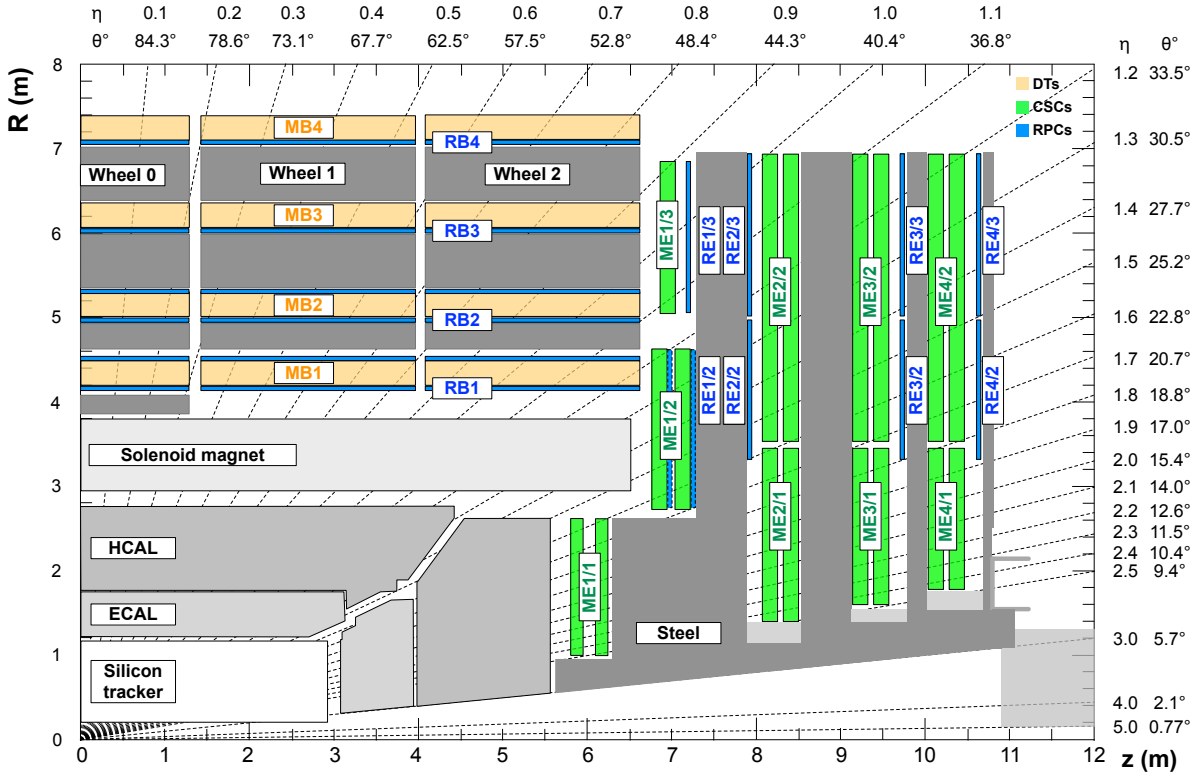


Figure 3.27: Placement of the RPC's with respect to DTs and CSC chambers is shown [19].

3.2.6 Luminosity Detectors

Among quantities such as detector acceptance, efficiency and background subtraction, accurate luminosity measurement plays a pivotal role in the CMS experiment's physics program. Five dedicated detectors, based on different occupancy measurements of individual CMS sub-detectors, are utilized by CMS to monitor and measure the number of pp inelastic collisions that occur while the LHC is running. They are: barrel DTs, HF, Pixel Luminosity Telescope (PLT) and Fast Beam Conditions Monitor (BCM1f). PLT and BCM1f are located close to the LHC beam pipe and use silicon pixel sensors and single-crystal diamonds, respectively. These dedicated detectors perform measurements online during data taking, as well as offline during the reconstruction step. Online luminosity measurements ensure that the CMS detector is able to collect data in an efficient manner, and that the CMS sub-detectors are sheltered from various challenges that could arise due to high instantaneous luminosity. An accurate measurement of the size of the dataset is ensured by the offline measurements.

Dedicated LHC runs, called Van der Meer (VdM) scans [52], are performed to calibrate luminosity measurements. During a VdM scan, the beams first produce head on collisions. Then, one of the two beams is moved in discrete steps in the x direction, until the hit rate reduces to zero. The same process is then repeated in the y direction in discrete steps. This allows measurement of the luminosity per colliding bunch.

3.2.7 Trigger and Data Acquisition

CMS detects a bunch crossing, provided by the LHC, every 25 ns. When translated in terms of frequency, this means CMS records collision data at a rate of 40 MHz. This rate is too high to be fully read out, processed and stored for offline analysis by the current capabilities of the available computing facilities. Another point to consider is that most of these events correspond to low energy QCD processes which are not of interest to most analyses in general. Therefore, in order to filter the collision data, CMS uses a trigger system. The broad aim

of the CMS trigger system is to select interesting signal events, while keeping a manageable output rate and CPU time.

The CMS trigger system is made up of two stages: Level-One (L1) Trigger and High-Level Trigger (HLT). These will be discussed in the next sections.

3.2.7.1 L1 Trigger

The L1 trigger is the first layer of the CMS trigger system, and its goal, in simple terms, is to reduce the event rate from 4 MHz to 100 kHz. It operates at the hardware level and is based on FPGAs and ASIC integrated circuits technology. By using information from the calorimeters and muon detectors, the L1 trigger selects the most interesting events in about 4 μs.

When particles produced in a collision pass through the various subdetectors, they generate signals that are collected in buffers in the front-end electronics of the subdetectors. Part of this information is sent to the L1 trigger boards, located in a service cavern next to the detector. The decision to keep or reject the event is made and this information is relayed back to the front-end electronics. This process takes place in about 3.2 μs, and during this time the total collision information is stored in the buffers. The L1 trigger has actually less than 1 μs to perform calculations needed to make the decision.

The L1 is divided into three parts: the calorimeter trigger, the muon trigger, and the global trigger. The calorimeter trigger is further divided into two stages: Layer-1 and Layer-2. Summarized data, received from ECAL and HCAL in packets called trigger primitives, is sent to Layer-1. Information such as the sum of ECAL and HCAL energies, energy calibration and computation of H/E ratio (ratio of HCAL to ECAL energy) is computed at this stage. This computation is done parallely, by processors that receive local information corresponding to different sections of the detectors. The information from Layer-1 is then sent to Layer-2, and computation of global information such as calorimeter based missing transverse energy, reconstruction of simple versions of jets, taus, electrons and photons is

performed at this stage.

Information about muons is provided by the muon trigger. The three muon subdetectors contribute to the muon trigger information. The front-end trigger electronics on the DTs and CSCs identify track segments from the hits registered in various detector planes of a single detector station. These track segments are then sent to the regional track finding processors, via optical fibres, which apply pattern recognition techniques to identify muon candidates and measure their momenta from their bending under the magnet. The combined muon trigger then provides information like the tranverse momentum of the muon candidates and their positions to the global trigger, which is the final L1 step, where this information can be accepted or rejected.

Once information from the calorimeter trigger and muon triggers has reached the global trigger, L1 selection algorithms stored in the electronics are used to select interesting events. Data corresponding to the interesting events is passed along to the High Level Trigger (HLT).

3.2.7.2 HLT Trigger

The HLT trigger is the final layer of the CMS trigger system, and it reduces the event rate from 100 kHz to 1 kHz. The HLT system is a computer farm that has access to the full event information and can perform calculations similar to the ones performed at the offline analysis level. Once an event reaches the HLT, it is sent to a separate processor and a final trigger decision is made, on an average, in less than 0.1 s. If an event passes the requirements for one of the configurable high level triggers, it is saved to storage and is reconstructed offline.

3.2.8 Physics object reconstruction

Figure 3.28 summarizes the general object reconstruction at CMS. Each particle creates a specific signature in the CMS detector: charged hadrons leave hits in the tracker and are stopped by the HCAL, neutral hadrons deposit their energy in HCAL without leaving any hits in the tracker, electrons and photons deposit their energy in the ECAL, and electrons

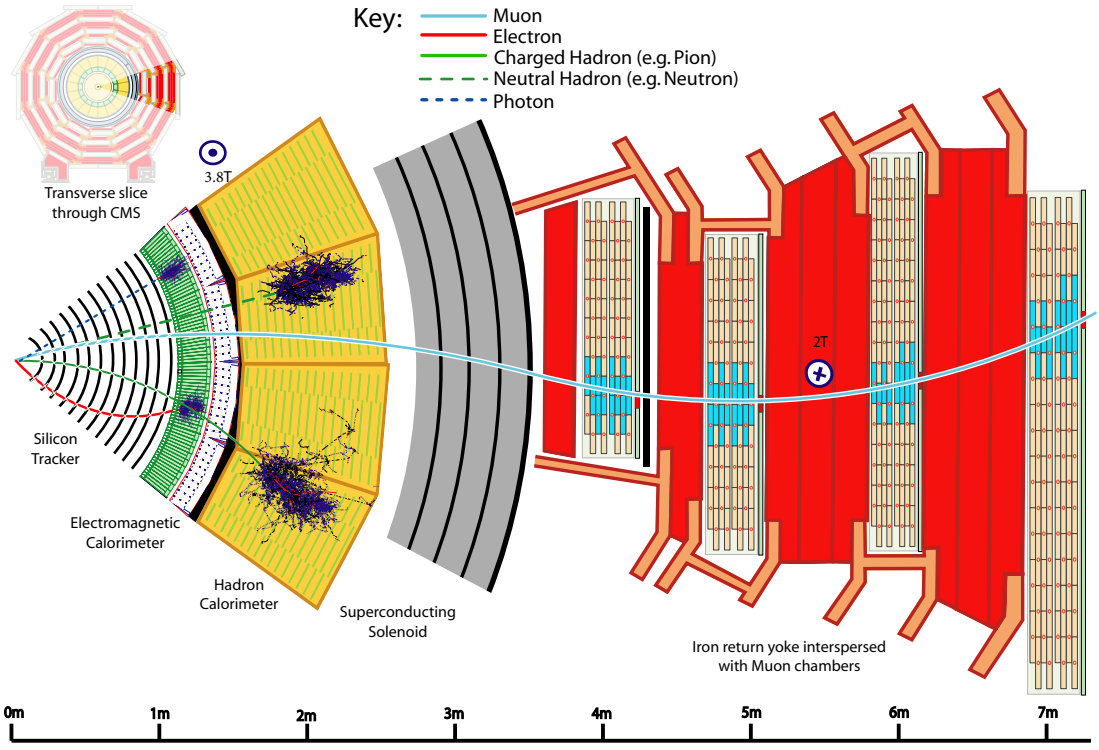


Figure 3.28: Transverse slice of the CMS detector [20].

also leave hits in the tracker, and finally muons leave hits in the tracker and the muon system without leaving any energy deposits in the calorimeters. Conventional reconstruction methods only utilize information from only one subdetector to reconstruct a particular physics object. On the other hand, the particle-flow (PF) reconstruction algorithm [53; 54], which was developed before the start of the LHC, creates a global description of the event using an optimal combination of information from all subdetectors.

The basic building blocks of particle flow are tracks, reconstructed using a tracking algorithm, and clusters of calorimeter energy deposits. These building blocks are joined together using an algorithm that will be detailed in section 3.2.8.1.

Once particles have been reconstructed using the PF algorithm, they can be further tuned to suit the needs of specific analyses. In order to reduce misidentification rate, more strict criteria are usually applied to reduce the background rate. Both, data and simulation, undergo PF reconstruction and this results in overall good agreement between them. In order to correct residual discrepancies in some quantities, event reweighting techniques are employed.

3.2.8.1 Particle flow

The particles reconstructed using the particle flow method are known as the PF candidates. PF candidates are constructed by combining information from the inner tracker, the calorimeters, and the muon system. Each PF candidate is assigned to one of the following object categories: muons, electrons, photons, charged hadrons, and neutral hadrons. Overlaps between different object selections are removed by the use of information from all subdetectors. The different ingredients of the PF algorithm are explained below:

- Iterative Tracking: The tracker provides excellent momentum resolution for charged hadrons up to a p_T of several hundred GeV. Additionally, the tracker also provides a precise measurement of the direction of charged particles. The tracker is, therefore, of utmost importance to the PF algorithm. Kalman filter based tracking strategy

[12] is used to ensure a tracking efficiency of $\approx$ 100%. The first step in the track finding algorithm involves application of a very tight criteria for the track seeds and reconstruction quality. This leads to a modest track efficiency with small fake rate. The next step involves removal of the hits assigned to those tracks, after which the tracking cycle is repeated two more times for the remaining hits with progressively looser track seeding criteria. The loose criteria ensures a high tracking efficiency while keeping a minimal fake rate. As a result of this iterative tracking, charged particles with p_T as low as 150 MeV, a minimum of three hits, and originating at a minimum distance of 50 cm from the beam axis, can be reconstructed with a fake rate of $\approx 1\%$.

- Calorimeter clustering: The calorimeters are tasked with providing measurements of the energy and direction of neutral hadrons and photons. Additionally, they should provide information on charged hadrons in case their track parameters are indeterminable with high precision. The clustering procedure is performed separately for the ECAL, HCAL and preshower, separately for barrel and endcaps. The first step involves identification of cluster seeds. Calorimeter cells whose energy deposits is above a certain threshold, and larger than their neighbouring cells are identified as cluster seeds. Cluster seeds from adjacent cells that pass a certain minimum energy threshold, chosen relative to the electronic noise level, are used to construct topological clusters. An iterative procedure is used to create PF clusters from the topological clusters. Each cluster seed creates one PF cluster. If multiple seeds are part of one large topological cluster, the energy of each calorimeter cell is shared among all PF clusters according to the distance between cluster and cell. The position of the PF cluster is computed as a weighted average of the position of the cells in an iterative manner. This iterative procedure continues until the cluster positions acquire stability.

- Link algorithm: In order to remove any possible double counting from different subsystems, a linking algorithm is devised to link the various elements originating from a single particle. The distance between linked elements is used to designate the quality of a given link. For

a charged particle, the link between its track and calorimeter cluster is established by extrapolating the track to the calorimeters, if the extrapolated track position falls within the cluster. In this case, the link distance is defined as the distance in the (η, ϕ) plane between the extrapolated track position and the cluster position. In case of electrons, the calorimeter clusters corresponding to bremsstrahlung photons are accounted for by extrapolating tangents to a track at each intersection between the track and a tracker layer towards the ECAL. The cluster is linked to the track if the extrapolated tangent position lies within a cluster. Calorimeter clusters are linked to each other when the position of a cluster in the higher granularity calorimeter falls within the boundaries of a cluster in the less granular calorimeter. Finally, the link between a charged particle track in the tracker and a muon track in the muon system is established when a global fit between the two returns an appropriate χ^2 value.

- Particle reconstruction and identification: Once links have been established between the various elements originating a single particle, they are converted into particles by the PF algorithm. The simplicity of the particle signature determines the order of the particle reconstruction: muons are reconstructed first, and neutral hadrons are reconstructed last. The following procedure is followed:

    1. A global muon is categorized as a PF muon if its momentum is similar to the momentum determined using tracker information only. The corresponding track information is then removed from further processing.

    2. Electrons are the next particles in line. They are reconstructed using a dedicated algorithm [55] that makes use of a Gaussian-Sum filter to refit the candidate electron track. This refitting takes the possible energy loss due to bremsstrahlung into account and follows the electron's trajectory to the ECAL. Certain identification criteria are applied to reduce the time required to perform the fitting procedure. The electron needs to pass a set of tracking and calorimeter require-

ments, in addition to having its track match with an ECAL cluster, in order to be fully identified. Once an electron fulfills these criteria, it is added to the PF electron collection. To remove the possibility of any double counting, the associated electron and ECAL clusters are removed before proceeding to further processing.

3. Tighter quality criteria, such as the requirement on relative uncertainty in the transverse mometum to be smaller that the relative calorimeter energy resolution for charged hadrons, are applied on the remaining tracks. The track momenta and calorimeter energies are compared to determine the presence of photons and neutral hadrons.

4. All charged hadron candidate tracks associated to HCAL clusters are found. In case a track is associated with more than one HCAL cluster, it is assigned to the closest one. The charged hadron candidate tracks associated with a given HCAL cluster are then matched with the ECAL clusters. If the total charged particle momentum is larger than the total reconstructed calorimeter energy, a relaxed muon reconstruction is performed. If on the other hand, the total track momentum is smaller than the total calorimeter energy, the tracks are considered to be consistent with charged hadrons. In this case, the tracks are added to the collection of PF charged hadrons.

5. When the track momentum is compatible with the calorimeter energy, the charged hadron energy and momentum are refit using both, the tracker and calorimeter information.

6. In case there is an excess of calorimeter energy relative to the track momentum, photons and neutral hadrons are reconstructed. Photons are reconstructed from ECAL clusters, and if this does not account for the full excess, neutral hadrons are reconstructed from the remainder calorimeter energy. ECAL and HCAL clusters without any matching tracks are finally reconstructed as photons and neutral hadrons, respectively.

7. The list of particles identified at the end of the PF sequence are used to contruct jets. By default, isolated electrons and muons are not included in jet reconstruction.

Chapter 4

The CMS Electromagnetic Calorimeter

The electromagnetic calorimeter of CMS (ECAL) is a hermetic, high resolution, and homogenous calorimeter made of 61200 lead tungstate ($PbWO_4$) crystals mounted in the central barrel ($|\eta| < 1.48$) part, closed by 7324 crystals in each of the two endcaps ($1.48 < |\eta| < 3.0$) [10]. The ECAL's primary aim is to measure electrons and photons originating from the interaction point. Electrons interact in matter primarily by producing bremsstrahlung radiation, and photons primarily produce pairs of electrons and positrons. Therefore, an incident photon or electron can transform into an electromagnetic shower on a very short timescale. Electromagnetic calorimeters can achieve resolution at the 1% level since, unlike the more complex hadronic showers, electromagnetic showers are more well behaved with minimal loss of energy to invisible excitations. The Higgs boson has an extremely narrow width (about 4 MeV), which means that the observed width of the Higgs resonance is dominated by the experimental resolution. Therefore, one of the main driving criteria in the ECAL design was the capability to have the best electromagnetic energy resolution as possible.

The preshower detector (ES), based on lead absorbers equipped with silicon strip sensors, completes the CMS ECAL system. It is placed in front of the endcap crystals, and covers the region $1.65 < |\eta| < 2.6$. The ES strips have a fine granularity of 2 mm which can resolve signals from high energy photons from the decays of neutral pions into two photons, when the separation angle between the two photons is small.

A diagram of the ECAL, with its various components highlighted, is shown in figure 4.1.

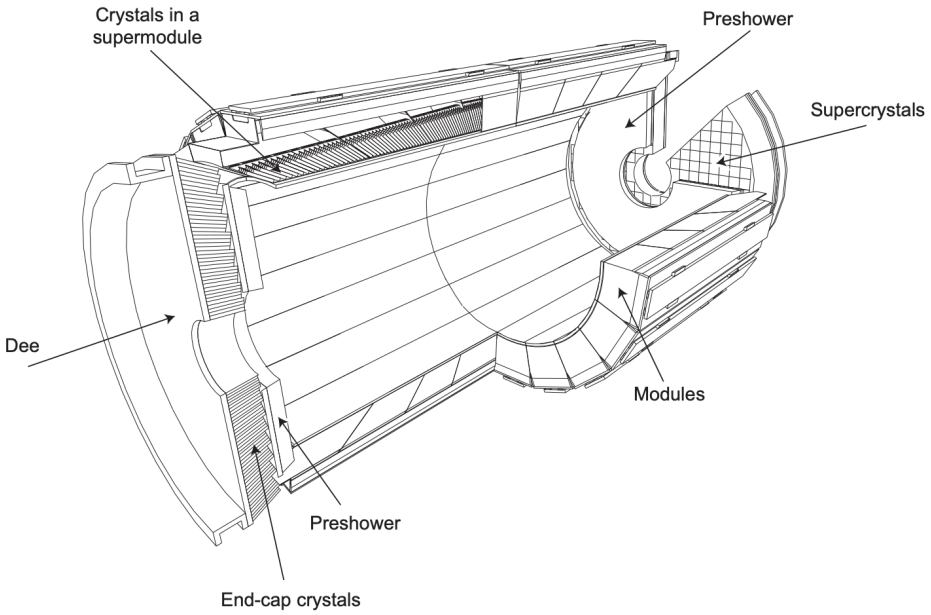


Figure 4.1: Schematic diagram of the CMS Electromagnetic Calorimeter [10].

4.1 Lead Tungstate crystals and ECAL electronics

The principle component of the ECAL are the lead tungstate ($PbWO_4$) crystals. They were chosen to serve, both as the absorber and the active material of the ECAL. Various characteristics of the $PbWO_4$ crystals make them the appropriate choice for operation at the LHC. $PbWO_4$ has a high density, 8.28g/cm^3, short radiation length (0.89 cm), and small Molière radius (2.2 cm), which drives the transverse development of the shower and enables fine granularity. Additionally, $PbWO_4$ has good transparency, good radiation resistance, and a scintillation time such that 80% of the light produced is emitted in 25 ns.

Both, barrel and endcap crystals, are shaped as trapezoidal prisms. They are of length 23 cm (22 cm) and correspond to 25.8 (24.7) radiation lengths in the ECAL barrel (endcap) region. The crystals have a pseudo-projective geometry, with the side facing the interaction point having an area of 22 $\times$ 22 mm^2 in the barrel region, expanding to 26 $\times$ 26 mm^2 at the rear end. The crystals are slightly larger in the endcap region, with an area of 28.6 $\times$ 28.6 mm^2 in the front and 30 $\times$ 30 mm^2 at the rear. Pictures of the barrel and endcap crystals

can be seen in figure 4.2.

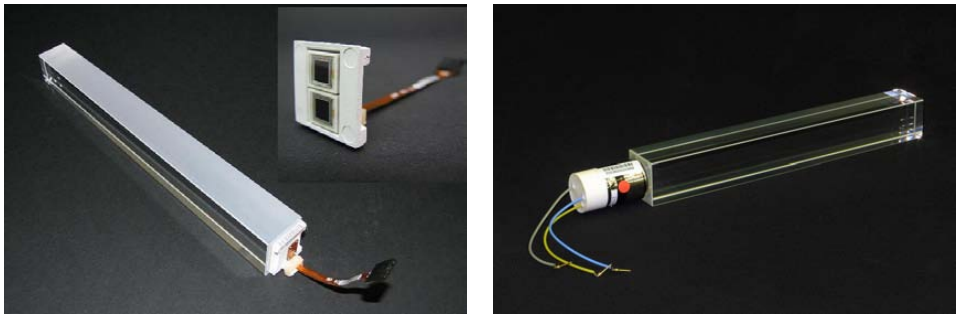
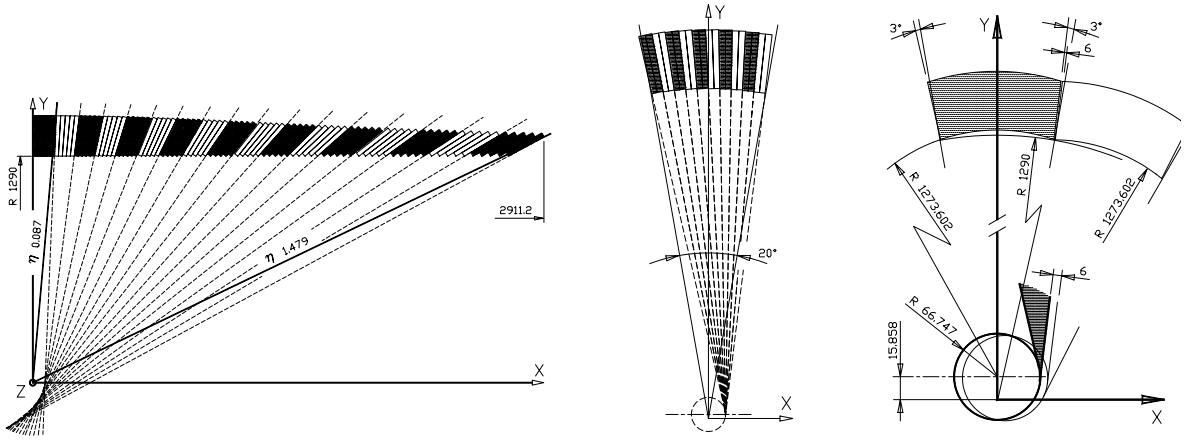


Figure 4.2: Pictures of the barrel (left) and endcap (right) crystals with attached photodetectors is shown [10].

The ECAL crystal axes are tilted at 3°with respect to the line from the nominal vertex position. This provides the crystals an off-pointing and pseudo-projective geometry. Figure 4.3 showcases the tilt of the crystals in the η and ϕ direction.



(a) η tilt (b) ϕ tilt

Figure 4.3: The tilt of CMS ECAL crystals is shown in the η (left) and ϕ (right) direction. The crystal longitudinal axes are inclined by 3°with respect to the line joining the crystal front face centre to the interaction point [21].

The barrel crystals are arranged in submodules (2×5 crystals in $\phi \times \eta$), which are then mounted into modules and grouped into supermodules (SM). Four modules, three with 10 $\times$ 4 submodules in $\phi \times \eta$ and one with 10×5 in $\phi \times \eta$, make up a supermodule. The barrel

section is structured with 36 identical supermodules, each covering half the barrel length and corresponding to $0 < |\eta| < 1.479$. Figure 4.1 displays the EB schema. The front faces of the barrel crystals are at a radius of 1.29 m from the beamline. The EB is located between the tracker and the hadron calorimeter barrel and its location within the CMS detector is depicted in figure 4.4.

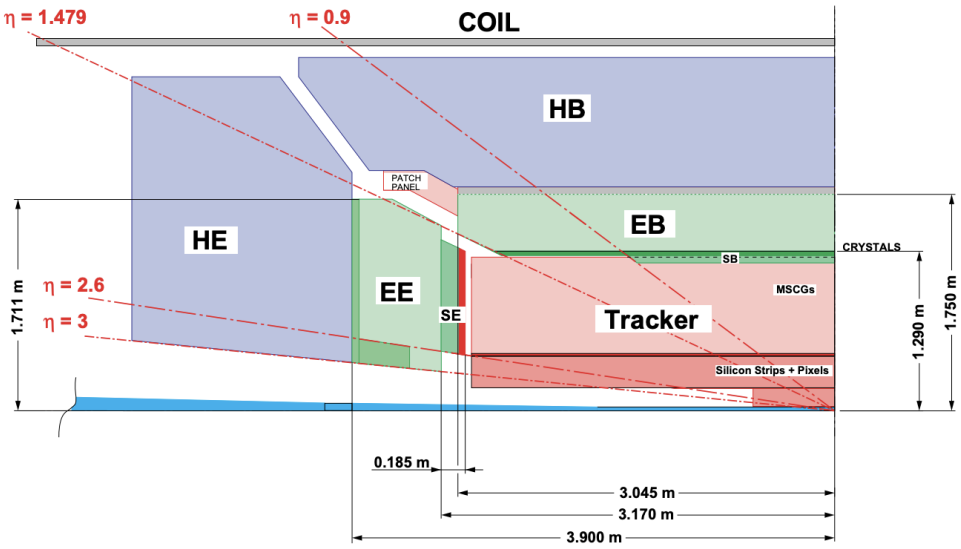


Figure 4.4: Longitudinal section of the central region of CMS [21].

In each endcap, the channels are split into two half's called 'Dee's', and the crystals are located at $|z| = 315.4$ cm. The geometrical design of the Dee's is depicted in figure 4.5. Figure 4.6 displays photographs of the supermodules and dees.

4.2 ECAL energy reconstruction

The reconstruction of photon and electron energy based on their energy deposits in the ECAL can be described by the following formula:

$$E_{e,\gamma} = \left[\sum_i (S_i(t) \times c_i \times A_i) \times G(\eta) + E_{ES}\right] \times F_{e,\gamma} \tag{4.1}$$

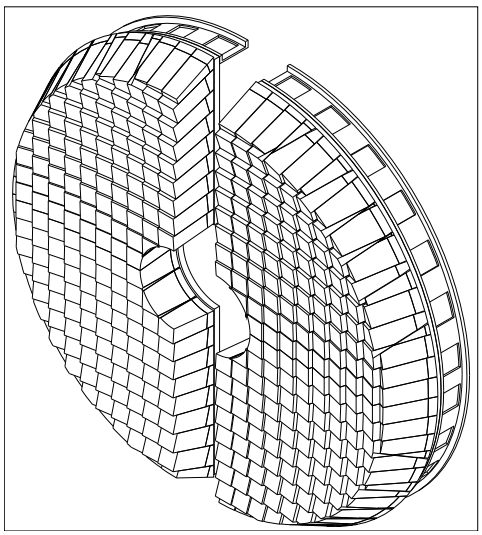
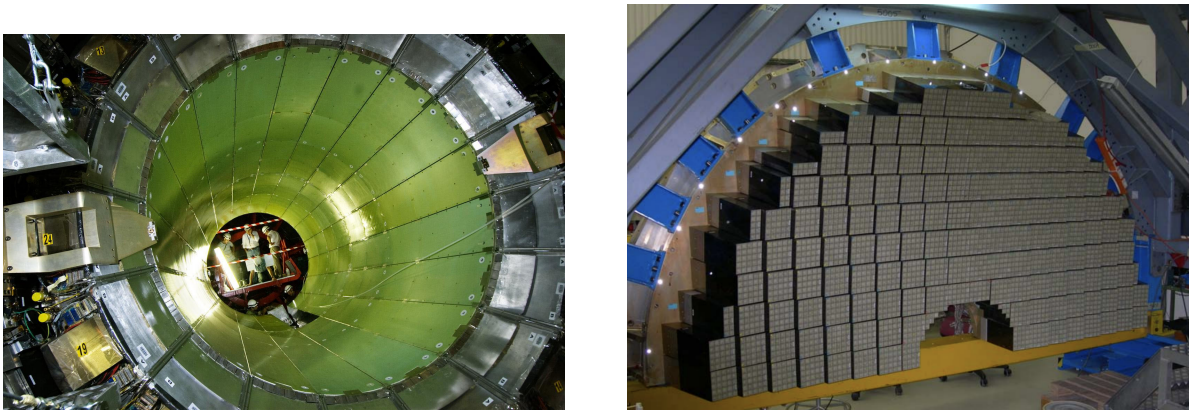


Figure 4.5: The ECAL endcap (EE) is divided in to 4 Dee's: 2 in EE+ and 2 in EE- [21].



(a) CMS ECAL supermodules

(b) CMS endcap Dee

Figure 4.6: Photographs of the CMS ECAL supermodules (left) and endcap dees (right) [22].

where A_i, c_i, $S_i(t)$ correspond to the per channel amplitude, intercalibration constant and light monitoring constant, respectively. $G(\eta)$ is the ADC to GeV absolute scale, E_{ES} is the preshower energy deposit, and $F_{e,\gamma}$ refer to the cluster corrections (different for photons and electrons). These are described in more detail in the subsequent sections.

4.2.1 Online pulse reconstruction

The electric signal generated by the photodetectors is amplified and shaped by a multi-gain preamplifier (MGPA), with gains 1, 6 or 12. This is followed by three integrated sampling Analog to Digital Converters (ADC), which operates at 40 MHz (equal to the LHC frequency.

During Run I of the LHC, an algorithm [56] was devised that estimates the signal amplitude A using a linear combination of the 10 samples S_i:

$$A = \sum_{i=1}^{N=10} w_i \times S_i \tag{4.2}$$

here, the weights w_i are calculated by minimizing the variance of A. However, another scintillation process might occur in the same crystal during the readout of one pulse, which is problematic for the in-time amplitude reconstruction and the energy deposits by particles originating from simultaneous proton collisions can distort the observed pulse shape. During Run II, the time spacing between colliding proton bunches was reduced to 25 ns. In order to mitigate the increased impact of out-of-time (OOT) pileup, an improved multifit algorithm was developed [57]. In the multifit method, the observed pulse shape is modeled as the sum of one in-time and up to nine out-of-time pulse amplitudes. The pulse shape is then reconstructed through minimization of the χ^2 variable defined as

$$\chi^2 = \sum_{i=1}^{N=10} \frac{\left(\sum_{j=1}^{M=10} A_j p_{ij} - S_i\right)^2}{\sigma^2_{S_i}} \tag{4.3}$$

where A_i are the amplitudes of up to M = 10 interactions, index j refers to a given bunch

crossing in which the pulse shape is generated, and index i refers to the time sample out of the total 10 samples recorded by the ADC. For each bunch crossing j, the pulse templates have the same shape and differ by 25 ns shift. Dedicated pedestal runs are performed to determine the total electronic noise S_i and the associated covariance matrix $\sigma^2_{S_i}$. These dedicated runs measure noise in the three gains (1, 6 and 12) of the MGPA in the absence of signal pulses.

Figure 4.7 displays an example of a fit in EB and EE for simulated events. The in-time pulse peaks at the sixth sample and the out-of-time pulses peak at different time samples and represent energy contribution from OOT.

4.2.2 Crystal response monitoring

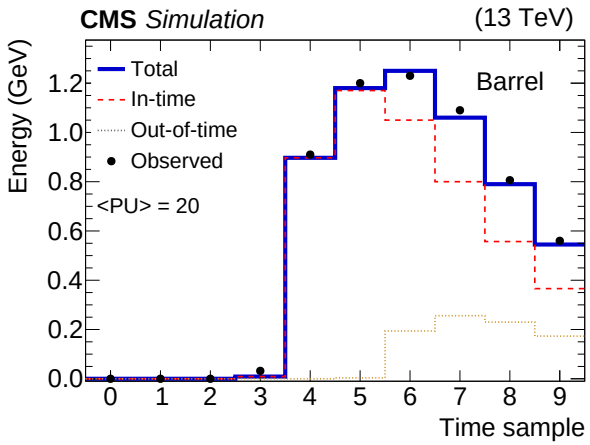
ECAL crystals experience a change in their transparency during periods of LHC operation due to color centres created in the $PbWO_4$ crystals from the ionizing radiation. The scintillation process is unaffected by the change in crystal transparency, but the amount of light detected by the photodetectors decreases. In order to mitigate the effects of these changes, time dependent corrections need to be applied to the reconstructed amplitude.

A light monitoring system [58] is employed to monitor the overall changes in the ECAL response. The system consists of lasers, operating at 447 nm, that inject light in each ECAL crystal. The laser light is then read by the standard ECAL readout. The measurement of change in transparency per crystal (R/R_0) is then used to calculate the ratio between the reconstructed amplitude and the injected light amplitude (S/S_0):

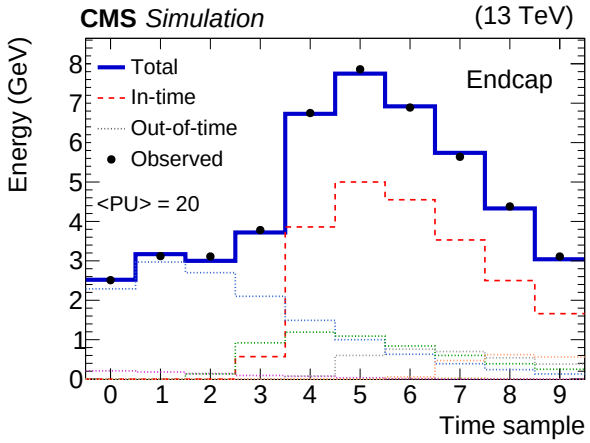
$$\frac{S}{S_0} = \left(\frac{R}{R_0}\right)^{\alpha} \tag{4.4}$$

α has been measured in test beams and has a value ≈ 1.5. The value of $\frac{S}{S_0}$ is then used as a correction factor to account for changes in the crystal response.

The light monitoring system works alongside the main ECAL DAQ system. During orbit



(a) Pulse shape in EB



(b) Pulse shape in EE

Figure 4.7: Examples of fitted pulses for simulated events with 20 average pileup interactions and 25 ns bunch spacing. Signals arising from a $p_T = 10$ GeV photon shower in the barrel (top) and in an endcap (bottom) are shown for individual crystals. Filled circles with error bars represent the 10 digitized samples, the red dashed (dotted multicolored) distributions represent the fitted in-time (out-of-time) pulses with positive amplitudes. The solid dark blue histograms represent the sum of all fitted contributions. The various colored dotted distributions represent the fitted out-of-time pulses with different bunch crossing. [23].

gaps, periods in which the LHC bunches are empty, the laser is emitted in various ECAL barrel and endcap regions, and one full reading is performed in about 40 minutes.

Figure 4.8 shows the changes in relative response as a function of time, for various pseudorapidity regions. Changes of up to 10% (50%) are seen in the ECAL barrel (endcap) region. The response changes up to 90% in the region closest to the beam pipe. The recovery of the crystal response is also visible during the LHC shutdown periods.

Figure 4.9 shows the impact of laser corrections on the ratio of the electron energy measured in ECAL barrel to the electron momentum measured in the tracker. The comparison of data points before the application of any laser correction (red) and after the application of laser correction (green) indicate that a stable energy scale is attained on applying the corrections.

4.2.3 Crystals inter-calibration

The variations of the measured energy among different ECAL channels are equalized by means of various inter-calibration (IC) procedures [59]. Different methods based on physics processes are employed to provide an IC constant for each channel. These methods compare an observable that is measured using exclusively ECAL information, with a reference value. The methods are:

- ϕ Symmetry: This method is based on the expectation that the energy flux is uniform along ϕ rings. The non-uniformities in this flux are corrected using the ICs. This method provides a lower precision compared to the other methods, and it was used up to 2016 to check the variation of the IC constants computed since Run I, since the data sample collected at 13 TeV was not large enough to compute a new set of IC constants using other methods. In 2017, a larger dataset was available and more precise methods were used to perform calibration in place of the ϕ symmtery method.

- $\pi^0 \rightarrow \gamma\gamma$: This method utilizes the position of the π^0 peak in the invariant mass

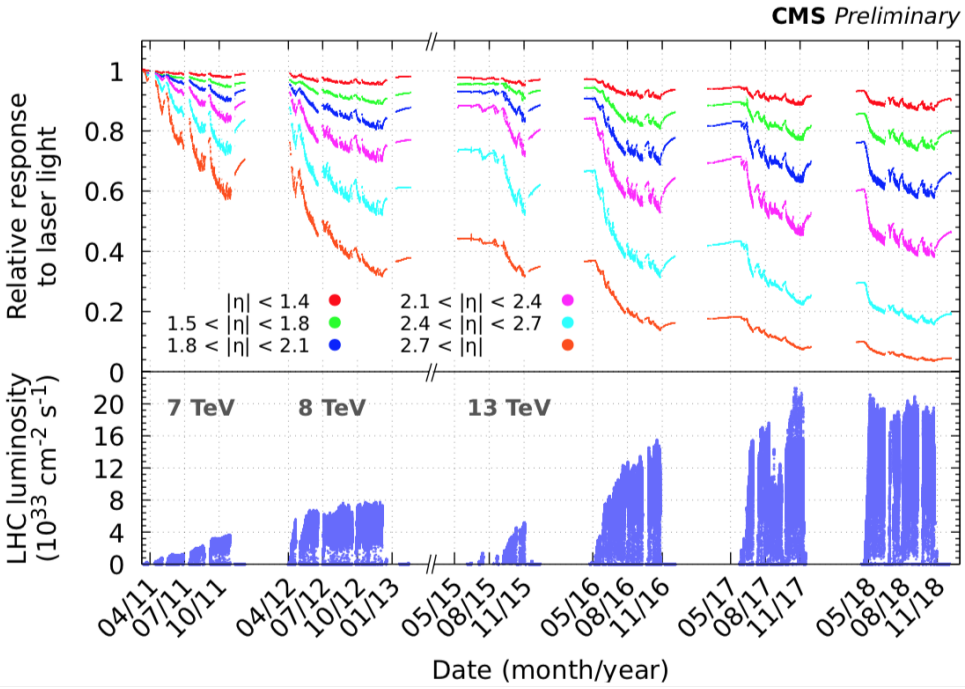


Figure 4.8: Relative response to laser light injected in the ECAL crystals, measured by the ECAL laser monitoring system. The response is averaged over all crystals and is shown in various bins of pseudorapidity, for the 2011, 2012, 2015, 2016, 2017 and 2018 data taking periods. The response change observed in the ECAL channels is up to 10% in the barrel and it reaches up to 50% in the endcap region. The response change is up to 90% in the region closest to the beam pipe. The recovery of the crystal response during periods without collisions is visible. These laser measurements are performed every 40 minutes and are used to correct the physics data. The bottom panel of the plot shows the instantaneous LHC luminosity delivered during this time period [24].

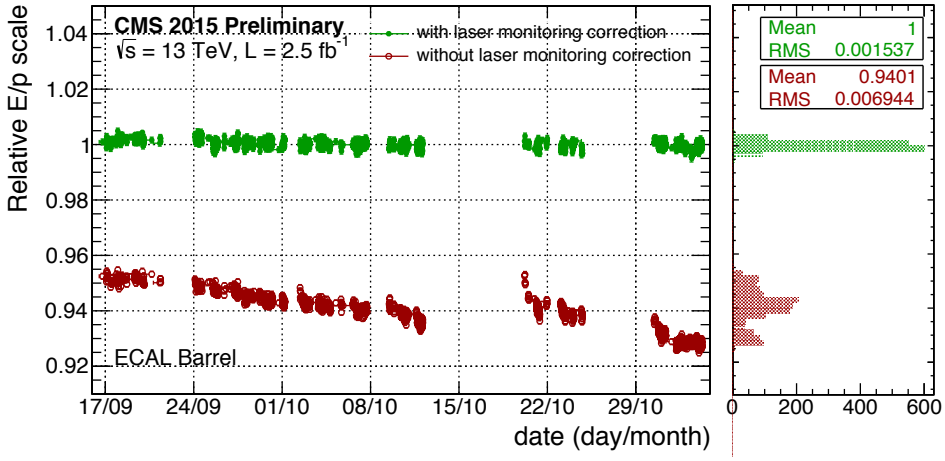


Figure 4.9: History plot for 2015 data of the ratio of electron energy E, measured in the ECAL Barrel, to the electron momentum p, measured in the tracker. The events correspond to $Z \to ee$ decays. The history plots are shown before (red points) and after (green points) corrections to ECAL crystal response variations due to transparency loss are applied, and a stable energy scale is achieved throughout the 2015 run after applying laser corrections.

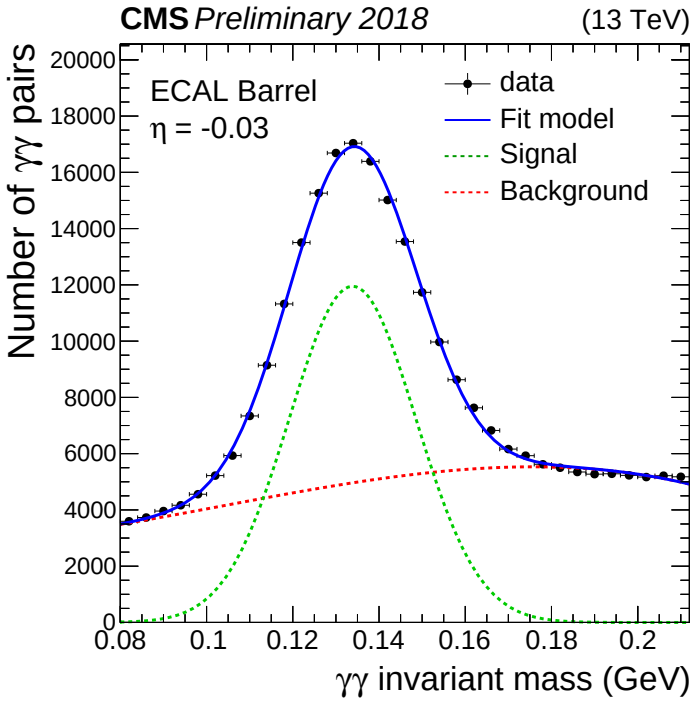
distribution of unconverted photon pairs, and assigns an IC constant to each channel. The principle idea of this method is to obtain a diphoton mass distribution for each channel, and extracting the position of the π^0 peak with a fit. The peak position is then equalized to the known π^0 mass and used to obtain a response correction. The dataset used in this method is obtained via a special HLT stream which saves limited ECAL information corresponding to the selected photon candidates. Figure 4.10 shows examples of fits to the diphoton mass distribution, one in EB and one in EE. These distributions utilize data collected in 2018. The resolution of the fits is maximized by varying the ICs iteratively. It has been observed that about 14 iterations are sufficient to reach an agreeable level of convergence.

- $Z \to ee$: This method is similar to the $\pi^0 \to \gamma\gamma$ method, but uses the invariant mass of the Z resonance in place of π^0.

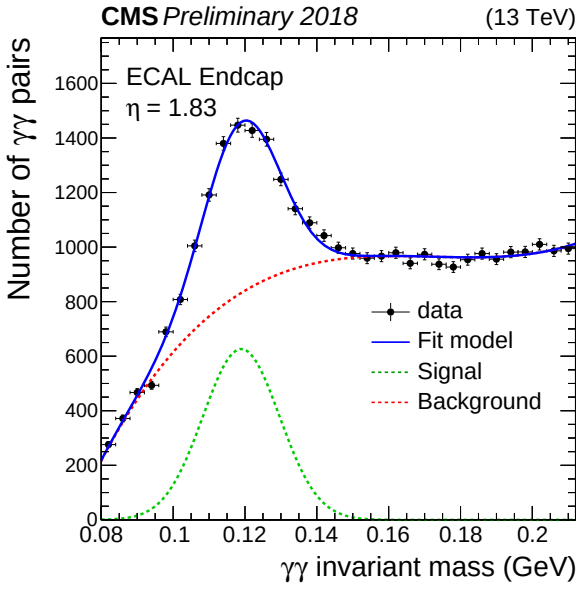
- E/p: This method makes use of prompt electrons from W/Z boson decays. It is based on the comparison of the electron energy measured in the ECAL and the elecrton

momentum measured in the tracker. The spread of the E/p distribution is minimized using an iterative method. The events used in this method are collected using standard CMS triggers and require the presence of at least one electron satisfying a predefined criteria.

The IC values, at a fixed value of η, obtained through the various methods are weighted by their respective precisions and combined. The residual mis-calibration of an inter-calibration method is calculated as the spread of the difference between the method's IC and the other method's IC at a fixed value of η. The residual mis-calibration in the 2018 dataset is presented in figure 4.11. The overall effect of performing inter-calibration on the di-electron mass distribution can be seen in the same figure.

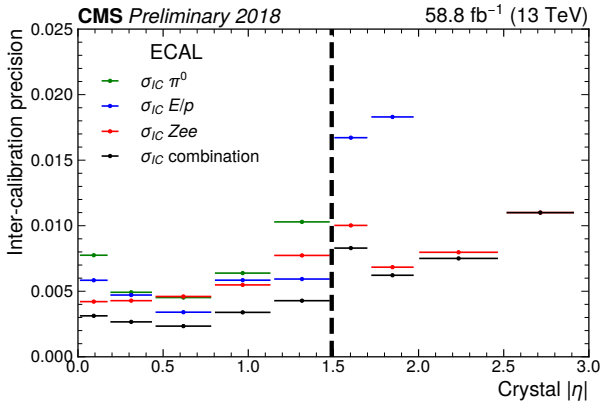


(a) Diphoton invariant mass in EB

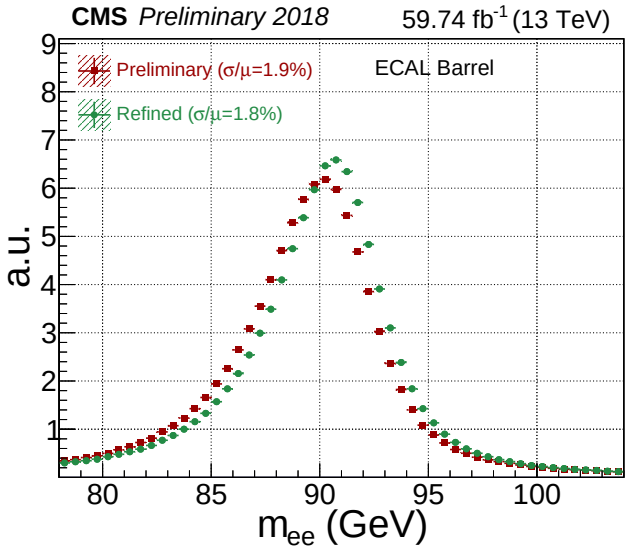


(b) Diphoton invariant mass in EE

Figure 4.10: Distributions of invariant mass of photon pairs with one photon depositing a fraction of its energy in a crystal of the EB at η = -0.03 (top), and of the EE at η = 1.83, in the π^0 mass range. A subset of data, corresponding to half of the total dataset collected in 2018, is used. These events are collected by CMS with a dedicated trigger at a rate of 7(2) kHz in the barrel (endcaps). A special clustering algorithm that only saves minimal information, in particular the energy deposits in ECAL crystals corresponding to a possible π^0 candidate, makes this high trigger rate possible to be used. The presence of preshower helps improve the mass resolution in EE. These events are used in the inter-calibration of ECAL crystals [25].



(a) 2018 Inter-calibration precision



(b) Dielectron invariant mass distribution in EB (2018 data)

Figure 4.11: (Top) Residual mis-calibration of the ECAL channel inter-calibration, as a function of the pseudorapidity is shown for the data recorded in 2018. The red, blue, and green points represent the residual mis-calibration of the intercalibration constants (IC) obtained with three different methods, and the black points represent the residual mis-calibration of the combination of the three methods. The IC combination is performed by weighting the different methods relatively to energy resolution performance as measured in $Z \to ee$ decays. (Bottom) The plot shows the di-electron invariant mass distribution for Z decay events with two calibration sets for the full 2018 dataset: the "preliminary" autumn 2018 calibration (RED) and a "refined" re-calibration performed in 2019 (GREEN). Both electrons are required to be in the ECAL Barrel, and the relative resolutions are quoted in the legend [25].

4.2.4 Absolute calibration

The absolute energy scale, G(η) from equation 4.1, is computed using the $Z \to ee$ events. $Z \to ee$ events are also used to set the η scale to ensure that different η regions have the same relative response.

4.2.5 High level calibrations

The cluster correction factor, $F_{e,\gamma}$, in equation 4.1 is obtained separately for electrons and photons. There is large amounts of material present in front of ECAL, up to $2X_0$ in the outer barrel region, as seen in figure 3.14 (left). This material leads to a high rate of bremsstrahlung radiation from photons, and high rate of photon conversions. To mitigate this effect, the energy of a supercluster is corrected using a multivariate analysis technique (MVA) trained on simulations. The MVA addresses the energy loss due to interactions with material in front of the ECAL, the energy leakage in gaps between the crystals, and the contribution from spurious energy deposits due to in-time pileup events. This MVA is trained separately for photons [60] and electrons [61].

4.3 ECAL energy resolution

4.3.1 Energy resolution of a calorimeter

The energy resolution of a calorimeter can be expressed as follows:

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c \tag{4.5}$$

where $\oplus$ denotes the sum in quadrature, E denotes the energy released by a particle in the calorimeter, and coefficients a, b, c represent the stochastic term, the noise term, and the constant term, respectively.

The development of the shower shape inside the detector represents the stochastic contribution to the energy resolution. In homogeneous detectors, such as ECAL, the shower development is almost fully contained within the active medium, and therefore the contribution of the stochastic term is usually small.

The electronic noise in the ECAL readout chain contributes to the noise term in Equation 4.5. The on-detector electronics are the prominent source of this noise. Due to various effects such as radiation damage, which has an impact on both the readout electronics as well as the crystal structure, the noise per crystal has increased with time. Nevertheless, for energies approximately above 100 GeV, the resolution is dominated by the constant term.

The constant term is independent of the incoming particle's energy. Several factors such as the non-uniformity of the longitudinal light collection, possibility of energy leak from the back of the calorimeter, random instabilities contribute to the constant term. It is of vital importance to keep the contribution of this term as low as possible so as not to degrade the resolution. This is achieved by the intercalibration methods and laser monitoring calibration.

During test beam studies, ECAL EB resolution has been studied for electrons, and the terms are directly measured to have the values: a=2.8% GeV^{-1}, b=0.124 GeV, and c=0.3%.

4.3.2 ECAL Energy Resolution during Run II

The ECAL energy resolution is extracted from an unbinned likelihood fit to $Z \to ee$ events, using a Voigtian (Landau convoluted with Gaussian) as the signal model. Figure 4.12 shows the relative resolution using the 2018 dataset as a function of pseudorapdity, η. The observed resolution gets worse at higher η in EB due to the larger amount of material present upstream of ECAL. The resolution is also degraded near the η cracks between ECAL modules, which are indicated by vertical lines in the plot. The energy resolution improves to $\approx 2\%$ for low bremsstrahlung electrons in the barrel region after a refined calibration using the full 2018 dataset is applied.

The excellent energy resolution provided by ECAL directly translates to improvements

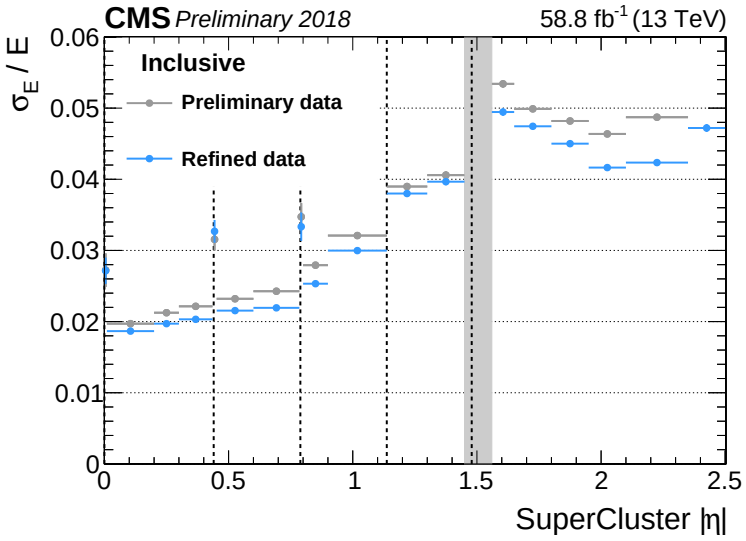


Figure 4.12: Relative electron (ECAL) energy resolution is shown as a function of pseudo-rapidity for the ECAL barrel and endcap for the 2018 data. Electrons from $Z \to ee$ events are used for these measurements. Comparison of the preliminary data (in grey) and refined data, with calibration using the full 2018 dataset, shows significant improvement in the energy resolution [25].

in physics analyses such as $h \to \gamma\gamma$, where an effective diphoton mass resolution of $\approx 1\%$ is achieved in the most sensitive analysis category [62].

This thesis describes an analysis with four photons in the final state, and the ECAL performance is a vital ingredient in the sensitivity of the results. The details of the analysis are discussed in Chapter 5.

4.4 ECAL Alignment

The primary goals of ECAL are to provide a precise energy and position resolution for electrons and photons in a large range of momenta. It is extremely important to have an accurate position measurement by the ECAL for photons, since it is the only way to measure their trajectory which in turn is critical for reconstructing the invariant mass, for example for the search of the Higgs Boson in the $h \to \gamma\gamma$ channel. A high accuracy is also required to correctly match the energy deposits in ECAL with the hits in the tracker, for both the

trigger and the offline electron reconstruction and identification. This matching helps to identify and remove fake electrons. Typically, a difference of at most 4 $\times 10^{-3}$ units in η and 20 mrad in ϕ between the position extrapolated from the tracker and that reconstructed by ECAL, defines the electron identification criteria. Simulation based studies [63] have shown that a position resolution of $\approx 10^{-3}$ units in η and 1.6 mrad in ϕ can be obtained for 35 GeV electrons using ECAL information only. The position of each supermodule (SM) has been determined at the time of installation in CMS, with an accuracy of about 1mm. However, it is necessary to perform further alignment of the ECAL with the tracker detector so as to improve the accuracy and reduce the error on the position measurement below the requirements for electron identification and diphoton invariant mass reconstruction. Therefore, an alignment technique was developed during Run 1 of the LHC [64]. I have been responsible for performing the alignment procedure during Run 2 of the LHC.

The alignment procedure makes use of the track measurement of isolated electrons that are produced by decays of the W or Z boson. The distance between the position measurement provided by the ECAL and that obtained from the track extrapolated on ECAL is minimized as a function of the alignment coefficients.

4.4.1 Alignment procedure

The electron position provided by the ECAL detector (supercluster position) is compared with the point of closest approach of the tracker track to the ECAL position. The impact position of an incoming electron drives the spatial distribution of the deposited energy, and the point of maximum shower activity in the ECAL crystals helps in the measurement of the position in ECAL. The position measurement can be defined as a weighted average of the positions of the crystals involved in the energy deposit:

$$\vec{X} = \frac{\sum_{xtal \in SC} w_{xtal} \vec{x}_{xtal}}{\sum_{xtal \in SC} w_{xtal}} \tag{4.6}$$

where

$$w_i = w_0 + \log\left(\frac{E_i}{\sum_{j\in SC} E_j}\right) \tag{4.7}$$

The tracker based position is the point of closest approach to the supercluster position, extrapolated from the innermost track position and direction. These distances along the η and ϕ directions between the ECAL position and extrapolated tracker position are used to build the following function:

$$\chi^2 = \chi^2_+ + \chi^2_- \tag{4.8}$$

The function $\chi^2_\pm$ depends on the lepton charge (electron or positron) and is calculated as:

$$\chi^2_\pm = \sum_{lepton} \frac{\left(\Delta\varphi - \left\langle \Delta\varphi^{MC}_\pm \right\rangle\right)^2}{\varepsilon^2_\varphi} + \frac{\left(\Delta\eta - \left\langle \Delta\eta^{MC} \right\rangle\right)^2}{\varepsilon^2_\eta} \tag{4.9}$$

where ε is the error associated with the SC position determination and depends on the electron energy [65]. The $\left\langle \Delta\varphi^{MC}_\pm \right\rangle$ and $\left\langle \Delta\eta^{MC}_\pm \right\rangle$ are Monte Carlo (MC) expectation values, and due to the interplay between the bending of the electron trajectory in the magnetic field and the tilt of the crystals, they are not expected to be zero. Since electrons and positrons bend oppositely in the magnetic field, their $\Delta\varphi^{MC}$ value is different. The values of $\left\langle \Delta\varphi^{MC}_\pm \right\rangle$ and $\left\langle \Delta\eta^{MC}_\pm \right\rangle$ are extracted from MC by applying selections similar to those applied on data from a W or Z sample, using ideal detector conditions, and they are found to have an insignificant dependence on η in the barrel region. The $\Delta\varphi$ and $\Delta\eta$ measured from data are shifted to match the MC values.

The $\chi^2_\pm$ is calculated and minimized with respect to the three-dimensional translations (and three Euler angles in the endcap) for each ECAL element (supermodules in the barrel and dees in the endcap).

Depending on the availability of the datasets, both $Z \to ee$ and $W \to ee$ events can be used for the ECAL alignment procedure. The kinematic and identification selections applied on the events mimic the online selections, but are kept a little tighter. In order to have a

good sample of electrons, it is required that the associated supercluster is composed of a singular basic cluster. This is known as the Golden electron selection, and it helps isolate the supercluster and bremsstrahlung effects from the elecrton position reconstruction.

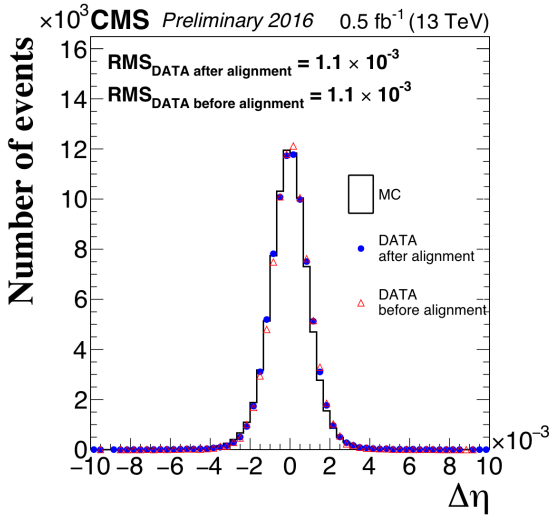
4.4.2 Alignment performance in Run 2

The ECAL alignment has been performed throughout the Run 2 data taking period; in 2016, 2017 and 2018.

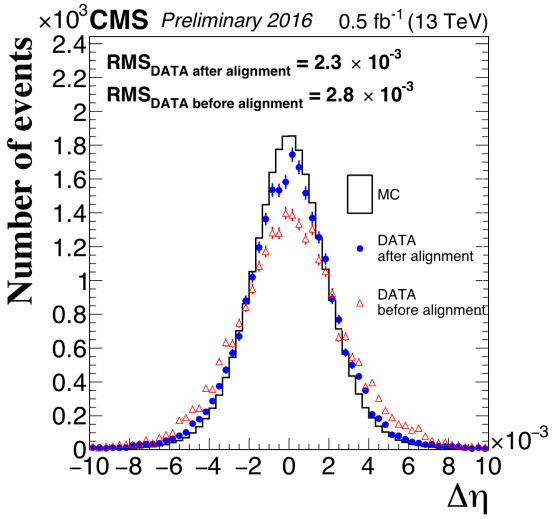
In 2016, 0.5 fb^{-1} of data from the $Z \to ee$ events were used for alignment purpose. The plots in figure 4.13 show the residual difference in η and ϕ after the alignment procedure has been performed. In this figure, red triangles correspond to data that use the 2015 ECAL alignment conditions, and blue circles correspond to data with the 2016 alignment conditions. The distributions for perfectly aligned Monte Carlo events are shown in black, and the improvement in data - Monte Carlo comparison, with the new 2016 alignment conditions, is clear from these plots. The $\Delta\phi$ distribution shown here is for electrons, and therefore has an asymmetric shape.

In 2016, the initial misalignment was found to be larger in the ECAL endcap minus crystals and this can be seen clearly from figure 4.14. The agreement between data and MC was greatly improved in the EE minus region after a new ECAL alignment condition was provided.

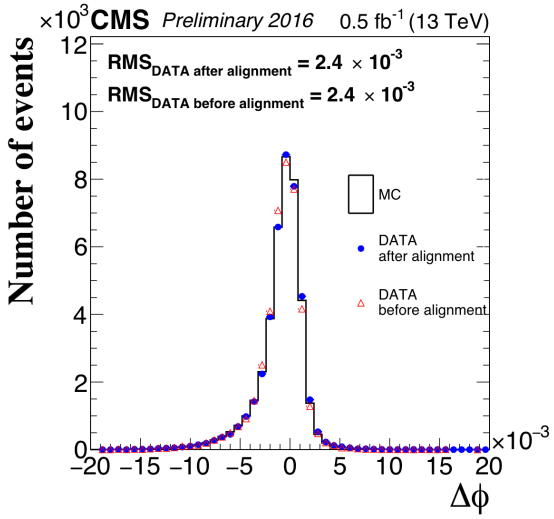
In 2017 and 2018, new ECAL alignment conditions were provided, and similar distributions of residual difference in η and ϕ between the position of the ECAL supercluster and the tracker based position for 2017 can be seen in figures 4.15 and 4.16.



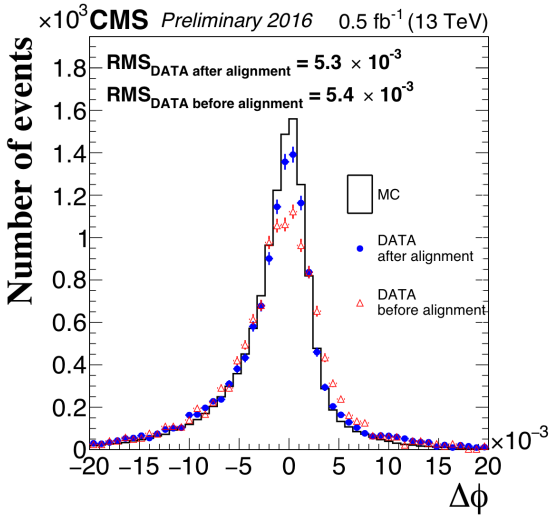
(a) $\Delta\eta$ in EB



(b) $\Delta\eta$ in EE

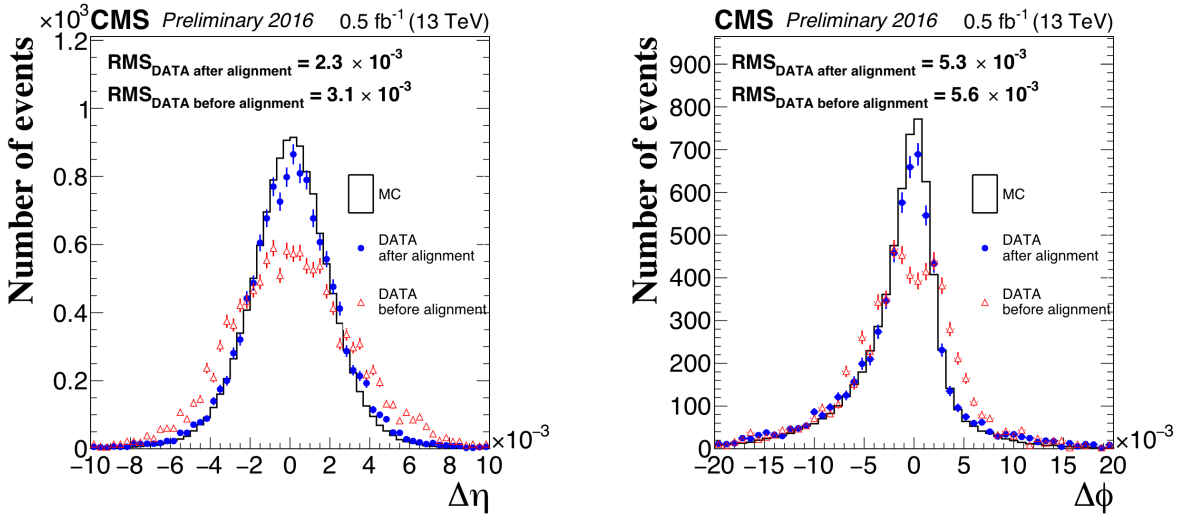


(c) $\Delta\phi$ in EB



(d) $\Delta\phi$ in EE

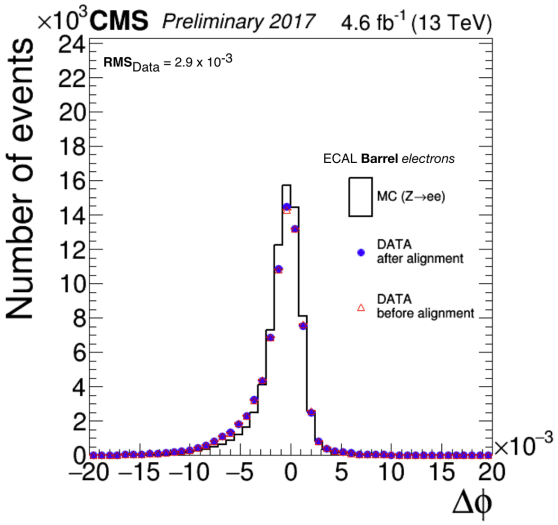
Figure 4.13: The residual difference in η and in ϕ and between the position of the ECAL supercluster and the tracker based position, which is the point of closest approach to the supercluster position, extrapolating from the innermost track position measurement is shown for the 2016 data taking year. In 2016, relative ECAL-tracker alignment precision of 5 $\times 10^{-3}$ rad in ϕ and 2×10^{-3} units in η in the endcap was achieved [26].



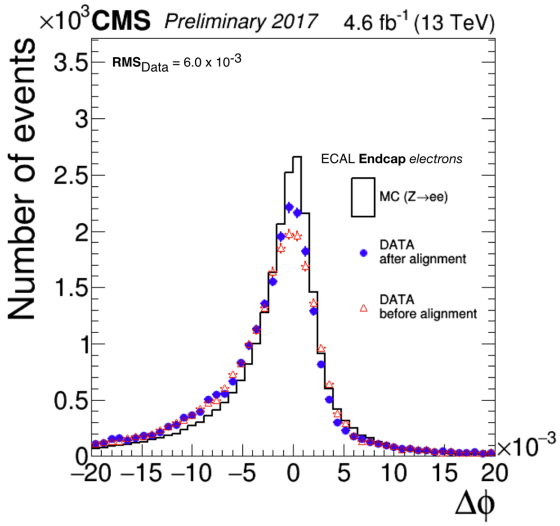
(a) $\Delta\eta$ in EE minus

(b) $\Delta\phi$ in EE minus

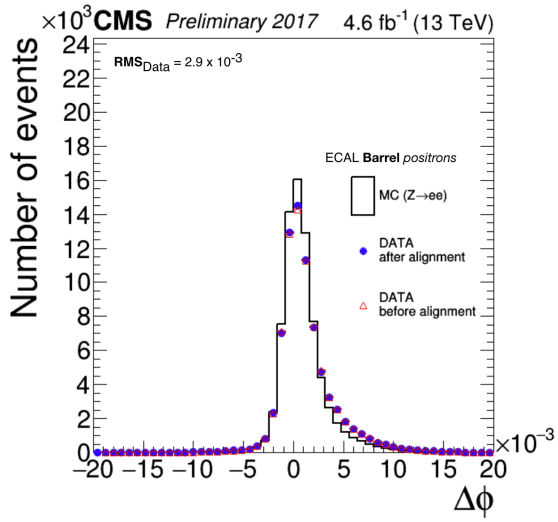
Figure 4.14: Relative alignment of the ECAL endcap minus crystals, where the initial misalignment was larger. The plots correspond to the data collected in 2016 [26].



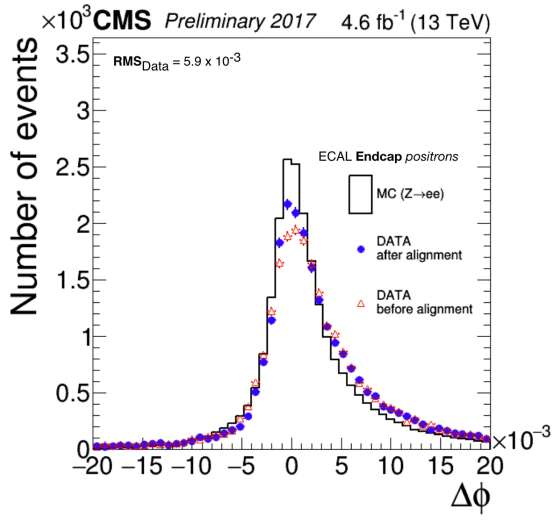
(a) $\Delta\phi$ in EB for electrons



(b) $\Delta\phi$ in EE for electrons

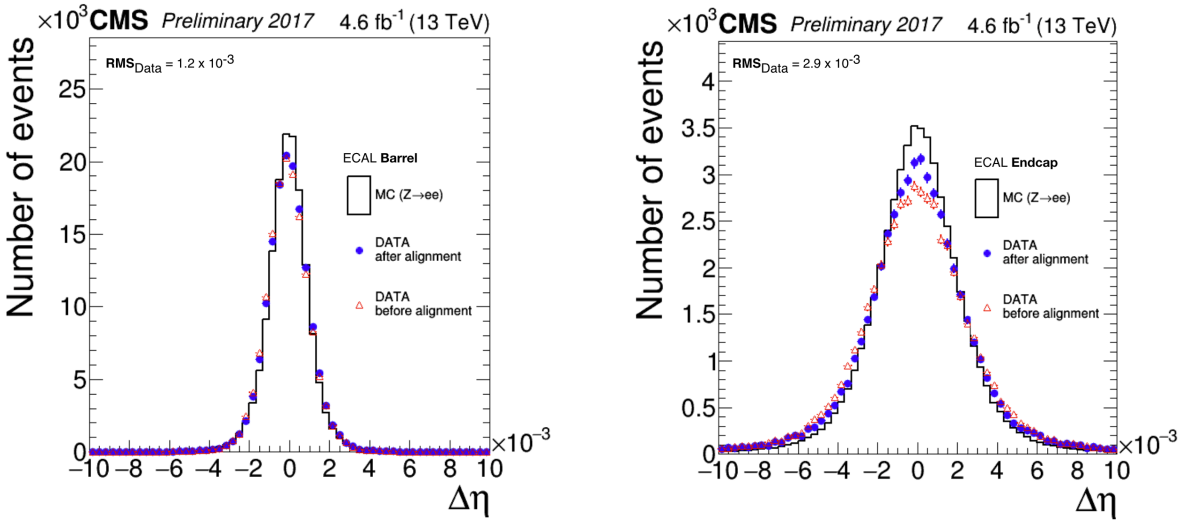


(c) $\Delta\phi$ in EB for positrons



(d) $\Delta\phi$ in EE for positrons

Figure 4.15: The residual difference in ϕ and between the position of the ECAL supercluster and the tracker based position, which is the point of closest approach to the supercluster position, extrapolating from the innermost track position measurement is shown for the 2017 data taking year. The plots show distributions for electrons and positrons separately [27].



(a) $\Delta\eta$ in EB

(b) $\Delta\eta$ in EE

Figure 4.16: The residual difference in η and between the position of the ECAL supercluster and the tracker based position, which is the point of closest approach to the supercluster position, extrapolating from the innermost track position measurement is shown for the 2017 data taking year [27].

The procedures detailed in Sections 4.2 and 4.4 describe the calibration and alignment performed in the ECAL sub-system, and are pivotal to the excellent energy and position resolution provided by ECAL.

The ECAL has given excellent performance during Run 2 and has aided CMS in reaching its physics goals. ECAL is undergoing various upgrades to mitigate the change in detector conditions during Run 3, and during the HL-LHC era. Details of these planned upgrades can be found in References [66] and [67].

Chapter 5

Search for light pseudoscalars in the four photon final state

This chapter describes an analysis of the CMS data performed on events from proton-proton collisions provided by the LHC at $\sqrt{s} = 13\,\text{TeV}$. The first CMS analysis in this final state, this analysis searches for exotic decays of the Higgs bosons into pseudoscalars that belong to BSM models such as those described in Section 2.2.4, with four photons in the final state.

Section 5.1 describes the analysis strategy in a nutshell, and Sections 5.2 and 5.3 describe the data and simulation samples utilized in the analysis, and the physics objects considered and selections applied on them, respectively. Section 5.4 details the strategy developed for estimation of background, and Section 5.5 explains the methodology used for further improving the signal discrimination from the background processes. Section 5.6 provides a look at the signal efficiency after all analysis selections are applied. Section 5.7 describes the methods used for statistical analysis of the data, and this is followed by the description of the systematic uncertainties which impact this analysis in Section 5.8. Finally, the results are presented in Section 5.9.

5.1 Strategy

This analysis searches for exotic decay of the 125 GeV Higgs boson to two pseudoscalars, with equal mass, each of which decay to two photons each. A Feynman diagram contributing

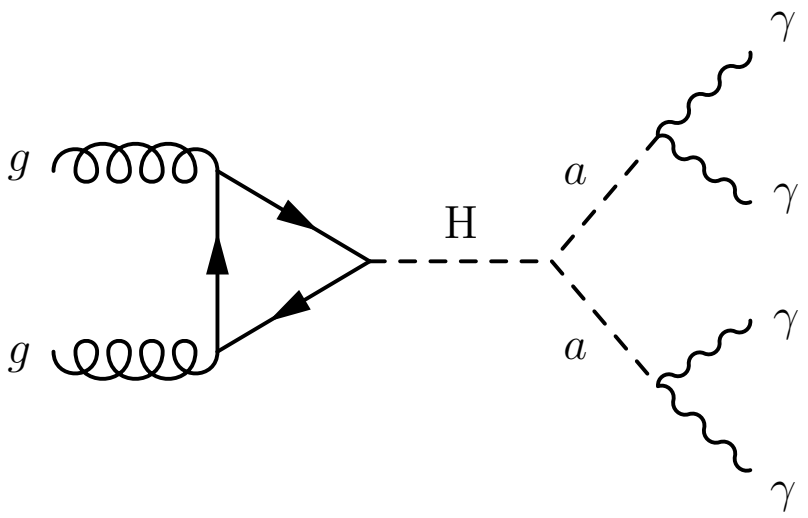


Figure 5.1: Feynman diagram of the $h \to aa \to \gamma\gamma\gamma\gamma$ process.

to this process at leading order (LO) is shown in Figure 5.1.

The topology of the final state is highly dependent on the mass of the pseudoscalar being probed, and the experimental signature of the final state is determined by the opening angle of the photons relative to the rest frame of the $a \to \gamma\gamma$ decay. The possible final state scenarios are as follows:

- Fully merged photons: Both pairs of photons coming from the $a \to \gamma\gamma$ decay are highly collimated and deposit their energy in the CMS ECAL such that they are reconstructed as a single photon object. In other words, both pairs of photons are considered to be merged and the final state consists of exactly 2 reconstructed photons in this case.

- Semi-merged photons: This scenario represents the case when one out of the two pairs of photons are highly collimated and their energy deposits in the ECAL are merged. The photons belonging to the second pair have a wide opening angle and deposit their energies in the ECAL such that they can be reconstructed as separate photon objects. In this scenario the final state consists of exactly 3 reconstructed photons.

- Fully resolved photons: Both pairs of photons have a wide opening angle and each of the four photons are reconstructed separately in the CMS ECAL. The final state consists of at least 4 reconstructed photons in this case.

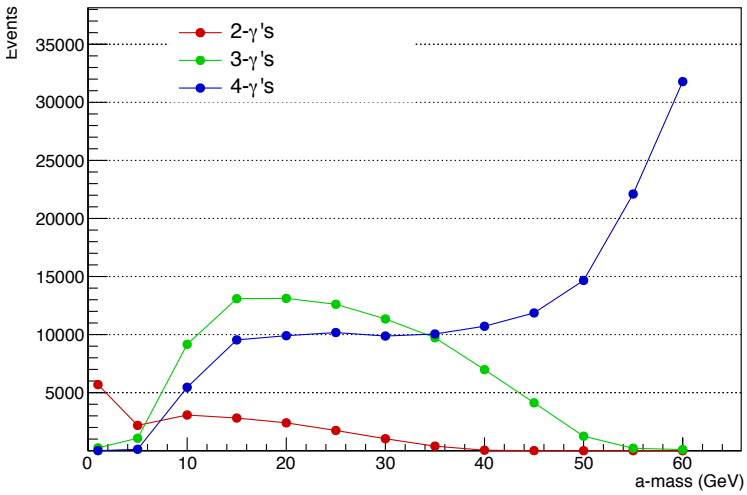


Figure 5.2: Number of events versus pseudoscalar mass are plotted for scenarios with two, three or four reconstructed events in the final state.

Figure 5.2 illustrates the of number of events, after certain analysis level selections are applied, for two, three or four reconstructed photons in the final state. The number of events are shown for pseudoscalars ranging in mass from 100 MeV to 60 GeV. As seen from the plot, for masses $m_a < 5$ GeV, the highest fraction of events fall under the category of two photons in the final state. In this scenario, the $a \to \gamma\gamma$ photons are highly collimated and deposit their energies in the CMS ECAL such that they are reconstructed as a single photon. For pseudoscalars with mass lying within the range $10 < m_a < 35$ GeV, the largest fraction of events have 3 photons in the final state. There are two possible scenarios which may lead to three photons in the final state:

- Two out of the four photons could be collimated and reconstructed as a single photon in ECAL.

- One photon out of the four is not reconstructed due to it failing online trigger selections or offline analysis level selections.

Both the above described scenarios, with two or three reconstructed photons in the final state, require development of novel photon reconstruction techniques and are not the subject of the analysis described in this thesis.

This analysis searches for pseudoscalars with masses which range from 15 to 60 GeV, and leads to at least four fully reconstructed photons in the final state.

For such pseudoscalars, the distribution of the invariant mass of the four-photon object, $m_{\gamma\gamma\gamma\gamma}$, peaks around the mass of the Higgs boson (125 GeV). The analysis strategy is, therefore, based on extracting the number of signal and background events using a parametric fit to the invariant mass of the four-photon system, $m_{\gamma\gamma\gamma\gamma}$.

In order to select events with good quality photons, events are required to have satisfied particular online trigger selections. Furthermore, additional selections are applied at the offline level.

An event classifier, which exploits the kinematic information of the photons and pseudoscalar objects, is trained to separate the signal from the background processes. An optimized selection based on the output of the classifier is used to define the final signal regions in the analysis.

Finally, a parametric fit to the signal simulation is done to construct the signal model, and a data driven strategy is used to construct the background model. These signal and background models are then used to set model independent limits on $\sigma(pp \to h) \times BR(h \to aa \to \gamma\gamma\gamma\gamma)$.

5.2 Samples

5.2.1 Data

This analysis is based on a total of $131.8\,\mathrm{fb}^{-1}$ of pp collision data collected by the CMS at $\sqrt{s} = 13$ TeV. Of this, $35.9\,\mathrm{fb}^{-1}$ was collected during 2016, $41.5\,\mathrm{fb}^{-1}$ during 2017, and 54.38 fb^{-1} during 2018.

Data used in this analysis must be vetted by CMS trigger system. The HLT path used in this analysis must be seeded by at least one electromagnetic candidate at the L1 trigger. At L1 stage of the trigger, at least one of the two photons is required to be present. However, since this can result in a high event rate, a p_T selection is applied. The threshold for this selection is typically set at 30 GeV. In order to increase efficiency, asymmetric HLT paths seeded by two electromagnetic candidates are also used. In 2016, the L1 p_T thresholds were 22 (15) GeV for the leading (subleading) candidates. In 2017 and 2018, these thresholds were set at 25 and 14 GeV, respectively.

At the HLT level, shower shape and photon isolation variables are included in the selection. p_T requirements on the two electromagnetic candidates are also applied at this stage, which were set at 30 (18) GeV on the leading (subleading) candidate in 2016, 2017 and 2018. A selection at 55 GeV on the diphoton invariant mass was also imposed in 2016 and 2017, and was removed during 2018 data taking.

The dataset names and names of the HLT paths used in this analysis are presented in Tables A.1 and A.8, respectively.

5.2.2 Simulation

The signal simulation used in this analysis is generated using MadGraph5_aMC@NLO v2.4.2, which is next-to-leading order in perturbative QCD [68]. The generated simulation is then interfaced with the pythia generator for showering and hadronization, with the tune CUETP8M1 [69] for 2016, and CP5 [70] for 2017 and 2018.

For each year, signal simulation is produced for pseudoscalars with masses that range from 15 and 60 GeV, in steps of 5 GeV. The names of simulated samples can be found in Tables A.2, A.3, and A.4.

Although this analysis makes use of data-driven methods for background modeling, background simulation samples are used to check the shape of the expected background and to optimize the analysis. The main backgrounds for the four photon final state come from either

well-isolated photons which result from the hard scatter or events where jets are misidentified as photons. The background contribution from prompt photons is simulated using SHERPA 2.2.4 [71], and the contribution that contains at least one misidentified photon is simulated using PYTHIA, with a Double-EM enriched filter applied during production to enhance the selection efficiencies.

The names of the datasets corresponding to the various background simulation used in this analysis are listed in Tables A.5, A.6, and A.7.

The response of the CMS detector is modeled using the GEANT4 [72] package. The simulated events include additional *pp* interactions within the same or nearby bunch crossings (pileup), as observed in the data.

5.3 Analysis Objects and Selections

The goal of this analysis is to identify events resembling Figure 5.1 in collisions in the CMS detector. In order to achieve this goal, events which satisfy certain selection criteria need to be identified and chosen to perform the analysis. This is done in two stages. First, events must be chosen by the trigger selections which are described in Section 5.2.1, then in order to be further considered in the analysis, photons need to pass a set of selections designed to reduce the contamination from objects faking photons. These selections are referred to as preselections, and are discussed in Section 5.3.1. Further selections on the vertex candidate are applied and are described in Section 5.3.2.

5.3.1 Photon Preselection

Events that have successfully passed the selections applied at trigger level, need to satisfy preselections which are similar to, but more stringent than the trigger requirements.

The triggers used in this analysis apply selections on the diphoton object, and therefore the preselection requirements are also applied on the diphoton object. In order to pass this

stage of the analysis, the event needs to contain at least one diphoton object that satisfies the preselection requirements.

Selections are applied on the following set of variables that are related to the shower shape and isolation of the photon object:

- R9: This is a shower shape variable, and is a defined as E3×3/E_{SC}. E3×3 refers to the energy sum of the 3 × 3 group of ECAL crystals surrounding the most energetic crystal in the ECAL supercluster (SC), before any energy corrections are applied.

- HoE: This is an isolation variable, and is defined as the ratio of the energy collected by the HCAL towers within a cone of radius 0.15 centred on the SC to the SC energy.

- $\sigma_{i\eta i\eta}(5 \times 5)$: This is a shower shape variable, and is defined as the standard deviation of the shower in η in terms of the number of crystal cells.

- pfPhoIso: This is an isolation variable, and is defined as the sum of the transverse energy with all particles identified as photons by the PF algorithm falling inside a cone of radius 0.3 around the photon candidate direction.

- TrackerIso: This is another isolation variable as is defined as the sum of the transverse momentum associated with all tracks falling in a cone of radius 0.3 around the photon candidate direction. Tracks which fall in an inner cone of radius 0.4 are not included in the sum.

- Pixel safe electron veto: The photon candidate is rejected if its supercluster is matched to a hit in the pixel detector.

The threshold of selections applied on these variables follow the changes in the trigger selections per year, and are summarised in Tables 5.1, 5.2 and 5.3 for 2016, 2017, and 2018, respectively.

		R9(5x5)	HoE	$\sigma_{i\eta i\eta}$ (5x5)	pfPhoIso	TrackerIso
Both photons in barrel	Barrel	>0.5	<0.07	< 0.0105	< 4 GeV	< 6 GeV
At least one in endcap	Barrel	>0.85	<0.07	< 0.0105	< 4 GeV	< 6 GeV
At least one in endcap	Endcap	>0.9	<0.035	< 0.0275	< 4 GeV	< 6 GeV

Table 5.1: 2016 Photon preselection

	R9(5x5)	HoE	$\sigma_{i\eta i\eta}$ (5x5)	pfPhoIso	TrackerIso
EB;R9 > 0.85	>0.5	< 0.08	<0.0105	< 4 GeV	< 6 GeV
EB;R9 $\leq$ 0.85	>0.5	< 0.08	<0.0105	< 4 GeV	< 6 GeV
EE;R9 > 0.9	>0.8	< 0.08	<0.035	< 4 GeV	< 6 GeV
EE;R9 $\leq$ 0.9	>0.8	< 0.08	<0.035	< 4 GeV	< 6 GeV

Table 5.2: 2017 Photon preselection

	R9(5x5)	HoE	$\sigma_{i\eta i\eta}$ (5x5)	pfPhoIso	TrackerIso
EB;R9 > 0.85	>0.5	< 0.08	<0.0105	< 4 GeV	< 6 GeV
EB;R9 $\leq$ 0.85	>0.5	< 0.08	<0.0105	< 4 GeV	< 6 GeV
EE;R9 > 0.9	>0.8	< 0.08	<0.035	< 4 GeV	< 6 GeV

Table 5.3: 2018 Photon preselection

Further selections are applied on the photon objects in events that pass the previously defined preselections. They are listed as follows:

- $N_\gamma \geq 4$; Event should contain at least four photons
- $p_T(\gamma_1) > 30$ GeV, $p_T(\gamma_2) > 18$ GeV, $p_T(\gamma_3) > 15$ GeV, $p_T(\gamma_4) > 15$ GeV; the p_T selections applied on the four photon objects are motivated by trigger level selections
- All photons are required to have $|\eta|$ <2.5 and not lie in the EB-EE gap, which corresponds to $1.442 < |\eta| < 1.566$
- All photons are required to pass the pixel safe electron veto
- $110 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV
- The photon MVA ID developed by the $h \to \gamma\gamma$ analysis is used in this analysis. The photon ID helps to discriminate between real photons and fake objects, such as jet fragments, that could pass the pre-selections, and uses input variables related to photon shower shapes and isolation. The details of the development of the photon MVA ID are available in reference [62]. There is no selection applied on photon MVA ID at this stage intentionally. This is because the photon MVA ID's are used as an input later for a categorization BDT, and no selection applied at this stage allows for availability of larger statistics for the training.

5.3.2 Vertex Selection

In the h→ $\gamma\gamma$ analysis, the choice of the interaction vertex from which the Higgs boson is produced has a large impact on the diphoton mass resolution, since it is determined according to the following equation:

$$m_{\gamma\gamma} = \sqrt{2E_{\gamma_1}E_{\gamma_2}(1-\cos\theta)} \tag{5.1}$$

here, E_{γ_1} and E_{γ_2} is the energy of each photon, and θ is the opening angle between them. Apart from accurate identification of the photons, the sensitivity of the analysis is contingent on determining the correction location of the interaction vertex, since that information determines the opening angle.

For similar reasons, determination of the interaction vertex is vital for this analysis as well. There are multiple proton-proton collisions at each bunch crossings, and multiple vertices are present in each event. In order to maximise the sensitivity of this analysis, a BDT is trained to identify the most likely interaction vertex.

The BDT is trained with simulated $h \to aa \to \gamma\gamma\gamma\gamma$ signal events. Variables related to the tracks recoiling against the $m_{\gamma\gamma\gamma\gamma}$ system are used as input for the training. The following variables are calculated for each possible vertex candidate:

- $\sum_i |\vec{p}_T^{\,i}|^2$-$\sum_i(\vec{p}_T^{\,i} \cdot \frac{\vec{p}_T^{\,\gamma\gamma\gamma\gamma}}{|\vec{p}_T^{\,\gamma\gamma\gamma\gamma}|})$
- $(|\sum_i \vec{p}_T^{\,i}| - \vec{p}_T^{\,\gamma\gamma\gamma\gamma})/(|\sum_i \vec{p}_T^{\,i}| + \vec{p}_T^{\,\gamma\gamma\gamma\gamma})$

where $\vec{p}_T^{\,i}$ refers to the p_T of the i-th track associated with a given vertex candidate and $\vec{p}_T^{\,\gamma\gamma\gamma\gamma}$ refers to the transverse momentum of the four-photon candidate.

In case more than one photon undergoes conversion into an electron, the following variables are also considered as input variables to the training:

- Number of conversions associated with the vertex candidate
- $\frac{|z_{PV} - z_{PV}^{conv}|}{\sigma_{conv}}$, where z_{PV} is the longitudinal position of the reconstructed vertex, z_{conv} is the longitudinal position of the vertex estimated using conversion tracks, and σ_{conv} is the uncertainty associated with z_{conv}

Since this analysis probes a wide range of pseudoscalar mass, a sum of the simulated signal corresponding to all the masses are used for training the BDT. The BDT assigns a score to each candidate vertex, and chooses the vertex with the highest score as the event vertex.

A similar vertex identification BDT training is performed in the h$\to \gamma\gamma$ analysis. Therefore, there are three vertex choices available: h$\to \gamma\gamma$ trained vertex [62], standard CMS vertex, and the newly trained $h \to aa \to \gamma\gamma\gamma\gamma$ vertex.

Figure 5.3 compares the efficiency of choosing the correct vertex as a function of $|\Delta$ Z $|$ for the three available vertex choices, where

$$|\Delta Z| = |Z_{GenVertex} - Z_{RecoVertex}| \tag{5.2}$$

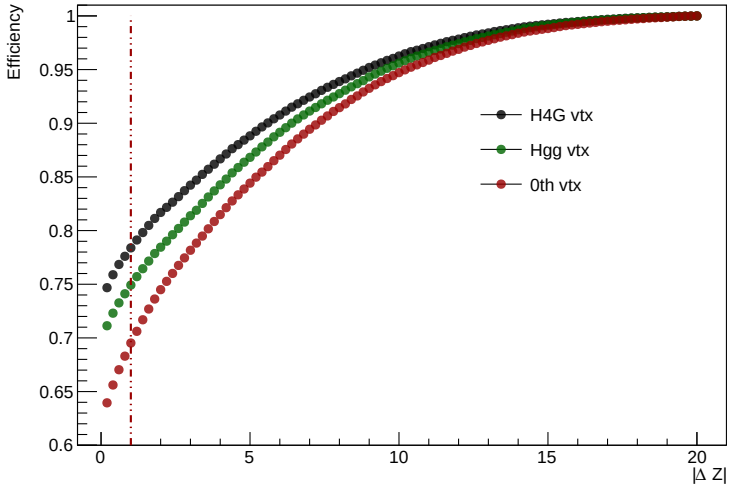
here $|Z_{GenVertex}|$ refers to the longitudinal position of the generated vertex, and $|Z_{RecoVertex}|$ is the longitudinal position of the reconstructed vertex.

The efficiency of choosing the event vertex with $|\Delta Z| < 1$ is higher than 80% when the trained vertex is used. Across the pseudoscalar mass point range, the trained vertex performs better than the standard CMS or the $h \to \gamma\gamma$ trained vertex. As a result of this, an improvement in the $m_{\gamma\gamma\gamma\gamma}$ resolution is expected.

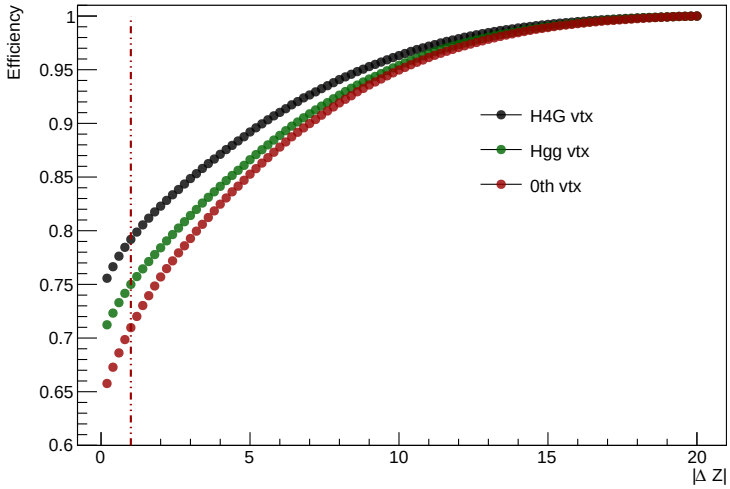
Due to changes in the CMS tracker detector during the Run 2 data taking period, it was observed that the shapes of the input variabes vary among the years. Therefore, a separate vertex identification BDT training was done for the three years. As seen from Figure 5.4, the performance of the vertex identification BDT is similar among the three years.

The score of the $h \to aa \to \gamma\gamma\gamma\gamma$ trained vertex BDT is not used an input to any additional BDT in the analysis. Therefore, it is not considered necessary to evaluate any additional systematics and related scale factors, with consequent validation on data.

The new dedicated vertex identifying algorithm, trained on the $h \to aa \to \gamma\gamma\gamma\gamma$ signal is used in this analysis to improve the signal resolution of the fitting variable $m_{\gamma\gamma\gamma\gamma}$. This improvement is illustrated in figure 5.5 which compares the signal resolution of a fit to the $m_{\gamma\gamma\gamma\gamma}$ distribution, for pseudoscalars of mass 15 and 60 GeV, using different vertex choices. For both pseudoscalar masses, this new training algorithm leads to an improvement of at least 1% in the $m_{\gamma\gamma\gamma\gamma}$ resolution, after applying standard analysis selections.

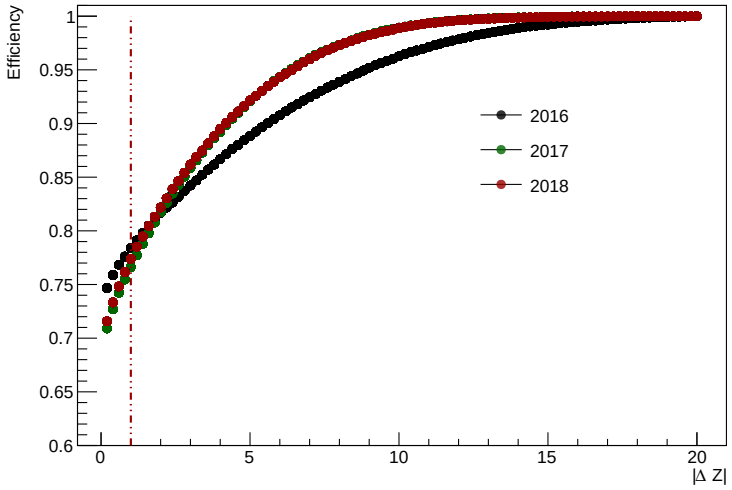


(a) m(a) = 15 GeV

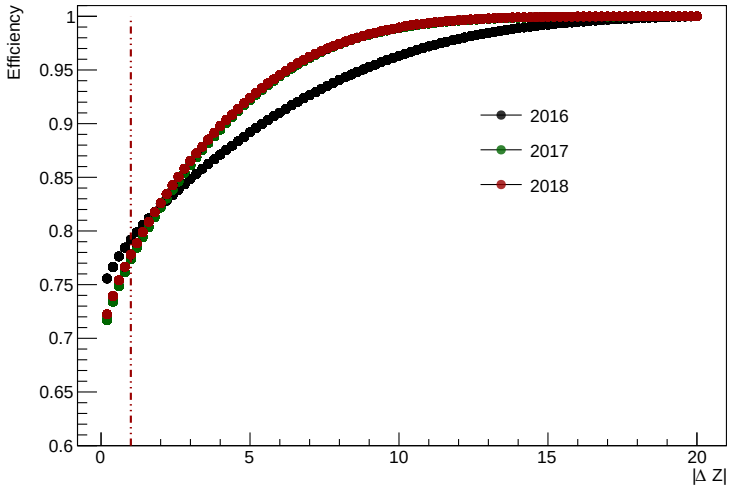


(b) m(a) = 60 GeV

Figure 5.3: Efficiency of choosing the correct interaction vertex as a function of $|\Delta$ Z $|$ is shown for m_a = 15 (top) and 60 GeV (bottom). A comparison between standard CMS vertex, $h \to \gamma\gamma$ trained vertex and $h \to aa \to \gamma\gamma\gamma\gamma$ trained vertex (with 2016 training) is shown.

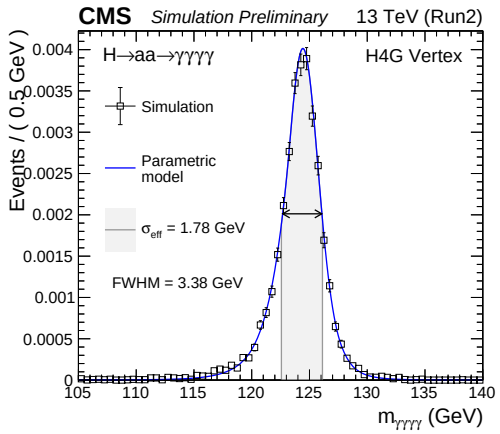


(a) m(a) = 15 GeV

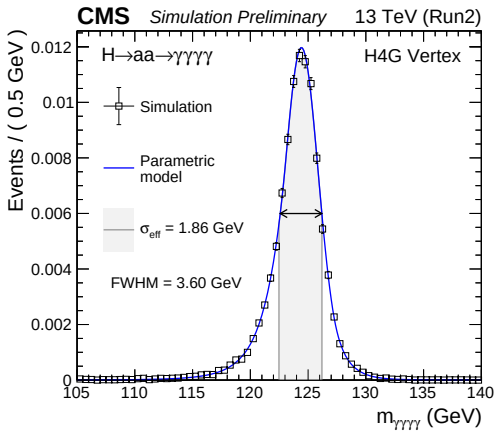


(b) m(a) = 60 GeV

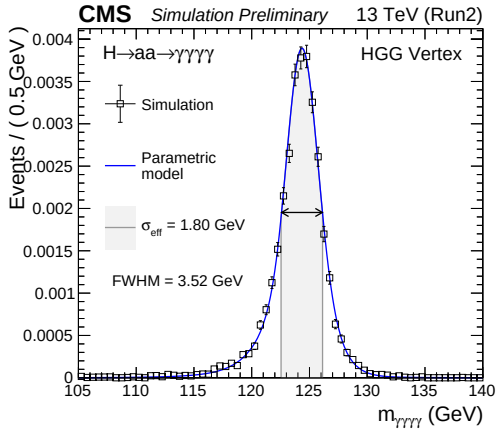
Figure 5.4: Efficiency of choosing the correct interaction vertex as a function of $|\Delta$ Z $|$. The efficiencies for 3 data taking years (2016, 2017, 2018) are compared for the $h \to aa \to \gamma\gamma\gamma\gamma$ trained vertex.



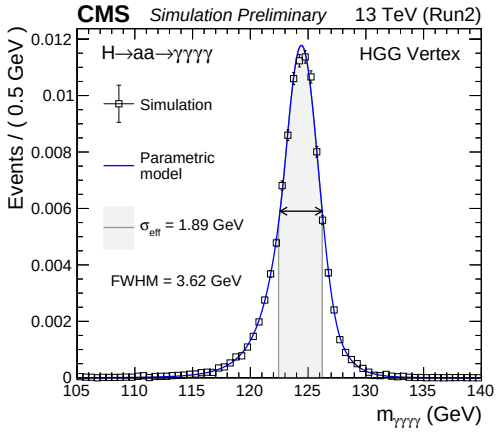
(a) m(a) = 15 GeV: $h \to aa \to \gamma\gamma\gamma\gamma$ Vertex



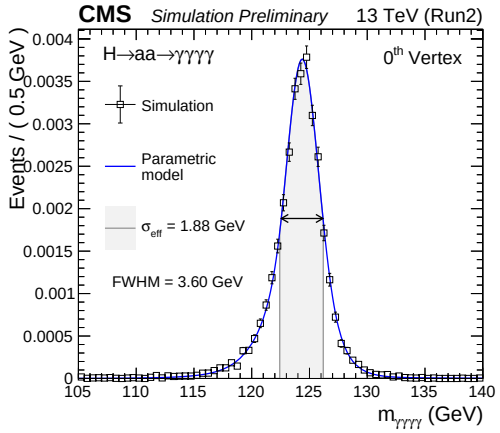
(b) m(a) = 60 GeV: $h \to aa \to \gamma\gamma\gamma\gamma$ Vertex



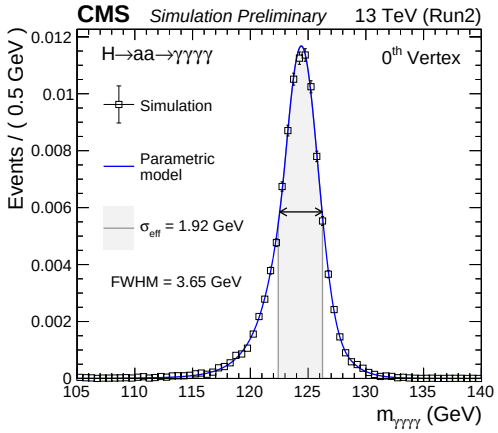
(c) m(a) = 15 GeV: $h \to \gamma\gamma$ Vertex



(d) m(a) = 60 GeV: $h \to \gamma\gamma$ Vertex



(e) m(a) = 15 GeV: CMS Standard Vertex



(f) m(a) = 60 GeV: CMS Standard Vertex

Figure 5.5: Distributions of$\mathrm{m}_{\gamma\gamma\gamma\gamma}$ variable are shown for the 3 available vertex choices: $h \to aa \to \gamma\gamma\gamma\gamma$ trained vertex, $h \to \gamma\gamma$ trained vertex and the 0^{th} vertex. The improvement in signal resolution with the use of $h \to aa \to \gamma\gamma\gamma\gamma$ trained vertex can be seen for both m(a) = 15 and 60 GeV.

5.3.3 Pseudoscalar candidate reconstruction

Once the four photons that have passed selections described in the previous sections are chosen, they are used to reconstruct the pseudoscalar candidate. In case of events with more than four photons, the first four photons with the highest p_T are chosen.

The following strategy is adopted to choose the pseudoscalar candidate: Four photons in the final state can be combined in to diphoton objects in three unique ways:

- $\gamma_1\gamma_2$, $\gamma_3\gamma_4$
- $\gamma_1\gamma_3$, $\gamma_2\gamma_4$
- $\gamma_1\gamma_4$, $\gamma_2\gamma_3$

For each combination of diphoton pairs ($\gamma_a\gamma_b$, $\gamma_c\gamma_d$), the difference between invariant mass of the diphoton candidates is evaluated:

$$\Delta M = |m_{\gamma_a\gamma_b} - m_{\gamma_c\gamma_d}| \tag{5.3}$$

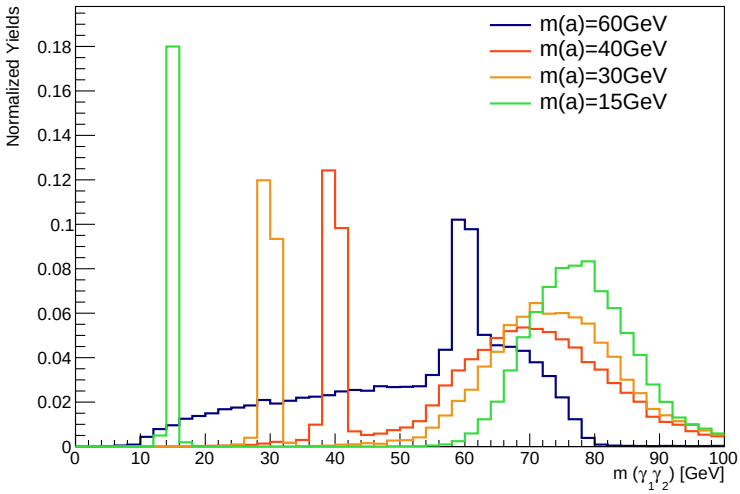
and the pair with the smallest value of Δ M is chosen as the correct pair of pseudoscalars.

It is evident from Figure 5.6 that this pairing method performs better than simply pairing the leading and sub-leading photons to construct the pseudoscalar object.

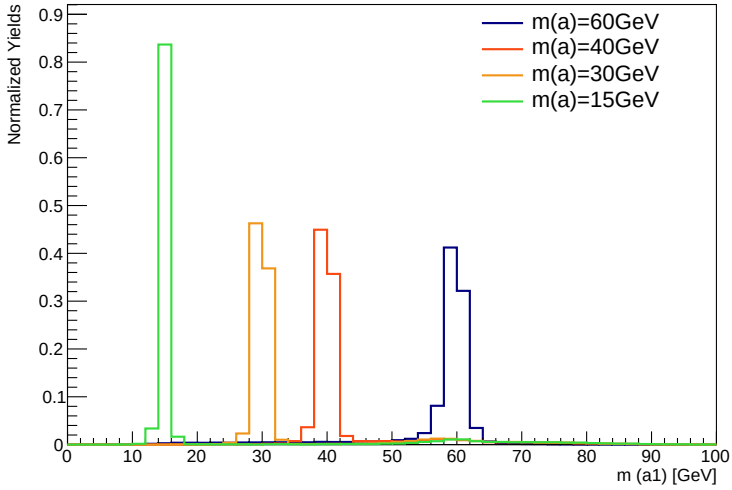
Although this analysis uses the $m_{\gamma\gamma\gamma\gamma}$ variable for evaluating the final result, it is essential to have good pseudoscalar mass resolution, since the invariant mass of the two pseudoscalars is used in training a further event categorization BDT.

5.3.4 Higgs candidate Reconstruction

The presence of at least four photons that have passed the selections described in the previous sections is required to select an event. In events with more than four photons, the first four photons with the highest p_T are chosen to construct the Higgs candidate. Figure 5.7 illustrates the Higgs mass distribution for pseudoscalars of various masses.



(a) Pseudoscalar candidate constructed using the first two highest p_T photons



(b) Pseudoscalar candidate constructed using the pairing strategy

Figure 5.6: Comparison of pseudoscalar candidate's mass resolution using two different methods. The pairing method provides a better mass resolution across the pseudoscalar mass range.

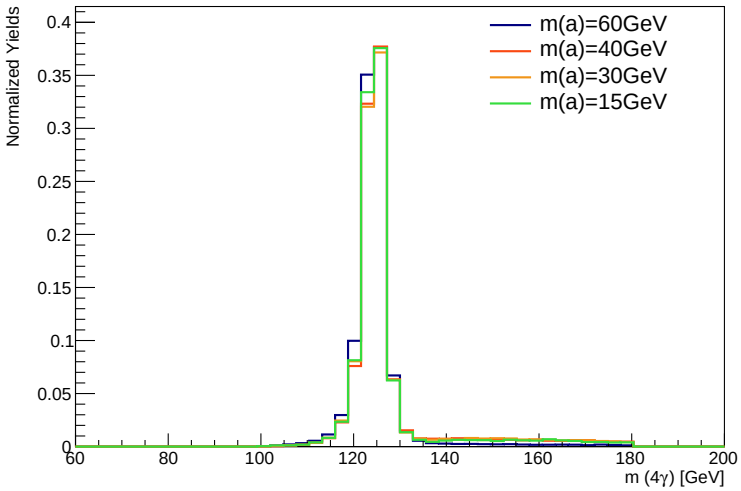


Figure 5.7: Distribution of the four-photon object is shown here for different pseudoscalar mass points. The distribution peaks around the mass of the Higgs boson (125 GeV).

5.4 Background estimation

This analysis searches for events with four photons in the final state, with $m_{\gamma\gamma\gamma\gamma}$ consistent with the Higgs boson mass. Any SM process that produces this signature is considered as a background to the search. The key goal of this analysis is to recognize these backgrounds, so that any excess of events can be identified. To this end, a BDT is trained to differentiate signal events from the background processes. Section 5.5 discusses the details of this BDT.

In this analysis, background estimation is completely data driven and is done by a fit to the $m_{\gamma\gamma\gamma\gamma}$ distribution. As a consequence, it is not essential to have a perfect agreement between data and background simulation for the analysis. However, since the final signal regions are defined by a selection on the BDT score, an accurate description of the background processes is beneficial for the optimal performance of the BDT.

The background simulation samples used in this analysis (QCD, γ+jets and $\gamma\gamma$+jets) suffer from low statistics and result in a poor description of events with fake photons. The QCD and γ+jets samples also have very large event weights. As a result of this, the agreement between data and background is sub optimal. The photon MVA ID scores of the four photons are very powerful in discriminating signal from background processes and are very important input variables to the categorization BDT training. Comparison of the data and background in distributions of MVA IDs of the leading and fourth photon, in Figures 5.8, 5.9 and 5.10 show that in all three years, 2016, 2017 and 2018, there is a large under prediction in the background simulation. These variables are chosen as an example for illustration purposes. Events shown in these distributions are selected after application of selections discussed in Section 5.3. The data and background simulation contribution belong to the $m_{\gamma\gamma\gamma\gamma}$ sideband region, defined as [$110 < m_{\gamma\gamma\gamma\gamma} < 115$ GeV, $135 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV]. In this analysis, the signal region is defined by the selection $115 < m_{\gamma\gamma\gamma\gamma} < 135$ GeV, and in order to remain unbiased to the presence of a signal, only events that belong to the sideband region are shown in these distributions.

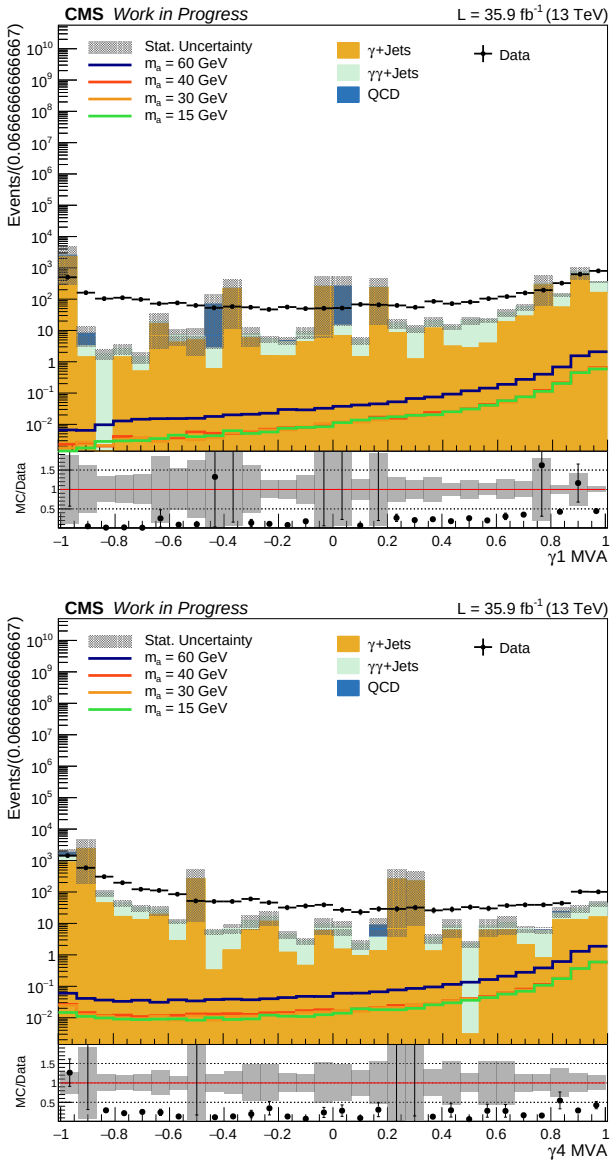


Figure 5.8: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2016 data and simulated samples. The data and background simulation belong to the sideband region.

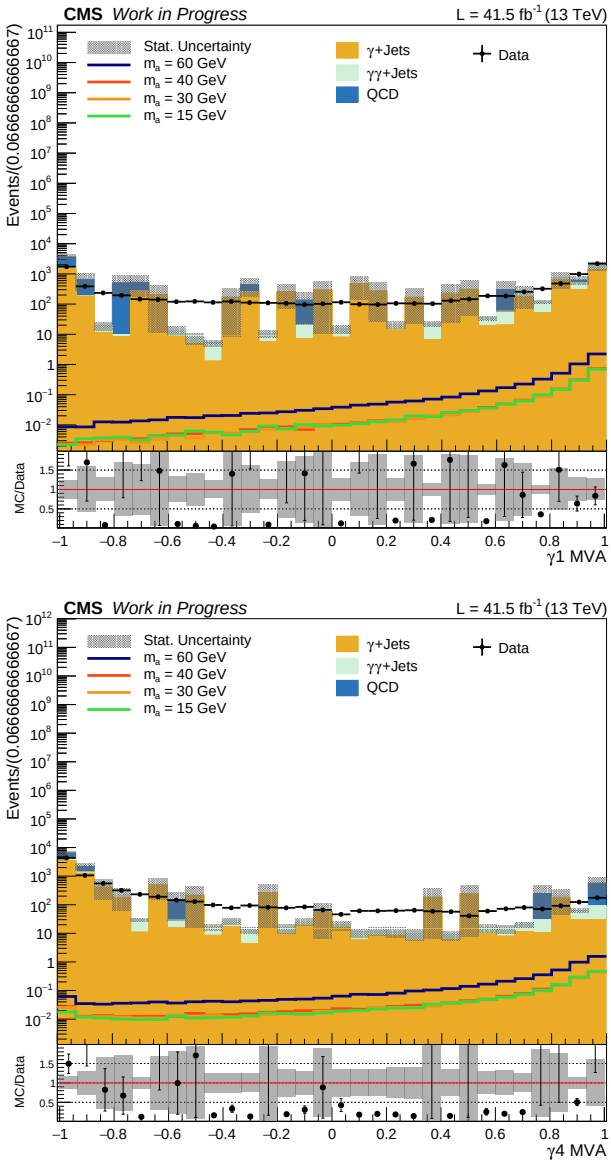


Figure 5.9: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2017 data and simulated samples. The data and background simulation belong to the sideband region.

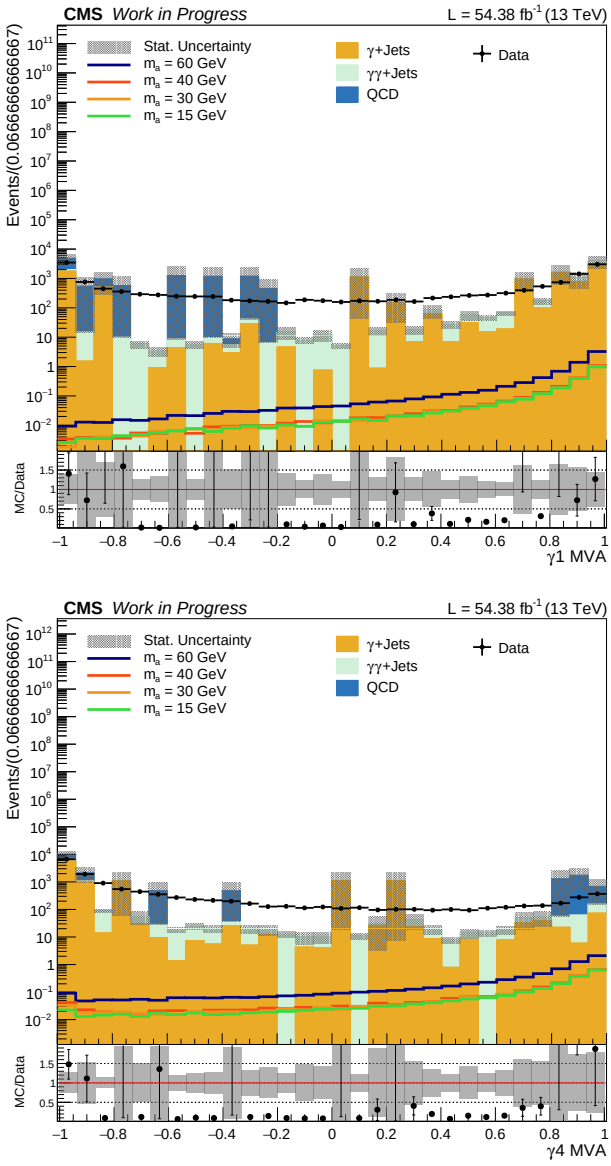


Figure 5.10: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2018 data and simulated samples. The data and background simulation belong to the sideband region.

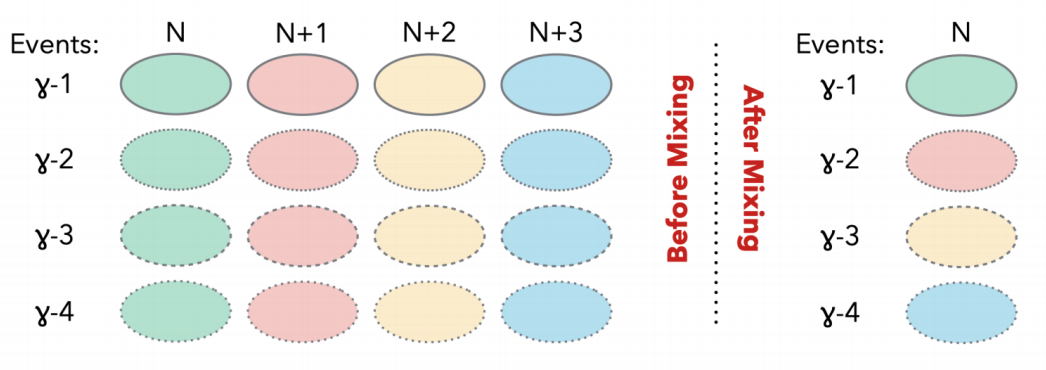


Figure 5.11: Schematic view of Event mixing: photons are shuffled between events to artificially create the background shape

As a solution to the poor description of background processes in the simulated samples, a data based background estimation method is devised for use in this analysis. The solution consists of obtaining a description of the background shapes by means of mixing photons between events. This method is referred to as event mixing, and involves replacing three out of the four photons in an event with photons from consecutive events. In other words, this procedure artificially creates the shape of the background by shuffling photons between events. A diagram of this scheme is presented in Figure 5.11.

The event mixing procedure can be summarized in the following steps:

- The preselection requirements on data are removed. This allows the production of a "mixed" data from real data that are unbiased.
- The second, third, and fourth photon in each event are replaced with photons from the next three consecutive events.
- Lastly, photon preselections are applied on the 'mixed' dataset.

This procedure provides a description of the background processes using a dataset of events that are similar to the original data, and are simultaneously insensitive to the presence of a possible signal.

The event mixing procedure is performed separately for 2016, 2017 and 2018. As a

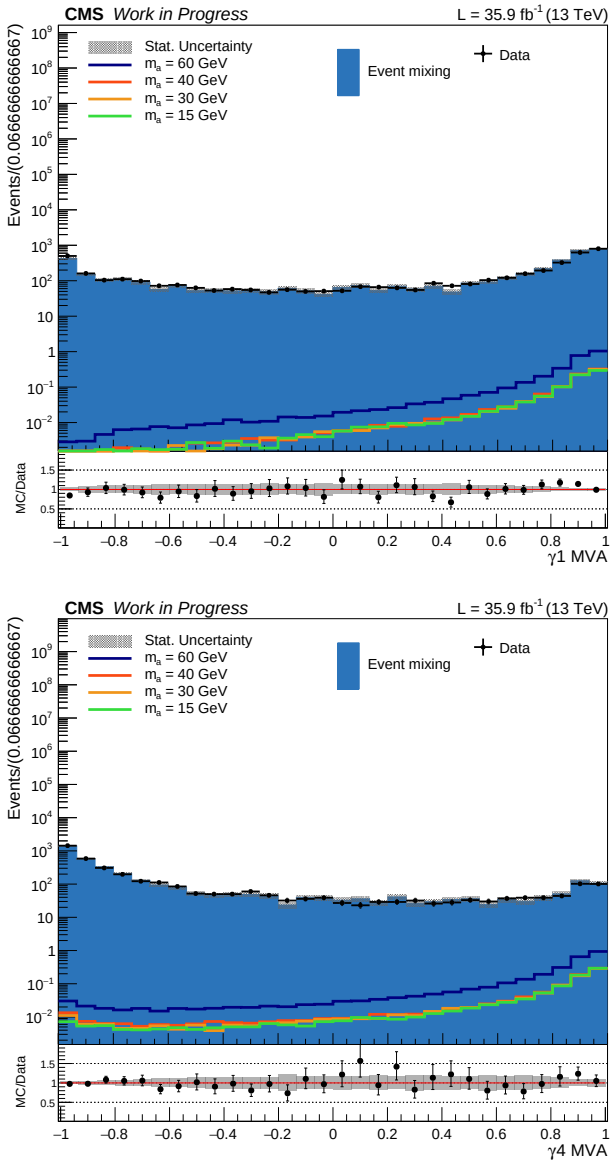


Figure 5.12: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2016 data and event mixing dataset. The data and event mixing dataset belong to the sideband region.

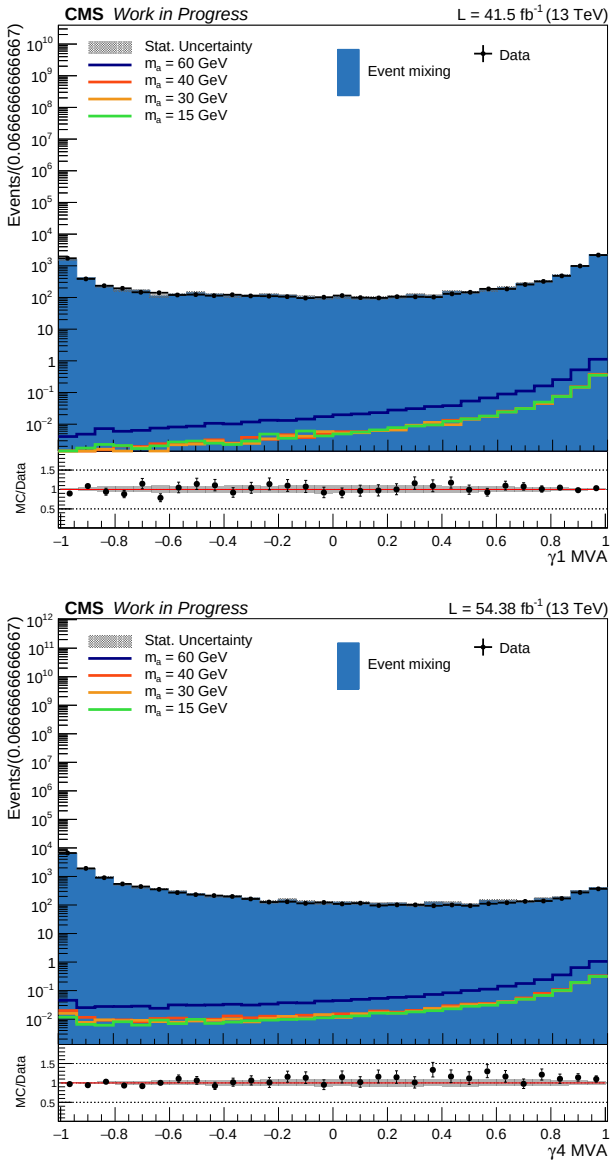


Figure 5.13: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2017 data and event mixing dataset. The data and event mixing dataset belong to the sideband region.

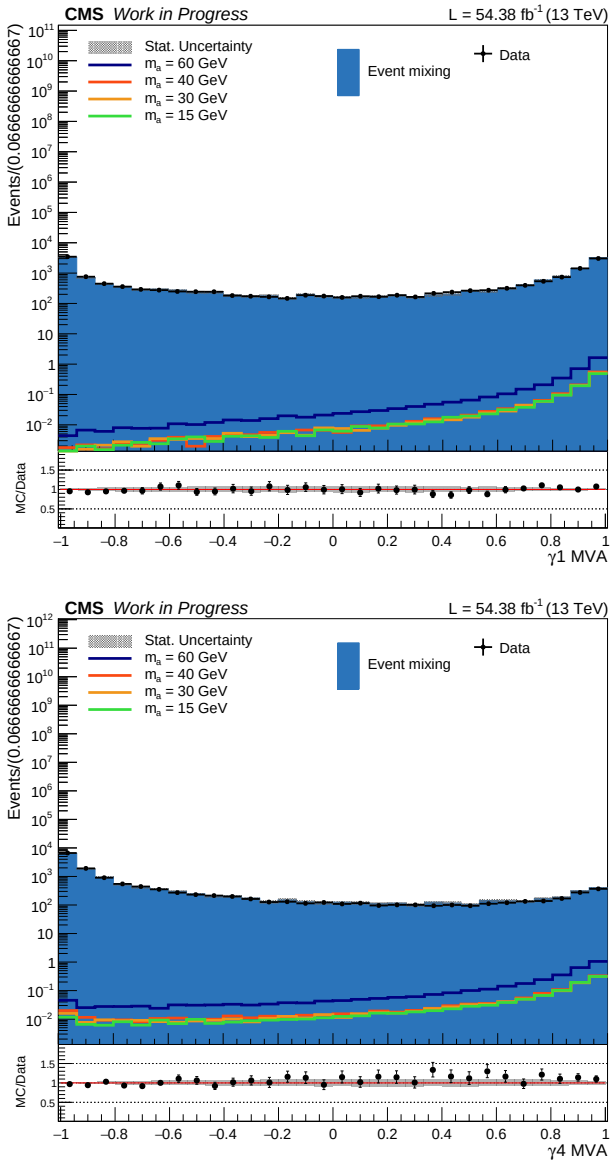


Figure 5.14: Photon ID MVA score distributions of the leading (top) and fourth (bottom) photon are shown for 2018 data and event mixing dataset. The data and event mixing dataset belong to the sideband region.

proof of concept, Figures 5.12, 5.13 and 5.14 show the distributions of MVA IDs of the leading and fourth photon, and the background contribution shown in these distributions is obtained from the event mixing dataset. A comparison of the level of agreement between the data and background in Figures 5.8 - 5.10 and Figures 5.12 - 5.14 serves as a check of the event mixing strategy's performance. The dataset obtained as a result of the event mixing procedure describes the shape of the background very well, and also results in an improved data-background agreement in all three years in the $m_{\gamma\gamma\gamma\gamma}$ sideband region.

This background estimation method is not designed to model background processes with at least two prompt photons ($\gamma\gamma$+jets). Extensive studies were performed which used the $\gamma\gamma$+jets simulated samples in addition with a data driven estimate of the γ+jets and QCD background processes. The data driven estimation, in these studies, involved creation of control regions by inverting selections on photon ID MVA scores. The overall lack of background statistics as a result of this method lead to poor categorization BDT performance and loss in the analysis sensitivity. Therefore, the event mixing procedure was deemed the best choice for background estimation. Future versions of the analysis can benefit greatly from production of dedicated background simulation samples.

5.5 Event categorization

A categorization BDT is trained to discriminate signal events, where four photons are produced from the decay of 125 GeV Higgs boson via two light pseudoscalars, from background events, where a four photon like experimental signture is seen from misidentified photons. The primary goal of designing this BDT is to assign a score to each event, that has a high value if the event is signal-like and a lower value if it is background-like.

Studies have been performed to identify variables that provide good discrimination between signal and background events. To train the BDT, simulated signal events corresponding to pseudoscalars ranging in mass from 15 to 60 GeV, in steps of 5 GeV are treated as signal. The dataset obtained as a result of event mixing procedure, described in Section 5.4 is treated as background. In order to maximize the BDT performance, simulated signal samples and event mixing dataset corresponding to the three years and added together, scaled by the appropriate luminosity for each year. To ensure that there is BDT output score is not over-trained and remains unbiased, the signal events is split into groups of odd and even (determined by their associated event number). Only odd events are considered while training the BDT, and the even events are used to construct the final signal models. All events are required to pass the analysis selection criteria. The input variables to the BDT are related to the kinematics of the four photons, and the pseudoscalars, as well as the identification variables related to the four photons.

In this analysis, simulated signal samples were only produced in steps of 5 GeV of the pseudoscalar mass (m_a). In order to provide final results in a finer granularity of m_a, the BDT is designed to be parametric. This is done by combining the simulated signal for all available m_a values when training the BDT, and by providing the generator level m_a information to the classifier. These two factors ensure that the output BDT score is parametric in m_a and allows for an interpolation of results.

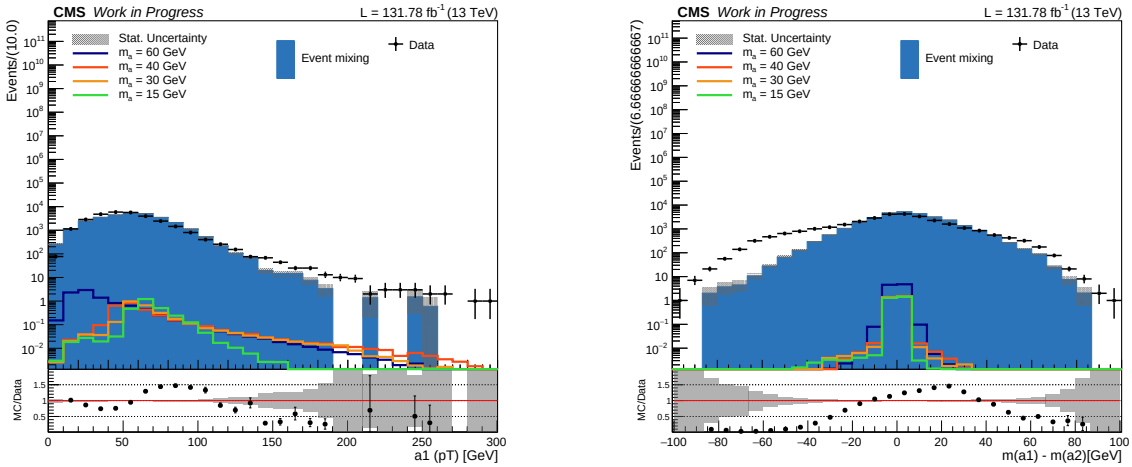
The list of input variables are as follows:

- Output score of the photon identification BDT for the four photons
- Transverse momentum of the two pseudoscalar candidates; p_T (a1), p_T (a2)
- Difference between the invariant mass of the two pseudoscalar candidates; m_{a1} - m_{a2}
- Difference between invariant mass of the pseudoscalar candidate at the reconstructed level and the generator level, divided by the $m_{\gamma\gamma\gamma\gamma}$ mass; (m_{a1} (Reco) - m_{a1} (Gen))/$m_{\gamma\gamma\gamma\gamma}$, ($m_{a2}$ (Reco) - m_{a2} (Gen))/$m_{\gamma\gamma\gamma\gamma}$
- Angular separation between the two pseudoscalar candidates divided by $m_{\gamma\gamma\gamma\gamma}$; $\Delta R(a1, a2)/m_{\gamma\gamma\gamma\gamma}$
- Cosine of the angle between the leading photon resulting from the decay of the leading pseudoscalar candidate and the direction of the $a \rightarrow\gamma\gamma$ decay; Cos $\theta_{a\gamma}$

Another motivation behind choosing these specific variables is their negligible correlation with the $m_{\gamma\gamma\gamma\gamma}$ variable, which is fitted to obtain final results in this analysis. This ensures that the trained BDT is not correlated with $m_{\gamma\gamma\gamma\gamma}$ and does not create any artificial signal like excess in the observed data.

Figure 5.14 shows an example of the distribution of two of the input variables, p_T (a1) (left) and m_{a1} - m_{a2} (right). These distributions indicate that the agreement between data and background are sub-optimal, which in turn could make the performance of the BDT sub-optimal. In order to improve the agreement in the data and background shape, an N-dimensional reweighting is performed. This is done by assigning a per-event weight that is computed using the ratio between background and data from the $m_{\gamma\gamma\gamma\gamma}$ sideband region. The variables used to compute this per-event weight are: Cos $\theta_{a\gamma}$, $\Delta R(a1, a2)$, p_T (a1), p_T (a2) and m_{a1} - m_{a2}. These per-event weights are computed separately for each year.

The effect of applying these event weights on the agreement between data and background is presented in Figure 5.14. The distributions of p_T (a1) (left) and m_{a1} - m_{a2} (right) variables, after the per-event weight is applied show a significant improvement in the agreement of the



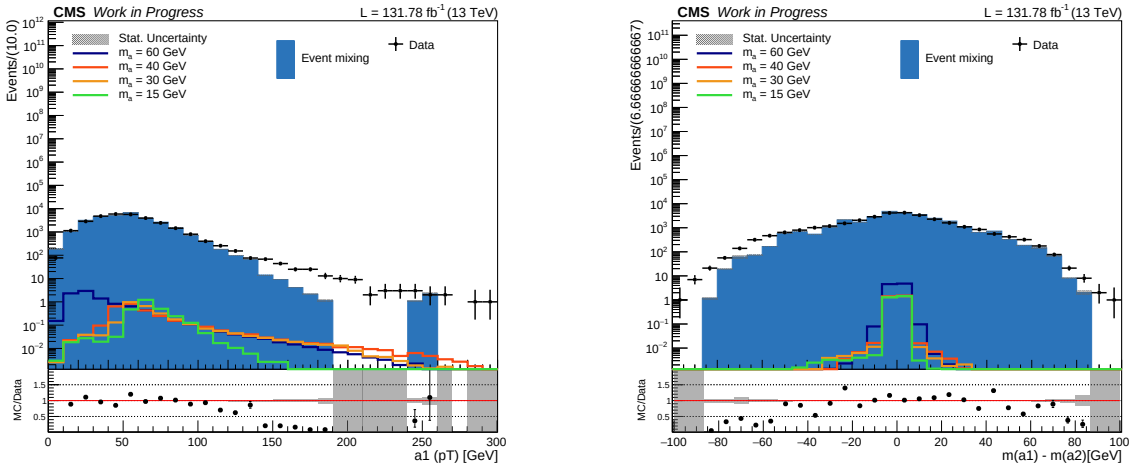
(a) p_T (a1) (b) m_{a1} - m_{a2}

Figure 5.14: Input variables p_T (a1) (left) and used m_{a1} - m_{a2} (right) in the training of the BDT. Contribution of event mixing dataset is shown in blue, and data are indicated by black points. The statistical uncertainty is shown by the grey band. The plots show data and simulated events corresponding to full run 2.

shapes between data and background, especially in the regions where most of events lie. This per-event weight is applied to the input variables prior to training the BDT.

The TMVA package is used to perform the BDT training, and the receiver operating characteristic (ROC) curve is used to gauge the performance of the training. The signal efficiency is plotted as a function of the background efficiency, for different values of selections applied on the BDT output score. A higher area under the curve corresponds to a high discrimination between signal and background. An area under the curve equal to 0.99 is obtained for this training.

The output BDT score is evaluated separately for each pseudoscalar mass point. This is done by setting the variable m_a (Gen) equal to the corresponding value. Figure 5.14 shows the output BDT score evalauted for m(a) = 15 GeV (left) and 50 GeV (right). The discrepancy between the output BDT scores of data and background have been traced to a difference in the correlations among the input variables between the two. This discrepancy is corrected by reweighting the background BDT score distribution to match data. It is worth



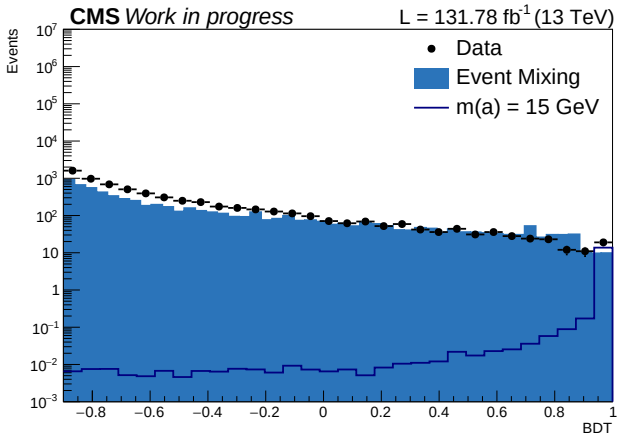
(a) p_T (a1) (b) m_{a1} - m_{a2}

Figure 5.14: Input variables p_T (a1) (left) and used m_{a1} - m_{a2} (right) in the training of the BDT, after the per-event weight is applied. Contribution of event mixing dataset is shown in blue, and data are indicated by black points. The statistical uncertainty is shown by the grey band. The plots show data and simulated events corresponding to full run 2. An improvement in the agreement between data and background is obtained as a result of applying the per-event weight.

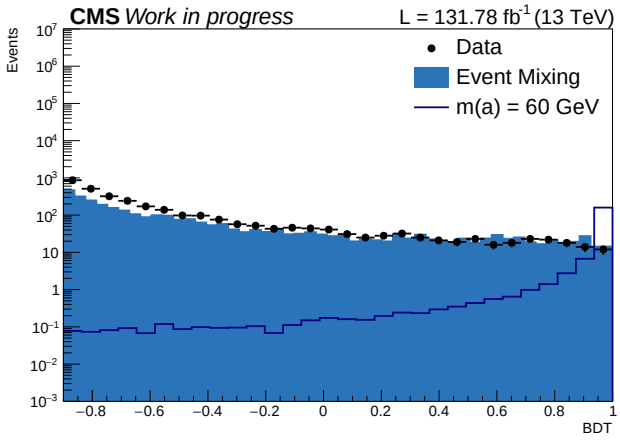
noting that this reweighting is only used for optimization of the analysis categories based on the BDT score, and not for evaluation of the final results. It is precisely for this reason that any uncertainty related to a remaining discrepancy, after reweighting, is not considered in the analysis since at most this will lead to a sub-optimal categorization of data events, and does not bias the final results in any form.

In order to maximize the sensitivity of the analysis, events are categorized according to the output of the BDT. This categorization is performed separately for each pseudoscalar mass point.

In the study of the event mixing procedure, a disagreement in the correlations between the input variables (used for the categorization BDT training)was observed between the event mixing dataset and data. It was also observed that the disagreement in correlations remains even after the application of per-event weight to the event mixing dataset. This discrepancy is the source of disagreement in the BDT scores between data and background



(a) m(a) = 15 GeV



(b) m(a) = 50 GeV

Figure 5.14: Score of the categorization BDT are shown for m(a) = 15 GeV (left) and 60 GeV (right). The black points represent the score evaluated on data, the blue histogram shows the score evaluated on event mixing dataset, and the blue line shows the score for the simulated signal.

(event mixing).

In order to correct for this disagreement, a reweighting of the BDT score is performed. The reweighting is performed using full Run 2 datasets, with datasets from each year scaled to appropriate luminosity. It should be mentioned here that the reweighting is only used for optimization of the analysis categories, and not for the evaluation of any final results. The BDT score distributions of the event mixing dataset and data from the sideband region are used to compute a per-event weight. This event weight is then applied to the event mixing dataset in both, the sideband as well as the signal region.

Once the categorization BDT score is obtained for all pseudoscalar masses, a category optimization procedure is performed for each pseudoscalar. This is done to define a BDT score boundary. The Approximate Mean Significance (AMS) is used as the figure of merit for this optimization, and its derivation is described in Ref. [73]. It is defined as:

$$AMS = \sqrt{2\left((S+B)\ln\left(1+\frac{S}{B}\right)-S\right)} \tag{5.4}$$

where S and B correspond to the number of signal and background events, respectively, in the $m_{\gamma\gamma\gamma\gamma}$ signal region defined as $115 < m_{\gamma\gamma\gamma\gamma} < 135$ GeV. For a fixed number of categories, the BDT score boundaries are optimized by maximizing the sum in quadrature of the AMS for each category. In addition, each category is required to have at least eight background events in the $m_{\gamma\gamma\gamma\gamma}$ sideband region. This requirement is imposed to ensure a robust background modeling.

For the purposes of category optimization, statistical fluctuations in the background distribution are minimized by a smoothing procedure. The 'SmoothSuper' method of TGraphSmooth is used for this procedure.

The categorization procedure has been performed as a function of up to five number of total categories. It was found that, for all pseudoscalar masses, a gain of less than 1% is observed in the total AMS when creating more than one category. Therefore, for all

pseudoscalar masses, only one analysis category is created. Events that do not fall in the chosen BDT based category are discarded.

Table 5.4 details the category boundary, the number of signal and background (event mixing) events in the signal region, as well the significance for each pseudoscalar mass. Since the BDT reweighting is evaluated from the sideband region and then applied to the background in signal region, a perfect data-background agreement in the signal region cannot be guaranteed. Therefore, some fluctuations in the BDT distribution of the background are expected and can explain the significance versus pseudoscalar mass trend seen in table 5.4.

m(a) [GeV]	Category	NSig	NBackground (Event mixing)	Significance
15	0.97, 1	5.06	4.97	1.99
20	0.96, 1	5.26	1.82	2.95
25	0.96, 1	5.19	4.00	2.21
30	0.93, 1	5.16	4.74	2.06
35	0.93, 1	5.16	4.74	2.06
40	0.96, 1	5.15	4.63	2.08
45	0.98, 1	5.44	3.02	2.55
50	0.94, 1	7.87	7.53	2.50
55	0.95, 1	12.05	8.180	3.54
60	0.95, 1	16.68	11.44	4.15

Table 5.4: Category boundaries, number of signal and background (event mixing) events in the signal region, and the significance of each category are shown for all pseudoscalars.

5.6 Selection Efficiency

As a summary of the various selections described in the previous sections, the cut flow of the analysis is as follows:

- Events containing at least one diphoton that passes trigger based pre-selections
- Events containing at least four photons
- Events containing at least four photons that pass the photon selections
- Events that pass the BDT score selection

Figure 5.15 shows the selection efficiency for all of the above stated selections as a function of the pseudoscalar mass. This efficiency is computed with respect to the total number of simulated signal events.

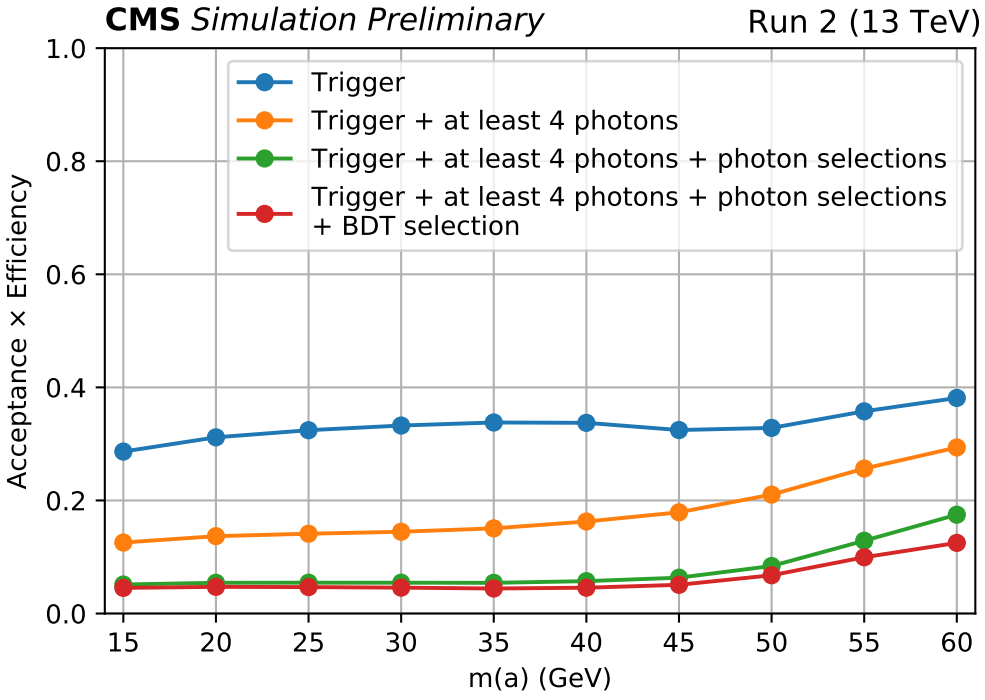


Figure 5.15: Signal efficiency of different selections applied in the analysis is shown as a function of pseudoscalar mass.

It is noteworthy that the efficiency at each step increases for pseudoscalar mass above 55 GeV. This is caused due to the diphoton mass requirement at the trigger level. Another drop in efficiency is caused by the p_T selections on the third and fourth photon. Future versions of this analysis can gain sensitivity by removal of the diphoton mass requirement in the online HLT paths, and by lowering the p_T requirements on the third and fourth photons.

5.7 Statistical Analysis of Data

The final results in this search are computed by a maxmimum likelihood fit of the signal and background models to the $m_{\gamma\gamma\gamma\gamma}$ distribution in the observed data. For each value of the parameter to be measured, the minimization of twice the negative log-likelihood (2NLL) is performed. As a result, the 2NLL curve is used to obtain the best-fit value and uncertainty on the parameter of interest. Statistical and systematic uncertainties are included in this fit. Systematic uncertainties are considered as nuisance parameters, where nuisance parameters are defined as the parameters that affect the fit but are not of interest for the final result. In this fitting procedure, the nuisance parameters are allowed to float freely and any deviation from their expected value leads to addition of a penalty term to the 2NLL. As a result, the width of the 2NLL curve increases and the overall uncertainty on the final result is accounted for.

Models of the signal and background $m_{\gamma\gamma\gamma\gamma}$ distributions are the two main ingredients for performing this fit. The signal model is built using simulation, with a separate model built for each pseudoscalar mass per year. The normalisation of the model is parametrized as a function of m_a. The background model is constructed using a data-driven technique [74] that makes use of various functional forms to represent the shape of the $m_{\gamma\gamma\gamma\gamma}$ distribution in the observed data.

Sections 5.7.1 and 5.7.2 detail the signal and background modeling procedure, respectively.

5.7.1 Signal Modeling

A separate signal model is constructed for each pseudoscalar mass using the corresponding simulated sample. Since data taking and reconstruction conditions vary per year, a separate signal model is created for 2016, 2017 and 2018.

The shape of the $m_{\gamma\gamma\gamma\gamma}$ distribution is modeled using a double-sided Crystal Ball (DCSB)

function, which is a modified version of the CB function with two independent exponential tails. It is defined as:

$$f\left(x;\mu,\sigma,\alpha_L,p_L,\alpha_R,p_R\right) = N\cdot\begin{cases} A_L\cdot\left(B_L-\frac{x-\mu}{\sigma}\right)^{-p_L}, & \text{for } \frac{x-\mu}{\sigma} > -\alpha_L \\ A_R\cdot\left(B_R+\frac{x-\mu}{\sigma}\right)^{-p_R}, & \text{for } \frac{x-\mu}{\sigma} > \alpha_R \\ e^{\frac{(x-\mu)^2}{\sigma^2}}, & \text{for } \frac{x-\mu}{\sigma} < -\alpha_L \text{ and } \frac{x-\mu}{\sigma} > \alpha_R \end{cases} \tag{5.5}$$

where

$$\begin{aligned} A_k &= \left(\frac{p_k}{|\alpha_k|}\right)^{p_k}\cdot e^{-\frac{\alpha^2}{2}}, \\ B_k &= \frac{p_k}{|\alpha_k|} - |\alpha_k| \end{aligned} \tag{5.6}$$

and L, R refer to the left and right tail, respectively.

The DCSB function has been determined to be a good choice to model the signal shape in this analysis. The parameters of the DCSB functions are determined by performing a fit to the simulated $m_{\gamma\gamma\gamma\gamma}$ distribution.

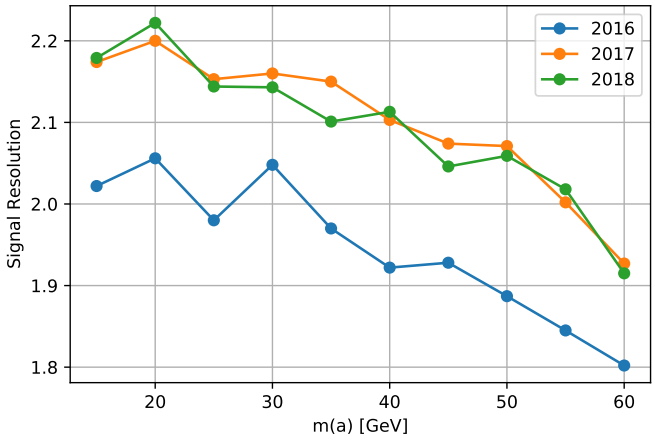


Figure 5.16: Resolution of the signal model is shown as a function of the pseudoscalar mass for the three years. This plot indicates that the resolution is expected to change by $\approx 1\%$ over 2 GeV.

For this analysis, the simulated signal samples were generated with a granularity of 5 GeV in m_a. In order to provide final results for intermediate pseudoscalar masses, an interpolation strategy is devised. To construct signal model at the intermediate masses, it

is first verified that the signal model parameters do not change significantly as a function of m_a. Figure 5.16 depicts the variation in signal resolution as a function of m_a, and illustrates that the resolution of the signal model is expected to change by only $\approx 1\%$ over 2 GeV. Other parameters of the signal model are also found to vary negligibly as a function of m_a, and therefore are not varied for signal model construction at the intermediate masses.

After a signal model has been created for each pseudoscalar mass, the normalization is determined by setting $\sigma(pp \to h)$ equal to 1 fb, and $\mathrm{BR}(h \to aa \to \gamma\gamma\gamma\gamma)$ equal to 1. Then, using the ratio of the total number of expected events obtained after all analysis selections are applied to the total number of generated events, the product of detector efficiency and analysis acceptance is modeled as a linear function of m_a. This variation is illustrated in Figure 5.17.

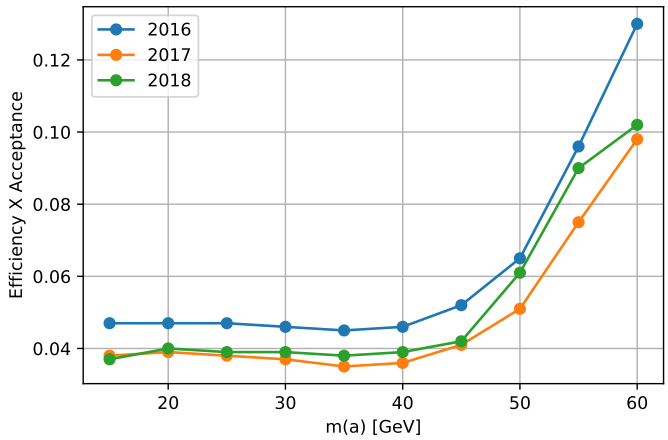
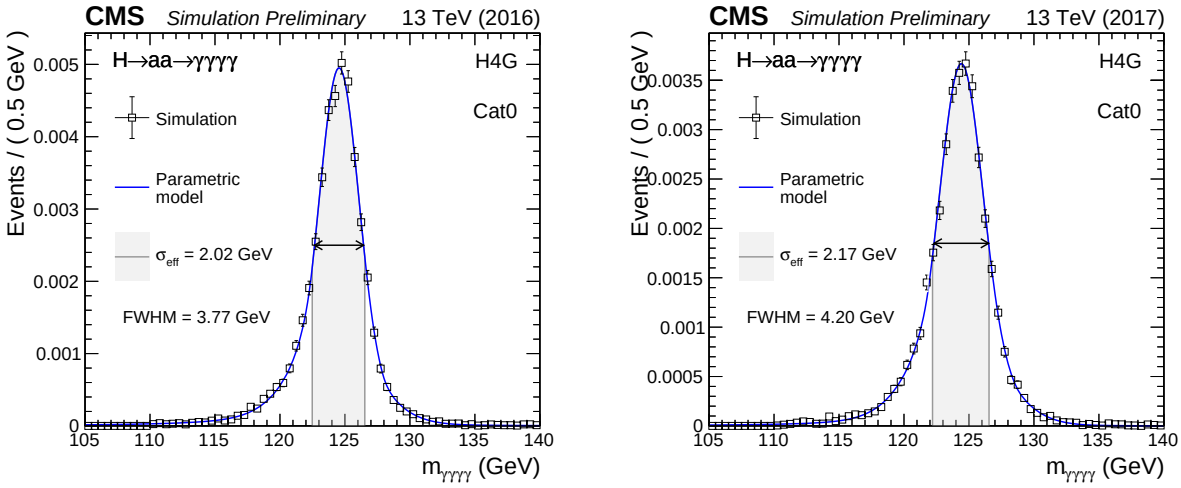


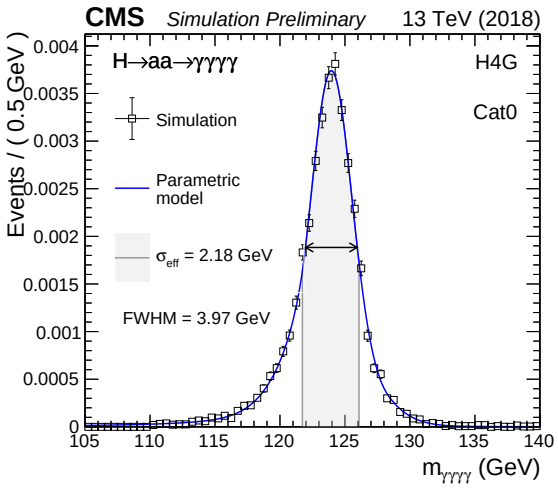
Figure 5.17: Efficiency times acceptance values are plotted as a function of the pseudoscalar mass for the three years. These values are used to obtain the efficiency times acceptance value, and to build a signal model, at the intermediate pseudoscalar mass points.

Lastly, this information allow creation of signal models for pseudoscalars ranging in mass from 15 to 60 GeV, with a granularity of 1 GeV.



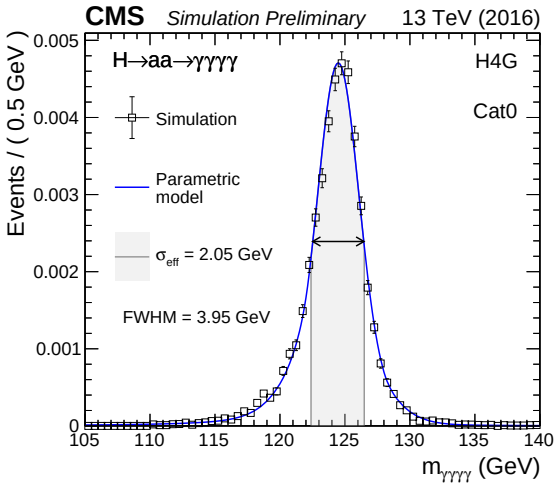
(a) 2016

(b) 2017

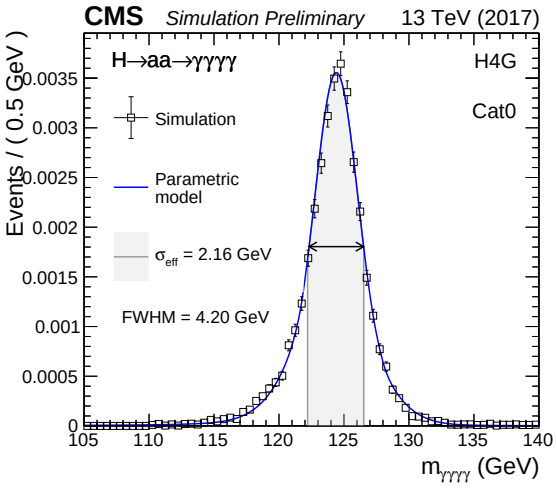


(c) 2018

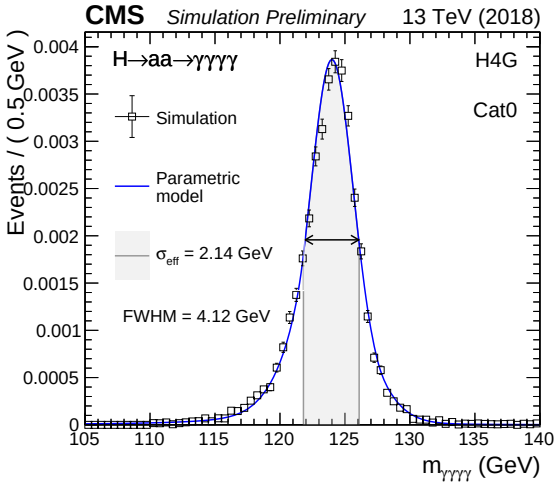
Figure 5.18: Signal shapes for $m_{\gamma\gamma\gamma\gamma}$ are shown for m(a) = 15 GeV. The open squares represent the simulated events and the blue line corresponds to the signal model. The σ_{eff} value, half the width of the smallest interval containing 68% of the $m_{\gamma\gamma\gamma\gamma}$ distribution, is also shown.



(a) 2016

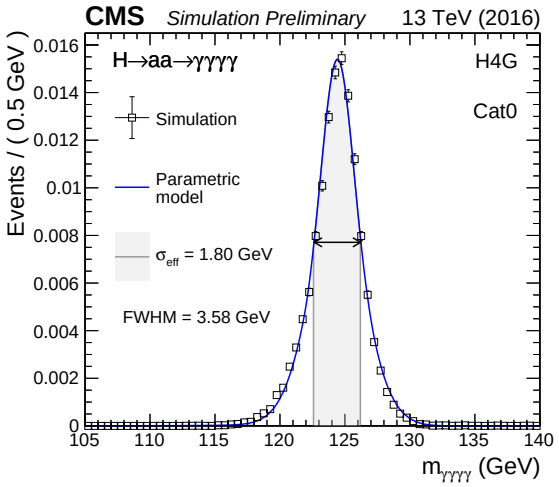


(b) 2017

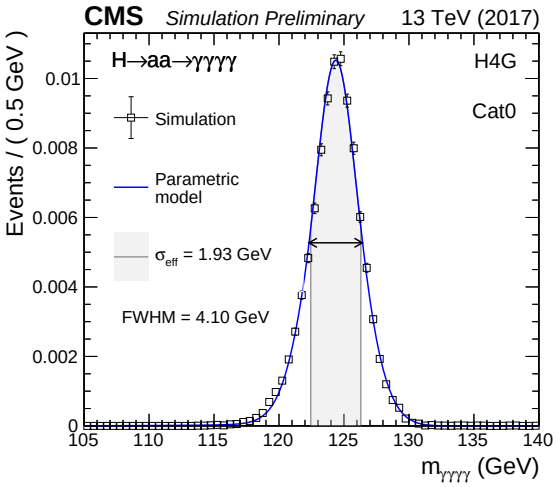


(c) 2018

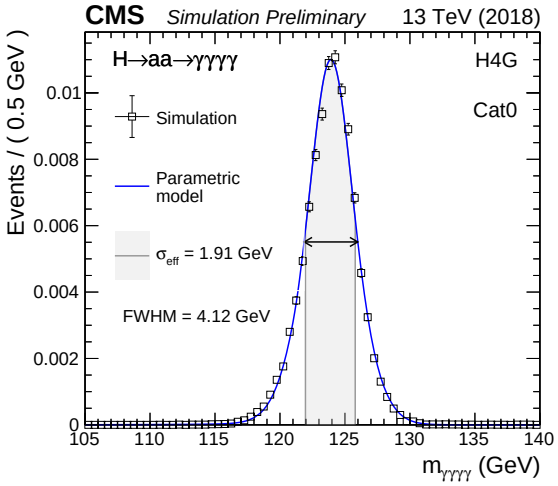
Figure 5.19: SSignal shapes for $m_{\gamma\gamma\gamma\gamma}$ are shown for m(a) = 30 GeV. The open squares represent the simulated events and the blue line corresponds to the signal model. The σ_{eff} value, half the width of the smallest interval containing 68% of the $m_{\gamma\gamma\gamma\gamma}$ distribution, is also shown.



(a) 2016



(b) 2017



(c) 2018

Figure 5.20: Signal shapes for $m_{\gamma\gamma\gamma\gamma}$ are shown for m(a) = 60 GeV. The open squares represent the simulated events and the blue line corresponds to the signal model. The σ_{eff} value, half the width of the smallest interval containing 68% of the $m_{\gamma\gamma\gamma\gamma}$ distribution, is also shown.

5.7.2 Background Modeling

The background model is built to describe the shape of the $m_{\gamma\gamma\gamma\gamma}$ distribution that results from background processes. Since the shape of this distribution is not known, different functional forms must be considered in the construction of the background model. The choice of function can result in different number of estimated events under the signal peak, and as a result affect the measured signal strength. This inherent uncertainty associated with the choice of the function needs to be accounted for in the final result. The discrete profile method, which applies an uncertainty on the background model by treating the choice of function used to model the background as a discrete nuisance parameter, is used in this analysis. It is described in reference [74].

The choice of the background function is considered to be a nuisance parameter, since it affects the measured signal strength but is not of any interest itself. In this analysis, four different families of functions that represent the $m_{\gamma\gamma\gamma\gamma}$ distribution in data are considered. For an N-parameter function with parameters p_0, p_1,....., p_N, the families of functions considered are:

- Sum of exponential functions: $f_N(x) = \sum_{i=0}^{N} p_{2i} \exp\left(p_{2i+1} x\right)$
- Sum of power law functions: $f_N(x) = \sum_{i=0}^{N} p_{2i} x^{-p_{2i+1}}$
- Bernstein polynomials: $f_N(x) = \sum_{i=0}^{N} p_i \begin{pmatrix} N \\ i \end{pmatrix} x^i (1-x)^{N-i}$
- Laurent series: $f_N(x) = \sum_{i=0}^{N} p_i x^{-4+L(i)}$, where $L(i) = \sum_{j=0}^{i} (-1)^j j$

A subset of functions from each family are used to build the background model. For each family, a test is first performed to fix the order. For each order, a likelihood fit is performed and a penalty value equal to the number of parameters in the function is added to the value of 2NLL. The order of the function is then increased until a minimum threshold of goodness of fit is reached. An F-test is performed for each order, which helps compute the improvement

in fit quality. As part of this test, the difference in 2NLL values is distributed as a χ^2 and a p-value is computed. If the p-value is found to lie below a certain threshold, the higher order function is considered suitable and is added to the set of considered functions. This process is repeated for until the lowest order suitable function is found for each family of functions.

In this analysis, the events from the full run 2 dataset (2016, 2017, and 2018) are combined to create a single background model. A separate background model for each value of pseudoscalar mass.

5.7.2.1 Bias Studies

As part of the background modeling procedure, after representative functions are chosen via the F-test, it must be ensured that the chosen functional form does not bias a possible signal strength measurement in the analysis. This can happen because the real background shape that is being fitted might not exactly be the functional form that is chosen. In this analysis, checking that the bias is less than 0.14, when the discrete profiling method is allowed to pick the optimal background fit function, is sufficient to establish that the chosen functional form does not introduce a bias in the signal strength measurement.

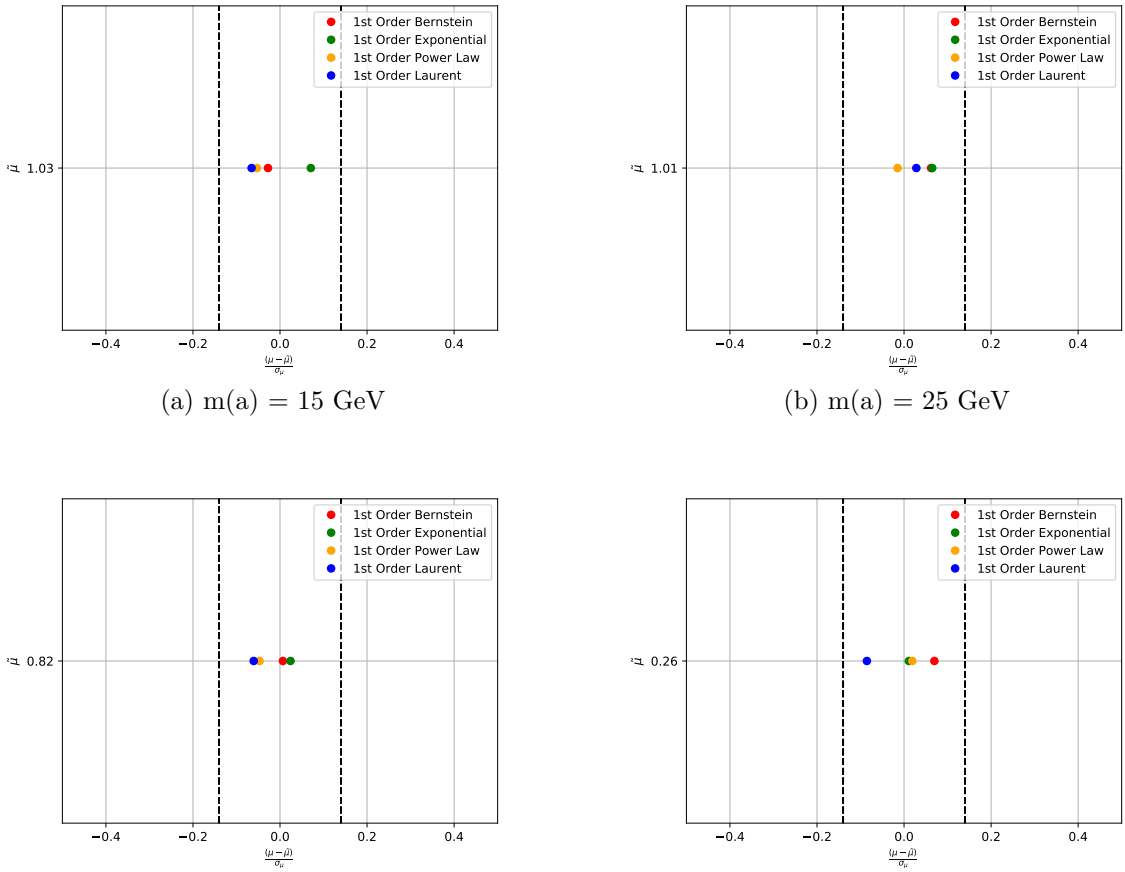
Bias is defined as the mean of the pull distribution, where the pull for an individual toy distribution is defined as

$$Pull = \frac{\mu - \tilde{\mu}}{\sigma}$$

where μ is defined as the generated value of the signal strength, $\tilde{\mu}$ is the fitted value of the signal strength and σ is the positive error on μ if $\tilde{\mu} > \mu$ and is the negative error if $\tilde{\mu} < \mu$.

After performing the F-test, the following functions are considered to model the background shape: 1st order bernstein polynomial, 1st order exponential, 1st order power law and 1st order laurent series. 2000 toys are generated for each choice of the background function, and then fitted using the discrete profiling method.

Since a unique background model is created for each pseudoscalar mass point, the bias study is performed separately for each mass point. For each pseudoscalar mass point, a pull value is evaluated by setting μ equal to the expected value of the signal strength.



(a) m(a) = 15 GeV

(b) m(a) = 25 GeV

(c) m(a) = 46 GeV

(d) m(a) = 60 GeV

Figure 5.21: Mean pull (x-axis) versus the generated signal strength (y-axis) plots are shown for m(a) = 15, 25, 46, 60 GeV. Different colors indicate the function used for generating the toy distributions. The envelope method was used for fitting the toy distributions. Vertical lines indicate a bias of -0.14 and 0.14, and for all pseudoscalar masses, the bias value is found to lie within these boundaries.

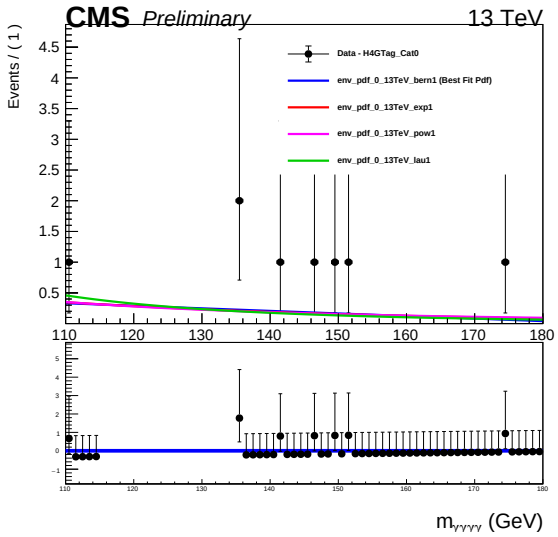
The results of the mean pull versus the generated signal strength are shown in figure 5.21 for a few representative pseudoscalar masses. The vertical lines in the plots indicate a mean pull of -0.14 and 0.14. When the toys are fitted using the envelope method, the mean of the pull distribution lies within the two vertical lines, i.e the bias is less than 0.14. Similar

results are seen for all pseudoscalar mass points, for toys generated using other background functions.

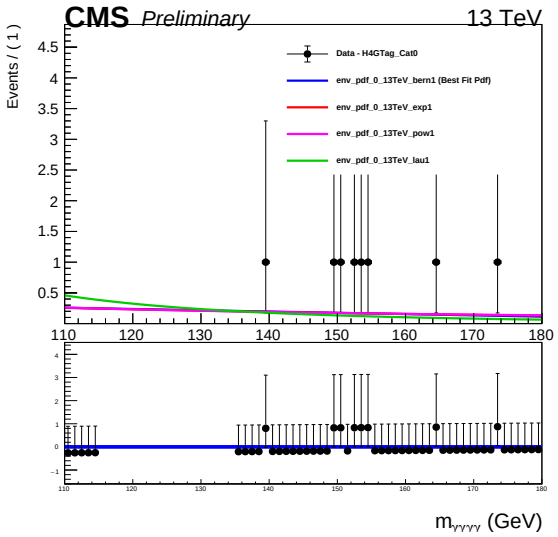
These bias studies establish that when the discrete profiling method is allowed to pick the optimal background fit function, the variation in the fit uncertainty is less than 1%, and therefore no additional related systematic uncertainty needs to be assigned to the background model.

The different functions chosen in construction of the background model are shown in Figure 5.22 for m_a values equal to 15, 30 and 60 GeV. In these figures, the data are blinded, meaning the data points are not shown in the signal region, $115 < m_{\gamma\gamma\gamma\gamma} < 135$ GeV.

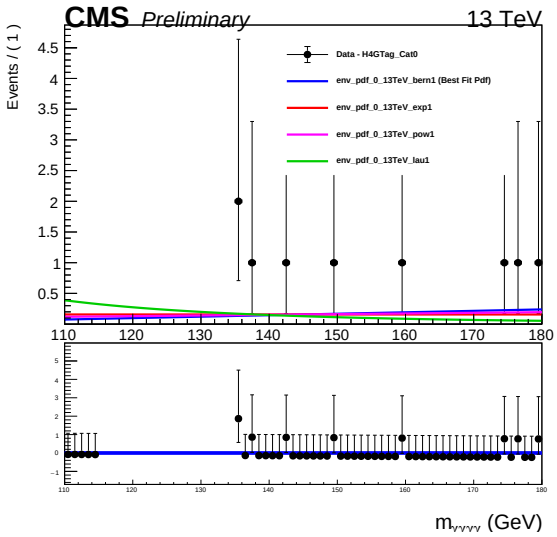
The background function in the final fit is chosen from the set of candidate functions, and parameters of each function are allowed to float in the fit. A penalty term, corresponding to each parameter in the chosen function, is added to the 2NLL.



(a) m(a) = 15 GeV



(b) m(a) = 30 GeV



(c) m(a) = 60 GeV

Figure 5.22: The set of candidate functions used for background modeling are shown for different pseudoscalar masses. The data are blinded in the region 115-135 GeV.

5.8 Systematic Uncertainties

In this analysis, the discrete profiling method is used to estimate the systematic uncertainty associated with the data-driven background estimation.

The systematic uncertainties which affect the signal model are of two kinds:

- Uncertainties which modify the shape of the distribution. They are built in to the signal model as nuisance parameters. These uncertainties are typically experimental in nature, are related to the energy of the photons, and affect the mean and width of the signal model.

- Uncertainties that leave the shape of the distribution unchanged are treated as log-normal variations in the event yield.

No theoretical uncertainties are considered in this analysis.

The various uncertainties considered in this analysis which affect the shape of the distribution are as follows:

- Photon energy scale and resolution: Corrections are applied to the photon energy scale in data, and to the energy resolution in simulation. The uncertainties related to these corrections are computed using $Z \to ee$ events.

- Non-linearity of the photon energy scale: Any remaining differences in the linearity of the photon energy scale between data and simulation are covered by this uncertainty.

- Shower shape corrections: This uncertainty is associated with the imperfect modeling of shower shapes in simulation, and is estimated by comparing the energy scale before and after any corrections are applied to the shower shape variables.

- Modeling of electromagnetic showers: This uncertainty is related to the electromagnetic showering process and its impact on the signal model is estimated by changing the

model which is used to generate bremsstrahlung energy spectrum created by an electron or photon.

- Non-uniformity of light collection: Within a given ECAL crystal, there is an uncertainty associated with the modeling of the light collection as a function of the emission depth. This uncertainty is estimated by comparing simulation with the longitudinal shower profile estimates.

- Modeling of material in front of the ECAL: The behavior of electromagnetic showers is affected by the amount of material present in front of the ECAL. This behavior may not be well modeled in simulation, and thus special samples with variations in amount of upstream material are used to compute the impact on the photon energy scale.

The uncertainties which only affect the event yield are as follows:

- Integrated luminosity: Estimated to be 2.5% (2016) [75], 2.3% (2017) [76], and 2.5% (2018) [77], and follows the recommendations from the CMS Lumi group. They are treated as uncorrelated across the three years.

- Photon identification BDT score: The output of the photon ID BDT, is different between data and simulation. The uncertainties that arise due to this difference is estimated by requiring the systematic variations to cover these discrepancies. The resulting uncertainty on the signal yield is estimated by propagating this uncertainty through all the analysis selections. The average size of this uncertainty is 1%.

- Trigger efficiency: Trigger selection efficiency is computed with the $Z \to ee$ events using the tag and probe technique, and has an average size of 1% in this analysis.

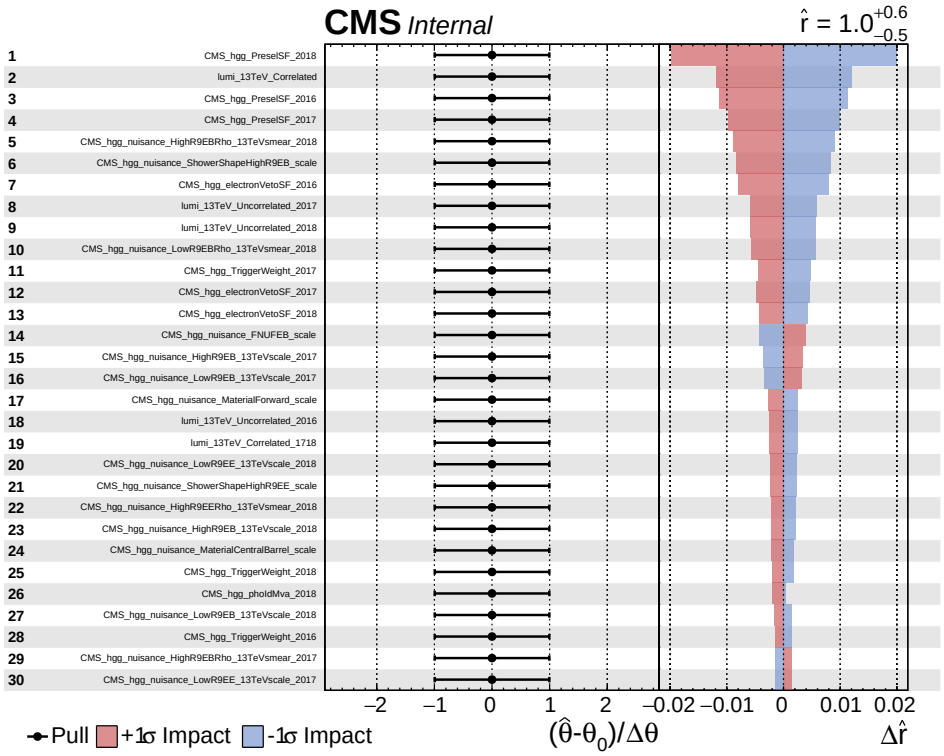
- Photon preselection: The ratio between the preselection efficiency measured in data and simulation is used to compute the uncertainty on the preselection efficiency, and has an average size of 1% in this analysis.

In this analysis, the signal modeling is performed independently for 2016, 2017 and 2018. By that means, the uncertainties affecting each year can be treated to be correlated or uncorrelated. In case of a correlated uncertainty, there is only one associated nuisance parameter in the fit and it affects the signal model corresponding to the three years. On the other hand, for uncorrelated uncertainties, a separate nuisance parameter is present in the fit for each year. The impact of an uncorrelated uncertainty is independent for each year.

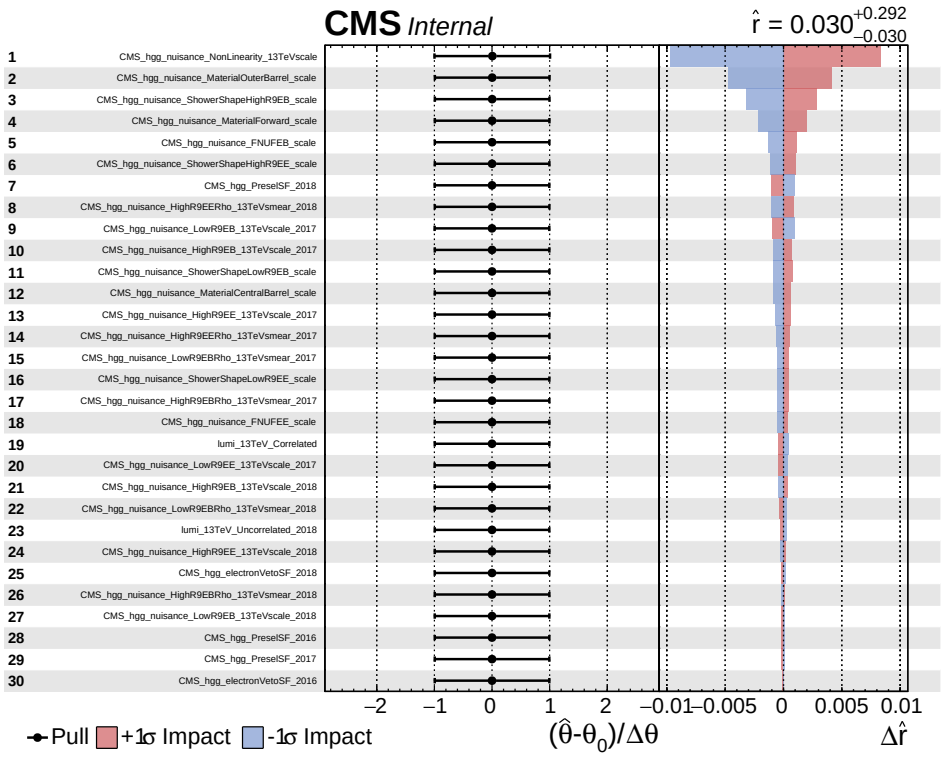
Since both, the data taking and reconstruction conditions vary between the years, the experimental uncertainties are treated as uncorrelated between the years. It is important to note here that the dominant source of uncertainty in this analysis is statistical in nature.

It is also useful to study the effects of these uncertainties on the signal strength, μ. In other words, this is referred to as computation of the 'impact' of each uncertainty. For a nuisance parameter θ, the impact on μ is defined as the shift $\Delta\mu$ that is obtained as θ is fixed and brought to its $\pm 1\ \sigma$ post-fit values, while keeping all the other nuisance parameters fixed in the fit. This provides a handle on understanding of the correlations between the nuisance paramaters and the parameter of interest, and also helps in determining the nuisance parameter with the highest effect on the uncertainty of μ.

This analysis makes use of the CMS Higgs Combine tool to compute the impacts. The expected and observed impacts on μ are evaluated for each pseudoscalar mass value. In order to calculate the expected impacts, it is assumed that the observed data perfectly matches the prediction, and the observed data in the signal region are used to calculate the observed impacts. Figure 5.23 compares the impact of the nuisance parameters in the expected (top) and observed (bottom) case. The nuisance parameters are ranked by their impact on the signal strength in a decreasing order.



(a) Expected impacts



(b) Observed impacts

Figure 5.23: The impact of nuisance parameters on the signal strength are ranked in a decreasing order and shown for the expected (top) and observed (bottom) scenarios. These impacts correspond to m(a) = 22 GeV, and similar impact calculation is done for all masses.

5.9 Results

Limits are calculated at 95% confidence level (CL) using the CL_S prescription [73; 78], by means of the CMS Combine tool. In this method of evaluating limits, the expected signal and background contributions to the signal region are used to construct a likelihood. The statistical and systematic uncertainties are used to create the nuisance parameters in the fit. The signal strength parameter, μ is left free to float during the fit to maximize the likelihood. Various iterations of fits are performed at discrete μ values, as a result of which a corresponding relative likelihood value is obtained. The relative likelihood can then be used to determine the probability of a background only hypothesis, p_b, by setting μ equal to 0. Similarly, the probability of a signal plus background hypothesis, p_{s+b} can be obtained by setting a non-zero value of μ. Typically, this is done by setting μ equal to 1. With this information in hand, the confidence limit can be constructed as:

$$CL_S = \frac{p_{s+b}}{1 - p_b} \tag{5.7}$$

The signal plus background hypothesis can be excluded at 95% if the CL_S value falls below 5%. In order to calculate the expected exclusion limits, it is assumed that the observed data perfectly matches the prediction. 1 and 2 σ uncertainty bands are obtained on this expected limit by variation of the nuisance parameters that induce a change in the likelihood. On the other hand, the observed limit makes use of the observed data in the signal region.

On unblinding the signal region in this analysis, no significant deviations from the SM expectations are seen. The unblinded $m_{\gamma\gamma\gamma\gamma}$ distributions corresponding to all values of pseudoscalar masses being probed in this analysis can be seen in Appendix B.

The results of this analysis are then statistically interpreted to set model-independent exclusion limits on $\sigma(pp \to h) \times BR(h \to aa \to \gamma\gamma\gamma\gamma)$. Figure 5.24 shows the result of this limit setting procedure as a function of pseudoscalar mass. In the figure, the black dashed lines indicate the median expected limit values and the $\pm$ 1 σ (2σ) uncertainty on these

values are indicated by the green (yellow) bands. The solid black lines represent observed limits that were calculated using the observed data. Across the pseudoscalar mass range, the observed limits are in agreement with the expected limits within two standard deviations of the expected limits, and no significant excess of events is seen.

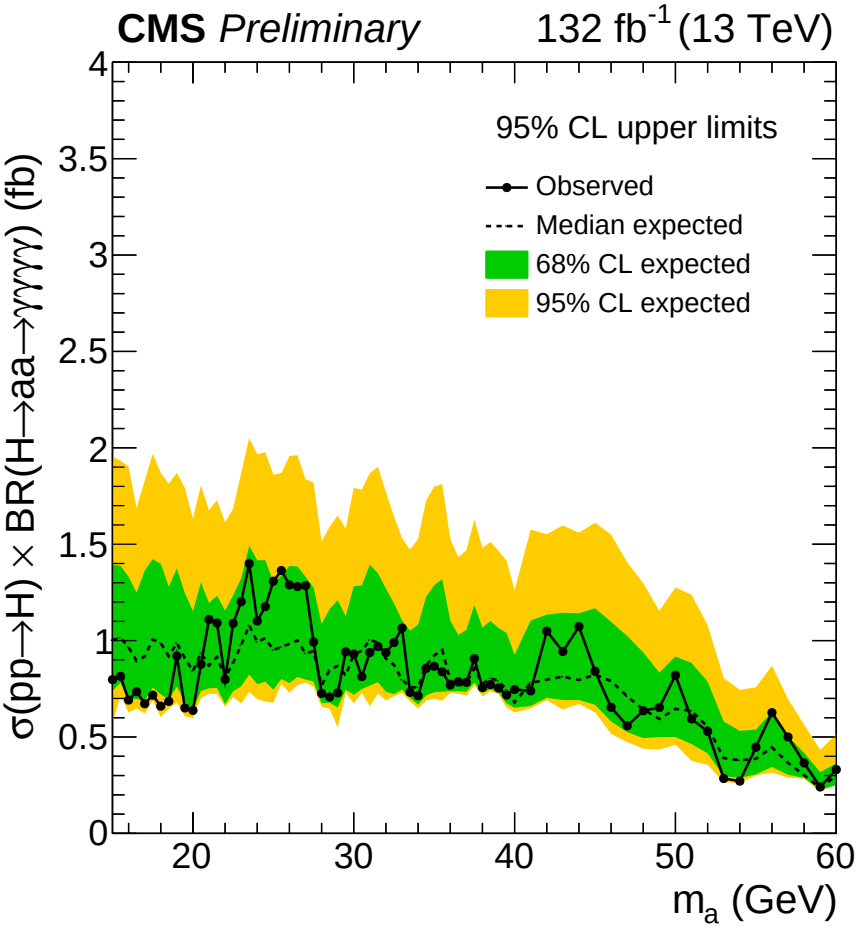


Figure 5.24: Model independent limits on $\sigma(pp \to h) \times BR(h \to aa \to \gamma\gamma\gamma\gamma)$ are presented in units of fb, as a function of the pseudoscalar mass. The black dashed lines represent the expected limits and the green (yellow) bands indicate the $\pm$ 1 σ (2σ) uncertainty on the expected values. The solid black lines show the observed limits, and are seen to be in agreement with the expected limits within the uncertainties.

These results, the first in the four photon final state by the CMS experiment, are found to be consistent with the Standard Model.

Chapter 6

Conclusions and Outlook

This thesis describes a search for exotic decay of the 125 GeV Higgs Boson to a pair of light pseudoscalars, with each pseudoscalar decaying to two photons each. Data used correspond to 132 fb^{-1} of $\sqrt{s} = 13\,\text{TeV}$ pp-collisions, collected by the CMS experiment in 2016-2018. Events containing four photon candidates are selected, and pseudoscalars ranging in mass from 15 to 60 GeV are probed. The signal is modeled by fitting a parametric function to the simulated events, and a data driven background estimation method is employed. The M($\gamma\gamma\gamma\gamma$) distribution is used as the fitting variable.

Limits are placed on the production cross-section of the Higgs Boson times the branching fraction of the Higgs to four photons via two light pseudoscalars. The expected and observed limits agree within two standard deviations of the expected limits, and no excess of events over the Standard Model prediction is seen across the pseudoscalar mass range.

This search is the first of its kind to be performed at CMS and can benefit from various improvements in the future. In particular, since the photons in the final state have p_T values lower than that for typical analyses at CMS, the sensitivity of this search can be improved by lowering the p_T thresholds at the Level-1 and/or the High Level Trigger stage of the CMS trigger system. Furthermore, removal of the selection on the invariant diphoton mass at the HLT level will increase the efficiency of selected events which, and also make the background modeling even more robust. Additionally, the development and employment of three photon triggers will also help enhance the signal efficiency and purity of this particular final state.

The range of pseudoscalar masses being probed can be further widened in the future versions of the analysis. In particular, for pseudoscalars with masses lying between 5 to 35 GeV, where the final state is dominated by three reconstructed photons, development of novel analysis techniques for reconstructing the Higgs mass using three photons will be advantageous and an interesting analysis to pursue.

CMS is not scheduled to collect more data until May 2022, and it will take several years to double the size of the Run 2 CMS dataset. In the meantime, there are some improvements that can be made to this analysis using the currently available Run 2 dataset. One of the most valuable improvement would be in the reconstruction of the pseudoscalar candidate. The invariant mass of the pseudoscalar provides a strong handle on discriminating the signal events from the background, and more sophisticated machine learning techniques which utilize the information related to the final state photons can improve the resolution of the pseudoscalar mass.

The performance of the CMS ECAL is a vital ingredient of this analysis. The excellent photon momentum and energy resolution provided by the ECAL during run 2 has direct implications on the sensitivity of this analysis. During Run 3 and beyond, the various planned upgrades of ECAL will prove to be crucial for this search.

Furthermore, statistically limited analyses such as the one described in this thesis, that provide a clean experimental signature, will continue to be important for the search of new physics in the upcoming eras of the LHC.

Finally, it is worth noting that the LHC has produced the largest high energy dataset available to date, which makes it a great machine to study rare and exotic processes that can lead us to new physics. When combining the currently available Higgs Boson measurements, 34% of its branching ratio is still unconstrained. Therefore, direct searches of the Higgs that are not predicted by the Standard Model remain pivotal to the LHC's physics goals.

The exotic Higgs decays program is phenomenologically and experimentally rich, and might just lead us to the elusive signs of new physics beyond the Standard Model.

Bibliography

[1] Commons W. 2020. File:standard model of elementary particles modified version.svg — wikimedia commons, the free media repository. [Online; accessed 5-April-2021].

[2] Olive K. 2014 Review of particle physics. *Chinese Physics C* **38**, 9, 090001. (doi:10.1088/1674-1137/38/9/090001).

[3] Aad G, Abbott B, Abdallah J, Abdinov O, Abeloos B, Aben R, AbouZeid OS, Abraham NL, Abramowicz H, et al. 2016 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. (doi:10.1007/jhep08(2016)045).

[4] Curtin D, Essig R, Gori S, Jaiswal P, Katz A, Liu T, Liu Z, McKeen D, Shelton J, Strassler M, *et al.* 2014 Exotic decays of the 125 gev higgs boson. *Physical Review D* **90**, 7. (doi:10.1103/physrevd.90.075004).

[5] Del Rosso A. 2017 Aerial view of the LHC and the four major experiments General Photo.

[6] Lefèvre C. 2008 The CERN accelerator complex. Complexe des accélérateurs du CERN.

[7] Team A. 1999 Diagram of an LHC dipole magnet. Schéma d'un aimant dipôle du LHC.

[8] CMS Public Lumi Results;
https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults .

[9] Sakuma T, McCauley T. 2014 Detector and Event Visualization with SketchUp at the CMS Experiment. *J. Phys. Conf. Ser.* **513**, 022032. (doi:10.1088/1742-6596/513/2/022032).

[10] Chatrchyan S, *et al.* 2008 The CMS Experiment at the CERN LHC. *JINST* **3**, S08004. (doi:10.1088/1748-0221/3/08/S08004).

[11] service photo. 2002 Fish-eye view of yoke. Vue oeil de poisson de la culasse. CMS Collection.

[12] Collaboration TC. 2014 Description and performance of track and primary-vertex reconstruction with the cms tracker. *Journal of Instrumentation* **9**, 10, P10009–P10009. (doi:10.1088/1748-0221/9/10/p10009).

[13] 2012 CMS Technical Design Report for the Pixel Detector Upgrade (doi:10.2172/1151650).

[14] Brice M. 2006 First half of CMS inner tracker barrel.

[15] Brice M, Vaillet L, Lazic L. 2008 Images of the CMS HCAL Barrel (HB). CMS Collection.

[16] Chatrchyan S, Khachatryan V, Sirunyan A, Tumasyan A, Adam W, Aguilo E, Bergauer T, Dragicevic M, Erö J, Fabjan C, *et al.* 2012 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. (doi:10.1016/j.physletb.2012.08.021).

[17] Sirunyan A, Tumasyan A, Adam W, Ambrogi F, Asilar E, Bergauer T, Brandstetter J, Brondolin E, Dragicevic M, Erö J, *et al.* 2018 Performance of the cms muon

detector and muon reconstruction with proton-proton collisions at s=13 tev. *Journal of Instrumentation* **13**, 06, P06015–P06015. (doi:10.1088/1748-0221/13/06/p06015).

[18] Hoch M. 2007 Muon Chambers. CMS Collection.

[19] Paolucci P, Hadjiiska R, Litov L, Pavlov B, Petkov P, Dimitrov A, Beernaert K, Cimmino A, Costantini S, Guillaume G, *et al.* 2013 Cms resistive plate chamber overview, from the present system to the upgrade phase i. *Journal of Instrumentation* **8**, P04005. (doi:10.1088/1748-0221/8/04/P04005).

[20] Barney D. 2015 CMS Slice .

[21] 1997 The CMS electromagnetic calorimeter project: Technical Design Report .

[22] Brice M. 2008 Images of the CMS ECAL Barrel (EB). CMS Collection.

[23] Sirunyan A, Tumasyan A, Adam W, Ambrogi F, Bergauer T, Dragicevic M, Erö J, Valle AED, Frühwirth R, Jeitler M, *et al.* 2020 Reconstruction of signal amplitudes in the cms electromagnetic calorimeter in the presence of overlapping proton-proton interactions. *Journal of Instrumentation* **15**, 10, P10002–P10002. (doi:10.1088/1748-0221/15/10/p10002).

[24] 2019 CMS ECAL Response to Laser Light .

[25] 2019 ECAL 2018 refined calibration and Preshower Run 2 performance plots .

[26] 2016 CMS ECAL first results with 2016 data .

[27] 2018 CMS ECAL with 2017 data .

[28] LHC Higgs Cross Section Working Group. 2016 Handbook of LHC Higgs cross sections: 4. deciphering the nature of the Higgs sector. *CERN* (doi:10.23731/CYRM-2017-002).

[29] Glashow SL. 1961 Partial-symmetries of weak interactions. *Nuclear Physics* **22**, 4, 579–588. (doi:https://doi.org/10.1016/0029-5582(61)90469-2).

[30] Salam A. 1962 Renormalizability of gauge theories. *Phys. Rev.* **127**, 331–334. (doi:10.1103/PhysRev.127.331).

[31] Weinberg S. 1967 A model of leptons. *Phys. Rev. Lett.* **19**, 1264–1266. (doi:10.1103/PhysRevLett.19.1264).

[32] Aaij R, Adeva B, Adinolfi M, Affolder A, Ajaltouni Z, Akar S, Albrecht J, Alessio F, Alexander M, Ali S, *et al.* 2015 Observation of $j/\psi p$ resonances consistent with pentaquark states in $\Lambda_b^0 \to j/\psi K^- p$ decays. *Phys. Rev. Lett.* **115**, 072001. (doi:10.1103/PhysRevLett.115.072001).

[33] Collaboration C. 2012 Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.* **B716**, 30–61. (doi:10.1016/j.physletb.2012.08.021).

[34] Collaboration A. 2012 Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.* **B716**, 1–29. (doi:10.1016/j.physletb.2012.08.020).

[35] Zwicky F. 1933 Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta* **6**, 110–127.

[36] Hinshaw G, Weiland JL, Hill RS, Odegard N, Larson D, Bennett CL, Dunkley J, Gold B, Greason MR, Jarosik N, *et al.* 2009 Five-year wilkinson microwave anisotropy probe observations: Data processing, sky maps, and basic results. *The Astrophysical Journal Supplement Series* **180**, 2, 225–245. (doi:10.1088/0067-0049/180/2/225).

[37] Clowe D, Bradač M, Gonzalez AH, Markevitch M, Randall SW, Jones C, Zaritsky D. 2006 A direct empirical proof of the existence of dark matter. *The Astrophysical Journal* **648**, 2, L109–L113. (doi:10.1086/508162).

[38] Wess J, Zumino B. 1974 Supergauge transformations in four dimensions. *Nuclear Physics B* **70**, 1, 39–50. (doi:https://doi.org/10.1016/0550-3213(74)90355-1).

[39] Gervais JL, Sakita B. 1971 Field theory interpretation of supergauges in dual models. *Nuclear Physics B* **34**, 2, 632–639. (doi:https://doi.org/10.1016/0550-3213(71)90351-8).

[40] Ramond P. 1971 Dual theory for free fermions. *Phys. Rev. D* **3**, 2415–2418. (doi:10.1103/PhysRevD.3.2415).

[41] Albornoz Vásquez D, Bélanger G, Godbole RM, Pukhov A. 2012 Higgs boson in the mssm in light of the lhc. *Phys. Rev. D* **85**, 115013. (doi:10.1103/PhysRevD.85.115013).

[42] Ellwanger U, Hugonie C, Teixeira AM. 2010 The next-to-minimal supersymmetric standard model. *Physics Reports* **496**, 1-2, 1–77. (doi:10.1016/j.physrep.2010.07.001).

[43] Branco G, Ferreira P, Lavoura L, Rebelo M, Sher M, Silva JP. 2012 Theory and phenomenology of two-higgs-doublet models. *Physics Reports* **516**, 1-2, 1–102. (doi:10.1016/j.physrep.2012.02.002).

[44] Haber HE, Kane GL. 1985 The Search for Supersymmetry: Probing Physics Beyond the Standard Model. *Phys. Rept.* **117**, 75–263. (doi:10.1016/0370-1573(85)90051-1).

[45] Kim JE. 1987 Light Pseudoscalars, Particle Physics and Cosmology. *Phys. Rept.* **150**, 1–177. (doi:10.1016/0370-1573(87)90017-2).

[46] Peccei RD, Quinn HR. 1977 CP Conservation in the Presence of Instantons. *Phys. Rev. Lett.* **38**, 1440–1443. (doi:10.1103/PhysRevLett.38.1440).

[47] Evans L, Bryant P. 2008 LHC machine. *Journal of Instrumentation* **3**, 08, S08001–S08001. (doi:10.1088/1748-0221/3/08/s08001).

[48] 1984 *LEP design report.* Geneva: CERN. Copies shelved as reports in LEP, PS and SPS libraries.

[49] Collaboration TA, Aad G, Abat E, Abdallah J, Abdelalim AA, Abdesselam A, Abdinov O, Abi BA, Abolins M, Abramowicz H, *et al.* 2008 The ATLAS experiment at the CERN large hadron collider. *Journal of Instrumentation* **3**, 08, S08003–S08003. (doi:10.1088/1748-0221/3/08/s08003).

[50] Alves AAugusto J, *et al.* 2008 The LHCb Detector at the LHC. *JINST* **3**, S08005. (doi:10.1088/1748-0221/3/08/S08005).

[51] Collaboration TA, Aamodt K, Quintana AA, Achenbach R, Acounis S, Adamová D, Adler C, Aggarwal M, Agnese F, Rinella GA, *et al.* 2008 The ALICE experiment at the CERN LHC. *Journal of Instrumentation* **3**, 08, S08002–S08002. (doi:10.1088/1748-0221/3/08/s08002).

[52] van der Meer S. 1968 Calibration of the effective beam height in the ISR. Tech. Rep. CERN-ISR-PO-68-31. ISR-PO-68-31, CERN, Geneva.

[53] 2009 Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET .

[54] 2010 Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV. Tech. rep., CERN, Geneva.

[55] 2015 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. (doi:10.1088/1748-0221/10/06/p06005).

[56] Adzic P, *et al.* 2006 Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter. *Eur. Phys. J. C* **46S1**, 23–35. (doi:10.1140/epjcd/s2006-02-002-x).

[57] Di Marco E. 2014 CMS electromagnetic calorimeter calibration and timing performance during LHC Run I and future prospects. Tech. rep., CERN, Geneva.

[58] Anfreville M, *et al.* 2008 Laser monitoring system for the CMS lead tungstate crystal calorimeter. *Nucl. Instrum. Meth. A* **594**, 292–320. (doi:10.1016/j.nima.2008.01.104).

[59] Chatrchyan S, *et al.* 2013 Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in *pp* Collisions at $\sqrt{s} = 7$ TeV. *JINST* **8**, P09009. (doi:10.1088/1748-0221/8/09/P09009).

[60] 2015 *Journal of Instrumentation* **10**, 08, P08010–P08010. (doi:10.1088/1748-0221/10/08/p08010).

[61] 2015 *Journal of Instrumentation* **10**, 06, P06005–P06005. (doi:10.1088/1748-0221/10/06/p06005).

[62] Sirunyan AM, *et al.* 2020 A measurement of the Higgs boson mass in the diphoton decay channel. *Phys. Lett. B* **805**, 135425. (doi:10.1016/j.physletb.2020.135425).

[63] Meschi E, Monteiro T, Seez C, Vikas P. 2001 Electron Reconstruction in the CMS Electromagnetic Calorimeter. Tech. rep., CERN, Geneva.

[64] 2013 Cms ecal alignment in the lhc run. CMS AN 2013/328.

[65] 2007 Position resolution at the 2006 ecal testbeam. CMS DM 2007/011.

[66] Collaboration C. 2017 The Phase-2 Upgrade of the CMS Endcap Calorimeter. Tech. Rep. CERN-LHCC-2017-023. CMS-TDR-019, CERN, Geneva. Technical Design Report of the endcap calorimeter for the Phase-2 upgrade of the CMS experiment, in view of the HL-LHC run.

[67] Collaboration C. 2017 The Phase-2 Upgrade of the CMS Barrel Calorimeters. Tech. Rep. CERN-LHCC-2017-011. CMS-TDR-015, CERN, Geneva. This is the final version, approved by the LHCC.

[68] Alwall J, Frederix R, Frixione S, Hirschi V, Maltoni F, Mattelaer O, Shao HS, Stelzer T, Torrielli P, Zaro M. 2014 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. (doi:10.1007/jhep07(2014)079).

[69] Khachatryan V, *et al.* 2016 Event generator tunes obtained from underlying event and multiparton scattering measurements. *Eur. Phys. J. C* **76**, 155. (doi:10.1140/epjc/s10052-016-3988-x).

[70] Sirunyan AM, *et al.* 2020 Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements. *Eur. Phys. J. C* **80**, 4. (doi:10.1140/epjc/s10052-019-7499-4).

[71] Gleisberg T, Höche S, Krauss F, Schönherr M, Schumann S, Siegert F, Winter J. 2009 Event generation with sherpa 1.1. *Journal of High Energy Physics* **2009**, 02, 007–007. (doi:10.1088/1126-6708/2009/02/007).

[72] Agostinelli S, *et al.* 2003 GEANT4 — a simulation toolkit. *Nucl. Instrum. Meth. A* **506**, 250. (doi:10.1016/S0168-9002(03)01368-8).

[73] Cowan G, Cranmer K, Gross E, Vitells O. 2011 Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C* **71**, 2. (doi:10.1140/epjc/s10052-011-1554-0).

[74] Dauncey P, Kenzie M, Wardle N, Davies G. 2015 Handling uncertainties in background shapes: the discrete profiling method. *Journal of Instrumentation* **10**, 04, P04015–P04015. (doi:10.1088/1748-0221/10/04/p04015).

[75] 2017 CMS Luminosity Measurements for the 2016 Data Taking Period. Tech. rep., CERN, Geneva.

[76] 2018 CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV. Tech. rep., CERN, Geneva.

[77] 2019 CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV. Tech. Rep. CMS-PAS-LUM-18-002, CERN, Geneva.

[78] Read AL. 2002 Presentation of search results: theCLstechnique. *Journal of Physics G: Nuclear and Particle Physics* **28**, 10, 2693–2704. (doi:10.1088/0954-3899/28/10/313).

APPENDIX A

Technical details related to data and simulated samples

Samples	Luminosity [fb^{-1}]
/DoubleEG/Run2016B-17Jul2018_ver1-v1/MINIAOD	
/DoubleEG/Run2016B-17Jul2018_ver2-v1/MINIAOD	
/DoubleEG/Run2016C-17Jul2018_v1/MINIAOD	
/DoubleEG/Run2016D-17Jul2018_v1/MINIAOD	35.9
/DoubleEG/Run2016E-17Jul2018_v1/MINIAOD	
/DoubleEG/Run2016F-17Jul2018_v1/MINIAOD	
/DoubleEG/Run2016G-17Jul2018_v1/MINIAOD	
/DoubleEG/Run2016H-17Jul2018_v1/MINIAOD	
/DoubleEG/Run2017B-31Mar2018_v1/MINIAOD	
/DoubleEG/Run2017C-31Mar2018_v1/MINIAOD	
/DoubleEG/Run2017D-31Mar2018_v1/MINIAOD	41.5
/DoubleEG/Run2017E-31Mar2018_v1/MINIAOD	
/DoubleEG/Run2017F-31Mar2018_v1/MINIAOD	
/EGamma/Run2018A-17Sep2018_v2/MINIAOD	
/EGamma/Run2018B-17Sep2018_v2/MINIAOD	54.4
/EGamma/Run2018C-17Sep2018_v2/MINIAOD	
/EGamma/Run2018D-17Sep2018_v2/MINIAOD	

Table A.1: Data MiniAOD samples used in the Run 2 analysis.

Dataset Name
/HAHMHToAA_AToGG_MA-15GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-20GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-25GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-30GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-35GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-40GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-45GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-50GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-55GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1
/HAHMHToAA_AToGG_MA-60GeV_TuneCUETP8M1_PSweights_13TeV-madgraph_pythia8/RunIISummer16MiniAODv3-94X_mcRun2_asymptotic_v3-v1

Table A.2: Simulated 2016 signal sample names

Dataset Name
/HAHMHToAA_AToGG_MA-15GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-20GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-25GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-30GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-35GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-40GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-45GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-50GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-55GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1
/HAHMHToAA_AToGG_MA-60GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1

Table A.3: Simulated 2017 signal sample names

Dataset Name
/HAHMHToAA_AToGG_MA-15GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-20GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-25GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-30GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-35GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-40GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-45GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-50GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-55GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/HAHMHToAA_AToGG_MA-60GeV_TuneCP5_PSweights_13TeV-madgraph_pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/

Table A.4: Simulated 2018 signal sample names

Dataset Name
/DiPhotonJetsBox_M40_80-Sherpa/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v1/
/DiPhotonJetsBox_MGG-80toInf_13TeV-Sherpa/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/
/QCD_Pt-30to40_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/
/QCD_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/
/GJet_Pt-20to40_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv3-PUMoriond17_backup_94X_mcRun2_asymptotic_v3-
/GJet_Pt-20toInf_DoubleEMEnriched_MGG-40to80_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/
/GJet_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/

Table A.5: Simulated 2016 background sample names

Dataset Name
/DiPhotonJetsBox_M40_80-Sherpa/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/
/DiPhotonJetsBox_MGG-80toInf_13TeV-Sherpa/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/
/QCD_Pt-30toInf_DoubleEMEnriched_MGG-40to80_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/
/QCD_Pt-30to40_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/
/QCD_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/
/GJet_Pt-20toInf_DoubleEMEnriched_MGG-40to80_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/
/GJet_Pt-20to40_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/
/GJet_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/

Table A.6: Simulated 2017 background sample names

Dataset Name
/DiPhotonJetsBox_M40_80-Sherpa/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/
/DiPhotonJetsBox_MGG-80toInf_13TeV-Sherpa/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/
/QCD_Pt-30toInf_DoubleEMEnriched_MGG-40to80_TuneCP5_13TeV_Pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v2/
/QCD_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/
/GJet_Pt-20toInf_DoubleEMEnriched_MGG-40to80_TuneCP5_13TeV_Pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15_ext1-v2/
/GJet_Pt-20to40_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15_ext1-v2/
/GJet_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCP5_13TeV_Pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/

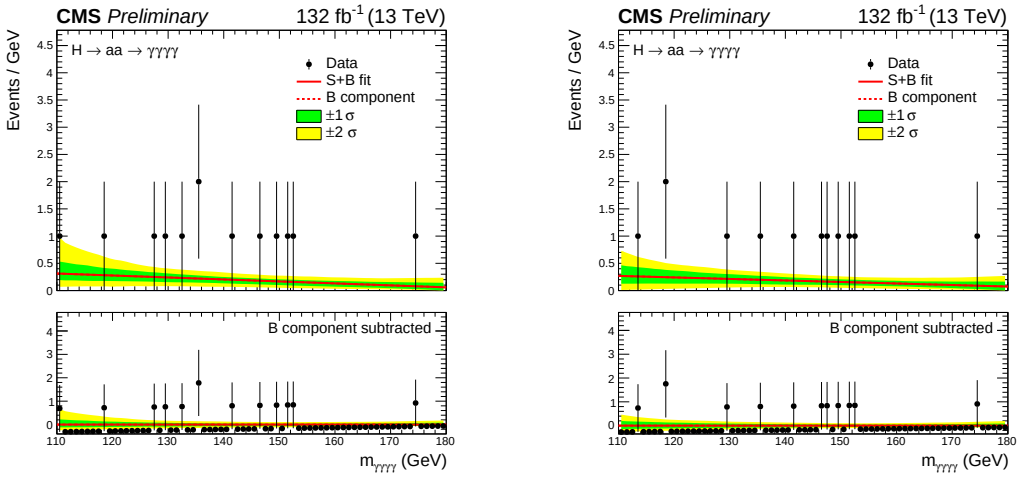
Table A.7: Simulated 2018 background sample names

HLT Path	Year
HLT_Diphoton30EB_18EB_R9Id_OR_IsoCaloId_AND_HE_R9Id_DoublePixelVeto_Mass55_v7,	
HLT_Diphoton30PV_18PV_R9Id_AND_IsoCaloId_AND_HE_R9Id_DoublePixelVeto_Mass55_v7	2016
HLT_Diphoton30PV_18PV_R9Id_AND_IsoCaloId_AND_HE_R9Id_PixelVeto_Mass55	2017
HLT_Diphoton30_18_R9IdL_AND_HE_AND_IsoCaloId_NoPixelVeto	2018

Table A.8: HLT paths used in the Run 2 analysis.

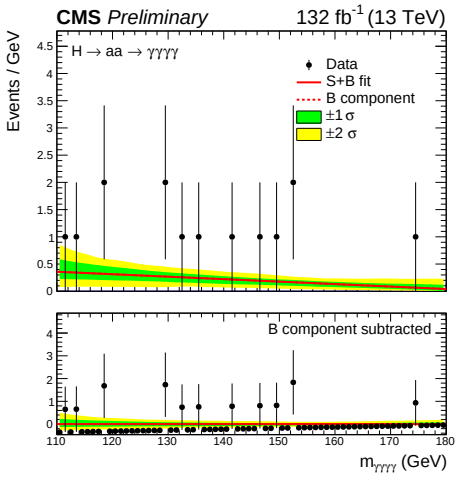
APPENDIX B

Observed four-photon invariant mass distributions

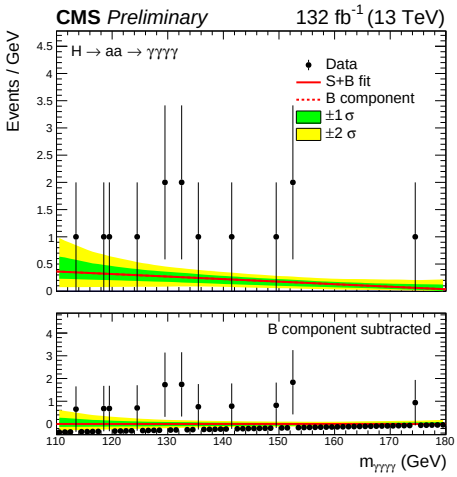


(a) m(a) = 16 GeV (b) m(a) = 17 GeV

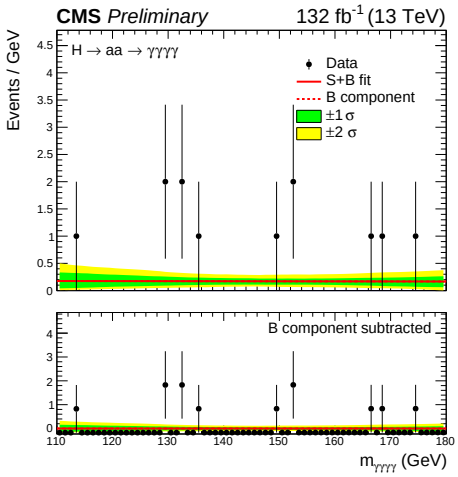
Figure B.1: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



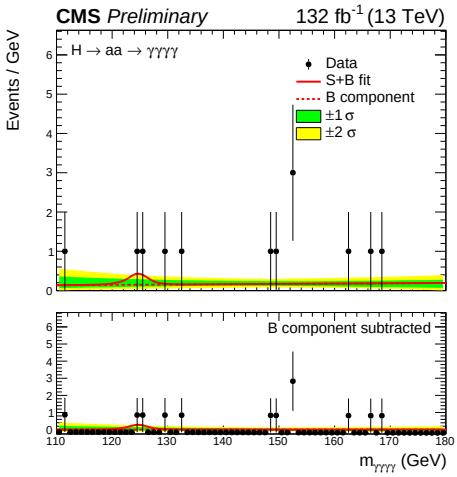
(a) m(a) = 18 GeV



(b) m(a) = 19 GeV

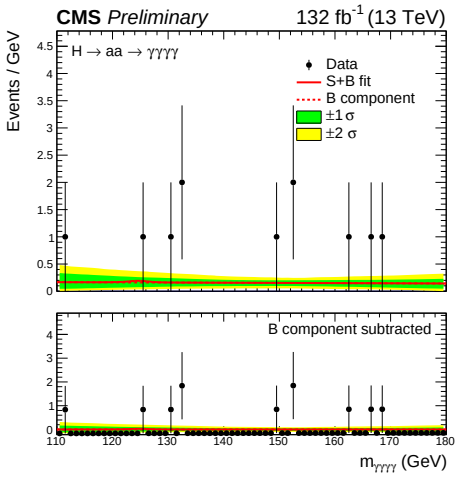


(c) m(a) = 20 GeV

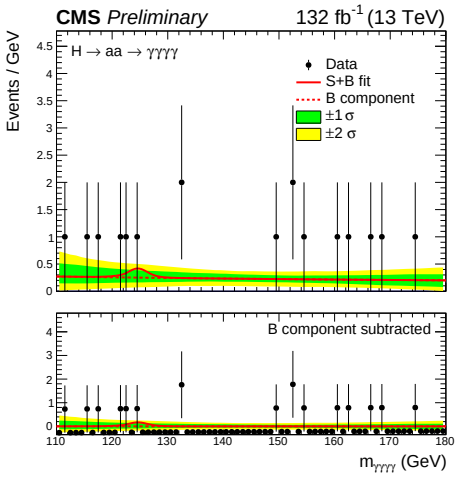


(d) m(a) = 21 GeV

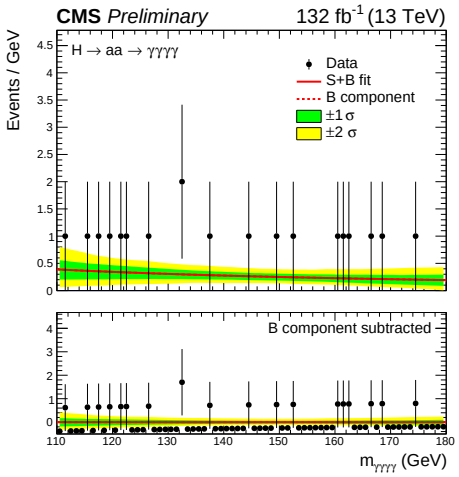
Figure B.2: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



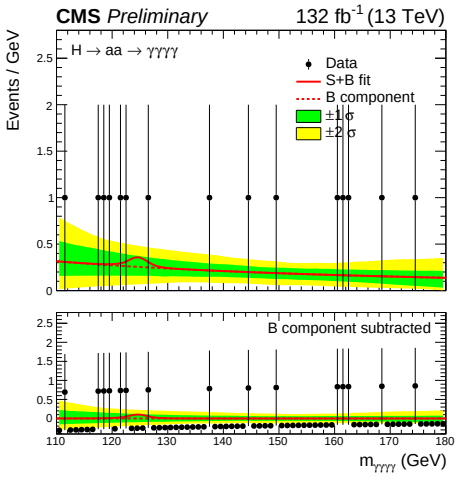
(a) m(a) = 22 GeV



(b) m(a) = 23 GeV

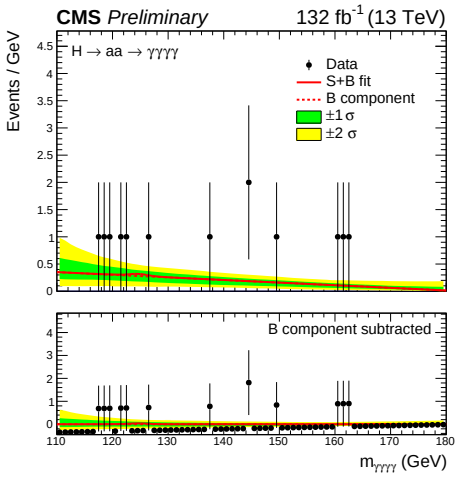


(c) m(a) = 24 GeV

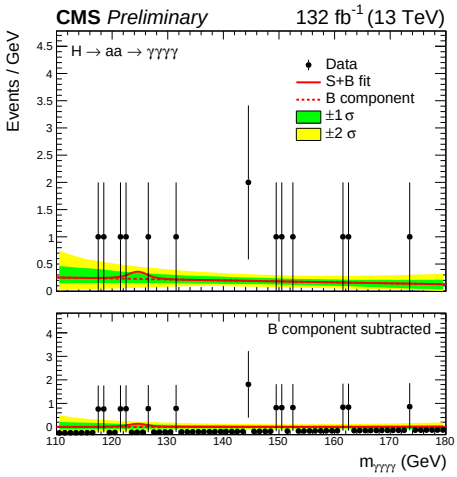


(d) m(a) = 25 GeV

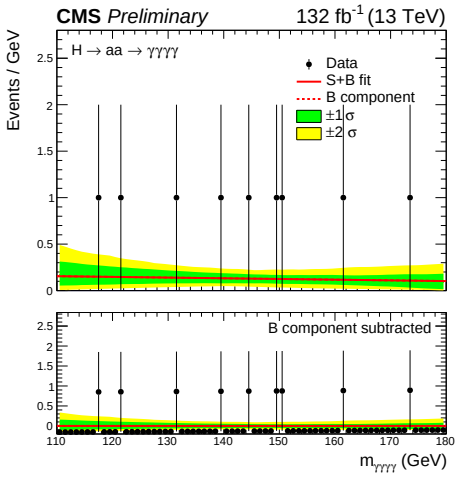
Figure B.3: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



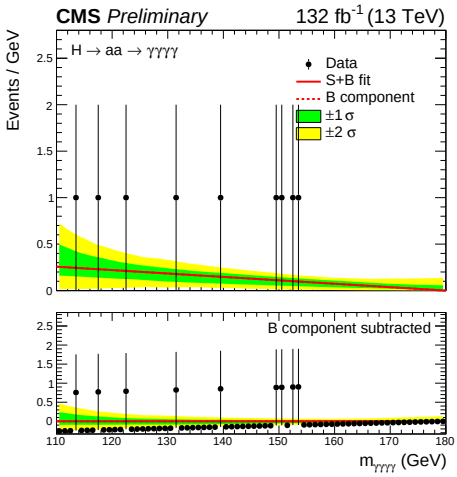
(a) m(a) = 26 GeV



(b) m(a) = 27 GeV

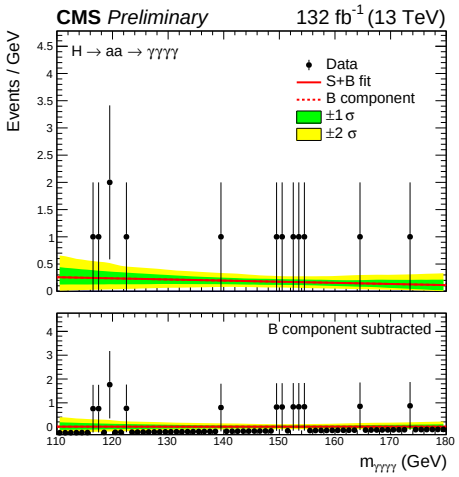


(c) m(a) = 28 GeV

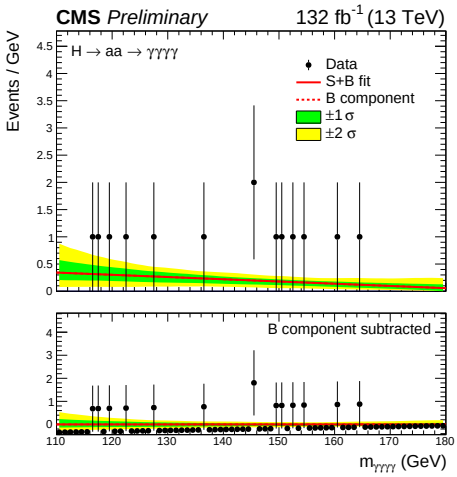


(d) m(a) = 29 GeV

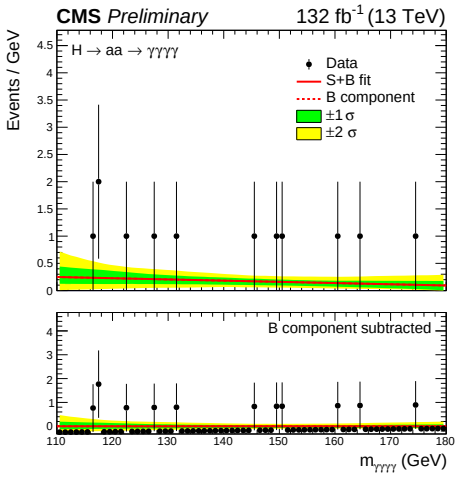
Figure B.4: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



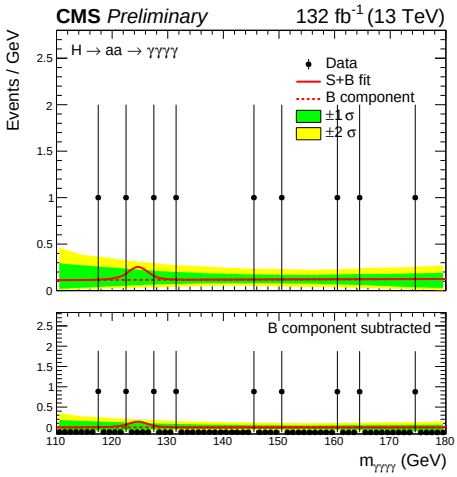
(a) m(a) = 30 GeV



(b) m(a) = 31 GeV

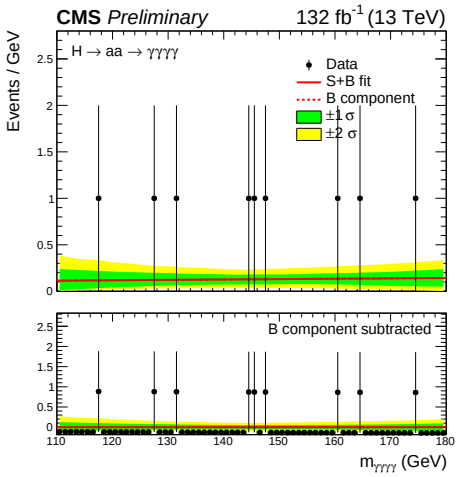


(c) m(a) = 32 GeV

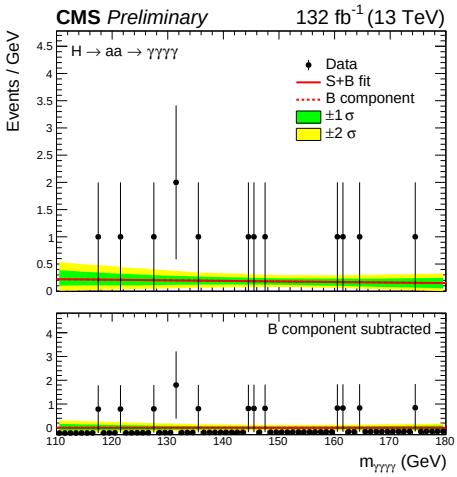


(d) m(a) = 33 GeV

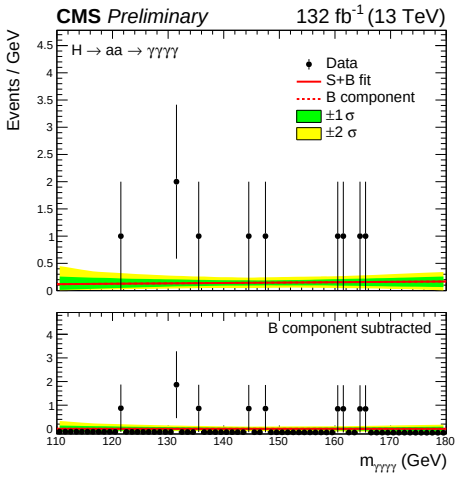
Figure B.5: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



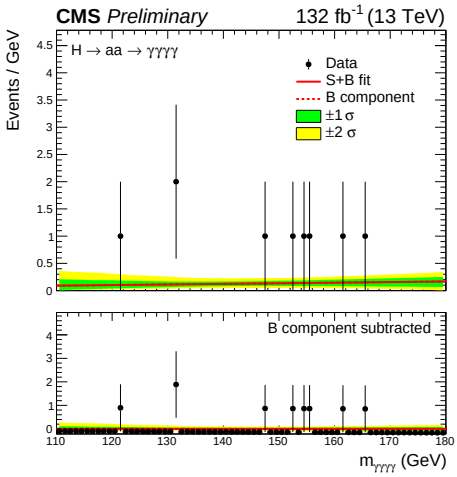
(a) m(a) = 34 GeV



(b) m(a) = 35 GeV

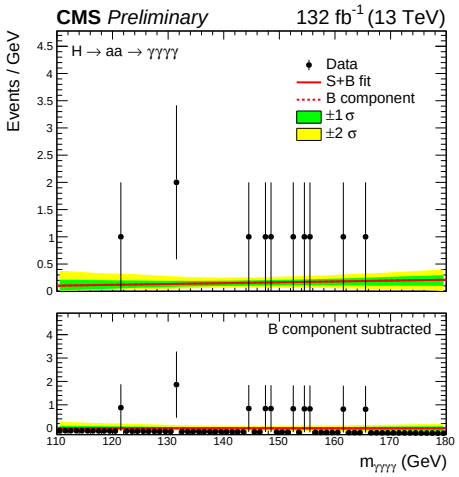


(c) m(a) = 36 GeV

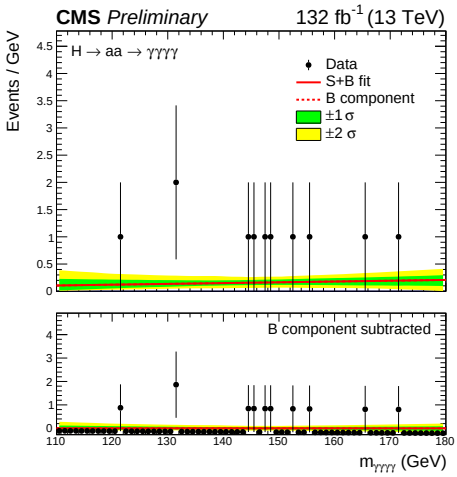


(d) m(a) = 37 GeV

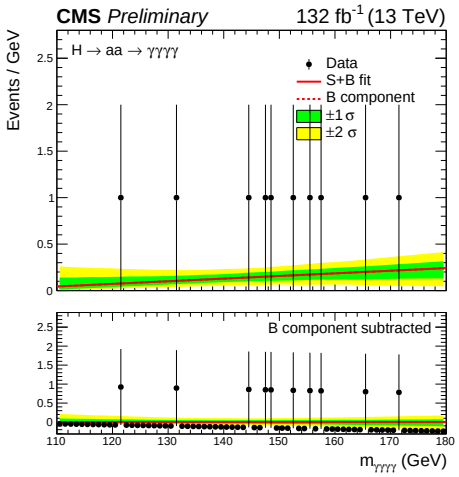
Figure B.6: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



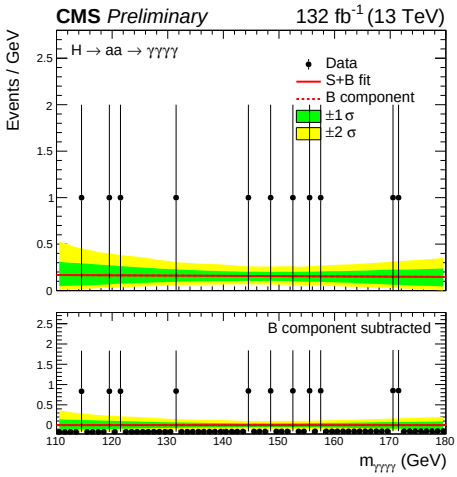
(a) m(a) = 38 GeV



(b) m(a) = 39 GeV

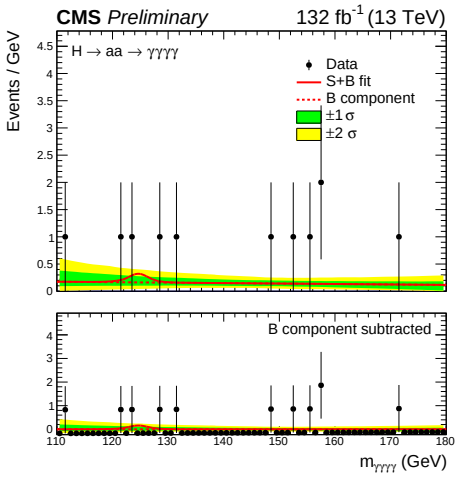


(c) m(a) = 40 GeV

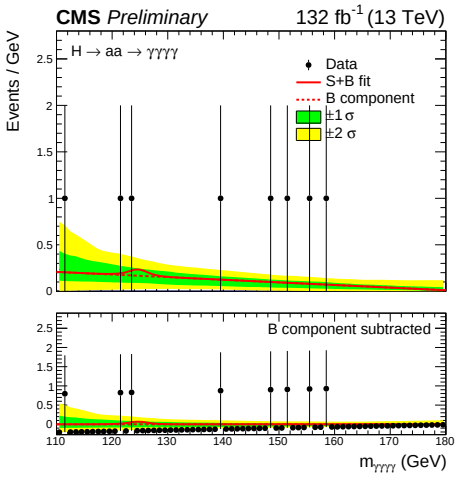


(d) m(a) = 41 GeV

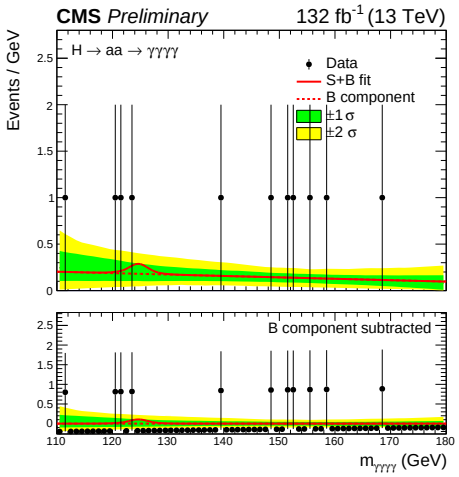
Figure B.7: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



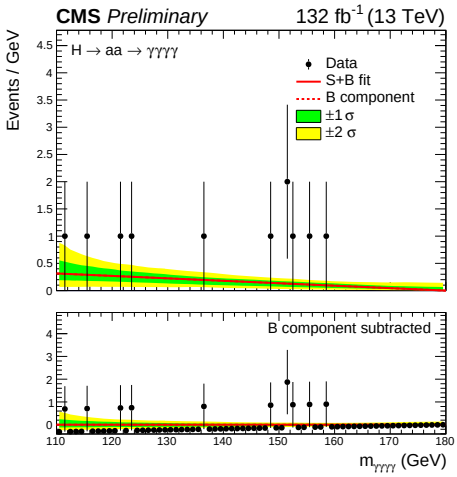
(a) m(a) = 42 GeV



(b) m(a) = 43 GeV

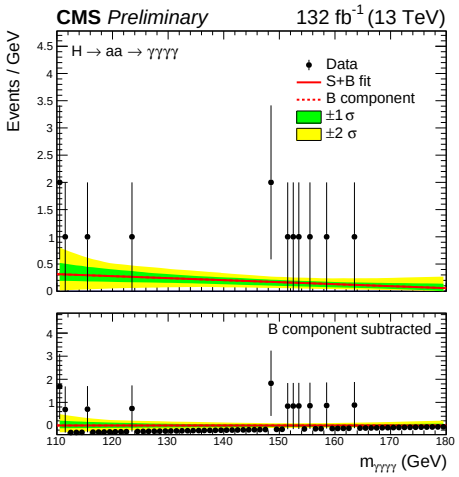


(c) m(a) = 44 GeV

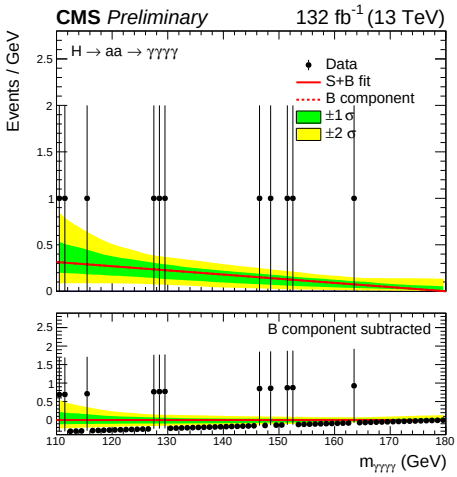


(d) m(a) = 45 GeV

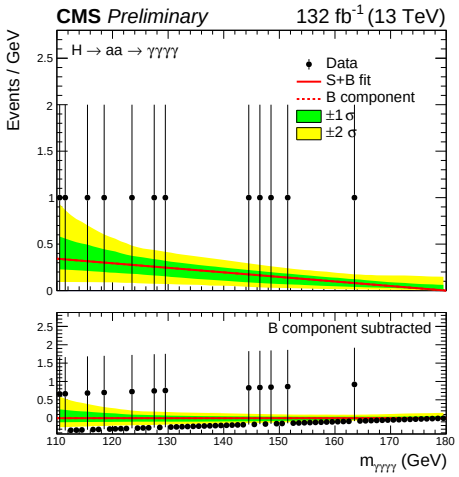
Figure B.8: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



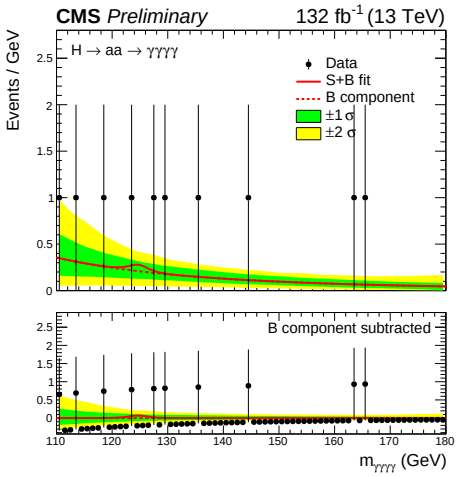
(a) m(a) = 46 GeV



(b) m(a) = 47 GeV

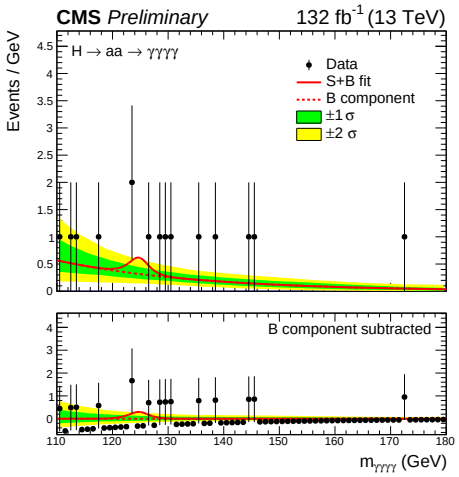


(c) m(a) = 48 GeV

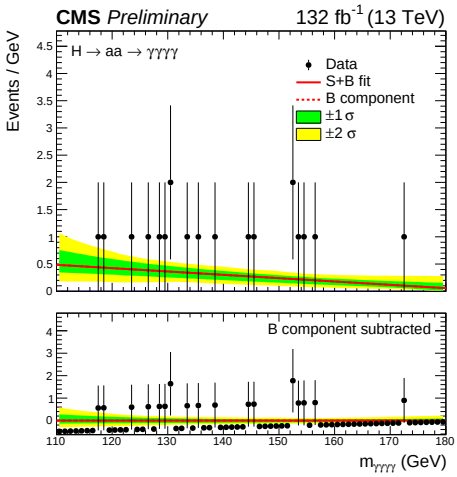


(d) m(a) = 49 GeV

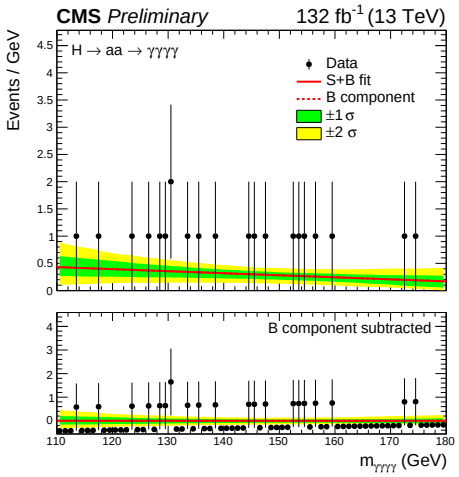
Figure B.9: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



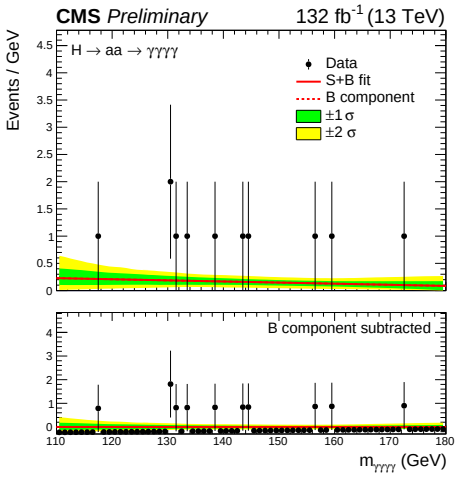
(a) m(a) = 50 GeV



(b) m(a) = 51 GeV

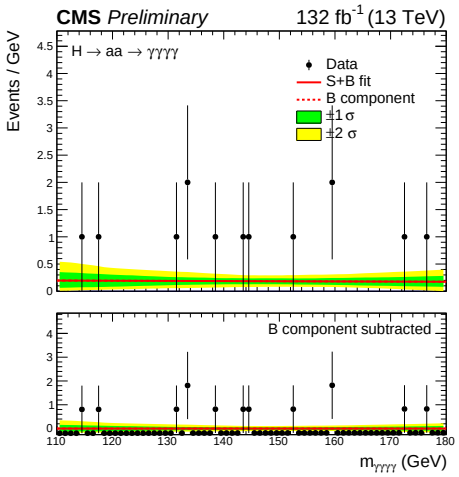


(c) m(a) = 52 GeV

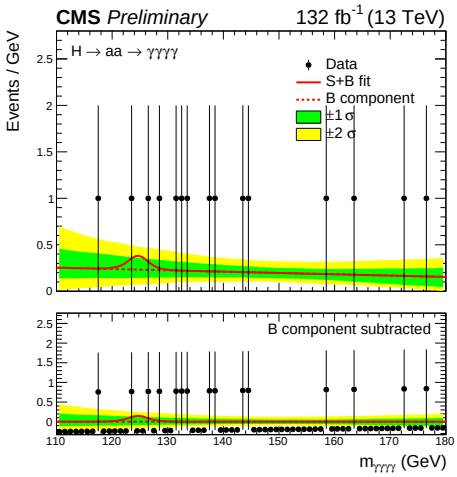


(d) m(a) = 53 GeV

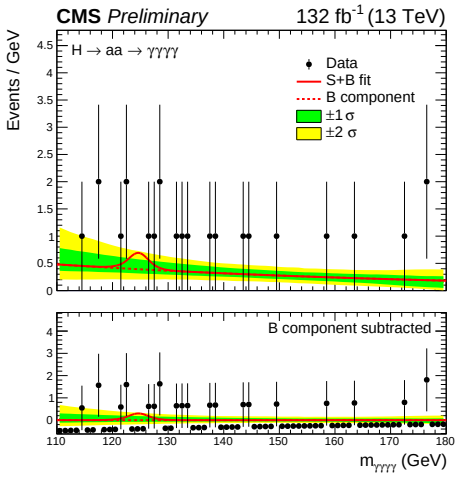
Figure B.10: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



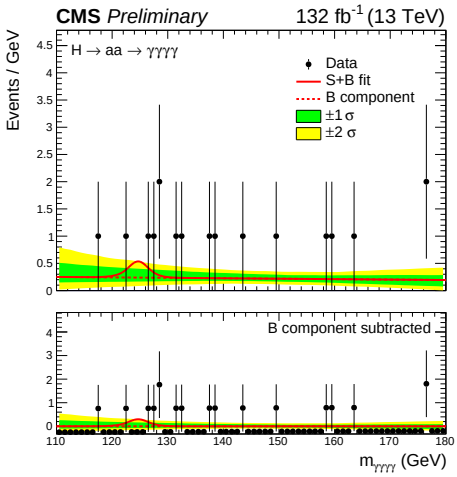
(a) m(a) = 54 GeV



(b) m(a) = 55 GeV

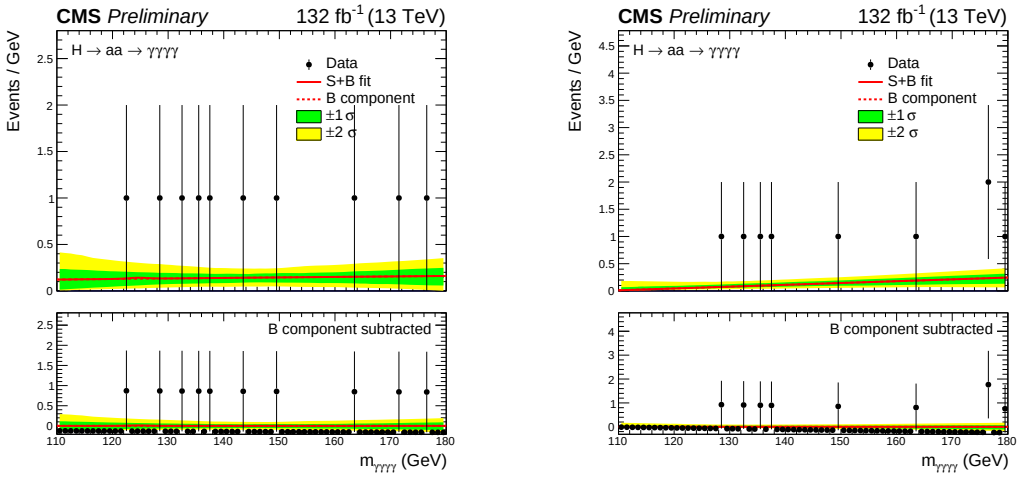


(c) m(a) = 56 GeV



(d) m(a) = 57 GeV

Figure B.11: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.



(a) m(a) = 58 GeV

(b) m(a) = 59 GeV

Figure B.12: Data points are shown in black, and the background fit to the data are shown in red. The one (two) standard deviation are shown in (green) yellow and indicate the uncertainties on the background fit.