Measurements of Higgs boson properties using the diphoton decay channel with the Compact Muon Solenoid experiment

Edward John Titman Scott

Imperial College London
Department of Physics

A thesis submitted to Imperial College London
for the degree of Doctor of Philosophy

Abstract

Measurements of Higgs boson production cross sections with the Higgs boson decaying into a pair of photons are reported. Events with two photons are selected from a sample of proton-proton collisions at a centre-of-mass energy of 13 TeV collected by the Compact Muon Solenoid detector at the Large Hadron Collider in 2016 and 2017, corresponding to a total integrated luminosity of 77.4 fb^{-1}. Cross sections for gluon fusion and vector boson fusion production, relative to the corresponding standard model predictions, are measured to be 1.15 ± 0.15 and $0.83^{+0.37}_{-0.31}$ respectively. These two production modes are further measured in kinematic regions within the simplified template cross section framework. All results are found to be in agreement with the standard model expectations.

To Mum and Dad

Declaration

The work contained within this thesis is my own. It was produced using existing work from, and in collaboration with, several individuals and the Compact Muon Solenoid (CMS) Collaboration. The details of the contributions relevant to each chapter are set out below; in all cases, the work of others has been referenced appropriately.

Chapter 1 introduces our current understanding of particle physics and the Higgs boson in my own words.

Chapter 2 explains the theory of the Standard Model and the Higgs mechanism, which is the work of others, in my own words. The status of the latest Higgs boson measurements is summarised, which includes my own work on the CMS $H \rightarrow \gamma\gamma$ analysis based on the 2016 dataset [1].

Chapter 3 describes the CMS detector, which was designed, built, and operated by others, in my own words.

Chapter 4 describes the High Granularity Calorimeter (HGCAL), the design and testing of which is the work of others within the CMS collaboration. The section on reconstruction is based on my own work, which was performed together with Lindsey Gray, Clemens Lange and Emilio Meschi [2]. The demonstration of the HGCAL's potential for a vector boson fusion analysis in the $H \rightarrow \gamma\gamma$ decay channel is my own work [3].

Chapters 5-8 describe the methodology and results of the CMS $H \rightarrow \gamma\gamma$ analysis which constitute the main part of this thesis and consist primarily of my own work. This analysis has also been documented in Ref. [4]. The detailed breakdown of each chapter is given below.

Chapter 5 describes how events are reconstructed at CMS. These techniques were developed and optimised by others within the $H \rightarrow \gamma\gamma$ analysis group and the CMS collaboration, but are summarised in my own words.

Chapter 6 describes the categorisation of events to maximise the sensitivity of the analysis. This is my own work, and is based on the strategies adopted in previous CMS $H \rightarrow \gamma\gamma$ analyses. The data-driven method for the training of the dijet boosted decision tree was initially developed by Yacine Haddad, and was implemented in this

analysis by Shameena Bonomally.

Chapter 7 describes the construction of the signal and background models, and the various uncertainties included in the analysis. This is my own work, although the techniques utilised were first developed by others.

Chapter 8 reports the final observed results and their uncertainties. This is also my own work, but the methods and tools used are originally the work of others.

Chapter 9 summarises the work done in this thesis and the implications of the results. This is in my own words and with my own considerations for the future development of this work.

Edward John Titman Scott

Acknowledgements

Many, many people have helped me through the past three and a half years, far more than I can list here. I am sincerely grateful to all of you; I have felt very lucky throughout my time at Imperial.

Firstly, thank you to the Imperial HEP group and STFC for together allowing me to do this PhD in the first place. There is a great environment both here and at CERN, and the opportunity to spend time out in Geneva was a fantastic one. The various conferences and travel opportunities have been terrific too.

This thesis contains the contributions of hundreds of members of the CMS Collaboration. Working as part of CMS has meant a lot to me, and the H $\rightarrow \gamma\gamma$ group in particular was lively, warm, and, at times, hilarious. I would like to thank the various CMS H $\rightarrow \gamma\gamma$ conveners for helping get us through the numerous approval processes over the years.

The IC H $\rightarrow \gamma\gamma$ has been a brilliant place to work, starting with my two supervisors. My experience as a student has been immeasurably improved by the kindness of both of you. Gavin, I cannot thank you enough for your incredible wisdom and patience; you have always kept me on track and focused on the bigger picture. Seth, you are the main reason this thesis was possible at all. The amount of help you have given me over the years is astonishing, as is your dedication. The humourous spirit with which you handled all that has kept me both sane and happy, somehow. I'd also like to thank Chris Seez for the help out at CERN, always delivered in a very much no-nonsense manner; I greatly enjoyed my HGCAL work because of you. Also thanks to Nick Wardle, for answering many stupid questions and always being up for a laugh.

To my friends at IC and from CERN: thank you so much. We have had a pretty intense time, but together it's been fun. I'm proud of us all, in advance. Thank you also to my friends at home, who mean so much to me. Your loyalty, support, laughter and provision of all-round decent times have been invaluable.

Mum and Dad: thank you for everything. Your love and support have been the most important things in my life. In my memory at least, this came more often in the form of driving me to and from Croydon for basketball practices than particularly

encouraging me to wonder about physics, but it had the same effect. You always encouraged me to pursue anything that I wanted to pursue, and made me believe it was worth doing so. I cannot thank you enough for that. To Jonny and Adam; I would not be who I am without you two. Growing up together has been more than anyone could ask for, and I amazed and proud of who you have become.

Finally, Ari. Spending the last ten (!) years with you has been the best thing I have ever done. Your patience, support, laughter, and passion mean the world to me. I love you more than anything.

Contents

List of Figures

List of Tables

"And once you have tasted flight,
you will walk the Earth
with your eyes turned skyward;
for there you have been,
and there you long to return"

Leonardo da Vinci

Chapter 1

Introduction

The standard model (SM) of particle physics is a theory which describes the fundamental structure of matter and its interactions [5–7]. Guided by the results of experiments studying the behaviour of high-energy particles, the SM unifies the electromagnetic and weak forces and places them in a coherent framework together with the strong force. The SM has been extraordinarily successful, with many of its predictions verified to unprecedented precision. A key aspect of the SM is the Higgs mechanism [8–10], which breaks the symmetry of the electroweak interaction, explains how particles obtain mass, and predicts the existence of a fundamental particle known as the Higgs boson. The Higgs boson was observed experimentally by the ATLAS and CMS Collaborations [11–14] in 2012 with data collected during Run 1 of the Large Hadron Collider (LHC) [15]. This discovery completed the particle content and modern understanding of the SM, and represents another great success of the theory.

The SM is however known to be incomplete as a theory of nature. Firstly, it does not include a description of the force of gravity. The SM also does not provide a suitable candidate for dark matter, which is required to explain certain astrophysical phenomena [16], and forms a key part of the modern understanding of the large scale structure of the universe [17]. Furthermore, neutrinos are treated as massless in the SM, whereas experimental measurements of neutrino oscillations confirm that they are not [18]. A wide range of extensions to the SM have been proposed which provide explanations for some or all of these phenomena [19]. These beyond standard model (BSM) theories can predict the existence of entirely new particles, and thereby modify the predictions made by the SM.

The results presented in this thesis test the SM predictions of Higgs boson production in the diphoton decay channel with the CMS experiment. Data collected as part of Run 2 of the LHC during 2016 and 2017 are analysed, yielding a dataset of $77.4\,\text{fb}^{-1}$. The CMS detector is a multi-purpose apparatus able to reconstruct sev-

eral types of high-energy particle. Its design is motivated in part by the objective of precisely measuring the properties of the Higgs boson. In particular, the performance of its electromagnetic calorimeter is excellent, which enables effective reconstruction of the photons arising from the decay of the Higgs boson and is vital for the success of the $H \to \gamma\gamma$ analysis. The final results of the analysis are measurements of Higgs boson production cross sections within the simplified template cross section (STXS) framework [20]. The STXS framework provides a coherent approach to performing measurements of Higgs boson couplings, minimising the effect of theoretical uncertainties whilst simultaneously permitting the use of advanced experimental techniques. Results within the STXS framework are presented which aim to test the SM as stringently as possible. Deviations from SM predictions may indicate the presence of BSM physics and guide the way to an improved understanding of nature, or otherwise tightly constrain the possible forms BSM theories can take.

This thesis presents the results of the CMS $H \to \gamma\gamma$ analysis using 2016 and 2017 data, which is documented in Ref. [4]. The structure of the thesis is as follows.

Chapter 2 describes the structure of the SM as a gauge field theory, before introducing the Higgs mechanism and its implications. The phenomenology of the Higgs boson and its decay into two photons is discussed, and the latest measurements of Higgs boson properties, including those within the STXS framework, are summarised.

Chapter 3 describes the LHC and the CMS detector itself. The role of each subdetector and how the various products of high energy proton-proton collisions are captured is explained.

Chapter 4 describes the planned upgrade program for the LHC to become the High-Luminosity LHC (HL-LHC) and the required changes to the CMS detector, focusing on the High Granularity Calorimeter (HGCAL). The motivation for and design of the HGCAL is presented, together with studies illustrating its performance in object reconstruction and physics analysis.

Chapter 5 describes how events recorded by the CMS detector are reconstructed. Emphasis is placed on the procedures used to form photon objects, and how they are subsequently combined with a vertex hypothesis to form diphoton candidates for the $H \to \gamma\gamma$ analysis.

Chapter 6 describes the procedure for categorising events in the $H \to \gamma\gamma$ analysis. Analysis categories targeting different production processes are defined, and the sensitivity of the analysis is increased by using machine learning techniques to discriminate between signal and background processes.

Chapter 7 describes the construction of models for the signal and background contributions to the diphoton invariant mass distributions of the final analysis categories.

The treatment of systematic uncertainties in the analysis is also detailed.

Chapter 8 describes the statistical fitting procedure used to extract the final results of the analysis and their associated uncertainties. Results of cross sections at different levels of granularity within the STXS framework are presented, and comparisons made to the corresponding SM predictions.

Finally, Chapter 9 discusses the conclusions drawn from the results of the analysis, and provides a perspective on future work to build on those results.

Chapter 2

Theory

2.1 Introduction

The standard model (SM) of particle physics is the current best description of nature at its most fundamental level. The SM incorporates the electromagnetic, weak, and strong forces in a single coherent framework, unifying the electromagnetic and weak interactions in doing so. The SM is a type of quantum field theory known as a gauge theory, which represents the fundamental constituents of matter and the forces between them as excitations of relativistic quantum fields. Many of its predictions have been experimentally verified to unprecedented precision [21]. The spontaneous breaking of gauge symmetry in the unified electroweak (EW) sector of the SM results in the prediction of the Higgs boson [8–10], the existence of which is now experimentally confirmed [12, 14]. In this chapter, the fundamental particles and forces of the SM are described, before its structure as a gauge field theory is explained. Spontaneous symmetry breaking and its relation to the Higgs mechanism is elucidated. The phenomenology of the Higgs boson and the consequences for experimental measurements are then detailed. Lastly, the latest precision measurements of the Higgs boson's properties are summarised and the simplified template cross section (STXS) framework is introduced.

2.2 Fundamental particles and forces

In the SM, particles other than the Higgs boson are divided into spin-half fermions and spin-one bosons. Fermions are the fundamental constituents of matter, and can either interact via the electromagnetic and weak forces (leptons), or via the electromagnetic, weak and strong forces (quarks). Both leptons and quarks exist in three distinct generations; the three particles comprising each family have identical proper-

ties except for mass, which increases across generations. In addition, each particle has a corresponding antiparticle, which has the same mass but whose charge and parity have the opposite sign to the original particle.

The interactions between fermions are mediated by a second class of fundamental particles, the gauge bosons. Three forces are represented by the SM gauge bosons: the electromagnetic force, the weak force, and the strong force. The mediator of the electromagnetic force is the massless photon, whilst the weak interaction occurs via the exchange of three massive particles, the W^+, W^-, and Z bosons. Due to the unification of these two forces into the electroweak sector of the SM, the photon, $W^\pm$, and Z bosons arise from combinations of the fundamental gauge fields. Gluons, which mediate the strong force, are massless.

The final particle in the SM is the Higgs boson, which is the only scalar (spin-zero) particle in the theory. The Higgs boson is a massive particle which arises due to spontaneous symmetry breaking in the electroweak sector, and whose existence is necessary to explain the masses of both bosons and fermions.

2.3 Gauge fields

The SM is realised as a particular type of quantum field theory (QFT) known as a gauge field theory. The dynamics and predictions of a given QFT can be derived from the Lagrangian ($\mathcal{L}$) using the Euler-Lagrange equations, in the same way as the equations of motion can be derived in classical field theory [22]. Typically, a Lagrangian is constructed with the aim of respecting the symmetries of the physical system it is attempting to describe. According to Nöther's theorem [23], each symmetry of a Lagrangian has a corresponding current which is conserved. Examples include the invariance of a Lagrangian under translations in time corresponding to the conservation of energy, and invariance under spatial translation to conservation of momentum. The theorem therefore illustrates that the symmetries of a Lagrangian are intimately linked with the conserved quantities and properties of the physical system described.

The defining feature of gauge field theories is that they require the Lagrangian to be invariant under a local gauge transformation. Here a local transformation means one which depends on spacetime co-ordinates, in contrast to a global transformation, which is constant. Elevating a global symmetry to a local one requires the introduction of additional fields, which facilitate interactions between particles and imply the existence of gauge bosons. This is illustrated below, starting with the Dirac Lagrangian which describes a free spin-half fermion [24, 25]:

$$\mathcal{L} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi, \tag{2.1}$$

where ψ is a Dirac spinor, $\bar{\psi}$ is its adjoint with $\bar{\psi} = \psi^\dagger\gamma^0$, and γ^μ are four 4×4 matrices which obey the anticommutation relation $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$ with $g^{\mu\nu}$ the Minkowski metric [25]. The Dirac Lagrangian is invariant under a global phase transformation corresponding to the $U(1)$ group, under which the field transforms as

$$\psi \to e^{ig\theta}\psi, \tag{2.2}$$

where θ and g are constant real numbers. This invariance is dependent on the transformed field commuting with the differential operator. Considering instead a local gauge transformation, meaning the transformation is a function of the spacetime co-ordinates

$$\psi \to e^{ig\theta(x)}\psi, \tag{2.3}$$

where $\theta(x)$ now depends on the spacetime co-ordinates x^μ. The Lagrangian is no longer invariant; instead there is a residual term remaining after the transformation

$$\mathcal{L} \to \mathcal{L} - g\bar{\psi}\gamma^\mu(\partial_\mu\theta)\psi. \tag{2.4}$$

In order to restore gauge invariance, a new field A^μ can be introduced which transforms as

$$A^\mu \to A^\mu + \partial^\mu\theta \tag{2.5}$$

This field is incorporated into the definition of the covariant derivative D^μ as

$$D^\mu = \partial^\mu + igA^\mu, \tag{2.6}$$

which transforms under the local gauge transformation in the desired way

$$D^\mu \to e^{ig\theta(x)}D^\mu. \tag{2.7}$$

With the addition of a free term for the field A^μ, the form of which is constrained by the requirements of being both Lorentz and gauge invariant, the Lagrangian can then be written as

$$\mathcal{L} = i\bar{\psi}\gamma^\mu D_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \tag{2.8}$$

where $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. This Lagrangian is now fully gauge invariant. A mass term for the boson, which would take the form $m_A A_\mu A^\mu$, is not included since it is forbidden by local gauge invariance. The Lagrangian in Eq. 2.8 can now be identified as a description of quantum electrodynamics (QED). The field A_μ corresponds to a photon, and g to the charge of the electron. The interaction vertex between photons

and electrons is contained in the trilinear terms such as $\mathcal{L} \supset -e\bar{\psi}\gamma^\mu A_\mu \psi$, and the electromagnetic field strength tensor is represented by $F^{\mu\nu}$. This illustrates the mechanism by which local gauge invariance introduces new interacting fields with a symmetry, in this case a $U(1)$ symmetry with one degree of freedom. It can be shown that the number of generated bosons is equal to the number of degrees of freedom, or equivalently the dimension, of the symmetry group [22]. The electroweak and strong sectors of the SM follow the same principle, with the symmetry groups being $SU(2) \times U(1)$ and $SU(3)$ respectively. How these symmetry transformations lead to the emergence of the desired SM properties is detailed in the following sections.

2.3.1 Strong interactions

The theory of strong interactions in the SM, known as quantum chromodynamics (QCD), is based upon the $SU(3)$ symmetry group. The symmetries of this group can be represented by traceless 3×3 unitary matrices; this implies there are eight independent matrices, or generators, of the group [26]. The covariant derivative is written as

$$D_\mu = \partial_\mu + ig_s \frac{\lambda^a}{2} A^a_\mu, \tag{2.9}$$

where λ^a are the Gell-Mann matrices. The eight fields A^a_μ correspond to gluons, the massless bosons which mediate the strong force. The QCD equivalent of electric charge, g_s, is known as colour charge. Three independent colour states exist; these are labelled red, green, and blue. Quarks are the only SM fermions which possess colour charge, and thus transform as a triplet under transformations in colour space.

In addition, it should be noted that $SU(3)$ is a non-Abelian group, meaning that its transformations do not commute. This introduces additional complexity to the theory, and has the direct consequence that gluons possess colour charge and can thus self-interact. This is manifest in the expression for the QCD field strength tensor, given by

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g_s f^{abc} A^b_\mu A^c_\mu, \tag{2.10}$$

where f^{abc} are the antisymmetric SU(3) structure constants defined by $[\lambda^a, \lambda^b] = if^{abc}\lambda^c$. The final form of the QCD Lagrangian in the SM is then

$$\mathcal{L}_{\text{QCD}} = \sum_f i\bar{\psi}_f(\gamma^\mu D_\mu - m_f)\psi_f - \frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu}, \tag{2.11}$$

where the index f denotes the six quark flavours and m_f the quark mass of that flavour. An important difference between QCD and QED is that the magnitude of the strong force increases in strength as a function of distance, rather than weakening.

Consequently particles with colour charge are never observed as free particles but are instead confined to colourless, composite bound states. This is the reason quarks and gluons are detected as hadronic showers of particles, known as jets.

2.3.2 Electroweak interactions

One of the key successes of the SM was the unification of the electromagnetic and weak forces. This was developed by Glashow, Weinberg and Salam [5–7] and their GWS model constitutes the formulation of the modern SM. Starting from the $SU(2)$ group, three fields are defined in the covariant derivative

$$D_\mu = \partial_\mu + i\frac{g}{2}W_\mu^i\sigma^i, \tag{2.12}$$

where W_μ^i are the three $SU(2)$ gauge fields and σ^i are the 2×2 Pauli matrices. The corresponding charge is known as weak isospin. Both quarks and leptons interact via the weak interaction and transform as doublets under weak gauge transformations. The weak interaction violates parity, meaning it is not invariant under an inversion in space co-ordinates. This can be encoded explicitly by introducing the parity operators

$$P_{L,R} = \frac{1}{2}(1 \mp \gamma^5), \tag{2.13}$$

where $P_{L,R}$ are the left and right handed parity operators respectively, and $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, which anticommutes with γ^μ and means that objects such as $\bar{\psi}\gamma^5\psi$ are odd under parity. In the SM, only the left-handed fermions couple to the W bosons via $SU(2)$ interactions, which enables parity violation. The formulation in Equation 2.12 would suggest that this is true of all three W^μ bosons; however it is known that the Z boson interacts with both left and right-handed fermions. The key insight of EW unification is that the three bosons arising from the $SU(2)$ group can be combined with a $U(1)$ boson to form the four physically observed bosons. This is achieved in the GWS model by constructing the $SU(2)_L \times U(1)$ group with weak isospin and weak hypercharge Y as the respective charges, and W_μ^i and B_μ as the respective fields in the covariant derivative

$$D_\mu = \partial_\mu + i\frac{g}{2}W_\mu^i\sigma^i + g'YB_\mu, \tag{2.14}$$

with g and g' real numbers. The left-handed fermions are represented by $SU(2)_L$ doublets ψ_L, whilst the right-handed fermions are singlets ψ_R. The physical charged bosons are given by

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \tag{2.15}$$

and the physical neutral bosons expressed as a rotation in the $SU(2)_L \times U(1)$ space with

$$\begin{aligned} Z_\mu &= \cos\theta_W W_\mu^3 - \sin\theta_W B_\mu, \\ A_\mu &= \sin\theta_W W_\mu^3 + \cos\theta_W B_\mu, \end{aligned} \tag{2.16}$$

where θ_W is called the Weinberg angle. The relevant right-handed parts of the EW Lagrangian can then be written

$$\mathcal{L}_{\text{EW}} \supset -\bar{\psi}_R g' \cos\theta_W Y_\psi \gamma^\mu A_\mu \psi_R. \tag{2.17}$$

By comparison with the electromagnetic Lagrangian, the relation $g' \cos\theta_W = |e|$ for the magnitude of the electron charge can be inferred. The hypercharge Y_ψ is identified as the value of the particle's electric charge. Finally, the left-handed terms include

$$\mathcal{L}_{\text{EW}} \supset -\bar{\psi}_L (g \sin\theta_W \frac{\sigma^3}{2} + g' \cos\theta_W Y_\psi) \gamma^\mu A_\mu \psi_L. \tag{2.18}$$

In order for the doublets representing the quarks and leptons to have charges which differ by one unit of charge, it is inferred that $g \sin\theta_W = e$. The hypercharges of the quarks and leptons are then required to be $+\frac{1}{6}$ and $-\frac{1}{2}$ respectively. The theory thus describes a photon which couples in the same way to left and right-handed fields, whilst allowing the weak interaction to violate parity, exactly as observed in nature.

This completes the description of how fermions and bosons interact in the SM. However an issue remains relating to the particle masses. As was described in Section 2.3, mass terms for the bosons are not invariant under local gauge transformations. This is not an issue for the description of the photon, but cannot be reconciled with the experimentally observed finite masses of the W and Z bosons. Furthermore, it can be seen that a mass term of the form

$$\mathcal{L}_{\text{mass}} = -m(\bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L) \tag{2.19}$$

is not gauge invariant, since the left and right-handed spinors transform in different ways. These considerations prevent the inclusion of mass terms for the quarks and leptons in the Lagrangian. Spontaneous symmetry breaking and the Higgs mechanism provide the resolution to this problem.

2.4 Spontaneous symmetry breaking

The SM grants masses to the W and Z bosons via spontaneous symmetry breaking in the EW sector. This is known as the Higgs mechanism, which was conceived concurrently by several people during the 1960s [8–10]. The mechanism can be illustrated using the Lagrangian for a locally gauge invariant complex scalar field ϕ with a fourth-order interaction term [26]

$$\mathcal{L} = (D^{\mu}\phi^{*})(D_{\mu}\phi) - m^{2}\phi^{*}\phi - \lambda(\phi^{*}\phi)^{2} - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \tag{2.20}$$

where λ is a positive real constant and the covariant derivative is as defined in Equation 2.6. Provided the m^2 term is positive, this potential has its minimum at $\phi = 0$. However the introduction of the quartic term permits the m^2 term to be negative, which results in a degenerate set of minima at non-zero values

$$|\phi| = \frac{\sqrt{-m^{2}}}{2\lambda} \equiv \frac{v}{\sqrt{2}}, \tag{2.21}$$

where the vacuum expectation value v has been defined. Although the potential itself is symmetric, the physical vacuum state is not; this is what is meant by the term spontaneous symmetry breaking. The shape of this potential is illustrated in Figure 2.1.

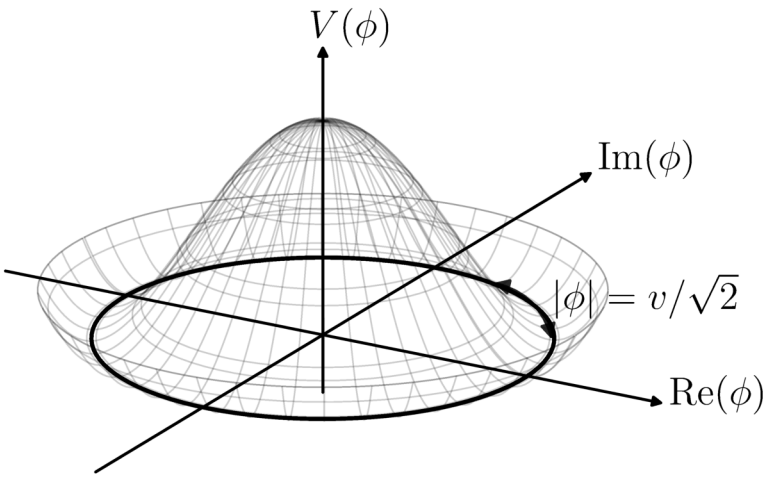


Figure 2.1: The shape of the quartic potential described in Equation 2.20. The potential itself is symmetric, but the symmetry is broken when a specific choice of ground state is made from the degenerate set of available minima.

Making the arbitrary choice of the ground state being in the real direction, it is

possible to expand about the minimum with a field of the form

$$\phi = \frac{1}{\sqrt{2}}(v + \phi_1 + i\phi_2), \tag{2.22}$$

where ϕ_1 and ϕ_2 are real scalar fields. Substituting this into the Lagrangian (Equation 2.20) yields the mass terms

$$\mathcal{L} \supset -2\lambda v^2 \phi_1^2 + 2g^2 v^2 A^\mu A_\mu. \tag{2.23}$$

Thus the gauge boson A_μ has acquired mass; spontaneous symmetry breaking of a Lagrangian invariant under local gauge transformations yields a massive gauge boson. This important result underlies the generation of masses for the SM vector bosons. Furthermore, there is now an additional massive scalar particle ϕ_1; this will shortly be identified as the Higgs boson.

In the SM, the scalar Higgs field transforms as an $SU(2)$ doublet H which has the required four degrees of freedom. Its potential is fourth order such that the Lagrangian can be written

$$\mathcal{L}_{\text{EW}} \supset (D^\mu H)^\dagger (D_\mu H) + \mu^2 H^\dagger H - \lambda (H^\dagger H)^2, \tag{2.24}$$

where μ and λ are positive constants and the covariant derivative is as defined in Equation 2.14. This potential also has a degenerate set of non-zero minima with

$$H^\dagger H = \frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}, \tag{2.25}$$

and the choice can be made to expand about the minimum in the neutral component of the doublet such that

$$H = \begin{pmatrix} 0 \\ \dfrac{v}{\sqrt{2}} + h(x) \end{pmatrix}, \tag{2.26}$$

where $h(x)$ is the physical scalar field of the theory. Parameterising the field in this way corresponds to choosing the unitary gauge. The covariant derivative acting on the Higgs field then results in, after the rotation described in Equation 2.16, the Lagrangian terms

$$\mathcal{L}_{\text{EW}} \supset \frac{g^2 v^2}{4} W_\mu^+ W^{-\mu} + \frac{(g^2 + g'^2)v^2}{8} Z_\mu Z^\mu. \tag{2.27}$$

Thus the W and Z bosons have acquired mass whilst leaving the photon massless. The Higgs mechanism is therefore able to explain the observed mass structure of the gauge bosons. Furthermore, the ratio of the masses of the and W and Z bosons is predicted to be equal to $\cos\theta_W$, which is also in agreement with experiment. This pattern of

symmetry breaking depends on the choice of direction for the vacuum expectation value and Higgs field; it has been chosen by construction to match the observed SM bosons and fermions.

In addition to explaining the masses of the gauge bosons, the Higgs mechanism enables mass terms for the fermions to be included in the Lagrangian. These terms are known as Yukawa terms [26]

$$\mathcal{L}_{\text{EW}} \supset \frac{g_f}{\sqrt{2}} v \bar{\psi}_L H \psi_R + h.c. \ , \tag{2.28}$$

where g_f is the Yukawa coupling of the fermion to the Higgs boson and $h.c.$ indicates the Hermitian conjugate. This combination is now gauge invariant, in contrast to the mass terms in Equation 2.19. Expanding the field H in terms of the vacuum expectation value and the field $h(x)$, the mass terms generated for the fermions are

$$m_f = \frac{g_f v}{\sqrt{2}} \tag{2.29}$$

The magnitude of the Yukawa coupling is not known a priori and must be measured experimentally. It can then be seen that there are also interactions between the fermions and the Higgs boson whose coupling strength is proportional to the fermion mass, as well as the Higgs boson mass.

Furthermore, the Higgs boson also has interaction terms with the gauge bosons. The relevant terms in the Lagrangian are

$$\mathcal{L}_{\text{EW}} \supset m_W^2 \left(\frac{2h}{v} + \frac{h^2}{v^2} \right) W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} \left(\frac{2h}{v} + \frac{h^2}{v^2} Z_\mu Z^\mu \right), \tag{2.30}$$

which describe the three point and four point interactions between the massive gauge bosons and the Higgs boson. Again the coupling strength is proportional to both the gauge boson mass and the Higgs boson mass. The Higgs boson mass terms in the Lagrangian together with its self-interactions are

$$\mathcal{L}_{\text{EW}} \supset -\lambda v^2 h^2 - \lambda v h^3 - \frac{1}{4} \lambda h^4, \tag{2.31}$$

where the Higgs boson mass can now be identified as $m_H = \sqrt{2\lambda} v$. The mass itself is unknown a priori, and must be measured experimentally. Given the Higgs boson mass, and considering all these interaction terms involving the Higgs boson, it is possible to infer the phenomenology of how the Higgs boson is produced and how it decays. This phenomenology and its consequences for experimental measurements of the Higgs boson's properties are discussed in the following section.

2.5 Properties of the Higgs boson

Discovering the Higgs boson, and thereby reaching observation of all elements of the standard model, was one of the main physics objectives of the LHC. This goal was achieved in 2012 when both the ATLAS and CMS experiments announced the discovery of a new particle resembling the Higgs boson [12, 14]. Several Higgs boson production modes and decay channels were analysed. Now that the existence of the Higgs boson is confirmed, the experimental focus is to measure its properties as precisely as possible. This enables stringent tests of the SM to be made, and in so doing either observe discrepancies from its predictions which indicate the presence of physics beyond the standard model (BSM), or place constraints on the permitted form of possible BSM theories. In this section the most important production and decay modes for Higgs boson measurements at the LHC are described, and current state-of-the-art results are summarised.

2.5.1 Higgs boson production at the LHC

The LHC produces high energy collisions between two circulating beams of protons. The production of the Higgs boson can be initiated in different ways, with cross sections that depend on the centre-of-mass energy ($\sqrt{s}$) of the proton-proton collision and the Higgs boson mass. For collisions at $\sqrt{s} = 13\,\text{TeV}$ and a Higgs boson mass of around 125 GeV, gluon fusion (ggH) is the dominant mode [20]. The ggH final state contains only the Higgs boson. This contrasts with other modes, such as vector boson fusion (VBF), which produces a Higgs boson together with two quarks that hadronise to form jets. These additional objects can improve the ability to discriminate the Higgs boson signal from background, meaning events which have similar detector signatures to the Higgs boson but arise from other SM processes. Other important production modes whose existence has been confirmed at the LHC include vector boson-associated production (VH) and production in association with a top quark-antiquark pair (ttH). Figure 2.2 shows the leading order Feynman diagrams for these four production modes, showing explicitly the additional objects present in events produced by modes other than ggH. The expected SM cross sections for each process, and for the rarer tH and bbH processes, are shown in Table 2.1. Experimental sensitivity to a given process depends not only on the signal cross section but also the amount of background. Observations of the ggH and VBF were made during Run 1 of the LHC [27–29], with the VH and ttH modes being observed more recently using Run 2 data [30–33].

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

Table 2.1: Cross section values for the main Higgs boson production processes at the LHC, for $\sqrt{s} = 13\,\text{TeV}$ and $m_H = 125\,\text{GeV}$. Values are taken from Ref. [20].

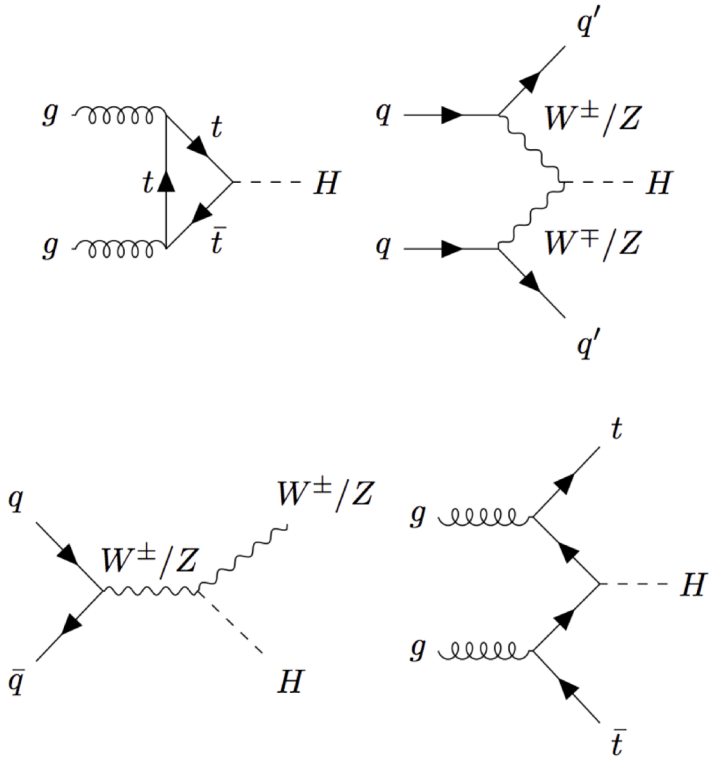


Figure 2.2: Feynman diagrams representing the four principal Higgs boson production modes at the LHC. In descending order of cross section, these are: ggH in the top left, VBF in the top right, VH in the bottom left, and ttH in the bottom right.

Decay mode	$b\bar{b}$	$W^{\pm}W^{\mp *}$	gg	$\tau^{+}\tau^{-}$	$c\bar{c}$	ZZ*	$\gamma\gamma$
Branching fraction	58.2%	21.4%	8.2%	6.3%	2.8%	2.6%	0.23%

Table 2.2: Branching fractions of the main Higgs boson decay modes. Values are taken from Ref. [20].

2.5.2 Higgs boson decay modes

The SM predicts that the Higgs boson has an extremely short lifetime [20]. The Higgs boson is therefore never observed directly, but instead its presence is inferred from its decay products. There are several permitted decay modes of the Higgs boson. Since its coupling strength to other particles is proportional to the mass of the decay product, it can decay directly to any massive particle[1]. In addition, loop diagrams permit the decay to massless particles, including the gluon and photon. The branching fractions for each of the seven principal decay modes is shown in Table 2.2. Observations of the $\gamma\gamma$, ZZ*, $W^{\pm}W^{\mp *}$, and $\tau^{+}\tau^{-}$ decay modes were made during Run 1 of the LHC [27–29], with the $b\bar{b}$ mode being observed more recently using Run 2 data [32–35].

It can be seen that the decay to $b\bar{b}$ has the largest branching fraction. However due to the large hadronic background at the LHC, this decay channel is very difficult to utilise experimentally. The channels with the greatest sensitivity, despite their relatively low branching fractions, are the diphoton ($\gamma\gamma$) and ZZ^{*} decays (where Z^{*} indicates a virtual, or off mass-shell, Z boson). The sensitivity of these channels depends on the relatively low background and the ability to precisely reconstruct the mass of the Higgs boson from the decay products. The analyses of the $\gamma\gamma$ and ZZ^{*} decay channels were the most important at the time of the discovery of the Higgs boson, and for the same reasons are now are ideally suited to the study of its properties. This analysis uses the diphoton decay channel. Three Feynman diagrams representing the largest contributions to the effective vertex between the Higgs boson and two photons are shown in Figure 2.3. As can be seen in the figure, the decay to two photons is therefore sensitive to the Higgs boson's coupling to other particles, including the top quark and the W boson.

2.5.3 Status of Higgs boson measurements

Since its discovery, remarkable progress has been made in measuring the properties of the Higgs boson. Several aspects of the SM Higgs boson structure can be tested experimentally. The first is the spin of the Higgs boson; it is the only fundamental

[1] Only decays to pairs of particles whose mass is less than half that of the Higgs boson are permitted kinematically. However the decay can still occur to virtual particles of greater mass, which then themselves decay.

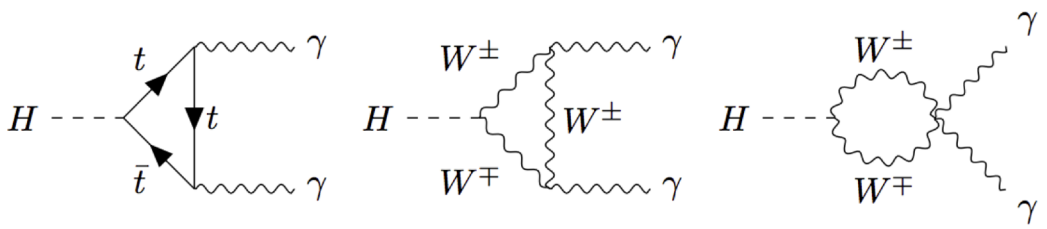


Figure 2.3: The Feynman diagrams with the largest contribution to the H $\rightarrow \gamma\gamma$ decay loop. The diagrams involving the top quark have opposite sign to those involving the W boson, leading to destructive interference.

scalar particle within the SM. Due to the observed decay to two spin-one photons, the Higgs boson must have even spin. Dedicated measurements by ATLAS and CMS have excluded both the spin-two hypothesis and the hypothesis of a Higgs boson with zero spin but negative parity (a pseudoscalar) [36, 37]. The possibility that the observed Higgs boson is a mixture of scalar and pseudoscalar states has not been ruled out, and remains an active subject of research.

Furthermore, the mass of the Higgs boson has also been measured precisely. The $\gamma\gamma$ and ZZ* modes are the two decay channels with sufficiently narrow mass resolution to contribute to this measurement. During Run 1, the combination of ATLAS and CMS analyses led to the measurement of the Higgs boson mass to be $m_H = 125.09 \pm 0.24\,\text{GeV}$ [38]. This has since been superseded by the measurements made by CMS in the ZZ* channel with Run 2 data, where a precision of better than 0.2% [39] is obtained. The observed value and its uncertainty is $m_H = 125.26 \pm 0.21\,\text{GeV}$.

The coupling of the Higgs boson to other SM particles can also be measured experimentally, by comparing the rate of Higgs boson production to the expectation from the SM. These measurements can be parameterised in different ways. For results based on Run 1 data, two different schemes were commonly used. The first utilises signal strength modifiers μ, which are defined as the ratio of the observed Higgs boson yield to the SM expectation. This can be defined inclusively, for all Higgs boson production and decay modes, or for individual combinations of production and decay modes. The general definition is written as

$$\mu_i^f = \frac{\sigma_i \mathcal{B}^f}{(\sigma_i)_{\text{SM}}(\mathcal{B}^f)_{\text{SM}}}, \tag{2.32}$$

where σ_i and $\mathcal{B}^f$ are the observed production cross section and decay branching fraction respectively, and the "SM" subscript refers to their respective SM predictions. The second is the so-called κ-framework [40], which is motivated by the leading-order

couplings of the Higgs boson to bosons and fermions. The coupling modifiers κ_j are defined such that

$$\kappa_j^2 = \frac{\sigma_j}{\sigma_j^{\mathrm{SM}}} \text{ or } \frac{\Gamma_j}{\Gamma_j^{\mathrm{SM}}}, \tag{2.33}$$

where in the SM the values of all the κ_j are equal to unity. In addition to including couplings to individual particles or groups of particles, effective coupling modifiers κ_g and κ_γ, which describe the loop processes of ggH production and diphoton decay respectively, can be defined. The combination of ATLAS and CMS Run 1 analysis documented in Ref. [29] presents results in terms of both signal strength modifiers and coupling modifiers.

An example of signal strength modifier measurements from the previous CMS $\mathrm{H} \to \gamma\gamma$ analysis [1] is shown in Figure 2.4. The modifiers for the ggH, VBF, ttH and VH production processes and the diphoton decay are shown, with the inclusive diphoton decay modifier overlaid. In this single decay channel with one experiment, the ggH signal strength is already measured to a precision of 20%. Combinations of results across different decay channels performed for optimal precision, with the inclusive Higgs boson signal strength now measured to a precision of around 10% [41, 42]. So far, all measurements are consistent with the SM predictions.

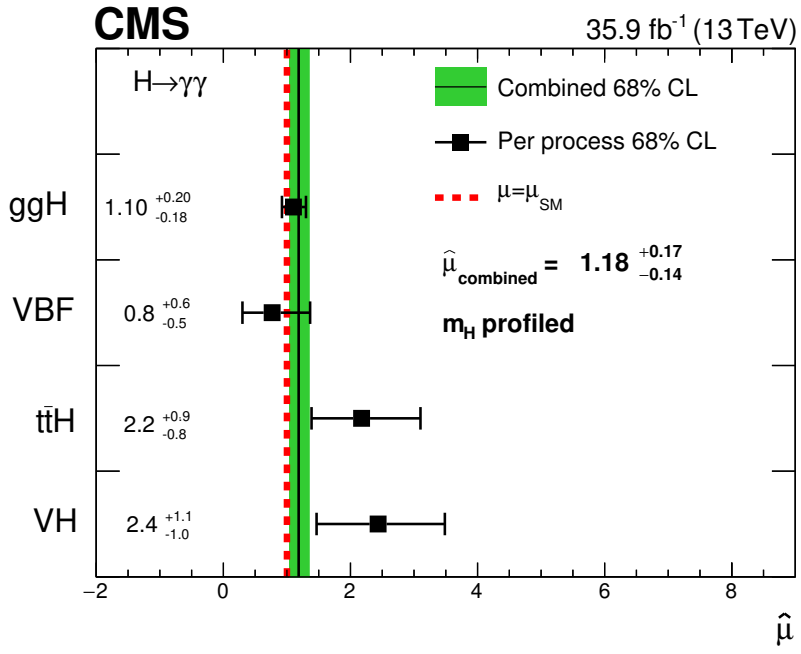


Figure 2.4: Signal strength modifier measurements in the $\mathrm{H} \to \gamma\gamma$ decay channel. The modifiers for each process (black points) are shown, with the SM Higgs boson mass profiled, compared to the overall signal strength modifier (green band) and to the SM expectation (dashed red line). Figure first shown in Ref. [1].

2.6 The simplified template cross section framework

The simplified template cross section (STXS) framework [20] provides a coherent approach with which to perform precision Higgs boson measurements. Its goal is to minimise the theory-dependence of Higgs boson measurements, whilst permitting the use of advanced analysis techniques to optimise the measurements' sensitivities. This also increases the reinterpretability of the measurements. It is designed to supersede the traditional signal strength modifier measurements.

In the STXS framework, theoretically-motivated kinematic regions based upon Higgs boson production modes are defined. These simplified regions, or bins, exist in varying degrees of granularity, following sequential "stages". Increasing the granularity of the signal bins provides additional information for different theoretical interpretations of the measurements, and enhances the sensitivity to possible signatures beyond the standard model. As the integrated luminosity of the datasets collected by the LHC increases, so does the experimental sensitivity, and measurements can progress to later stages within the STXS framework.

At the so-called STXS stage 0, the bins correspond closely to the different Higgs boson production mechanisms. An additional requirement is placed on the Higgs boson rapidity $|y_H|< 2.5$, which reduces the theoretical uncertainty that would otherwise arise when extrapolating measurements to the full phase space, a large part of which is not accessible experimentally. The experimental acceptance of $H \rightarrow \gamma\gamma$ events with $|y_H|> 2.5$ is negligible. The stage 0 bins are illustrated in Figure 2.5. The principal difference from the production processes used in the signal strength measurements, such as those shown in Figure 2.4, is that VH production where the vector boson decays hadronically is grouped together with VBF production to define an "electroweak qqH" bin.

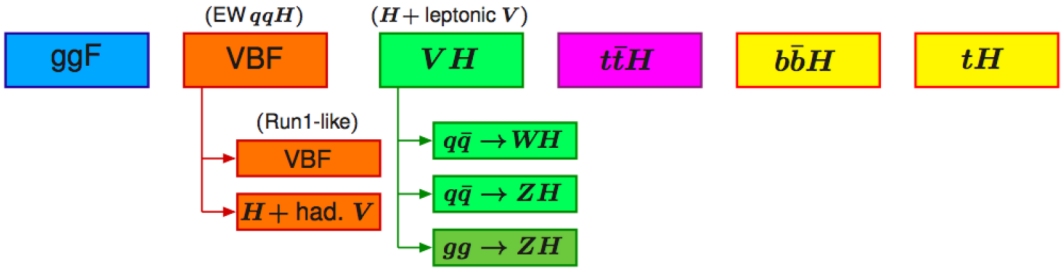


Figure 2.5: The STXS stage 0 bins. The bins are designed to closely follow the Higgs boson production processes used for measurements during Run 1 of the LHC. Figure taken from Ref. [20].

Measurements of stage 0 cross sections in the H $\rightarrow \gamma\gamma$ decay channel were performed in the previous CMS H $\rightarrow \gamma\gamma$ analysis [1]. The results are shown in Figure 2.6, which is similar to Figure 2.4. Aside from the change in signal bin definitions and the absence of an inclusive measurement, the main difference is in the treatment of the theoretical uncertainties. When performing a measurement of μ, a fit is performed for the signal strength modifier ratio $\sigma/\sigma_{\mathrm{SM}}$. This requires that the theoretical uncertainty on the SM prediction, which enters via the denominator, be included. In contrast if the fit parameter is just the observed cross section σ, the measurement does not depend on the overall SM prediction and this uncertainty does not need to be included. It is instead considered as the uncertainty on the SM prediction, as displayed in Figure 2.6. This separation of the theoretical uncertainties means that the measurement is less dependent on the theoretical prediction and remains useful even if improved theoretical predictions are obtained in the future.

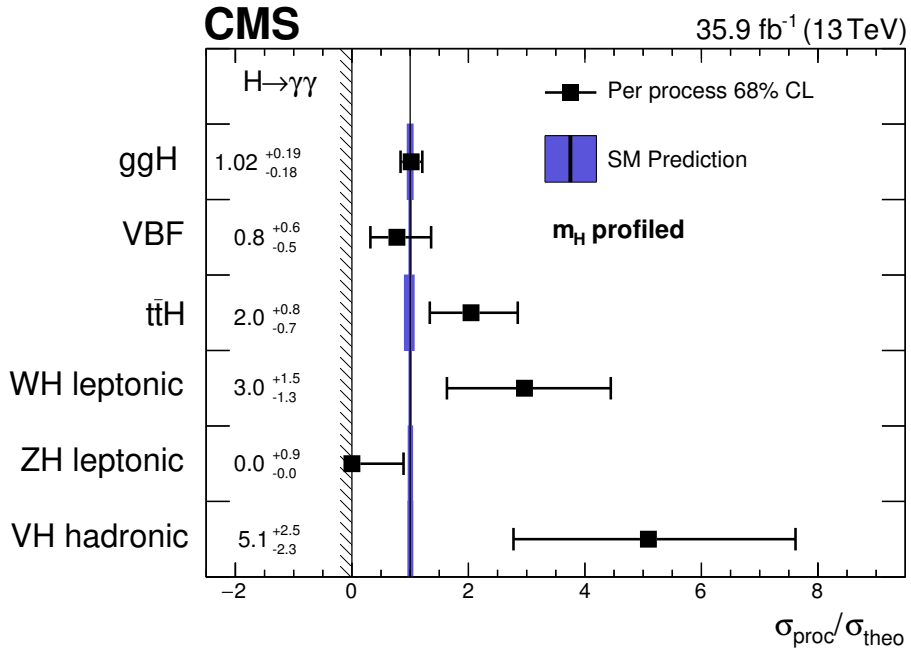


Figure 2.6: Normalised cross sections measured for each stage 0 bin (black points) in the STXS framework, with the SM Higgs boson mass profiled, compared to the SM expectations and their uncertainties (blue band). The signal strength modifiers are constrained to be non-negative, as indicated by the vertical line and hashed pattern at zero. Figure first shown in Ref. [1].

At stage 1 of the STXS framework, a further splitting of bins into different kinematic regions is performed. This provides additional information for different theoretical interpretations of the measurements, and enhances the sensitivity to possible signatures beyond the standard model. Measurements at stage 1 of the framework have

already been reported by the ATLAS Collaboration [43–45]; this analysis comprises the first CMS measurement of STXS stage 1 regions in the diphoton channel, covering the gluon fusion (ggH) and vector boson fusion (VBF) production modes. The definitions of the different bins are shown in Figures 2.7 and Figures 2.8 for ggH and VBF production respectively. For ggH production, bins are defined using the transverse momentum of the Higgs boson and the number of jets in the event. In addition there are two bins for events with two jets where the event kinematics resemble those of typical VBF events. For VBF production, the bins are defined using the kinematics of the characteristic dijet system. The ultimate goal of this analysis is to measure the cross sections of these bins as precisely as possible. Further detail of the exact bin definitions and the methods used to discriminate between them is given in Chapter 6.

2.7 Summary

The SM is a gauge field theory which describes all known fundamental particles and the forces which govern their interactions, with the exception of gravity. Its predictions have been tested to extremely high levels of precision, and in 2012 the last remaining unobserved particle in the SM, the Higgs boson, was discovered. The Higgs boson is a crucial piece of the SM as its existence is required to explain the origin of the masses of both gauge bosons and fermions. Once the mass of the Higgs boson, which has been measured experimentally, is known, its interactions with other particles in the SM are fully determined. Performing precision measurements of the Higgs boson and its properties is therefore one of the foremost priorities in current particle physics research. The STXS framework provides a coherent approach to performing these measurements, minimising the theoretical dependence of experimental results whilst permitting the use of advanced analysis techniques. This analysis aims to measure STXS bins in the $H \rightarrow \gamma\gamma$ decay channel at various levels of granularity, thereby testing the SM and its predictions as stringently as possible.

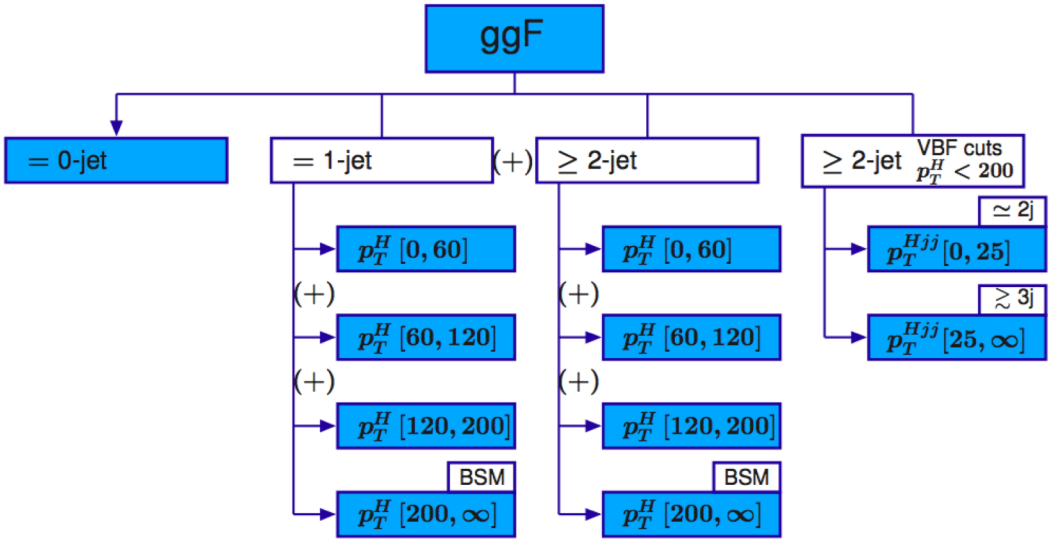


Figure 2.7: The STXS stage 1 bins for the ggH production mode. The inclusive ggH process is subdivided into bins according to the transverse momentum of the Higgs boson and the number of jets in the event. Figure taken from Ref. [20].

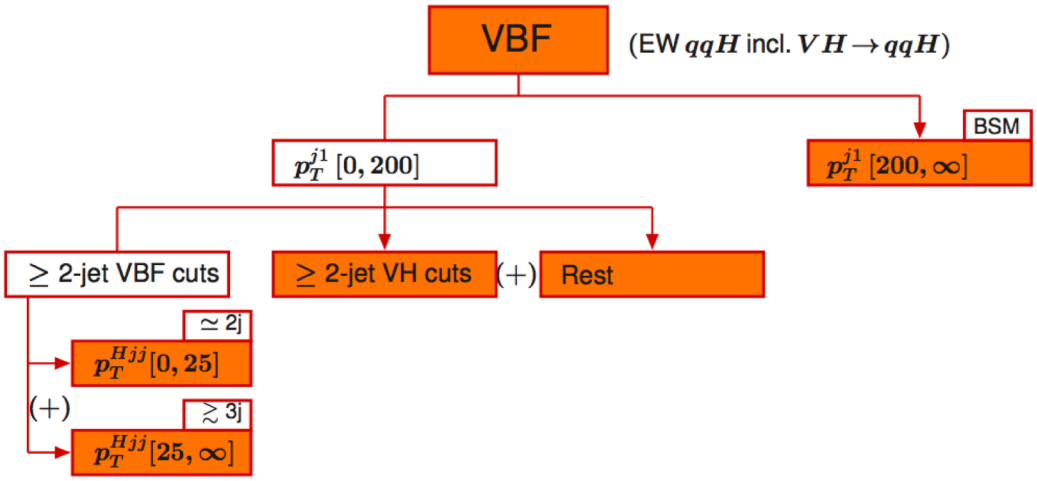


Figure 2.8: The STXS stage 1 bins for the VBF production mode. Both VBF events and VH events where the vector boson decays hadronically are included. The inclusive processes are subdivided according to the kinematics of the Higgs boson and the jets in the event. Figure taken from Ref. [20].

Chapter 3

The Compact Muon Solenoid

3.1 Introduction

The Large Hadron Collider (LHC) at CERN is a 27 km hadron accelerator and collider with a design centre-of-mass energy of up to 14 TeV [46]. Its purpose is to provide collisions sufficient in energy and number to precisely probe physics at the electroweak scale. Four detectors, one of which is the Compact Muon Solenoid (CMS), are situated around the LHC ring. CMS is a general-purpose detector designed to reconstruct a wide range of physics objects. Together, the LHC and CMS facilitate a wide-ranging programme of physics studies, from SM precision measurements to dark matter searches, and including exploration of electroweak symmetry breaking via production of the Higgs boson. This chapter will describe the design and operation of both the LHC and the CMS detector.

3.2 The Large Hadron Collider

The LHC is situated in the tunnel that previously housed the Large Electron Positron collider (LEP) [47], around 100 m below ground across the French-Swiss border near CERN. It is the final machine in a series of accelerators which form the CERN accelerator complex [48]; these act as the injection system for the LHC. The two counter-circulating beams cross at four interaction points, at each of which a detector is situated. Directly opposite CMS is a second general purpose detector, ATLAS [11], whose physics objectives are identical to those of CMS but whose design and operation are independent. Two further detectors, LHCb [49] and ALICE [50], focus on flavour and heavy-ion physics respectively. Three smaller experiments are also situated at the LHC; these are MoEDAL, TOTEM and LHCf [51–53], which between them perform studies of forward physics and searches for magnetic monopoles. Whilst the LHC is capable of

producing heavy ion collisions in addition to proton-proton collisions, the remainder of this section will be dedicated to its proton-proton operations. A schematic of the full CERN accelerator complex and the LHC experiments is shown in Figure 3.1. The operation of the LHC and the data it has accumulated so far are described in detail below.

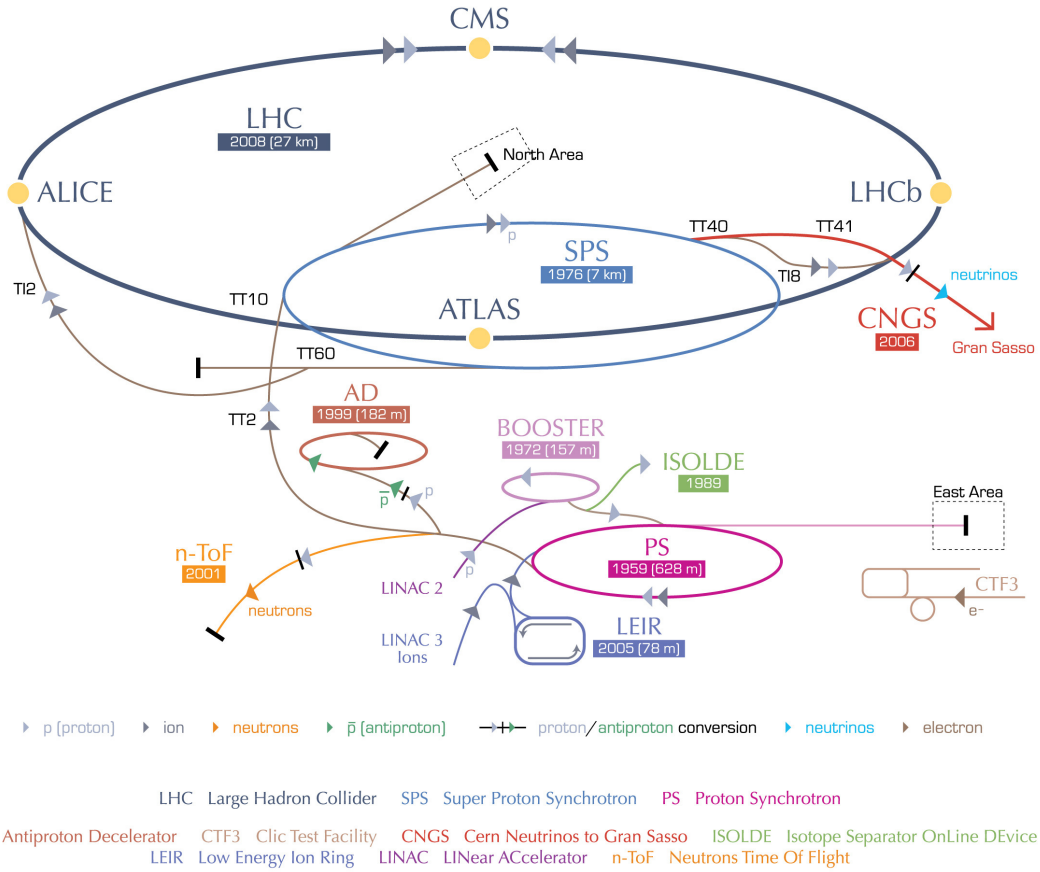


Figure 3.1: The LHC and its experiments within the CERN accelerator complex. Also shown is the network of accelerators through which the protons pass before reaching the LHC, including the PS and the SPS. Figure taken from Ref. [48].

The LHC design parameters allow for proton-proton collisions to occur at a maximum centre-of-mass energy of 14 TeV with an instantaneous luminosity of up to $1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. Prior to being injected into the LHC beam pipes, bunches of protons are accelerated in several stages. First, protons are obtained by stripping the electrons from hydrogen atoms using a strong electric field. The protons produced are then accelerated up to an energy of 50 MeV by Linear Accelerator 2 (LINAC 2). LINAC 2 leads into the Proton Synchrotron Booster (PSB), where the protons reach an energy of 1.4 GeV before passing into the Proton Synchrotron (PS). Here the beam

is accelerated up to 25 GeV before being transferred to the Super Proton Synchrotron (SPS). The SPS is the last step before the protons enter the LHC itself at an energy of 540 GeV. Therefore the LHC is responsible for the final acceleration from 540 GeV to the full energy. Thus far, the highest energy reached for stable operation is 6.5 TeV per beam; the full design energy of 7 TeV per beam is expected to be achieved in future, beginning in 2021.

The key components of the LHC are its 1,232 main dipole magnets, 392 main quadrupole magnets and 16 radiofrequency (RF) cavities. Superfluid helium cools the dipole magnets to 1.9K, at which temperature they produce the 8.3T magnetic field required to keep the beams in circular orbit. Quadrupole magnets are used principally to focus the beams near the interaction points, which increases the probability of a high-energy proton-proton collision. The RF cavities deliver an accelerating field of 5MV/m at a frequency of 400MHz, and furthermore maintain the shape of the 2808 proton bunches per beam. Collisions occur at the four interaction points, where bunches are induced to collide at a frequency of 25ns.

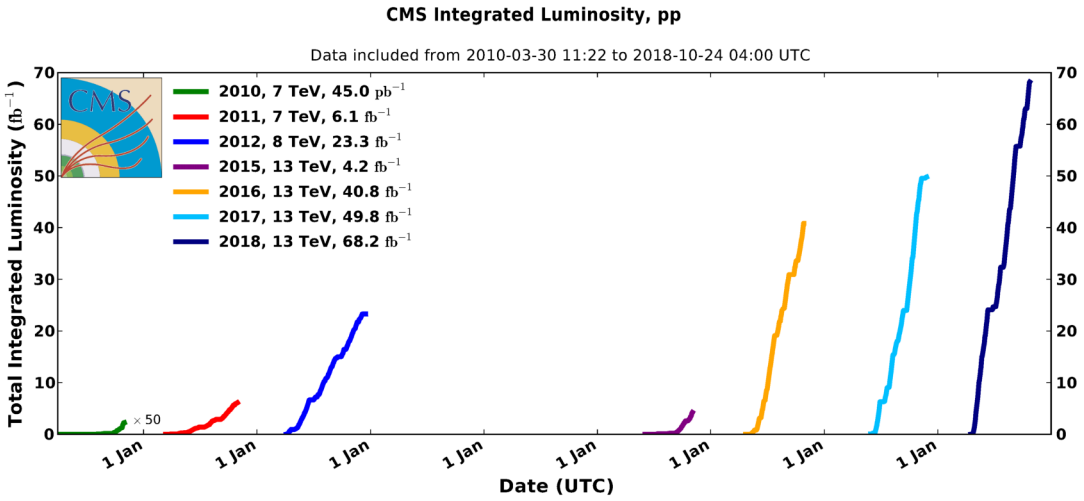


Figure 3.2: The integrated luminosity and centre-of-mass energy for each year of LHC operation. Figure taken from Ref. [54].

The operation of the LHC to date has comprised two separate runs. Run 1 commenced in 2010 with $\sqrt{s} = 7\,\text{TeV}$, continuing into 2011 at the same centre-of-mass energy to give a total of $6.1\,\text{fb}^{-1}$ of data. In 2012 this was increased to 8 TeV, and a total of $23.3\,\text{fb}^{-1}$ of data were collected. Analyses based upon this Run 1 dataset were able to discover the Higgs boson [12, 14]. After a shutdown for upgrades to the machine, Run 2 ran from 2015 to 2018 at a constant 13 TeV centre-of-mass energy. The LHC was able to exceed its design luminosity in each of the years, eventually levelling the luminosity at $2 \times 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ for the majority of 2018 operations. A number

of additional inelastic proton-proton collisions in addition to a collision of interest, known as pileup, occur at a rate which increases with the instantaneous luminosity. The mean number of pileup events was 23 per bunch crossing in 2016, and 32 in both 2017 and 2018 data-taking. Figure 3.2 summarises the data collected in each year, demonstrating the large increase in integrated luminosity obtained during Run 2. The values shown in the figure refer to the integrated luminosity delivered to the CMS detector. The integrated luminosity actually recorded by the detector is slightly lower, with typical efficiencies of slightly more than 90%.

3.3 The CMS detector

CMS is a general purpose detector [13] designed to capture the full range of objects required to achieve the goals of the LHC physics programme. The apparatus must be able to cope with high pileup at a very high rate; around 1000 charged particles are produced every 25 ns. Fine spatial and temporal granularity is required to obtain the sufficiently low occupancy necessary to perform well in these conditions. The resulting quantity of detector channels presents challenges for the detector electronics. Furthermore, the detector and its electronics must withstand high doses of radiation and fluence.

In addition to these technical constraints, CMS has four key physics requirements. These are summarised as:

- **Muon identification and resolution:** muons must be identified and well-measured over a wide range in energy and angle, with good dimuon mass resolution and the capability to accurately determine their charge. This is very important for Higgs boson measurements, particularly in the ZZ decay channel.

- **Charged particle measurement:** good momentum resolution along with efficient reconstruction and triggering on τ leptons and b quarks. Excellent jet reconstruction is essential in the majority of physics measurements.

- **Electromagnetic energy resolution:** good energy resolution and isolation for electrons and photons, whilst maintaining high acceptance. These properties are vital for Higgs boson measurements, especially in the diphoton decay channel.

- **Missing energy resolution:** accurate missing energy calculation and mass resolution of dijet objects. Missing energy is the key signature in many analyses searching for supersymmetry.

The design of CMS fulfils each of these requirements. The hermetic detector is cylindrical in shape, housing the eponymous 4 T solenoidal magnet in the central section. Its other key components include the silicon tracker, homogeneous crystal electromagnetic calorimeter (ECAL) and a sampling hadronic calorimeter (HCAL). The tracker and calorimeters lie inside the bore of the magnet coil; the muon detection system is instead interleaved with the return yoke. Each of these subsystems is displayed in Figure 3.3, and described in detail in the remainder of this chapter.

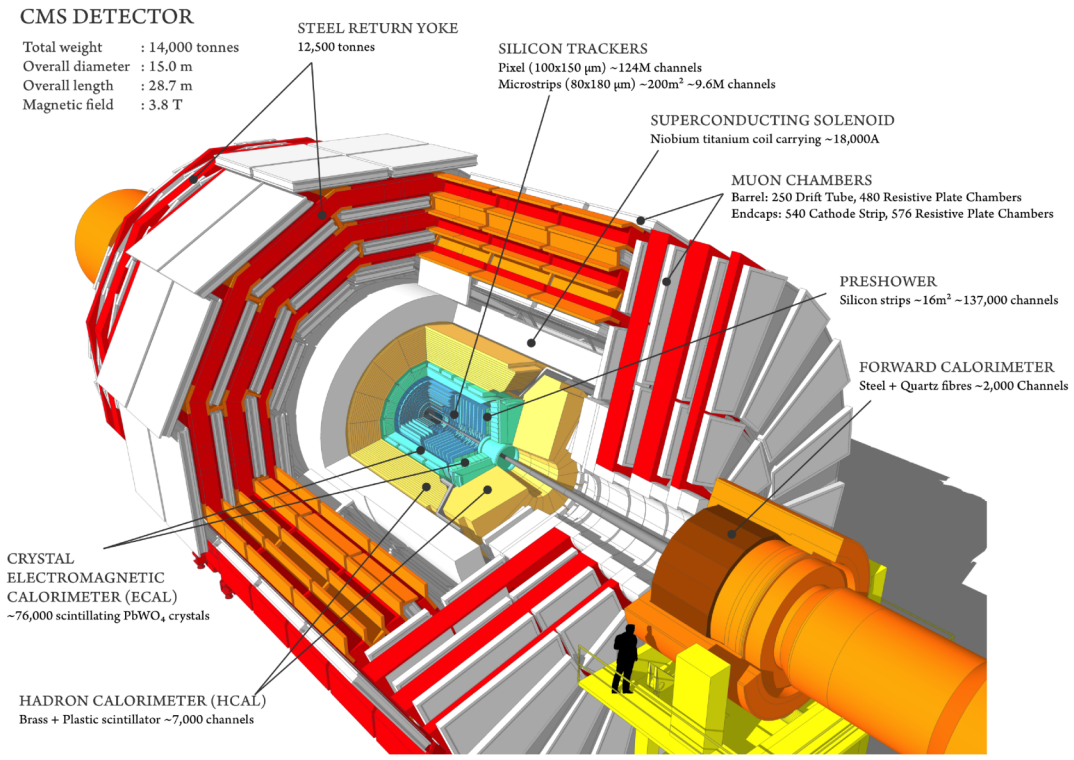


Figure 3.3: A schematic view of the CMS detector. Part of the detector is cut away in order to display each subsystem. Figure taken from Ref. [55].

A cylindrical coordinate system is used to describe events within CMS, with its centre at the nominal proton-proton interaction point. The central part is referred to as the barrel, with two endcaps closing the cylinder. The positive z-axis is defined to be parallel to the beampipe, anti-clockwise when viewed from above; this defines a right-handed coordinate system with the positive x-axis in the direction of the centre of the ring and the positive y-axis pointing vertically upwards. The polar angle θ is measured from the z-axis, and the azimuthal angle ϕ from the projection onto the plane perpendicular to the z-axis. Other common quantities are: pseudorapidity, defined as $\eta = -\ln\tan(\theta/2)$; momentum in the direction transverse to the plane of the LHC, p_T;

and the magnitude of the negative vector sum of particle momenta in the transverse plane, E_T^{miss}. High pseudorapidity values, corresponding to a direction close to the beampipe, are referred to as being forward.

3.3.1 Solenoid

The central feature of the CMS detector is the superconducting solenoid, whose 220 tonne cold mass is maintained at 4.5 K. The coil of the magnet measures 6 m in diameter, is 13 m long, and provides a maximum field strength of 4 T, storing a total energy of 1.6 GJ. During normal operation, however, the current is reduced to approximately 95% of its capacity, resulting in a field strength of 3.8 T. This ensures the longevity of the solenoid. The magnetic flux is returned by a 10,000 tonne yoke, composed of three layers in both the barrel and endcaps.

The extremely high bending power of the CMS magnet enables precise measurement of charged particle tracks , and hence excellent momentum resolution up to very high particle momenta. Its layout drives the design of the rest of the detector.

3.3.2 Tracking

The innermost subsystem of CMS is the silicon tracker, which is 5.8 m long and 2.5 m in diameter [13, 56]. Proximity to the beampipe implies the tracker must withstand a high radiation dose over its expected 10-year lifetime without degradation in performance. It is responsible for reconstructing the curved tracks produced when charged particles are deflected by the magnetic field. Furthermore, it is used to identify interaction vertices; this applies both for locating the primary vertex of the hard interaction and for identifying secondary vertices, produced for example in decays of b quarks. The tracker must be sufficiently granular to disambiguate tracks not only from the primary vertex but from additional pileup vertices. Its response must also be fast enough to correctly assign tracks to the correct bunch crossing. Finer granularity and precise timing capability require a high density of on detector electronics, and corresponding cooling capability. This results in an increased material budget in front of the detector calorimetry, which adversely affects the achievable energy resolution; a compromise between the two objectives is necessary. The final all-silicon design satisfies the requirements above, and is displayed in Figure 3.4.

The CMS tracker design comprises an inner pixel detector and a strip detector, covering the pseudorapidity range $|\eta| < 2.5$. The pixel detector provides precise two-dimensional (2D) space measurements in the ϕ and z directions. In the barrel, there are three modules arranged in cylindrical layers at radii of 4.4, 7.3, and 10.2 cm. Two disc layers are present in each endcap, ensuring the presence of at least three tracking points

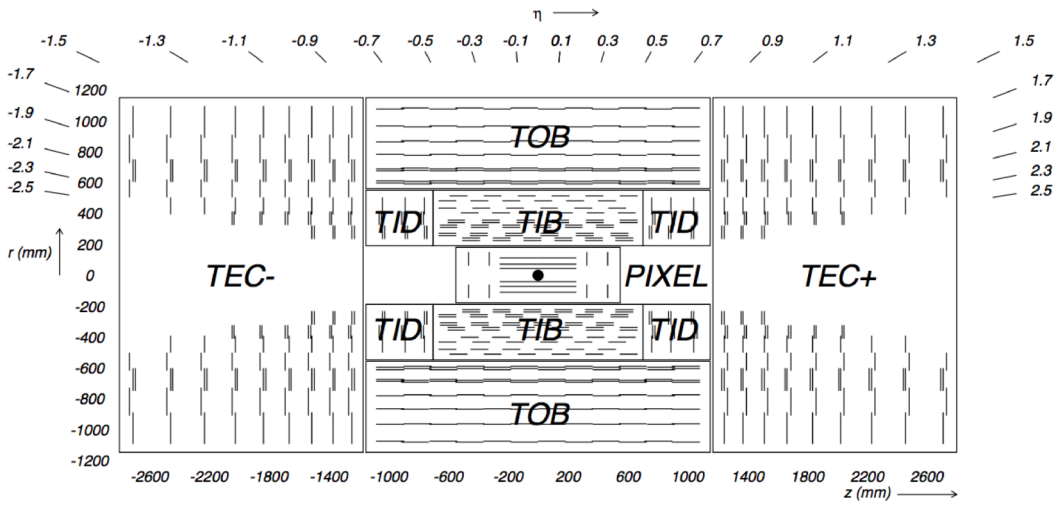


Figure 3.4: The various sections of the CMS tracking system. The black circle represents the nominal interaction point. Abbreviations are described in the following text. Figure taken from Ref. [13].

across almost the full range in η. Each cell measures 100 µm × 150 µm, resulting in a total of 66 million pixels covering an area of 1 m^2. This results in an optimal position resolution of 10 and 20 µm in the transverse and longitudinal direction respectively.

Beyond the pixel detector, the strip tracker provides one-dimensional measurements using 320 and 500 μm thick silicon micro-strip sensors. The strip tracker consists of three separate subcomponents, as illustrated in Figure 3.4. The Tracker Inner Barrel and Disks (TIB and TID) together consist of four barrel layers and three endcap disks, covering a radius between 20 and 55 cm. The TIB/TID provide as many as four ϕ measurements, depending on the track pseudorapidity. Adjacent to the TIB/TID is the Tracker Outer Barrel (TOB), which extends up to 116 cm in radius and 118 cm in z. A further 6 ϕ measurements are made in the TOB. In both the TIB/TID and the TOB, the first two layers have a second strip module mounted back-to-back with the first, permitting a 2D measurement in $\phi - z$. Finally the Tracker EndCaps (TECs) complete the CMS tracking apparatus. Each has 9 disks providing a ϕ measurement, with the first, second and fifth containing a second module to facilitate 2D $\phi - z$ measurements. In sum therefore, the tracker performs approximately nine measurements in the range $|\eta| < 2.5$, of which around four are 2D. It achieves a momentum resolution of better than 2% for high p_T (100 GeV) charged particle tracks up to $|\eta| = 1.6$. In the H $\rightarrow \gamma\gamma$ analysis, the tracker plays an important role in identifying the diphoton interaction vertex. The low material budget also maintains the excellent intrinsic photon energy resolution of the ECAL.

During the year-end technical stop between 2016 and 2017, an upgraded pixel detector was installed [57]. The motivation for the upgrade was to maintain or improve existing performance despite the greater than expected instantaneous luminosities provided by the LHC. The original pixel detector suffered from data loss at high occupancy, lower efficiency at high pileup, and degradation in performance due to radiation damage. Key improvements in the upgraded version include a fourth layer, a new readout chip, slightly reduced material budget and greater radiation hardness. The installation was successful and smooth operation was reached during 2017 data-taking. Performance in several areas, such as b-tagging, is enhanced as a result. However the overall impact on the $H \rightarrow \gamma\gamma$ analysis is relatively small.

3.3.3 Electromagnetic calorimeter

The CMS ECAL is a homogeneous detector which uses lead tungstate ($PbWO_4$) crystals as the active material [13, 58]. A total of 61,200 crystals cover the barrel area (EB), with 7,324 in each of the endcaps (EE). Readout is performed by silicon avalanche photodiodes (APDs) and vacuum phototriodes (VPTs) in the barrel and endcap regions respectively. An additional preshower detector is present in front of the endcaps, in order to improve π_0 rejection. The total coverage in pseudorapidity reaches $|\eta| = 3$; full energy resolution is maintained to $|\eta| = 2.5$. The ECAL is the most important sub-detector in the $H \rightarrow \gamma\gamma$ analysis, and the ECAL design itself was driven by the need to precisely reconstruct the two photons produced in Higgs boson decays. Hence this section is dedicated to describing the ECAL in some detail. The overall structure is shown in Figure 3.5, which illustrates the multiple subsystems composing the ECAL.

The choice of $PbWO_4$ was motivated by the need for the active material to be fast, dense, and radiation hard. The short response time is necessary to collect the energy deposit before the next bunch crossing; $PbWO_4$ crystals emit approximately 80% of the total scintillation light in 25 ns. Their high density ($8.28\,\mathrm{g/cm^3}$) facilitates the compactness and granularity of the detector without compromising on the containment of electromagnetic showers. A radiation length of 0.89 cm means the 22 cm barrel (23 cm endcap) crystals cover a $24.7\,X_0$ ($25.8\,X_0$) in the longitudinal direction. Lateral containment is also excellent, with a Molière radius[1] of 2.2 cm.

The EB extends to $|\eta| = 1.479$, as shown in Figure 3.6. Each crystal has a tapered shape which covers 0.0174×0.0174 in $\eta - \phi$, which is around $22 \times 22\,\mathrm{mm^2}$ at the front face and $26 \times 26\,\mathrm{mm^2}$ at the back. The geometry is "pseudo-projective"; the axis of each crystal is rotated by approximately 3° in both η and ϕ relative to the vector from

[1]The Molière radius is defined as the radius containing on average 90% of a shower's total energy deposition

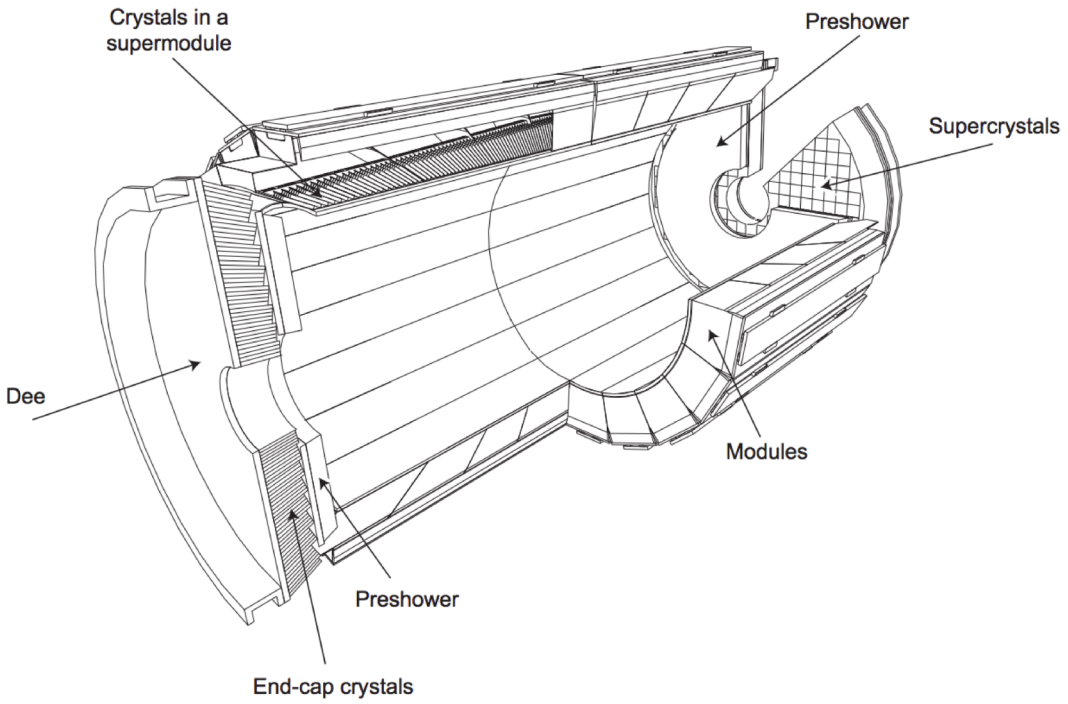


Figure 3.5: A schematic view of the CMS ECAL showing the various subsystems. Part of the detector is cut away for clarity. Figure taken from Ref. [58].

the interaction point. This ensures that a particle cannot follow a trajectory entirely within a crack in the detector. Crystals are then grouped into supermodules, each covering 20° in ϕ.

Each EE comprises two so-called "Dees", built from 5-by-5 supercrystals, reaching $|\eta| = 3$. The crystals in the EE are slightly shorter and wider than in the EB, and the offset angle varies between 2° and 8°. The preshower detectors (ES) in front of the endcaps are sampling calorimeters, each composed of two layers of lead with silicon strip sensors behind. Measurements of both energy and transverse shower shape are made by the ES. The setup as a whole is displayed using one quadrant of the detector in Figure 3.6.

To reconstruct the shower energy, the photodetectors amplify the scintillation light to produce around 4,500 photoelectrons per GeV. After digitisation by 12-bit analogue-to-digital converters (ADCs), ten consecutive amplitude measurements are stored in a buffer. Once a trigger is received, these recordings are sent to the off-detector electronics. There the original energy deposition in the crystal can be inferred with knowledge of the typical pulse shape of each channel.

The intrinsic energy resolution (σ_E/E) of the ECAL can be modelled by the equa-

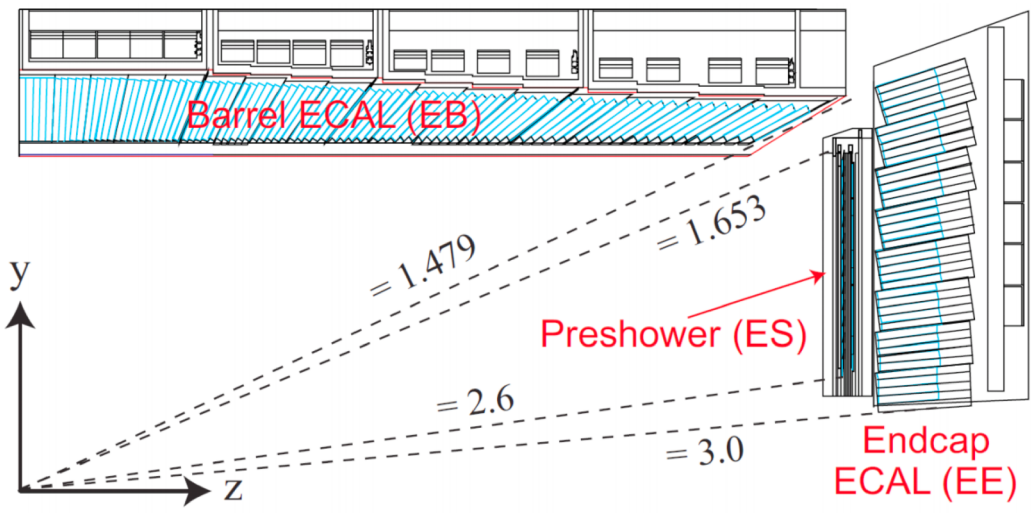


Figure 3.6: The structure in pseudorapidity of the ECAL. Figure taken from Ref. [58].

tion

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2, \tag{3.1}$$

where S is the stochastic term, N is the noise term and C is the constant term. For energy in units of GeV, the values in beam tests are found to be: $S = 2.8\%$, $N = 12\%$, and $C = 0.3\%$ [13]. Here energy values are computed simply summing 5×5 arrays of crystals. In real conditions, further corrections are applied to achieve similar performance to that in beam tests.

A series of calibrations are required in order to generate accurate shower energy estimates. Furthermore, not all the energy from a given shower will be contained within one crystal. Typically showers are spread in the ϕ direction due to the effect of the magnetic field on particles produced via secondary interactions. Therefore nearby crystals are merged into "superclusters" (SC), which are then used to estimate the energy of the candidate object using the equation [58]:

$$E_{e/\gamma} = F_{e/\gamma}\, G\, \Sigma_i S_i(t) C_i A_i + E_{ES} \tag{3.2}$$

where the individual channel amplitudes (A_i), corrected for variations in time ($S_i(t)$) and by channel (C_i), are summed and multiplied by the global ADC-to-GeV conversion factor (G). After any contribution from the preshower detector is added (E_{ES}), a final correction ($F_{e/\gamma}$) for effects due to upstream material, geometry, and imperfect clustering is applied. The $S_i(t)$ monitor the radiation-induced change in transparency

over time by measuring the response of each crystal to 440 nm laser light every 40 minutes. The intercalibration factors C_i reflect the need to have an equal response from each crystal in the detector. These are computed using both the $\phi-$symmetry of the detector and the known mass of π^0 and η diphoton decays. The global scale G is derived by requiring the measured dielectron mass of the Z boson to match its known true value. Finally, the particle-dependent $F_{e/\gamma}$ corrections are estimated using a multivariate regression which mostly accounts for the varying material budget in front of the ECAL [58]. After this lengthy procedure, the final energy resolution on the high p_T, unconverted photons which drive the $H \rightarrow \gamma\gamma$ sensitivity is approximately 1%.

3.3.4 Hadronic calorimeter

Surrounding the ECAL is the HCAL, a brass and scintillator sampling calorimeter covering the entire region $|\eta| < 5$ [13, 59]. The HCAL measures the energy of neutral hadrons, and plays an important role in the reconstruction of jets and E_T^{miss}. Four subdetectors form the full calorimeter, as illustrated in Figure 3.7.

The barrel (HB) covers the central region $|\eta| < 1.3$. Wavelength-shifting fibres contained within scintillator tiles channel emitted light to photodetetors. Division into 16 η sectors and 16 wedges in ϕ results in a granularity of 0.087×0.087. The HB has 14 brass layers surrounded by a layer of steel at the front and back, which totals between 5.8 and 10.6 interaction lengths (λ_I) depending on η [2]. This is not sufficient to completely contain hadronic showers, due to the space restriction imposed by the magnet coil. Consequently an additional outer hadron calorimeter (HO) is present beyond the solenoid. The HO uses the solenoid coil as an absorber, with the same scintillator technology as the HB. Once the HO is taken into account, the minimum depth becomes $11.8\,\lambda_I$, which significantly reduces energy leakage.

The hadronic endcaps (HE) reach across a wide pseudorapidity range, covering $1.3 < |\eta| < 3.0$. This more forward region requires highly radiation hard material, for which reason brass is chosen. The HE design minimises the size of the barrel-endcap transition region, and results in a depth of at least $10\,\lambda_I$. Its granularity matches the HB up to $\eta = 1.6$, but reduces to 0.17×0.17 beyond that.

Finally, the very forward region contains the HF (forward calorimeter). Its design is driven primarily by the need to withstand an extremely high radiation dose. Good performance must also be maintained, since the HF plays an important role in forward jet reconstruction, which occurs for example in VBF $H \rightarrow \gamma\gamma$ events. The chosen

[2] An interaction length is defined as the mean free path, or the mean distance travelled, of a hadronic particle before it undergoes an inelastic nuclear interaction in a given material.

active material is quartz fibre, with a steel absorber structure. The fibres generate light via Cherenkov radiation, which is detected by photomultipliers. Fibres are bundled together with a granularity of 0.175×0.175. Two different lengths of fibre are utilised; around half run the full length of the HF, whilst the other half begin at a depth of 22cm from the HF front face. This enables electromagnetic showers to be disambiguated from hadronic showers.

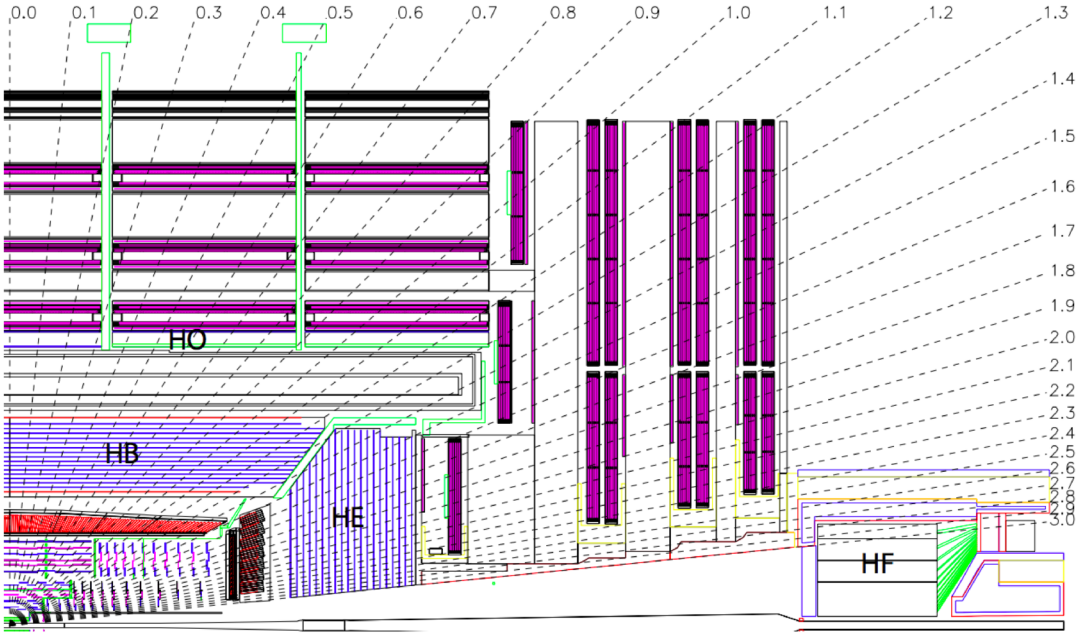


Figure 3.7: The structure in pseudorapidity of the HCAL. Figure taken from Ref. [13].

3.3.5 Muon system

The muon system at CMS is composed of three separate gaseous chamber detectors [13, 60]. The system is required to identify muons, precisely measure their momentum, and inform the triggering decision. Good muon performance is crucial to the success of CMS, particularly for the very powerful $H \to ZZ^* \to 4\mu$ decay mode. Interleaved with the return yoke of the CMS magnet, the muon system comprises a cylindrical barrel system and two endcap planes, covering a total area of 25,000 m^2.

The barrel drift tube (DT) covers the central region up to $|\eta| = 1.2$ with four layers, known as stations, between layers of the return yoke. Drift chambers containing a mixture of Ar and CO_2 gases are used; this is possible due to the low rate and uniform magnetic field in the barrel region. Each station has eight chambers providing a measurement of the ϕ direction and four measuring the z coordinate. At high p_T, the identification efficiency using the DT alone is over 95%.

In the endcap region there is both a higher muon rate and higher neutron background, which necessitates the use of a more radiation hard technology. The four Cathode Strip Chamber (CSC) stations overlap with the DT from $0.9 < |\eta| = 1.2$. Beyond this, a muon will pass through at least three CSCs between $\eta = 1.2$ and $\eta = 2.4$. The CSCs have fast timing capability, which enables them to inform the trigger and also correctly identify the bunch crossing with efficiency greater than 99%.

The final component of the muon system is the Resistive Plate Chamber (RPC) system. The six RPC layers are embedded within the DT but operate independently. Position resolution is coarser, but the timing resolution is fast. This facilitates a complementary trigger input, as well as helping to remove track ambiguities arising from multiple hits in a single chamber.

Overall, the muon system alone achieves a momentum resolution of around 10% for central muons up to $p_T = 200\,\text{GeV}$. This worsens to between 15% and 40% for 1 TeV muons, dependent on pseudorapidity. Once combined with the information from the tracker however, this improves to 5% in the barrel and 10% in the endcaps.

3.3.6 Trigger system

The rate of information produced at CMS is extremely high; the bunch crossings occurring at rate of 40 MHz each contain approximately 1 MB of data [61]. It is unfeasible to read out 40 TB of data per second, so a drastic reduction in the event rate must be applied. This is achieved by ignoring the large majority of events where the protons collide at a relatively low angle and no high energy objects are produced as a result. The system responsible for this rate reduction is known as the trigger.

The CMS trigger runs in two sequential stages: first the Level-1 Trigger (L1T) and then the High-Level Trigger (HLT). Since the maximum output rate of the detector electronics is 100 kHz, the L1T is responsible for a factor 400 decrease in the number of events selected. The L1T must make a decision on an event within approximately $3.2\,\mu s$, and therefore only uses a coarse readout of data from the calorimeters and muon system. The full resolution interpretation of an event is buffered until the final trigger decision is made. Simple algorithms, for example the energy sum from arrays within a calorimeter, are used to identify high energy objects of interest. The computations necessary are performed mostly by customised, reprogrammable electronics known as Field Programmable Gate Arrays (FPGAs). Once a decision has been made, if the event passes the L1T selection, the full readout is transmitted to the HLT.

The HLT is a software system which runs on a single farm of commercially available processors. It must further reduce the event rate to 1 kHz, which corresponds to a factor of 100. As the HLT has access to the full event information, sophisticated

reconstruction algorithms similar to those used offline are employed. It must maintain selection efficiency whilst maintaining an acceptable rate and CPU-time. Typical quantities evaluated to inform HLT decisions include invariant mass, object isolation, and track information. Finally, once the HLT has accepted an event, the full readout is saved to disk. The event is then processed by the full offline CMS software, which produces objects ready for physics analysis.

3.4 Summary

The CMS detector is hermetic detector with a 3.8 T solenoidal magnet at its core. Its tracking system, electromagnetic and hadronic calorimeters, and muon system together enable highly efficient, precise reconstruction of a wide range of final-state objects. Furthermore, its trigger system enables the rate of collisions requiring further processing to be reduced to a manageable level. The lateral structure of CMS is summarised in Figure 3.8, which shows the characteristic paths followed by the different types of particles produced in LHC proton-proton collisions and how they interact with the different subdetectors. The importance of each component of the detector in reconstructing the full range of particles present in CMS events is illustrated.

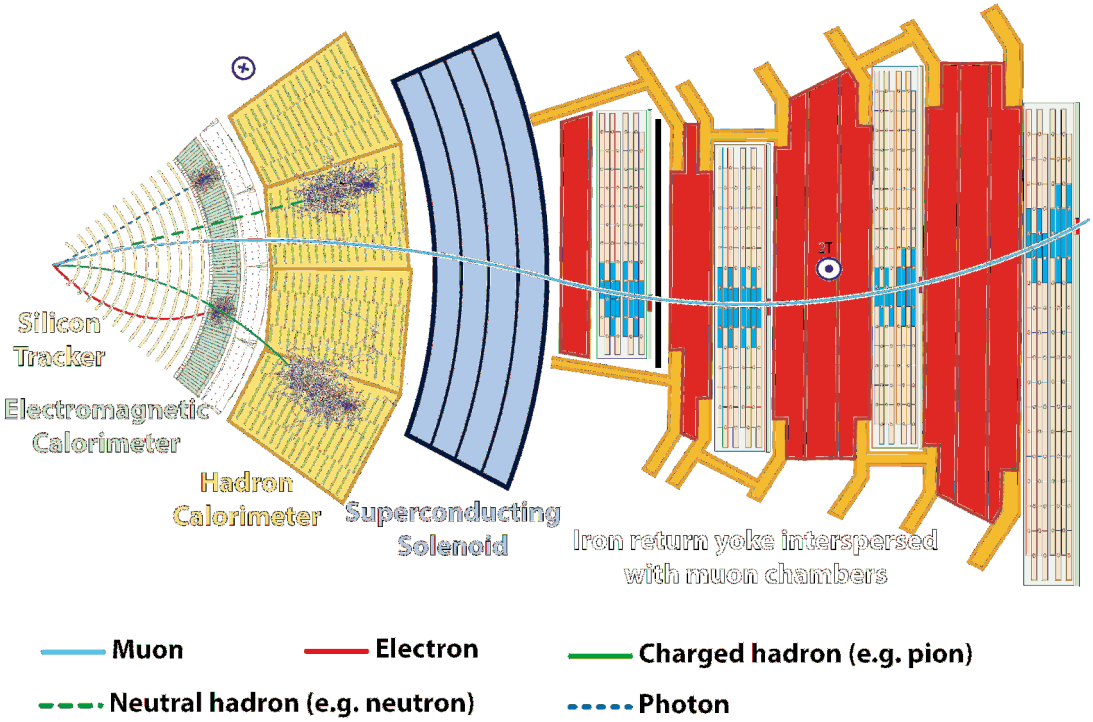


Figure 3.8: A cross-sectional view of a slice of the CMS detector, illustrating the characteristic signatures of different types of particle. Figure taken from Ref. [62].

Chapter 4

The High Granularity Calorimeter

4.1 Introduction

In order to fully exploit the physics potential of the LHC, the accelerator will be operated until around 2040 in order to collect as much data as possible. During this period, various upgrades are planned which are designed to maximise the instantaneous luminosity delivered to the experiments, including CMS. As a result of this increase, and due to accumulated radiation damage, parts of the CMS detector will need to be upgraded. A key part of this upgrade program is the replacement of the electromagnetic and hadronic calorimeter endcaps with the high granularity calorimeter (HGCAL). In this chapter, the motivation for and design of the HGCAL is described. The development of reconstruction software is detailed, and the detector's physics potential is demonstrated using the example of the VBF H $\rightarrow \gamma\gamma$ analysis.

4.2 The High Luminosity LHC

Run 2 of the LHC is now complete, with over 190 fb^{-1} of data delivered at $\sqrt{s} = 13\,\text{TeV}$. This exceeds the original target of 150 fb^{-1}, and was achieved partly because the machine was eventually operated at twice its nominal instantaneous luminosity. The instantaneous luminosity reached values as high as $2 \times 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ during 2018. The second long shutdown (LS2) commenced following the completion of Run 2 and will last until 2021. During this time various improvements to the LHC will be made, including a substantial upgrade to the injection system. The machine will also be readied for operation at the increased energy of 7 TeV per beam. However these upgrades will

not substantially affect the conditions experienced by the LHC experiments; the peak instantaneous luminosity is not envisaged to increase beyond $2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. As such no major changes are required to the CMS detector during LS2, although various improvements are planned; these include upgrades to the muon system and HCAL barrel. Therefore the expectation for Run 3, which commences in 2021 with two years of high-availability data-taking in 2022 and 2023, is that a further $150\,\mathrm{fb}^{-1}$ of data will be accumulated.

Beyond Run 3, the usefulness of running the LHC with its current parameters decreases. In order to reduce the statistical error on physics measurements by a factor of two, high-availability operation for more than ten years would be required. Therefore a major upgrade to the LHC is planned, referred to as the Phase 2 upgrade, to maximise its physics reach. The resulting High Luminosity LHC (HL-LHC) [63] will have a nominal levelled instantaneous luminosity of $5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, permitting a total of $3000\,\mathrm{fb}^{-1}$ of data to be collected by the mid-2030s. The current planned schedule of the future running of the LHC and HL-LHC is summarised in Figure 4.1.

4.3 The CMS upgrade

At the HL-LHC, the mean pileup per bunch crossing is expected to be 140; however an additional 50% beyond the nominal value is allowed for in the HL-LHC design, which would result in mean pileup values of up to 200. This constitutes a major change, and the environment will be significantly harsher than with the current LHC conditions. These conditions pose serious challenges to the detectors in terms of radiation tolerance and reconstruction in high pileup. In order to maintain or improve upon the excellent performance exhibited in Run 2, a suite of upgrades to the CMS detector is planned. The key aspects of the CMS Phase 2 upgrade can be summarised as follows [65, 66]:

- **Tracker:** the tracker will suffer significant radiation damage and must be entirely replaced for Phase 2. The upgraded tracker will have finer granularity, increased coverage in the forward region, and be much lighter, resulting in a reduced material budget. Furthermore, the design will allow track information to be included in the L1 trigger decision.

- **Endcap calorimeters:** the calorimeter endcaps (both ECAL and HCAL) will also be radiation-damaged by the end of Run 3, and will therefore be replaced. The replacement design, known as the high granularity calorimeter (HGCAL), will have both electromagnetic and hadronic sections. Fine segmentation in each of the longitudinal and transverse directions will facilitate precise measurements of showers in three dimensions.

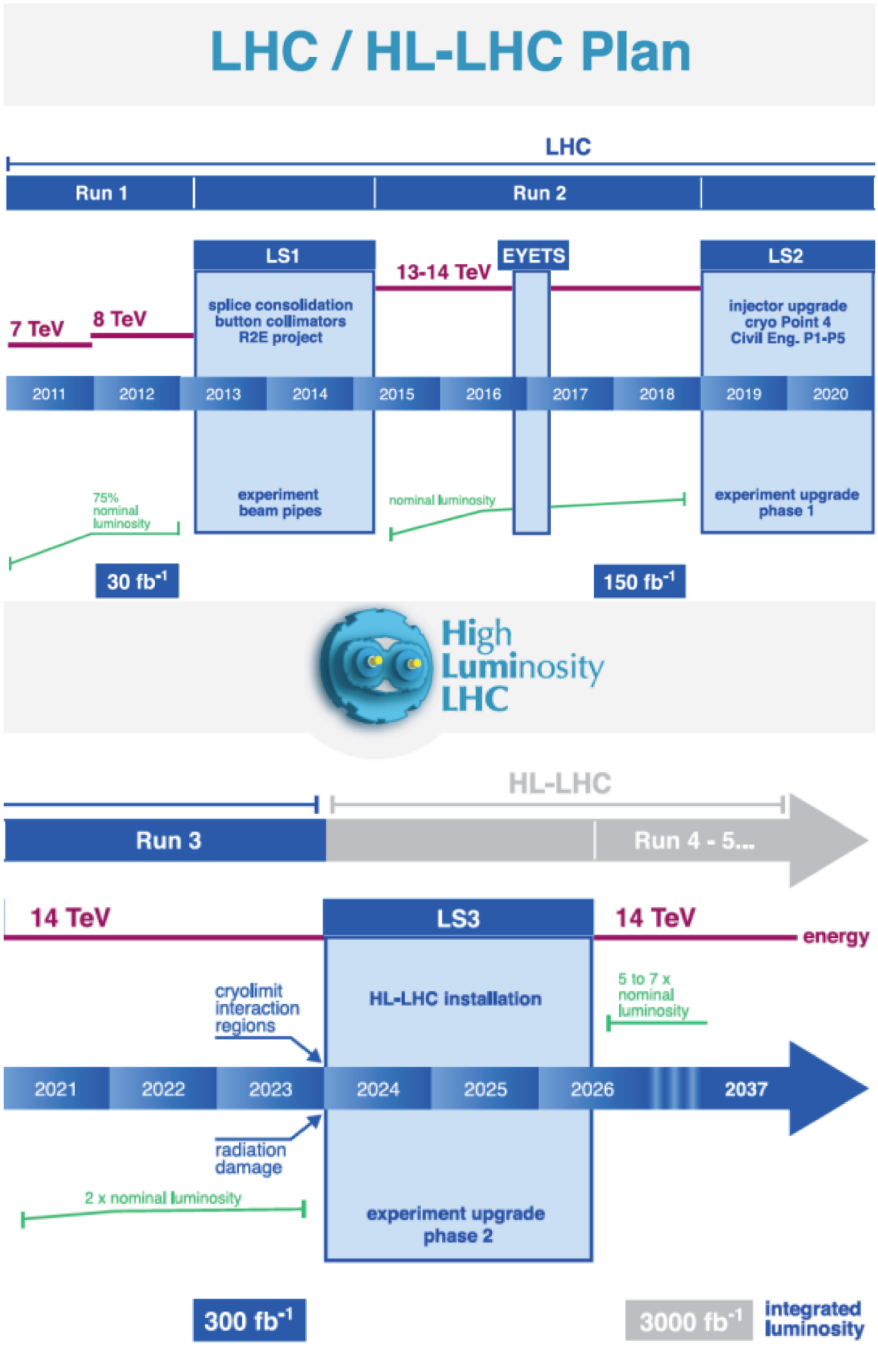


Figure 4.1: The planned schedule for the operation of the LHC and its high-luminosity upgrade. Figure taken from Ref. [64].

- **ECAL barrel:** the current ECAL detector electronics are not capable of meeting the stringent HL-LHC trigger requirements. The level of noise in the silicon avalanche photodiodes which detect the scintillation light will also increase due to

radiation damage. Therefore the electronics will be optimised and the operating temperature of the system reduced.

- **Timing:** precision timing measurements of objects will be made with the HGCAL and the upgraded ECAL. In addition, a dedicated timing layer providing precise timing information for minimum ionising particles will be added to the barrel and each of the endcaps. By enabling the reconstruction of vertex times, the capability for pileup rejection will greatly improve.

- **Muon endcaps:** the CSC muon system will be enhanced with additional stations. This will increase the forward coverage and maintain muon acceptance at the L1 trigger.

- **Trigger and data acquisition:** due to the tracker upgrade, the latency of the L1 trigger at Phase 2 is increased to 12.5 µs. Combined with upgrades to the front-end electronics of various subdetectors, this enables an increase of the L1 trigger acceptance rate to 500 kHz. Consequently the data acquisition system must also be upgraded to handle the increase in event rate, event size, and the complexity of high PU reconstruction.

Full details of the CMS upgrade scheme can be found in Refs [3, 66–71]. The remainder of this chapter is dedicated to describing the HGCAL in further detail, based on Ref. [3].

4.4 Requirements for the HGCAL

The primary challenge which drives the design of the HGCAL is the need to sustain physics performance in the extremely high pileup conditions foreseen at the HL-LHC. Simulations indicate that the HGCAL will be required to withstand up to 2MGy of total radiation dose, together with a maximum fluence of around $10 \times 10^{16}\,\mathrm{n_{eq}/cm^2}$. The inhomogeneous distribution of this dose is shown in Figure 4.2, which illustrates how the radiation dose is greatest near the beampipe. Studies performed in recent years have shown that silicon sensors and the associated electronics retain acceptable performance after exposure to this level of radiation, and have hence been chosen as the most reliable active material for the majority of the calorimeter. The remaining parts of the detector in lower-radiation regions will instead use cheaper plastic scintillator tiles with silicon photomultiplier (SiPMs) readout. These layers of active material are interleaved with the copper-tungsten alloy absorber in the front section of the calorimeter, and stainless steel absorber further back.

To maintain performance throughout the operation of the HL-LHC it is necessary to inter-calibrate cells to the level of a few per-cent. This is possible provided the

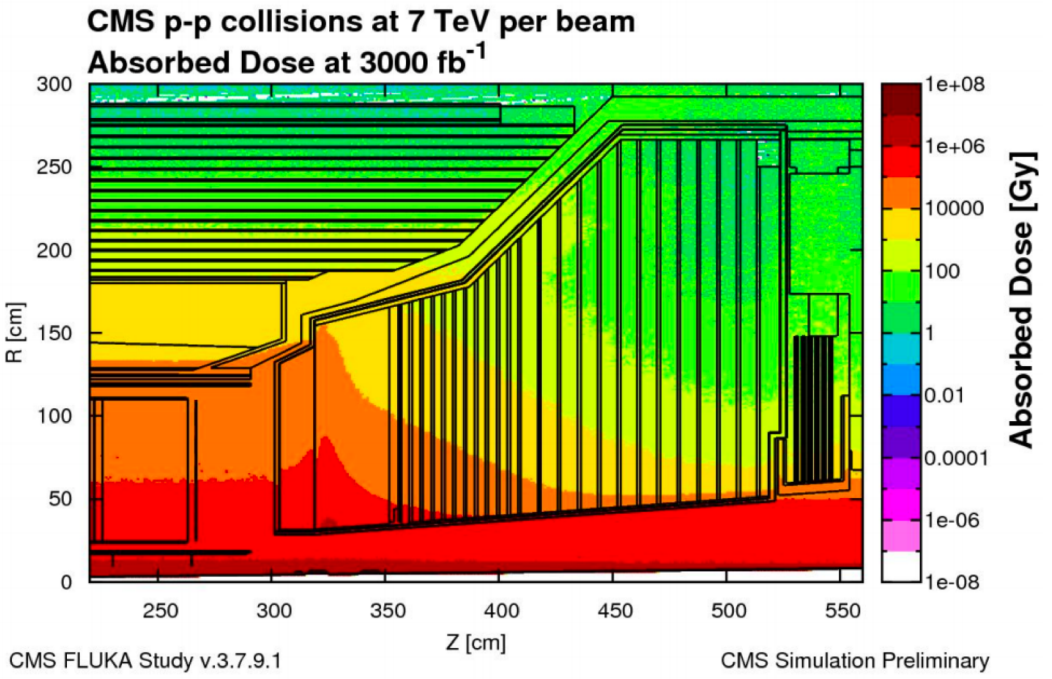


Figure 4.2: The dosage of ionising radiation the HGCAL is expected to absorb after a total integrated luminosity of 3000 fb^{-1}, as a function of detector radius and depth. The colour scale indicates the magnitude of the absorbed dose. Figure taken from Ref. [3].

signal-to-noise ratio (S/N) for minimum-ionising particles (MIPs) is sufficiently high. For this to be achieved after 3000 fb^{-1} of data have been collected, small silicon cells with low capacitance are required, which results in high lateral granularity. Fine lateral granularity also has many benefits for physics performance, including the ability to separate nearby showers, identify narrow jets such as those originating from the quarks produced in vector boson fusion (VBF) events, and minimise the amount of pileup entering energy measurements. To take advantage of this fine segmentation the calorimeter is required to be dense, thereby preserving the compactness of showers in the transverse direction. Similarly, fine longitudinal granularity facilitates precise energy measurements, as well as enabling discrimination between different types of shower using the depth profile. These features are particularly important within the CMS particle flow reconstruction paradigm [72]. A detector producing three-dimensional images of showers would provide powerful separation of electromagnetic, charged hadronic and neutral hadronic components, facilitating more precise energy measurements, particularly of jets.

In addition, the HGCAL should be able to perform precise timing measurements enabling excellent pileup rejection, and provide an input to the L1T decision. The

proposed design specification that captures all of these desired features is described in the following section.

4.5 Design

An overview of the HGCAL design is presented in Figure 4.3. The HGCAL is composed of an electromagnetic and a hadronic section, called the CE-E and CE-H respectively, covering the pseudorapidity range $1.5 < |\eta| < 3.0$. The CE-E comprises 28 layers with hexagonal silicon sensors as the active element. The total depth, including the neutron moderator layer at the front, is 34 cm, which corresponds to approximately 26 radiation lengths (X_0) and 1.7 interaction lengths[1] (λ). Three different thicknesses of silicon sensors are used, with thickness decreasing as a function of fluence. Absorbers are made of copper-tungsten alloy and copper plates are used for cooling. All layers of the CE-E are used for energy measurements, but alternate layers give inputs to the L1 trigger primitive formation.

The CE-H is formed of 12 layers with 35 mm thick stainless steel absorber and another 12 where the absorber thickness is 68 mm, contributing an additional $9\,\lambda$ in depth. The active medium in the CE-H varies as a function of depth and radius, and is determined by the radiation level. In regions of sufficiently low fluence (those which are nearest the back of the detector and furthest from the beam-pipe), plastic scintillator tiles are used with SiPM readout. The exact threshold between the scintillator and silicon is determined by the S/N required to measure the MIP response, which is decreased by exposure to radiation. Further detail on the design specifications of the HGCAL can be found in Ref. [3]. The results detailed in this chapter use the same detector geometry and central CMS software (version 9.3.7) as described in Ref. [3].

4.6 Reconstruction

4.6.1 Electromagnetic objects

The HGCAL's intrinsic performance measuring the energy of electromagnetic showers is modelled using a dedicated simulation with pileup corresponding to an average of 200 interactions per bunch crossing. Energy deposits in a radius of 26 mm around a single unconverted photon are summed to estimate the energy resolution, which is shown in Figure 4.4. The left plot shows that the resolution is approximately constant as a function of η, and robust against pileup, with only a very small degradation in res-

[1] A radiation length is defined as the mean distance over which a high energy electron will have its energy reduced to a fraction e^{-1} of the initial value.

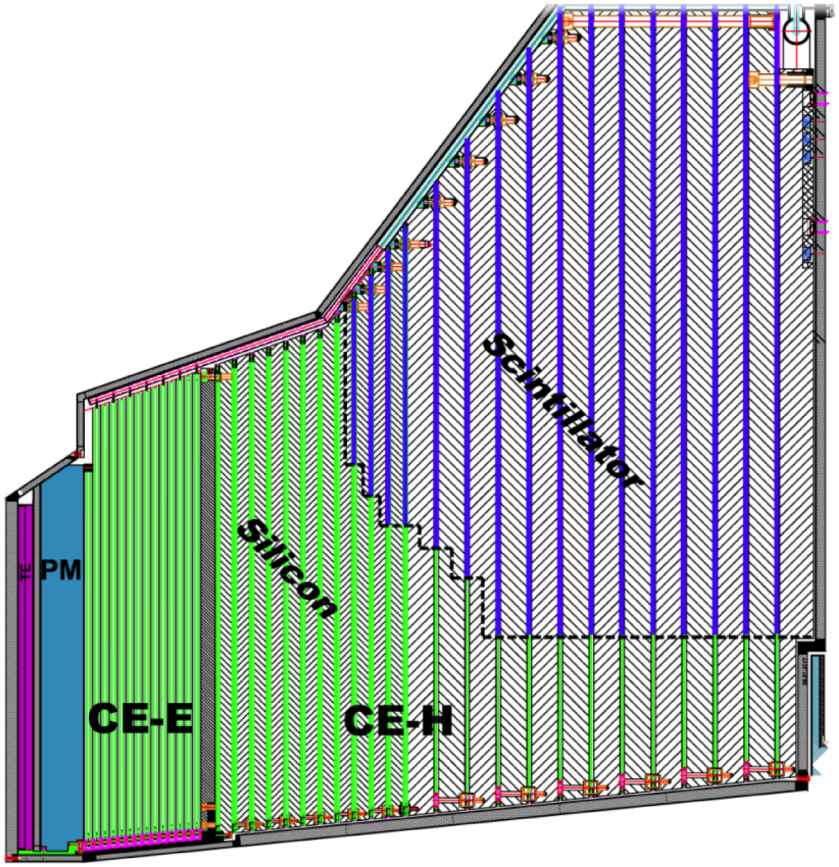


Figure 4.3: A schematic of the HGCAL's longitudinal structure. The endcap timing layer (TE, purple) is in the front of the HGCAL, nearest to the nominal interaction point. A layer of polythene that moderates the flux of neutrons (PM, light blue) also sits in front of the HGCAL. The active material of the electromagnetic part of the calorimeter (CE-E) is entirely composed of silicon (green), whilst the hadronic part (CE-H) uses a mixture of silicon and plastic scintillator (dark blue) as active material. Adapted from [3].

olution between PU 0 and PU 200. The intrinsic performance is further demonstrated by using this method to reconstruct unconverted photon pairs from simulated $H \rightarrow \gamma\gamma$ decays, where both photons are contained within the fiducial region of the HGCAL and the vertex location is assumed to be known exactly. The resulting diphoton mass distribution is shown in Figure 4.5, with resolution of around 1.8 GeV. This value is comparable to the expected resolution of the upgraded CMS barrel calorimeter, representing a substantial improvement relative to Run 2, where the endcap resolution is significantly worse than the barrel.

The HGCAL provides more detailed shower information than existing CMS detectors, and it is envisaged that eventually a sophisticated four-dimensional particle

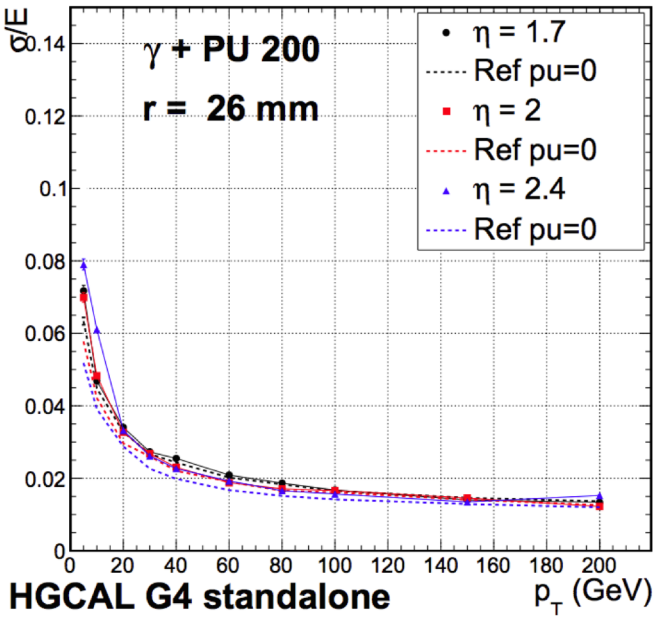


Figure 4.4: The intrinsic energy resolution for single photons in PU 200. The energy is estimated by summing all deposits within a 26 mm radius of the generated particle axis. Figure taken from Ref. [3].

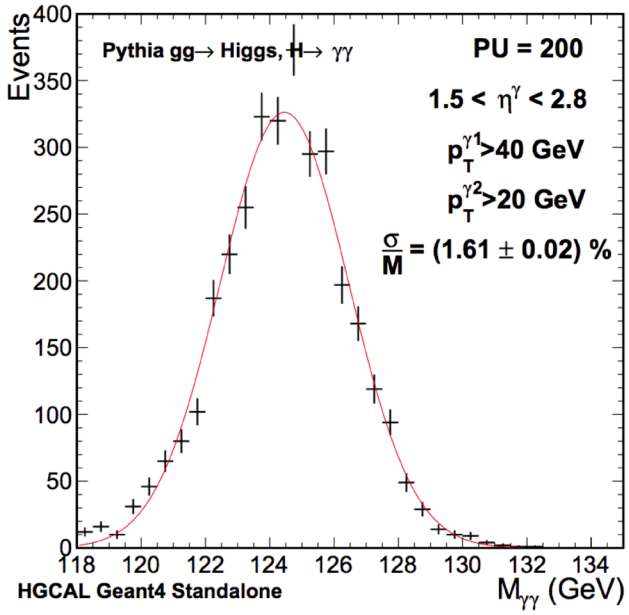


Figure 4.5: The intrinsic diphoton mass resolution of the HGCAL in simulated H $\rightarrow \gamma\gamma$ events where both photons are within the fiducial region of the HGCAL. Figure taken from Ref. [3].

flow approach will be used to incorporate as much of this information as possible. In the meantime, more straightforward approaches to reconstruction have been developed, in order to understand which approaches are feasible and to produce object and physics-level results that demonstrate the potential of the detector.

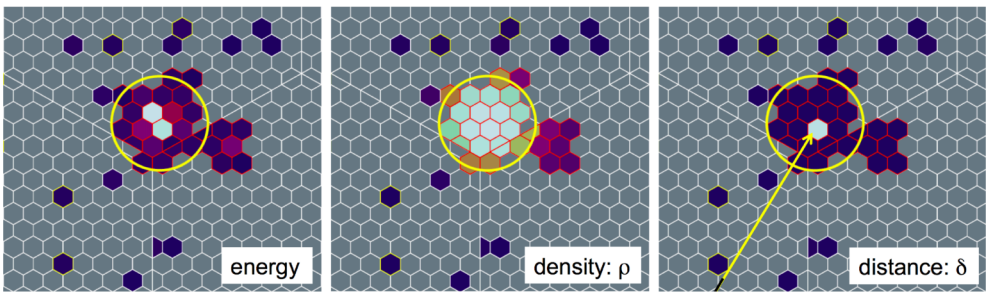
The current method begins by clustering hits in each two-dimensional (2D) layer independently, using an imaging algorithm [73]. The algorithm proceeds as follows, and is illustrated in Figure 4.6:

1. Construct an energy density map of all hits above a threshold E_c. The threshold value is defined as a function of the noise resolution (σ_{noise}), which depends on the cell type. The density is defined simply as the energy sum of all hits within a critical distance δ_c, whose value is chosen to be similar to the Molière radius.

2. For each hit, calculate the distance to the nearest hit with higher density.

3. Assign hits with both density and distance parameters greater than threshold values as cluster centres. The density threshold (ρ_c) can be defined relative to the maximum density in the event, or as a function of the noise resolution. The distance threshold is equal to δ_c.

4. Form clusters by assigning each hit to the same cluster as the nearest hit with higher density.



(a) Energy **(b)** Density, ρ. **(c)** Distance, δ.

Figure 4.6: Illustration of the procedure used to form layer clusters for electromagnetic objects in the HGCAL. The event considered is a single photon with $p_T = 35\,\text{GeV}$. The colour scale represents only the difference between hits, and is in arbitrary units. The yellow circle indicates the size of the Molière radius and the yellow arrow points to the hit chosen as a cluster centre. It can be seen that many hits in a genuine cluster have a high density, whereas all but one will have a low distance to the nearest higher energy hit; requiring high values of both these parameters is therefore an appropriate way to define a cluster centre.

There are therefore several parameters in the process which can be optimised: E_c, δ_c, and ρ_c. To form a suitable object on which to optimise, a more realistic

object must be formed. Therefore the 2D clusters, or layer clusters, produced are associated together in depth to form so-called multiclusters. An axis is defined by the highest energy 2D cluster, then any layer clusters within a certain distance of this axis are added to the multicluster. Using this procedure with sensible parameters, an unconverted photon will have almost all of its energy contained within a single multicluster. Optimisation can then be performed by minimising the energy resolution of unconverted photons simulated at a range of p_T values. Further detail on the 2D clustering and multiclustering procedures is given in Ref. [2].

Finally, electromagnetic objects are formed using a superclustering procedure very similar to the one utilised in Run 2, where showers which have been spread out in the ϕ direction by the magnetic field are collected together. To perform this step, multiclusters are used as inputs to the existing Run 2 algorithm [74]; no re-optimisation is performed. Despite this, high p_T photons reconstructed in this way have a resolution of below 2%, not far from the intrinsic resolution values shown in Figure 4.4. This method of electromagnetic object reconstruction was used for all the physics results in Ref. [3], including the study described in Section 4.7.

Electrons defined in this way are used to test the ability of the HGCAL to discriminate between signal ($Z \to e^+e^-$) and background (QCD multijet) processes. Lateral and longitudinal shower shape variables, along with tracking information, are used as inputs to a classifier. Two examples are shown in Figure 4.7: the longitudinal development of the shower is indicated by the layer at which 10% of the total energy in the CE-E has been deposited, and the lateral spread of the shower evaluated using the pseudorapidity difference between the electron multicluster and the extrapolated track. For both variables, the distributions are robust to the increase in pileup and show good discrimination between the signal and background processes. For a 95% signal efficiency, the background efficiency is 1% for electrons with $p_T > 20\,\text{GeV}$, comparable to the Run 2 value. An improvement in performance is seen when the lateral and longitudinal shape variables are added to a classifier using solely track information, demonstrating the value of the HGCAL granularity to the object identification.

4.6.2 Hadronic objects

The multiclustering procedure is not found to work sufficiently well for hadronic showers. This is due to the fact that hadronic showers are less well-contained, and also have more variation in transverse and longitudinal structure than electromagnetic showers. Instead, single hadrons were reconstructed using a so-called megaclustering procedure, where layer clusters within a truncated cone are combined to form the object. For these more dispersed hadronic showers, the resolution substantially improves once the

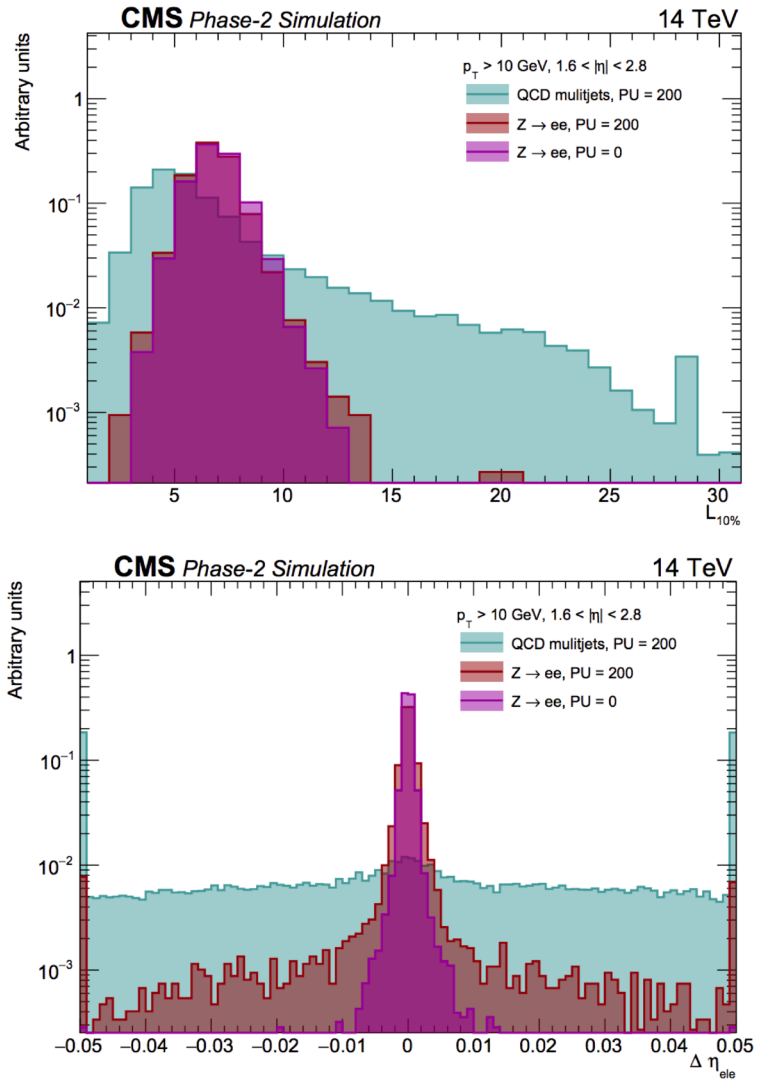


Figure 4.7: Shower shape variables used for electron identification. The longitudinal development of the shower is indicated by the upper plot, which shows the layer at which 10% of the total energy in the CE-E is deposited. The lower plot shows the pseudorapidity difference between the electron multicluster and the extrapolated electron track, which is a measure of the lateral spread of the shower. Figures taken from Ref. [3].

contribution of pileup is subtracted. The pileup subtraction was implemented by removing the total energy of a similar cone randomly rotated in ϕ. The energy resolution for a single pion with $p_T = 25\,\text{GeV}$ before and after the subtraction is shown in Figure 4.8. The megaclustering algorithm is shown to yield adequate energy resolution of around 20%. It is also robust against pileup; Figure 4.9 shows the modest worsening between PU 0 and PU 200 that decreases quickly as a function of p_T. This method

of reconstruction is not yet incorporated in the central CMS reconstruction software. For the physics results in Ref. [3], a realistic truth-information-driven approach was instead used for hadronic objects.

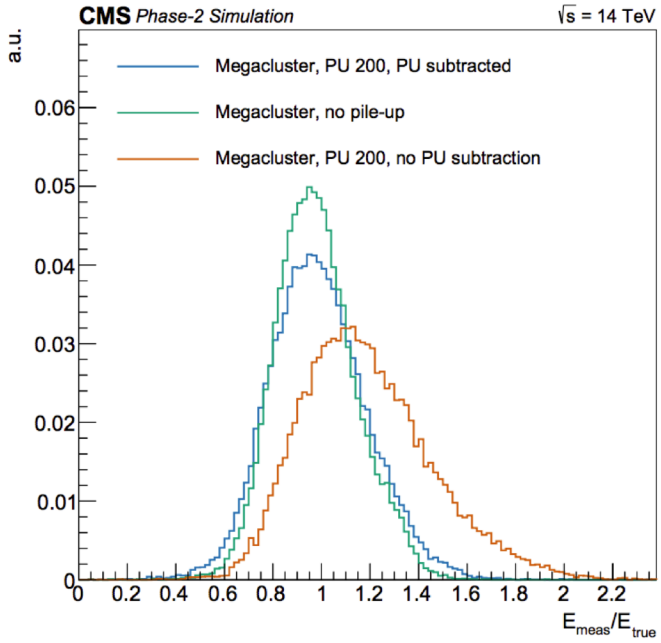


Figure 4.8: The distribution of the energy resolution for single $p_T = 25\,\text{GeV}$ pions reconstructed using the megaclustering algorithm, before and after pileup subtraction. Figure taken from Ref. [3].

4.6.3 Future development

The reconstruction methods described above are preliminary and designed only to show the potential performance of the HGCAL subdetector. There is a wide range of possibilities that will be explored before the eventual installation of the HGCAL. Firstly, it is possible to extend the 2D clustering algorithm to three dimensions, enabling the construction of multicluster-like objects in one step. This could potentially improve performance by including correlations between layers and including the longitudinal shower shape, rather than treating each layer as independent. Furthermore, the image-like data produced by the HGCAL are well-suited to the use of machine learning algorithms, particularly those based on neural networks (NNs). It is likely that eventually the initial pattern recognition step identifying showers will use NNs, and this is already under study.

Secondly, as mentioned above, the CMS particle flow algorithm has not yet been optimised to include all of the new information provided by the upgraded detector.

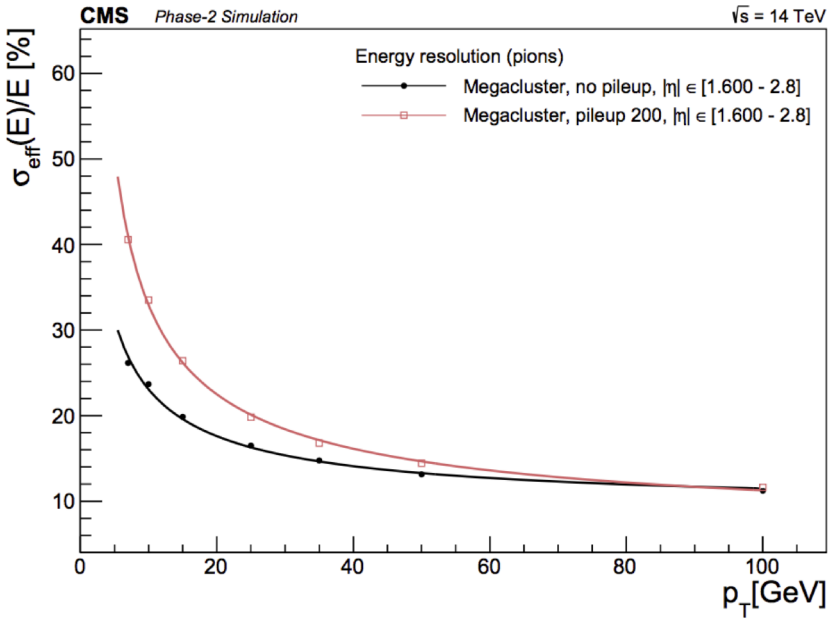


Figure 4.9: The mean energy resolution for single $p_T = 25\,\text{GeV}$ pions reconstructed using the megaclustering algorithm, after pileup subtraction, as a function of p_T. Figure taken from Ref. [3].

So far all studies have focussed on the calorimeter system alone; dedicated integration of tracking information, probably with an iterative approach, will bring substantial improvements. This has already been demonstrated by the effectiveness of the particle flow approach in Runs 1 and 2 [72].

Finally, there is the exciting possibility of performing event reconstruction in four dimensions using timing information. Both the upgraded ECAL and the HGCAL will provide precise timing information for showers, which will be used both for rejecting out-of-time pileup and for the separation of overlapping and adjacent showers. In addition, the timing information from the MIP timing detector brings many new possibilities. It enables the four-dimensional reconstruction of vertices and tracks, which has already been shown to improve the vertex identification in H → γγ events at PU 200 from 40% to 80% [66]. The added benefit of fully integrating the timing information into the particle flow algorithm in a coherent way remains to be seen.

4.7 Physics performance in the H → γγ decay channel

The performance of the HGCAL and the reconstruction techniques developed for its use are tested by evaluating their impact on CMS physics analyses. In this section, a study of the impact on the CMS H → γγ analysis is described; the diphoton channel

will continue to be very important for characterising the Higgs boson's properties at the HL-LHC. Rather than repeating the analysis in full, specific aspects of the analysis which will be affected by the HGCAL upgrade are investigated.

In Run 2, the sensitivity of the H $\rightarrow \gamma\gamma$ analysis is driven by barrel-barrel diphoton pairs. The effect of the HGCAL greater acceptance (to $|\eta|= 3$) relative to the current ECAL endcaps (to $|\eta|= 2.5$) is to increase the number of available diphotons by approximately 12%. Furthermore, the diphoton mass resolution in the HGCAL is found to be very similar to that in the barrel. This is an improvement upon the Run 2 endcap performance, which has significantly worse resolution than the barrel, and also increases the usefulness of the greater acceptance of forward photons.

A more substantial benefit of the HGCAL upgrade will be the ability to separate quark-initiated jets from jets which originate from gluon emission. This is illustrated in a study of the ggH and VBF production modes in the H $\rightarrow \gamma\gamma$ decay channel, using events simulated under HL-LHC conditions with the upgraded CMS detector. In the Run 2 analysis, boosted decision trees (BDTs) are used extensively to discriminate between different types of physics processes. A BDT is an example of a supervised machine learning method, which takes a set of labels corresponding to target truth classes together with a set of dependent variables and outputs a set of scores indicating how likely an event is to belong to a given class. A more detailed description of BDTs, how they are trained and how they are used in physics analyses is given in Section 6.2. One of the BDTs used in the Run 2 analysis, known as the dijet BDT, is used to discriminate between ggH and VBF using photon and jet kinematic variables. Here, the possibility of exploiting the granularity of the HGCAL by using jet shape variables as BDT inputs in addition to the kinematic variables is explored.

Jets initiated by gluons tend to be more dispersed than quark-initiated jets, which are relatively highly collimated and contain fewer particles [75]. Therefore variables relating to the jet shapes can be used to discriminate between the two. Since jets in ggH events tend to be gluon-initiated, and the VBF jets quark-initiated, this can in turn provide discrimination between the two production processes.

The impact of the HGCAL is evaluated by comparing the performance of two different BDTs. Three jet shape variables are used to construct a BDT referred to as the jet shape BDT. A second BDT, known as the dijet BDT, uses additional kinematic variables including those of the photons; its inputs are identical to the dijet BDT used in the Run 2 analysis (see Chapter 6 for details), but with the three jet shape variables added. The same training procedure as that used in the Run 2 analysis was followed, where the VBF events are treated as signal, with the ggH as background. The performance of the two classifiers is illustrated in Figure 4.10. In each case, a

receiver operating characteristic (ROC) curve is constructed. The ROC curve shows the selection efficiency for signal VBF and background ggH events for all values of the BDT output score. The area under the ROC curve is used a measure of the BDT performance; the area is unity for perfect discrimination between signal and background, and equals one half if the BDT has no discriminating power. The area under the ROC curve is 0.71 for the jet shape BDT, and 0.79 for the dijet BDT. For comparison, the value for the Run 2 dijet BDT was 0.75. This demonstrates that the additional information provided by the HGCAL translates directly into an improvement at the analysis level.

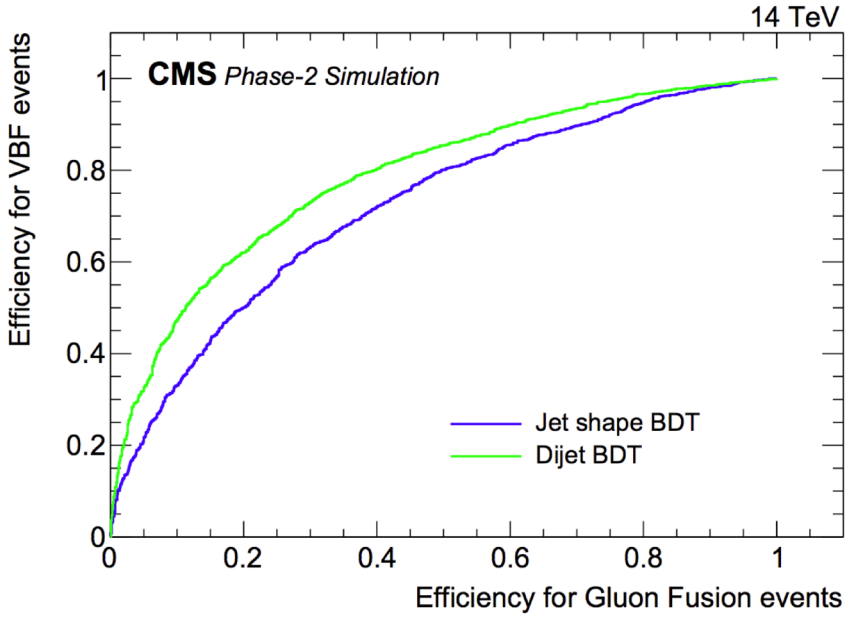


Figure 4.10: The selection efficiency for VBF and ggH events for two different BDTs. Figure taken from Ref. [3].

Separately, another BDT is trained to reduce the amount of background entering analysis categories. The VBF events are treated as signal, and are trained against H → γγ backgrounds including events with two real photons, events with one real photon and one jet misidentified as a photon, and events with two jets misidentified as photons. This classifier uses only the kinematics of the two photons as inputs. It is less powerful than the Run 2 equivalent because not all of the inputs used in the Run 2 version are available. These include the photon identification score, the mass resolution estimates, and the vertex probability estimate.

A 2D scan was used to choose cut values on both the dijet and background BDTs, and generate the working points in Table 4.1. Three categories were chosen in order to facilitate comparison with the Run 2 results presented in Ref. [1]. The working points

Event Categories	SM 125 GeV Higgs boson expected signal			Bkg
	Total	ggH	VBF	per GeV
WP 0	750	25.4 %	74.6 %	678
WP 1	1275	35.9 %	64.1 %	876
WP 2	1926	45.8 %	53.2 %	1353
Summed WP	3951	38.7 %	61.3 %	2907
Run 2 summed WP	3878	42.0 %	58.0 %	1984

Table 4.1: The signal and background yields for three working points, and the sum thereof. The dijet BDT cut is varied, with a fixed cut on the background BDT. The expected number of events given is for 3000 fb^{-1} of collected data. The Run 2 WP contains the sum of selected events in all three VBF categories of Ref. [1], extrapolated to 3000 fb^{-1}. Table taken from Ref. [3].

were chosen to maximise the total expected significance, using the simple metric of the sum in quadrature of the ratio of signal events to the square root of signal plus background events for each working point [76]. The number of signal events selected, the fraction of ggH and VBF events, and the background per GeV are shown. Also included is a Run 2 working point, consisting of all the events entering the analysis' three VBF tags, for comparison.

These results show that performance comparable to that in Run 2 is achieved despite the increase in pileup. The expected amount of background is higher than in Run 2, but it is expected that this could be reduced by a classifier using further photon quality variables.

4.8 Beam tests

To validate the design of the HGCAL and ensure its behaviour is well-modelled by simulation, beam tests have been conducted at both CERN and Fermilab in 2016 and 2017. Prototype silicon modules representative of those in both the CE-E and CE-H were built, with plastic scintillator tiles modified from an existing detector developed by the CALICE Collaboration [77]. At Fermilab, the available electron beams were of relatively low energy (up to 32 GeV), and so a test configuration with thickness of around 15 X_0 was used. A set of sixteen silicon modules, arranged in sixteen layers, was used. At CERN, energies of up to 250 GeV were available but with only eight modules. The chosen setup therefore comprised eight layers placed between 5 and 27 X_0.

Comparisons to simulation of the measured electron energy resolution and shower shape are shown in Figure 4.11. The variable E1/E7 is defined as the ratio of the energy in the most energetic cell (E1) to the sum of the energies of the most energetic cell and

its six surrounding cells (E7); this indicates the lateral extent of the shower. Also shown is the relative energy resolution for a range of energy values, using both the CERN and Fermilab setups. In both cases, excellent agreement is observed between data and simulation; the observed data distributions match those predicted by simulation to within 5%. These beam test results are particularly important because they constitute the first demonstration that the HGCAL behaves as predicted by simulation.

Furthermore, additional tests confirm the expected intrinsic timing capabilities of the silicon sensors, with the timing resolution measured to be less than 30 ps. The timing performance of the silicon was also measured to be a function of S/N only, meaning it does not degrade with increasing radiation exposure.

4.9 Summary

The HGCAL is a key part of the CMS detector upgrade planned as part of the HL-LHC programme, which is expected to collect $3000\,\mathrm{fb}^{-1}$ of data. The design of the HGCAL is driven by the need to sustain the excellent performance of the CMS detector in the harsh conditions of the HL-LHC, where the mean pileup is expected to be between 140 and 200 per bunch crossing. For this reason the HGCAL is extremely resistant to radiation, with silicon active material in the front and forward regions where the dose is expected to be greatest. Its fine lateral and longitudinal granularity will ensure high physics performance throughout the lifetime of the HL-LHC, and provides a vast amount of information for use in event reconstruction. Preliminary reconstruction methods have been developed for electromagnetic objects, and used in realistic simulations of physics analyses to demonstrate the potential of the HGCAL. Furthermore, initial beam tests of HGCAL modules confirm that the simulation provides an accurate model of the subdetector.

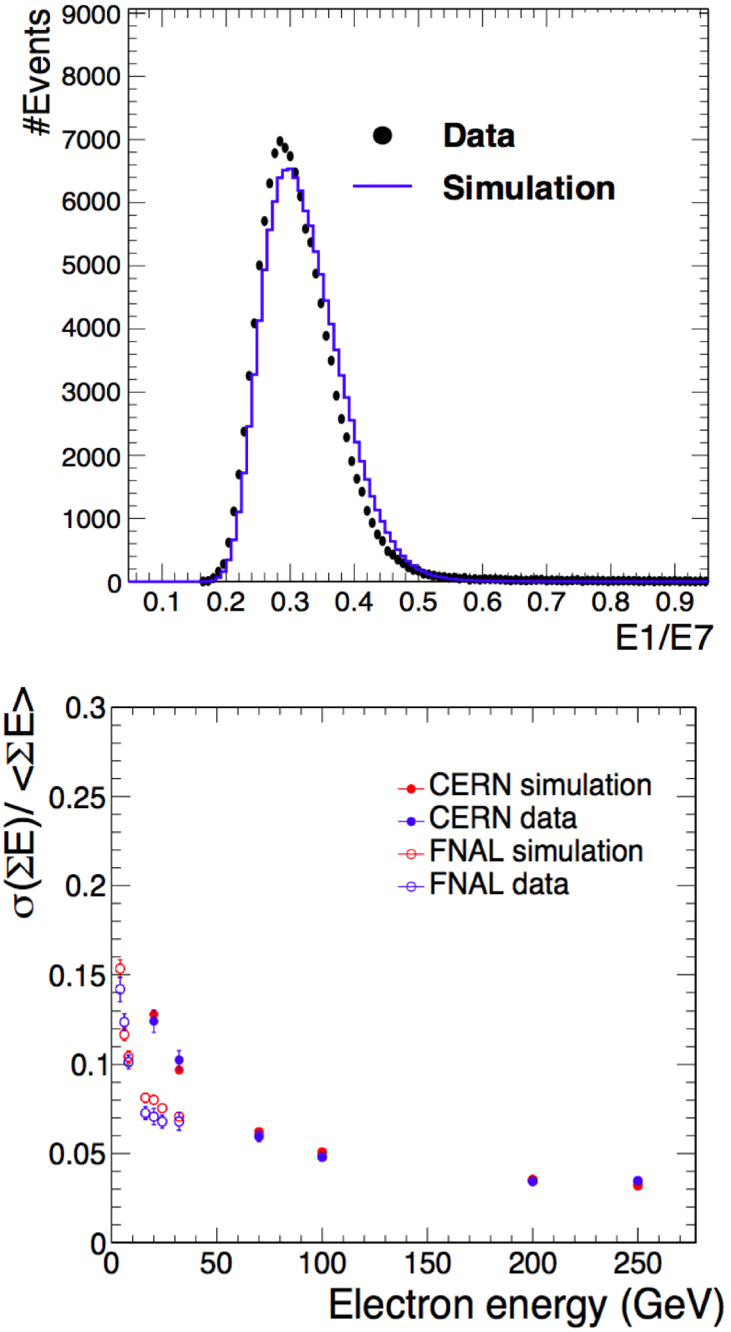


Figure 4.11: The upper plot shows the distribution of the energy ratio E1/E7 (defined in the text) for 100 GeV electrons using the CERN beam test setup. Data points are shown in black and the blue histogram represents the simulation. The lower plot shows the relative electron energy resolution as a function of energy, for both the CERN (solid circles) and Fermilab (hollow circles) setups. Simulated points are shown in red, with data points in blue. Figures taken from Ref. [3].

Chapter 5

Event Reconstruction and Selection

5.1 Introduction

Measurements of Higgs boson properties can be made by directly using the diphoton invariant mass distribution. Photon pairs resulting from Higgs boson decays produce a narrow signal peak, centred at the value of the Higgs boson mass, on top of the smoothly falling background distribution produced by other SM processes. The Higgs boson mass is inferred from the two photons by constructing the diphoton invariant mass, given by the following expression

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

where E_{γ_1} and E_{γ_2} is the energy of each photon, and θ is the opening angle between them. The sensitivity of the analysis is maximised when the reconstructed diphoton mass peak is as narrow as possible, thereby minimising the diphoton mass resolution. This requires the two photons to be correctly identified and their positions and energies accurately measured. Furthermore, the location of the interaction vertex from which the photons originated must be established in order to calculate the opening angle. Additional objects in the event, including jets and leptons, are further used to improve the sensitivity of the analysis and measure different Higgs boson production processes.

This section describes the official CMS procedure for reconstructing physics objects using the particle flow algorithm [72]. In addition, the photon and vertex identification techniques specific to the H $\rightarrow \gamma\gamma$ analysis are detailed. The approach used is almost identical between the 2016 and 2017 datasets; any differences are highlighted in the text.

5.2 Particle flow

The global event description at CMS is formed using the particle flow (PF) algorithm [72]. The goal of PF is to optimally combine the information of all the CMS subdetectors, enabling the best possible identification and energy measurements for all types of objects. Inputs to the PF algorithm are tracks originating from the tracker and muon system, and calorimeter clusters from the ECAL and HCAL. CMS is able to benefit from the PF approach due to its strong magnetic field, alongside the fine segmentation and hermeticity of the tracker, calorimeters, and muon system. Together these allow different types of objects to be separately identified, and the energy measurement to be driven by the subdetector with the best resolution.

Tracks are reconstructed from hits in the tracker using multiple iterations of a combinatorial track-finding procedure [56]. Each iteration proceeds in the following way. First, track seeds comprising two or three hits are chosen, defining the initial track parameters. Then an extrapolation is performed along the expected track paths, adding any additional hits consistent with the path hypothesis. Next the track parameters are re-estimated, and the track candidate collection is pruned based on quality criteria. All the selected hits are then removed from consideration in the following iterations. In this way, the first iteration identifies the most obvious tracks, normally those with high p_T and near to the interaction point. The complexity of the subsequent iteration is reduced since many hits have been removed. This therefore permits lower thresholds to be used and tracks with lower p_T or highly displaced tracks to be found. In addition, an independent procedure is used to construct muon tracks from hits in the muon system.

In the calorimeters, a clustering algorithm is used to collect together energy deposits belonging to each shower. The procedure in the ECAL is described here, since it is an important input to the $H \rightarrow \gamma\gamma$ analysis; the HCAL procedure is similar. The clustering algorithm begins by selecting cluster seeds, which have an energy above a threshold and higher than any neighbouring crystal [74]. So-called topological clusters are then constructed iteratively by adding deposits which share a side or corner with one already in the cluster, provided their energy exceeds a threshold dependent on the noise level. If a crystal could be included in more than one topological cluster, its energy is shared between them assuming a Gaussian shower shape. Finally, topological clusters are grouped into so-called superclusters (SCs). This step is designed to recover energy lost via bremsstrahlung; radiated showers typically have very similar η values but are spread out in the ϕ direction by the magnetic field.

Given these tracks and calorimeter clusters as inputs, the particle flow algorithm forms collections of candidates for five types of particle:

- **Muons:** a path extrapolated from the tracker is consistent with a muon track. The energy is inferred from the curvature of the track.
- **Electrons:** an ECAL SC is present and associated with a track from the inner tracker. The energy is measured using a combination of the track p_T and the SC energy.
- **Photons:** an ECAL SC is present and no track is associated with it. The energy is measured using the ECAL SC only.
- **Neutral hadrons:** matched ECAL and HCAL clusters with no associated track. Energy measurement is the sum of the cluster energies.
- **Charged hadrons:** a track is matched to ECAL and HCAL clusters. The track curvature is used together with the calorimeter deposits to calculate the energy.

In this way, the central CMS software provides analyses with physics objects ready for use in measurements. The remainder of this chapter describes how these objects are used in the $H \to \gamma\gamma$ analysis.

5.3 Samples

5.3.1 Data

This analysis is based on a total of $77.4\,\text{fb}^{-1}$ of p-p collision data collected by CMS at $\sqrt{s} = 13\,\text{TeV}$. Of this, $35.9\,\text{fb}^{-1}$ was collected during 2016 and $41.5\,\text{fb}^{-1}$ in 2017.

Data selected for use in the analysis must first be selected by the CMS trigger system, which reduces the total event rate to an acceptable level. This two-step system requires each high-level trigger (HLT) path to be seeded by at least one electromagnetic candidate at the level one trigger (L1T). A higher efficiency is obtained if only one of the two photons is required to be present at level one; however since this results in a high event rate, a stringent p_T selection is required. The threshold is dependent on instantaneous luminosity and the photon isolation, but is typically set at 40 GeV. This results in loss of efficiency in $H \to \gamma\gamma$ events, so asymmetric HLT paths seeded by two electromagnetic candidates are also used. During 2016 data-taking, the p_T thresholds at L1 were 22 and 15 GeV for the leading and subleading candidates respectively. In 2017, this was increased to 25 and 14 GeV respectively.

At the HLT, a clustering procedure similar to that used offline is performed. This allows shower shape and photon isolation variables to be included in the selection criteria. These include the ratio of energy in a 3×3 grid of crystals to that of the SC (R_9), and the ratio of energy in the HCAL behind the SC to the SC energy. Together with asymmetric requirements on the photon p_T and a cut on the diphoton mass, these form the HLT selection. In 2016, the HLT p_T thresholds were 30 and 18 GeV respectively; this selection was tightened to 30 and 22 GeV respectively in 2017.

The efficiency of the trigger selection is measured using the tag and probe technique in $Z \to e^+e^-$ events where the electrons are reconstructed as photons. In this method, dielectrons consistent with decays from the Z boson are selected, and a tight identification requirement is placed on one electron (known as the tag). The second electron (the probe) must then pass a very loose requirement, and the efficiency of the additional HLT selection criteria can be measured. The trigger efficiency is over 97% in the barrel and above 95% in the endcaps. In simulation, its effect is accounted for by applying weights binned in p_T, η and R_9.

5.3.2 Simulation

Various types of events are simulated with Monte Carlo (MC) techniques for use in training event classifiers, producing the signal model, and for performing validation of the analysis.

The signal simulation used to produce the final signal model is generated using MadGraph5_aMC@NLO [78], which is next-to-leading order in perturbative QCD. This is interfaced to PYTHIA8 [79] to perform the parton showering and hadronisation, using the tune CUETP8M1. Events from ggH production are then reweighted to agree with the next-to-next-to-leading order program NNLOPS [80]. Additional independent signal samples are produced using POWHEG [81]. These POWHEG samples are used to train the various classifiers used to define the analysis categories; it is preferred over MadGraph5_aMC@NLO for this purpose because it does not contain events with negative weights. This also ensures that the events used to train the classifiers are independent from the events used to construct the final signal model of the analysis.

Simulated background events are not used to produce the final results of the analysis, but are used to train several of the classifiers used in the analysis. The predominant source of background is the irreducible SM production of two prompt (real) photons, which is simulated using the SHERPA [82] program. There are two types of reducible background: γ + jet events where the jet is misidentified as a photon ("prompt-fake" events), and events where two jets are misidentified as photons ("fake-

fake" events). These are both modelled using PYTHIA, with filters applied to increase the number of events containing jets with high electromagnetic energy fractions. The Drell-Yan events used extensively in the analysis validation are simulated using MadGraph5˙aMC@NLO [78].

The CMS detector itself is modelled using GEANT4 [83]. This simulation includes additional pileup interactions. The distribution of the number of vertices in simulation is reweighted to agree with data, where an average of 23 and 32 pileup interactions were observed in 2016 and 2017 data-taking respectively,

5.4 Photon reconstruction

5.4.1 Overview

Photons to be used in the H $\rightarrow \gamma\gamma$ analysis are selected from the PF set of photon candidates. The energy of each photon candidate is estimated from the SC, which includes deposits from the many particles comprising the electromagnetic shower. In some cases, photons will interact with detector material upstream of the ECAL and produce an electron-positron pair; these are known as "converted" photons. Converted photons will also deposit energy in the ECAL preshower detector, which is included in the SC energy estimate. A correction is then made to this SC energy using a multivariate regression technique. After that, data and MC are brought into agreement by applying additional scale and smearing corrections to the photon energies. The details of the energy estimation procedure are described in Section 5.4.3. Once the photon energy has been established, a set of preselection criteria is applied to obtain the final set of photons considered in the analysis. One of these criteria is a requirement placed on the output score of the photon identification BDT, which is trained to reduce the contamination from other objects which mimic real photons. The full set of preselection requirements, including the photon identification BDT, are described in Sections 5.4.4 and 5.4.5.

5.4.2 Variable definitions

Different types of variables are used in the photon reconstruction. They can be divided into shower shape and isolation variables. The ECAL shower shape information is used both to correct the energy and discriminate between real and fake photons. Isolation variables can be used to reduce the mis-identification of jets and other objects imitating a real photon signal. The full set of variables used in this section and their definitions are given below.

Shower shape variables:

- $E_{2\times2}/E_{5\times5}$: the ratio of the energy in the 2×2 grid containing the most energetic cells in the SC to the energy in the 5x5 grid centred on the SC seed.
- $cov_{i\eta i\phi}$: the covariance of the single crystal η and ϕ values in within the 5x5 grid centred on the SC seed.
- $\sigma_{i\eta i\eta}$: the standard deviation of the shower in η in terms of number of crystal cells.
- R_9: $E_{3\times3}/E_{SC}$ where $E_{3\times3}$ is the energy sum of the 3×3 grid surrounding the SC seed and E_{SC} is the energy sum of the SC before corrections.
- σ_η: the standard deviation of single crystal η values within the SC, weighted by the logarithm of the crystal energy.
- σ_ϕ: the standard deviation of single crystal ϕ values within the SC, weighted by the logarithm of the crystal energy.
- σ_{RR}: the standard deviation of the shower spread in the x-y plane of the preshower detector (endcap only).

Isolation variables:

- H/E_{SC}: the ratio of the energy in the HCAL cells behind the SC and the SC energy.
- $\mathcal{I}_{ph}$: photon isolation, defined as the sum of the transverse energy of the particles identified as photons falling inside a cone of radius $R = 0.3$ around the photon candidate.
- $\mathcal{I}_{tk}$: track isolation, which is the sum of the transverse momenta of all tracks in a cone of radius $R = 0.3$ around the photon candidate (with tracks in an inner cone of size $R = 0.04$ not included in the sum).
- $\mathcal{I}_{ch}$: charged hadron isolation, the sum of the transverse momenta of charged particles inside a cone of radius $R < 0.3$ around the photon candidate.

Miscellaneous:

- *Electron veto*: rejects the photon candidate if its supercluster is matched to an electron track.
- ρ: the median energy density per unit area in the event.
- E_{SC}: the energy of the supercluster before corrections.
- E_{true}: the true photon energy.

5.4.3 Photon energy

The photon supercluster energy is computed from the calibrated ECAL SC, as described in Section 3.3.3. However, this energy estimate is imperfect for several rea-

sons. Firstly, the SC may not capture all the energy of an electromagnetic shower and therefore underestimate the photon energy. This can occur when the photon interacts with material before the ECAL and therefore begins to shower before reaching the ECAL, resulting in both energy lost in the material and a dispersed shower which is not contained by the SC. In addition, leakage can occur into the gaps between crystals and modules in the ECAL, as well as through the back of the ECAL if the showering begins near the end of a crystal.

The energy estimate provided by the uncorrected SC (E_{SC}) can be significantly improved by the use of a multivariate regression. This provides an additional energy correction to each reconstructed photon, so that the photon energy in simulation agrees with the true photon energy. The training objective for the regressor is to learn the parameters of the E_{true}/E_{SC} probability distribution function, which is parameterised as a Gaussian core with power law tails on each side. There are a large number of inputs to the regressor which account for different effects. These include information relating to the shower shape, the position of the SC, the seed crystal, and the energy density in the event.

The shower shape and position variables enable the difference in shower containment due to variations in the ECAL geometry, and the showering induced by upstream material, to be corrected. Variables relating to the seed crystal position within the SC and energy ratios formed using the seed account for the local shower containment. The energy density and number of vertices accounts for any additional energy included due to pileup.

The most probable value of the probability distribution function returned by the regression is used to correct the supercluster energy for each photon. The distribution is also used to infer the per-photon uncertainty on the energy. This resolution estimate is then used in the classification of events, as described in Chapter 6.

After the regression, additional corrections are applied to account for differences between data and simulation. The first of these is to correct the energy scale in data; the detector response changes to damage from radiation during LHC operation and subsequent recovery during LHC down time. This can cause the measured energy to both drift slowly and change discontinuously over time. Corrections are derived using $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events where the electrons are reconstructed as photons, using the known Z boson mass. These corrections are extracted differentially in time, and binned in the photon η (two bins in each of barrel and endcap) and R_9 (two bins). Their magnitude varies from 0.1% to 0.3% in the barrel and from 0.2% to 2.0% in the endcaps.

Finally, the energy resolution in simulation is corrected to match that in data. This is implemented as a Gaussian smearing using the same bins as the energy corrections.

The size of the corrections range from approximately 0.1% to 2.7%. Once all the corrections have been applied, the photon energy resolution ranges from around 1.0% to 2.5% in the barrel and from 2.5% to 3.5% in the endcaps. Validation of the energy after corrections is shown in Figure 5.1, where data are compared to simulation in $Z \to e^+e^-$ events with electrons reconstructed as photons. Full details of the entire energy correction procedure can be found in Ref. [74].

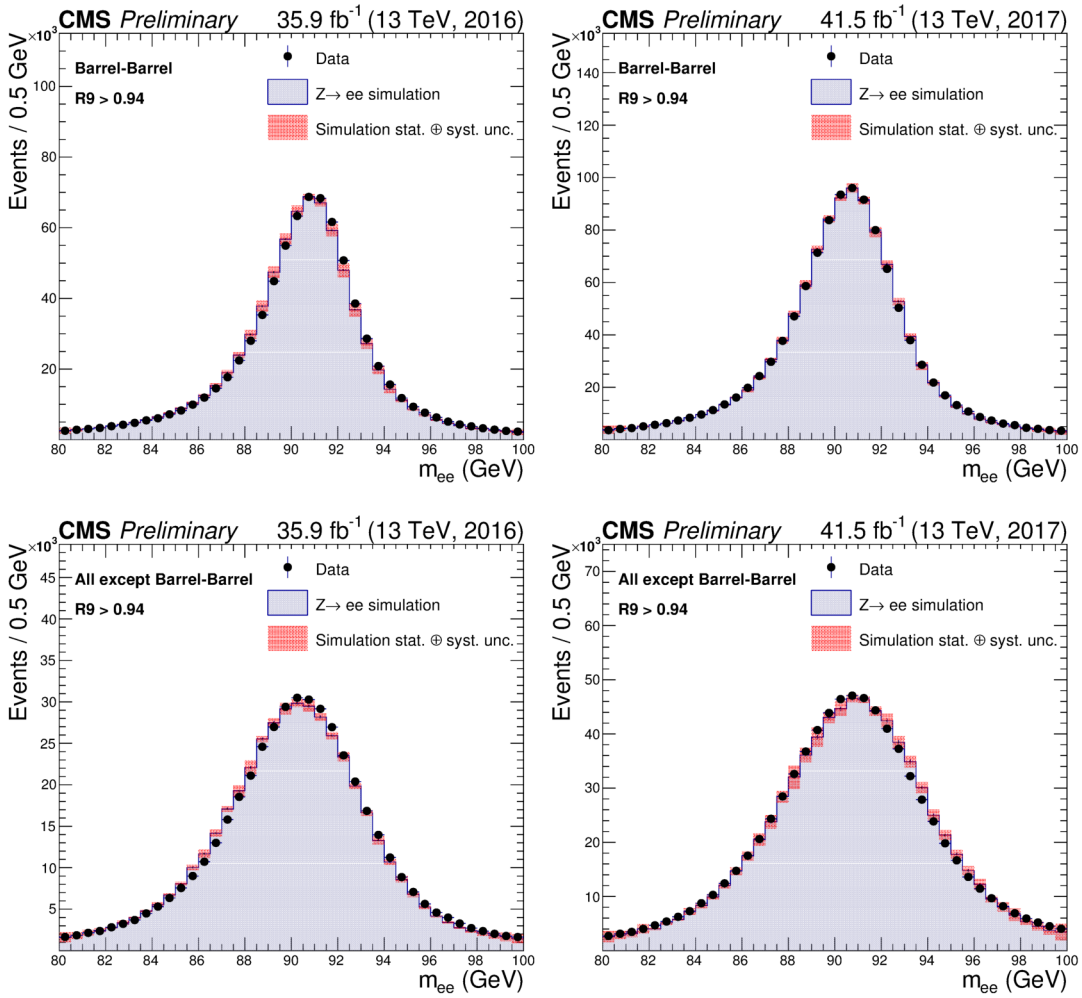


Figure 5.1: Comparison of the dielectron invariant mass distributions in data and simulation (after energy smearing) for $Z \to e^+e^-$ events where electrons are reconstructed as photons. The R_9 variable is required to be greater than 0.94. The simulated distributions are normalised to the integral of the data distribution in the range $87 < m_{ee} < 93$ GeV to highlight the agreement in the bulk of the distributions. The comparison is shown for events where at both electrons are in the ECAL barrel (top), and least one electron is not in the ECAL barrel (bottom). The plots on the left show data and simulation from 2016, with 2017 data and simulation shown in the plots on the right [4].

5.4.4 Photon preselection

To be further considered in the H $\rightarrow \gamma\gamma$ analysis, photons must pass a set of criteria designed to reduce the contamination from objects faking photons. This preselection is similar to, but more stringent than, that required at trigger level. The full set of requirements is:

- to be within the ECAL acceptance; this means $|\eta| < 2.5$ and excluding the endcap transition region $1.44 < |\eta| < 1.57$.
- passing the electron veto
- $p_T > 30(20)$ GeV for the leading (subleading) photon in 2016. The thresholds are 35 and 25 GeV respectively in 2017, due to the tighter trigger requirements.
- either $\mathcal{I}_{\text{ch}} < 20$ GeV, or $\mathcal{I}_{\text{ch}}/p_T < 0.3$ for photons with $R_9 < 0.8$.
- an additional set of R_9 and η-dependent shower shape and isolation requirements, summarised in Table 5.1.

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

Table 5.1: Summary of photon preselection requirements. The values are chosen to be slightly more stringent than the equivalent trigger requirements [1].

The efficiency of these preselection criteria, aside from the electron veto, is measured using Z $\rightarrow e^+e^-$ events. To measure the electron veto efficiency, Z $\rightarrow \mu^+\mu^-\gamma$ events are used. Efficiencies range from over 94% for high R_9 photons in the barrel, to around 50% for low R_9 photons in the endcaps. Good agreement is observed between data and simulation, with differences between data and simulation in the range 0.1% to 0.3% in the barrel and 0.5% to 1% in the endcap.

5.4.5 Photon identification

An additional part of the photon preselection is the photon identification BDT. Its purpose is to discriminate between real photons and photon candidates passing the preselection which originate from other objects, such as jet fragments. The BDT is trained using simulated γ + jet events passing the preselection described above, with real photons treated as signal and fake photons as background. The inputs are similar to the those used in the energy regression, including shower shape and isolation vari-

ables. In addition, pileup-sensitive variables ρ and the number of interaction vertices, together with photon kinematic variables, are used as inputs.

A loose selection on the photon identification BDT at -0.9 completes the analysis preselection. The output score on preselected diphoton events is shown in Figure 5.2. Additional validation using $Z \to e^+e^-$ events is demonstrated in Figures 5.3. In both cases, good agreement between data and simulation is observed.

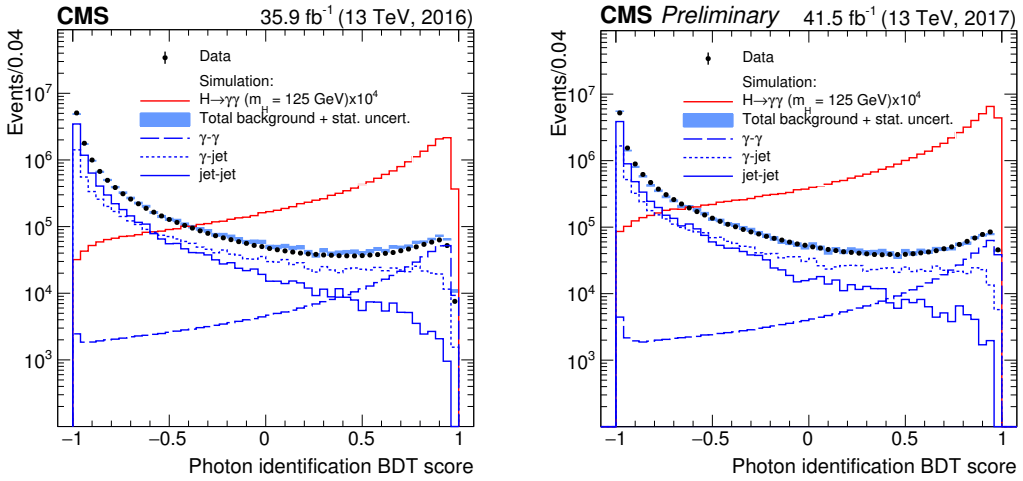


Figure 5.2: Distribution of the photon identification BDT score of the lowest scoring photon of diphoton pairs with an invariant mass in the range $100 < m_{\gamma\gamma} < 180\,\text{GeV}$, for events passing the preselection in data (black points), and for simulated background events (blue histogram). Histograms are also shown for different components of the simulated background. The sum of all background distributions is scaled up to data. The red histogram corresponds to simulated Higgs boson signal events. The left figure shows 2016 data and simulation [1], with 2017 data and simulation shown on the right [4].

5.5 Vertex reconstruction

As can be seen in Equation 5.1, the only parameters affecting the diphoton mass are the photon energies and the opening angle between them. Since there are multiple p-p interactions at each bunch crossing, multiple vertices are present in each event, spread along the z-axis. The choice of the vertex from which the diphoton originated has a direct impact on the angle θ, and therefore the diphoton mass resolution. Provided the chosen vertex is within approximately 1 cm of the true interaction vertex, the mass resolution is dominated by the photon energy resolution and the contribution from the angle is negligible. A BDT is therefore trained to identify the most likely diphoton interaction vertex.

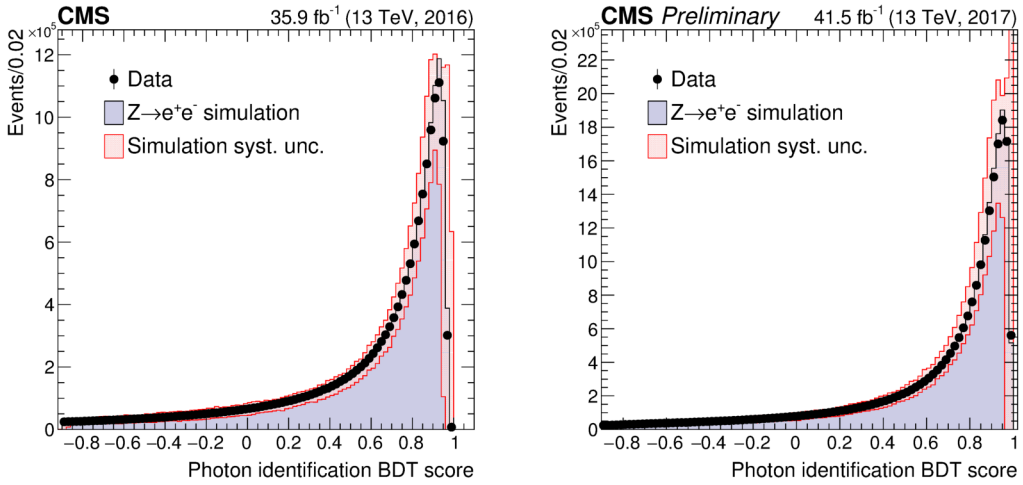


Figure 5.3: Distribution of the photon identification BDT for $Z \to e^+e^-$ events in data and simulation, where the electrons are reconstructed as photons. The systematic uncertainty applied to the shape from simulation (hashed region) is also shown. The left figure shows 2016 data and simulation [1], with 2017 data and simulation shown on the right [4].

5.5.1 Vertex selection

The vertex identification BDT is trained with simulated $H \to \gamma\gamma$ signal events, with the objective of selecting the true interaction vertex. Since photons do not produce tracks, its inputs are variables related to tracks recoiling against the diphoton system. The variables are calculated for each candidate vertex, and are defined as follows:

- $\sum_i |\vec{p}_{T,i}|^2$,
- $-\sum_i \vec{p}_{T,i} \cdot \frac{\vec{p}_{T,\gamma\gamma}}{|\vec{p}_{T,\gamma\gamma}|}$,
- $|\sum_i \vec{p}_{T,i}| - p_T^{\gamma\gamma} \,/\, |\sum_i \vec{p}_{T,i}| + p_T^{\gamma\gamma}$,

where $\vec{p}_{T,i}$ is the p_T of the i-th track associated with a given vertex and $\vec{p}_{T,\gamma\gamma}$ is the vector representing the diphoton transverse momentum.

If one or more of the photons have converted into electrons in the tracker, two more variables are used in addition to those above:

- the number of conversions,
- the pull $|z_{\text{vtx}} - z_e|/\sigma_z$ between the longitudinal position of the reconstructed vertex, z_{vtx}, and the longitudinal position of the vertex estimated using conversion track(s), z_e, where the variable σ_z denotes the uncertainty on z_e.

The vertex with the highest vertex identification BDT output score is then chosen as the H $\to \gamma\gamma$ interaction point. This choice is considered to be the "correct" vertex if the distance from the true interaction vertex is less than 1 cm, and is used to define the vertex choice efficiency. The inclusive efficiency is approximately 82% in 2016 conditions, decreasing only slightly to 81% in 2017 conditions. This shows that the vertex identification procedure is robust to increases in pileup. It also represents a substantial improvement compared to the default CMS vertex, which is correct in around 74% of events.

The performance of the vertex identification BDT is validated using Z $\to \mu^+\mu^-$ events. The tracks from the muons are ignored in order to mimic the diphoton system. Comparison between data and simulation in these events is shown in Figure 5.4; good agreement between the two is observed.

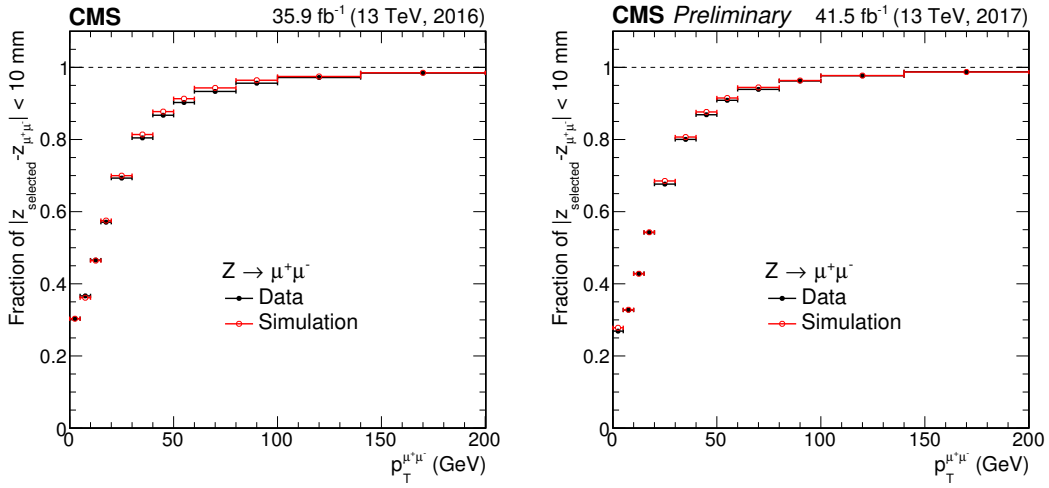


Figure 5.4: Validation of the H $\to \gamma\gamma$ vertex identification algorithm on Z $\to \mu^+\mu^-$ events omitting the muon tracks. Simulated events are weighted to match the distributions of pileup and location of primary vertices in data. The left figure shows 2016 data and simulation [1], with 2017 data and simulation shown on the right [4].

5.5.2 Vertex probability

A second vertex-related multivariate discriminant (vertex probability BDT), is designed to estimate, event-by-event, the probability for the vertex assignment to be within 1 cm of the diphoton interaction point and thus considered "correct". The vertex probability BDT is trained on simulated H $\to \gamma\gamma$ events using the following input variables:

- the number of vertices in each event,

- the values of the vertex identification BDT score for the three most probable vertices in each event,
- the distances between the chosen vertex and the second and third choices,
- the transverse momentum of the diphoton system, $p_T^{\gamma\gamma}$,
- the number of photons with an associated conversion track.

The vertex choice efficiency, together with the estimated vertex probability, in simulated H $\rightarrow$ $\gamma\gamma$ signal events is shown in Figure 5.5. The simulated performance is shown as a function of both the p_T of the diphoton system and of the number of vertices in the event. The figures demonstrate that the average vertex probability estimate accurately reflects the true vertex efficiency.

5.6 Jet reconstruction

The collimated collection of particles produced by the hadronisation of quarks or gluons are referred to as jets. Around 60% of the average jet energy comes from charged hadrons, whose energy is well-measured by the tracker, and another 30% from photons, whose energy is measured precisely by the ECAL. The remaining 10% of energy from neutral hadrons is measured with worse resolution by the HCAL. Jets are important in the H $\rightarrow$ $\gamma\gamma$ analysis because they enter the definition of the STXS signal bin definitions, are used to categorise the events targeting those bins, and are used to differentiate between production modes.

At CMS jets are formed using the anti-k_T clustering algorithm [84] with distance parameter $R = 0.4$. It is an iterative algorithm that defines a distance metric dependent on the p_T and angular parameters of candidate clusters of energy deposits, and groups together these clusters starting with the nearest two objects. Once the shortest distance is between an object and the beampipe, the process stops and the clustered objects are defined as a jet. This is then repeated until all clusters have been collected into jets.

The inputs to the anti-k_T algorithm are PF candidates with charged hadron subtraction (CHS) [85]. CHS reduces the contribution from pileup by identifying any charged hadrons associated with vertices other than the primary vertex. Since the tracker coverage extends only to $|\eta|= 2.5$, jets further forward than this do not have pileup removed via CHS. Once the jets have been constructed, a set of corrections are applied to correct the jet energy scale and resolution in both data and simulation. The correction procedure can be summarised as [85]:

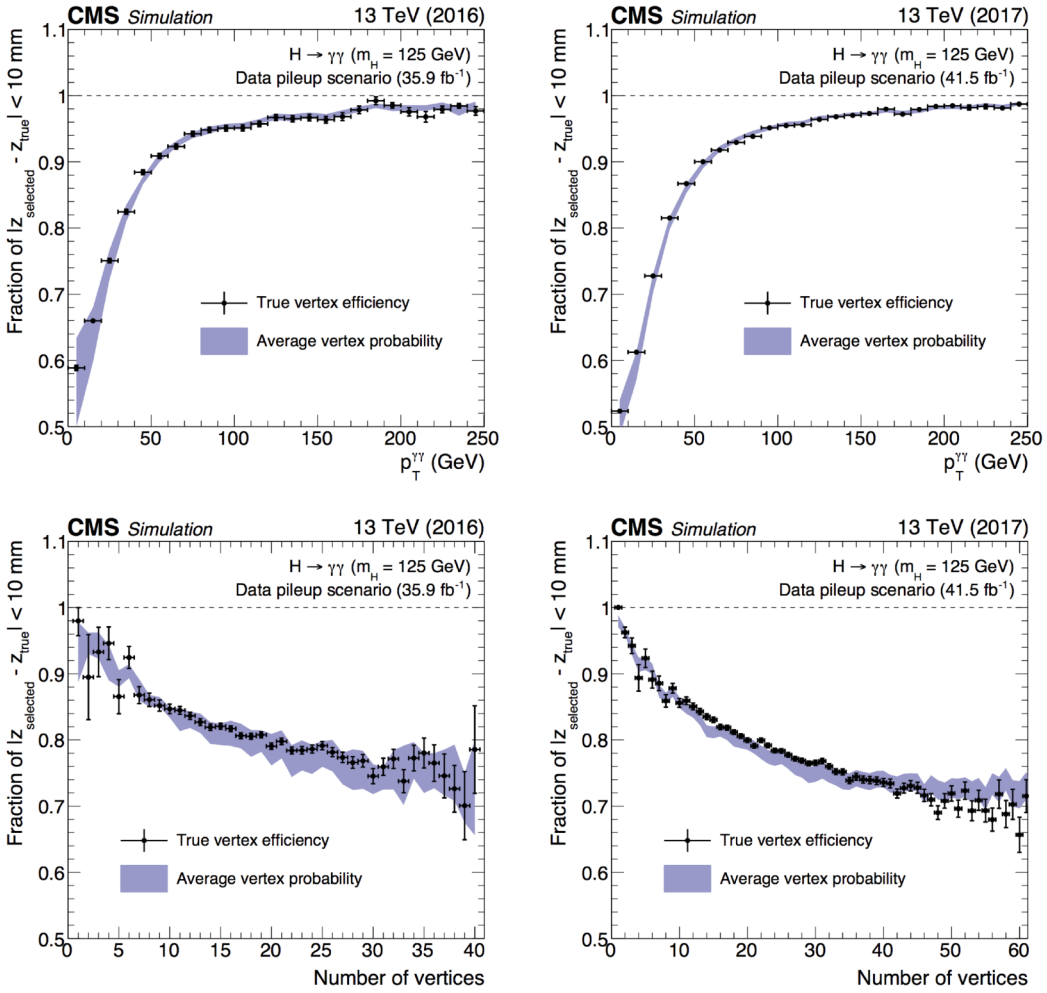


Figure 5.5: Comparison of the true vertex identification efficiency and the average estimated vertex probability as a function of the reconstructed diphoton p_T (top) and of the number of primary vertices (bottom) in simulated H $\rightarrow \gamma\gamma$ events with $m_H = 125\,\text{GeV}$. Events are weighted according to the cross sections of the different production modes and to match the distributions of pileup and location of primary vertices in data. The error bars and uncertainty bands include the contribution from both the statistical uncertainty on the simulated events and the differences between data and simulation in the vertex efficiency. The plots on the left show simulation under 2016 conditions [1], with simulation under 2017 conditions shown in the plots on the right [4].

- corrections are made for the pileup contribution to jets, as a function of the event energy density, and the jet p_T, η and area. Any remaining differences between data and MC are derived after these corrections using pileup-only events.

- the reconstructed jet energy is corrected to agree with the true jet energy in

both data and MC. This is performed as a function of p_T and η to account for variation in response across the detector.

- residual differences between data and MC are corrected in two steps. First η-dependent corrections are applied using dijet events, where the scale of a jet is corrected using a jet of similar p_T in the reference, well-measured $\eta < 1.3$ region. The p_T-dependent corrections are derived using events with a jet recoiling against a photon or Z boson. This corrects the absolute scale of the jet.

Additional selection can be applied to these calibrated jets to further reject jets originating from pileup. Pileup jets are typically more dispersed than real jets, since they are often composed of multiple overlapping jets with relatively low p_T values. A pileup jet identification BDT is therefore trained using jet shape variables, and additional track variables related to the interaction vertex of the jet. The output score of this BDT is used to reject pileup, with thresholds dependent on the p_T and η of jets. For a jet to be considered in the analysis, it must pass the pileup jet identification criterion and have $p_T > 30\,\text{GeV}$ and $|\eta| < 4.7$.

5.7 Reconstruction of other objects

Additional objects can be used to specifically target rarer Higgs boson production modes, such as VH and ttH. Whilst not used directly in this analysis, they are important in both Refs [1] and [86].

5.7.1 Muons

Muons can be produced in the decays of W or Z bosons arising in VH events. The reconstruction of muons can be seeded either by tracks found in the muon system, or by tracks from the inner tracker that are extrapolated to the muon system, or both. Identification criteria ensuring the muon is isolated, applied by checking that the p_T and tracks and calorimeter deposits are consistent with the p_T of the muon track, reduces the mis-identification of other charged hadrons. The energy is inferred from the curvature of the inner track for $p_T < 200\,\text{GeV}$. Above this value, additional information from the muon tracks is used to accurately fit the track and improve the precision of the p_T measurement. A full description of the muon reconstruction is given in Ref. [60].

5.7.2 Electrons

Similarly to muons, electrons can be present in leptonic VH decays. Electrons are reconstructed in a very similar way to photons, but requiring, instead of vetoing, a matching track. They can be seeded either by calorimeter deposits or tracks. The energy measurement is performed using a BDT similar to that photons, but which also incorporates track information and further accounts for energy lost by bremsstrahlung. Further detail on the electron reconstruction is available in Ref. [87].

5.7.3 Missing transverse momentum

The presence of neutrinos in an event can be inferred by a global imbalance in the vector sum of the p_T of all objects, resulting in so-called "missing" transverse momentum (E_T^{miss}). The E_T^{miss} computation uses the corrected p_T values for objects, and with all jet energy corrections applied. It can be used to identify the neutrinos resulting from W boson decay in VH events.

5.8 Summary

The CMS detector utilises information from all of its subdetectors in order to optimally reconstruct different types of particles, using a technique known as particle flow. Photon candidates produced by the particle flow algorithm are subjected to a set of selection requirements in order to enter the $\text{H} \to \gamma\gamma$ analysis. These requirements are based on the shape of the electromagnetic shower deposited in the detector and how isolated that shower is relative to other particles in the event. The photon energies are corrected using a multivariate regression, which optimises the resulting diphoton mass resolution of photon pairs selected for the analysis, and thereby improves the analysis sensitivity. A BDT known as the photon identification BDT is trained to reduce the rate at which jet fragments are misidentified as photons. Other types of objects, including jets, are used to discriminate between different Higgs boson production modes.

Chapter 6

Event Categorisation

6.1 Introduction

The analysis event selection consists of the preselection described in Chapter 5, and in addition requires the two leading preselected photon candidates to have $p_T^{\gamma 1} > m_{\gamma\gamma}/3$ and $p_T^{\gamma 2} > m_{\gamma\gamma}/4$ respectively, with an invariant mass in the range $100 < m_{\gamma\gamma} < 180\,\text{GeV}$. The requirements on the scaled photon transverse momenta prevent distortions at the lower end of the mass spectrum. Both photons must also satisfy the pseudo-rapidity requirement $|\eta| < 2.5$ and must not be in the barrel-endcap transition region $1.44 < |\eta| < 1.57$; the reduced containment in this transition region worsens the photon energy resolution. The above η requirement is applied to the photon supercluster position, and the requirement on the photon p_T is applied after the vertex assignment.

The $H \to \gamma\gamma$ analysis depends on the ability to distinguish the narrow signal peak from the smoothly falling background in the diphoton mass distribution. Selected events are therefore subject to further categorisation, in order to increase the ratio of the number of signal events to the number of background events (S/B). This enhances the sensitivity of the analysis, reducing the expected uncertainties on the measured quantities.

Analysis categories are also constructed to target events in which the Higgs boson was produced by a specific production mechanism. This is achieved using the information provided by additional objects in the event, alongside the two photons arising from the Higgs boson decay. As well as facilitating measurements of cross sections corresponding to individual production mechanisms, these dedicated categories also enable the S/B to be improved.

In the previous $H \to \gamma\gamma$ analysis using the 2016 dataset [1], dedicated categories targeting the VBF, ttH, and VH modes were constructed, with the remaining so-

called "Untagged" categories composed mostly of ggH events. Here a similar approach is employed, but with additional divisions targeting individual stage 1 bins for the ggH and VBF processes.

The dataset collected by CMS in 2017 has already been used in an analysis targeting ttH production, which is described in Ref. [86]. In order to ensure that the set of data events included in this analysis is orthogonal from those included in Ref. [86], a veto is applied to events selected for categories targeting ttH production. The criteria used to select ttH include two dedicated BDTs, one targeting events where at least one W boson decays leptonically and the other preferentially selecting fully hadronic W boson decays. Input variables to the two BDTs include photon, lepton, and jet kinematics, information relating to b-tagging, and missing transverse momentum. The veto is implemented by applying these criteria used to construct the ttH categories, and then removing them from further consideration. Around fifteen events are expected to be removed as a result, most of which would otherwise have populated the ggH 2J categories with higher p_T^H values. Furthermore, there are no dedicated analysis categories for VH production; the number of events in the diphoton decay channel with this dataset is insufficient to measure any of the individual stage 1 bins [20].

The categorisation targeting ggH is based on the reconstructed diphoton transverse momentum ($p_T^{\gamma\gamma}$) and the number of jets in the event. A BDT referred to as the diphoton BDT is then used to reduce the amount of background. The VBF analysis categories make use of the same diphoton BDT to reduce the number of background events. Additionally, a BDT targeting the kinematics of the characteristic VBF dijet system, known as the dijet BDT, is utilised to reduce the contamination from ggH events.

Due to conditions differing between the two years, the analysis is optimised separately for the 2016 and 2017 datasets. A simultaneous fit to the categories from both years is then performed to estimate the values of the parameters of interest and their uncertainties (described in Chapter 8). The following section provides an introduction to BDTs and how they are used in physics analyses. The remainder of the chapter then describes in detail the training of the diphoton and dijet BDTs, and their use in the category optimisation process for both ggH and VBF events.

6.2 Boosted decision trees

In the CMS H $\rightarrow \gamma\gamma$ analysis, BDTs are used for several purposes. Generally, the purpose of the BDT is to discriminate between one signal-like and one background-like process. The BDT provides a per-event output score which indicates how signal-like the event is, and a criterion can be applied on this output score when selecting or categorising events. This section gives a brief explanation of what BDTs are and how they are trained.

A BDT is an example of a machine learning algorithm which takes as input a set of features, a set of training parameters, and a loss function, and from those inputs returns an output score [88]. The so-called feature vector $\vec{x}$ is a set of real values for each event which can be used to discriminate between signal and background processes. Examples of features used in the analysis include kinematic variables such as p_T or η. In addition, when training the BDT, the target outcome y must be provided. The events used for training typically come from simulation, where the truth process is known. For all the cases considered in this analysis, y simply takes the value 1 for signal events and 0 for background events. The training parameters $\vec{w}$, also referred to as hyper-parameters, are parameters which control the behaviour of the learning procedure. The loss function L defines the metric on which the learning procedure is optimising. The output score for events being evaluated is denoted Y.

A BDT is an ensemble of decision trees (DTs), which are so-called base learners that recursively split the input feature space into distinct regions by applying binary partitions at each node. Once a given branch reaches its final node where no further splitting is performed, an output value is assigned to that region. The training procedure for an individual DT involves considering many possible configurations for the tree and computing the loss function for each one. The point at which the training process terminates depends upon the hyperparameter values. For example, there may be a restriction on the tree depth, meaning the number of binary partitions permitted for each branch. In this way an optimal configuration is chosen, and any given input for evaluation (without the target label y) will be placed in a single region and assigned the corresponding score.

The boosting procedure combines individual DTs into a more powerful learner. An iterative training process enables the existing models to be improved upon; each new learner corrects the previous version. The final BDT output function then consists of a weighted sum of individual DTs. A type of boosting method known as gradient boosting trains each successive tree on the residuals, or errors, of the existing function, thereby attempting to correct the mistakes made by existing trees. This is equivalent to subtracting the derivative of a squared error loss function; this is the reason the

method is given the name gradient boosting. This gradient boosting procedure can be generalised to use pseudo-residuals, which are given by the derivative of an arbitrary differentiable loss function. The loss function chosen here is distinct from that used to train the individual DTs. Once a new tree is trained on these pseudo-residuals, it is then added to the existing function, itself a weighted sum of trees. The weight of the new tree is assigned by minimising the loss of the new summed function. This procedure is repeated until a given performance threshold or other stopping criterion is met.

The performance of a given BDT training is evaluated on a sample independent of that used to train it, in order to ensure that the features learned by the BDT generalise beyond the specific training dataset used. If the BDT's performance is greater on the training dataset than the so-called test dataset, it is said to have overfitted, or been overtrained. This means it has placed too much importance on statistical fluctuations on the training dataset which will not be present in other samples. The hyperparameters chosen for a given training will affect whether or not the BDT is overtrained. For example, the learning rate reduces the weight assigned to each tree in the iterative learning process, which is analogous to the step size in gradient descent methods. A smaller learning rate will reduce the chance of overfitting but also increase the time taken for the training to converge. For this reason, an optimisation procedure is normally used to choose the hyperparameters in such a way that maximises performance without overtraining the BDT.

6.3 Gluon fusion categorisation

6.3.1 Signal bin definitions

At stage 1 of the STXS framework, the gluon fusion process (ggH) is divided into a total of eleven particle level bins. The events are split first by the number of jets, defined at particle level using the anti-k_T algorithm with radius parameter 0.4 and jet $p_T > 30\,\text{GeV}$ [84]. All stable particles except for the decay products of the Higgs boson are included in the anti-k_T clustering. There are zero (0J), one (1J), and greater than or equal to two (2J) jets bins. In events with at least one jet, a further splitting by the value of the transverse momentum of the Higgs boson (p_T^H) is performed. Four bins are defined by boundaries placed at 60, 120, and 200 GeV. These bins are denoted as low, medium (med), high, and BSM, respectively. Finally, a separate ggH region is dedicated to the vector boson fusion-like phase space, for which a pair of jets (a dijet) with invariant mass $m_{jj} > 400\,\text{GeV}$ and difference in pseudorapidity $\Delta\eta > 2.8$ is required. The dijet is formed from the two leading jets in the event. This region is split

Region	Definition	Fraction	Cross section (pb)
0J	Exactly zero jets, any p_T^H	60.0%	26.49
1J low	Exactly one jet, $p_T^H < 60$ GeV	15.4%	6.79
1J med	Exactly one jet, $60 < p_T^H < 120$ GeV	10.4%	4.61
1J high	Exactly one jet, $120 < p_T^H < 200$ GeV	1.7%	0.76
1J BSM	Exactly one jet, $p_T^H > 200$ GeV	0.4%	0.16
2J low	$\geq$ two jets, $p_T^H < 60$ GeV	2.9%	1.26
2J med	$\geq$ two jets, $60 < p_T^H < 120$ GeV	4.5%	2.00
2J high	$\geq$ two jets, $120 < p_T^H < 200$ GeV	2.3%	1.00
2J BSM	$\geq$ two jets, $p_T^H > 200$ GeV	1.0%	0.43
VBF-like 2J	$\geq$ two jets, $p_T^H < 200$ GeV, $\lvert\Delta\eta\rvert > 2.8$, $m_{jj} > 400$ GeV, $p_T^{Hjj} < 25$ GeV	0.6%	0.27
VBF-like 3J	$\geq$ two jets, $p_T^H < 200$ GeV, $\lvert\Delta\eta\rvert > 2.8$, $m_{jj} > 400$ GeV, $p_T^{Hjj} > 25$ GeV	0.9%	0.38

Table 6.1: The particle level definition of each ggH stage 1 bin and the corresponding fractional and absolute cross sections. The fractions are estimated from simulated ggH H $\rightarrow \gamma\gamma$ events within the region $|y_H| < 2.5$. Details of the simulated samples can be found in Chapter 5. Each bin is exclusive; events passing the VBF-like selection enter the VBF-like region and are not considered for the other ggH 2J bins.

into two-jet-like ($p_T^{Hjj} < 25\,\text{GeV}$) and three-jet-like ($p_T^{Hjj} > 25\,\text{GeV}$) bins, where p_T^{Hjj} is defined as the transverse momentum of the Higgs boson plus dijet system. Each bin is exclusive; events passing the VBF-like selection enter the VBF-like region and are not considered for the other ggH 2J bins. A table summarising the definition of each bin, its cross section, and the fraction of the total ggH cross section is shown in Table 6.1. The inclusive ggH cross section is 48.52 pb at $m_H = 125.09\,\text{GeV}$, computed to an accuracy of three loops in perturbative QCD and next-to-leading order (NLO) in EW perturbations [20, 89, 90]. Approximately 44.2 pb of this is within $|y_H| < 2.5$.

6.3.2 Categorisation strategy

In order to measure the stage 1 bins individually, categories must be constructed which differentiate between them. Thus the reconstruction level event categorisation is designed to target all of the bins to which some sensitivity can be obtained in the H $\rightarrow \gamma\gamma$ decay channel with this dataset. In this case, the majority of the ggH stage 1 bins can be measured individually. The exceptions are the two VBF-like bins, which are difficult to separate from true VBF production. Therefore events not entering the categories targeting VBF production can be assigned to one of nine target ggH stage 1 bins. This assignment is performed using the reconstructed diphoton transverse

momentum ($p_T^{\gamma\gamma}$) and the number of jets – which are respectively the detector level equivalents of the particle level quantities p_T^H and number of jets used to define the bins. The same boundaries are used as for the particle level bins; 0, 60, 120 and 200 GeV in $p_T^{\gamma\gamma}$ for 1J and 2J events, and inclusive in $p_T^{\gamma\gamma}$ for 0J events. Once a target signal bin has been assigned, the diphoton BDT is used to further divide the events into up to three categories. Requirements are placed on the output score of the diphoton BDT, with the category with the highest threshold referred to as "Tag 0", the next highest as "Tag 1", and so on. This reduces the amount of background and therefore improves the analysis' sensitivity to each ggH stage 1 bin.

6.3.3 Diphoton BDT

The diphoton BDT is trained to discriminate signal events (where two photons are produced from the decay of a Higgs boson) from background events (where two photons are produced by other SM processes). The goal of the classifier is to assign high scores to signal-like events; requirements can then be placed on the output score of the classifier to increase the S/B for the analysis categories. The criteria for an event to be signal-like include having signal-like photon kinematics, and having high scores from the photon identification BDT. In addition, events with a good diphoton mass resolution are preferred since this reduces the amount of background present under the signal peak.

To train the diphoton BDT, simulated events from the ggH, VBF, ttH and VH production modes are treated as signal. The event generator used for training the BDT (POWHEG) is different from that used to construct the final signal model. The POWHEG sample is preferred for the BDT training because it does not contain any events with negative weights. It is checked that the input and output distributions from POWHEG agree with those from MadGraph5_aMC@NLO within uncertainties. Simulated background events include the contributions from prompt-prompt, prompt-fake, and fake-fake sources. For both signal and background, events are weighted in accordance with the SM process cross section. Furthermore, events from the QCD sample are down-weighted by a factor of 25; the limited number of events in this sample results in very high per-event weights, which causes the classifier to over-estimate the importance of individual events. All events are required to satisfy the analysis selection criteria. The input variables to the classifier are the kinematic properties of the diphoton system, a per-event estimate of the diphoton mass resolution, and the photon identification scores of each photon. The set of input variables is chosen such that the value of the Higgs boson mass cannot be inferred; it is for this reason that the photon momenta are divided by the diphoton mass. The full list of input variables is

as follows:

- the transverse momentum of the two leading photons, divided by the diphoton mass, $p_T^{1,2}/m_{\gamma\gamma}$;
- the pseudorapidity of the two leading photons, $\eta^{1,2}$;
- the cosine of the angle between the two photons in the transverse plane, $\cos(\Delta\phi)$;
- the output score of the photon identification BDT for the two leading photons;
- the per-event relative mass resolution estimate, under the hypothesis that the mass was computed using the correct primary vertex $\sigma_{rv}/m_{\gamma\gamma}$;
- the per-event relative mass resolution estimate, under the hypothesis that the mass was computed using an incorrect primary vertex $\sigma_{wv}/m_{\gamma\gamma}$;
- the per-event probability estimate that the correct primary vertex was selected, p_{vtx}, which is the output of the vertex probability BDT described in Chapter 5;

The per-event relative mass resolution under the correct vertex hypothesis depends only on the photon energy measurements performed by the ECAL. It is calculated by propagating the photon energy resolution estimates, assuming the resolutions are independent and Gaussian distributed:

$$\sigma_{rv}/m_{\gamma\gamma} = \frac{1}{2}\sqrt{\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2} \tag{6.1}$$

where both the energy ($E_{1,2}$) and energy resolution ($\sigma_{E_{1,2}}$) estimates are obtained from the regression described in Chapter 5. Under the incorrect vertex hypothesis, an additional term must be added to account for the worsening in the diphoton mass resolution due to the uncertainty of the vertex position. The distance between the correct and incorrect vertices is assumed to follow a Gaussian distribution with a width dependent on the size of the beamspot. The contribution from the vertex (σ_{vtx}) can then be computed analytically using the measured positions of the photons in the detector and combined in quadrature with σ_{rv} to calculate σ_{wv}:

$$\sigma_{wv}/m_{\gamma\gamma} = \sqrt{\left(\frac{\sigma_{rv}}{m_{\gamma\gamma}}\right)^2 + \left(\frac{\sigma_{vtx}}{m_{\gamma\gamma}}\right)^2} \tag{6.2}$$

An example of the distribution for one of the input kinematic variables, the lead photon $p_T/m_{\gamma\gamma}$, is shown in Figure 6.1. The background, split into its prompt–prompt, prompt–fake, and fake–fake components, is compared to the data sidebands (where the

region with $115 < m_{\gamma\gamma} < 135\,\text{GeV}$ is excluded). Good agreement between the two is observed.

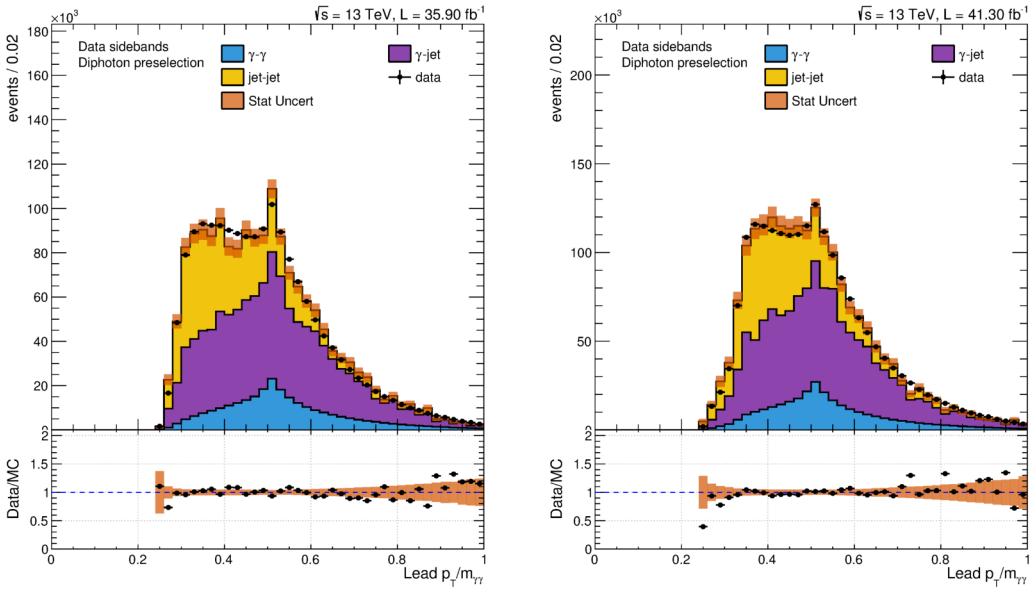


Figure 6.1: Distribution of the leading photon $p_T/m_{\gamma\gamma}$ in background (stacked histogram) and data (black points) events. The statistical uncertainty is shown by the orange band. The left plot shows 2016 data and MC, with 2017 data and MC on the right.

None of the input variables above encode the preference for events with good mass resolution. For this purpose, an additional weight is applied to signal events which increases the relative importance of events with better values of the diphoton mass resolution. The weight applied, w_{res}, is given by the formula

$$w_{res} = \frac{p_{vtx}}{\sigma_{rv}} + \frac{1 - p_{vtx}}{\sigma_{wv}} \tag{6.3}$$

This ensures that the classifier output score is higher for events which have a relatively low expected diphoton mass resolution. Finally, the signal and background samples are divided into a training and a test set, containing 70% and 30% of the total number of events respectively. With this configuration of input variables, signal and background events, and event weights, the classifier is trained and its performance evaluated using the XGBoost software package [91].

The performance of the training is evaluated using a receiver operating characteristic (ROC) curve, where the signal efficiency is plotted as a function of the background efficiency, with each point corresponding to a specific threshold value placed on the classifier output score. The area under the ROC curve is used to gauge the perfor-

mance of a given classifier; a higher area corresponds to more effective discrimination between signal and background. For the 2016 and 2017 trainings, the areas under the respective ROC curves are both equal to 0.85.

This metric is utilised to compare the performance of the classifier training with different values of the BDT's so-called hyper-parameters. These hyper-parameters are parameters of the BDT, which are not learned but instead affect how the learning algorithm behaves. The most important hyper-parameters affect the extent to which the algorithm learns the specific detail of the training sample provided. To check for potential overtraining, the training score as measured by the area under the ROC curve of the training set is compared with the score from the independent test set. The statistical variation of the ROC score can be estimated by varying the random seed for the training and by choosing a different subset of events for the training. If the difference between the test and training set scores is significantly higher than this statistical variation, the classifier has been overtrained. The default classifier hyper-parameters as defined in the XGBoost package [91] display no overtraining. A coarse hyper-parameter optimisation procedure is then performed, in which each hyper-parameter is varied individually. No significant improvement without overtraining is found, and therefore the default training parameters are used.

An additional parameter which can be optimised is the relative weight of the signal and background samples. With the default sample weights corresponding to the SM sample cross section, the two classes are highly imbalanced. This imbalance can cause suboptimal learning in the classifier. To check this, the classifier is trained in a scenario where the signal event weights are increased by a uniform factor, such that the total sum of weights for signal and background is equal. This is purely a technical change to the training designed to improve the learning outcomes of the classifier. When evaluating the performance, the default event weights are used. Equalising the total weights in this way results in a modest improvement in performance. The areas under the ROC curves become 0.87 for both the 2016 and 2017 datasets. The improvement is relatively small, but is however larger than the variation in training performance induced by either changing the random seed used for training or choosing different subsets of the input data for training and testing. Therefore we choose to use this training scenario for the final classification. The hyper-parameter optimisation procedure was repeated and no significant improvement was obtained.

Validation of the diphoton BDT is performed using $Z \to e^+e^-$ events, where simulated Drell-Yan events are compared with data. This validation is important because although the background model used in the analysis is entirely data-driven, the signal model is taken from simulation. Therefore it is necessary to ensure that there is

reasonable agreement between data and simulation for signal-like objects, for both the inputs to the diphoton BDT and the output score itself. In this $Z \to e^+e^-$ control region, the electrons are reconstructed as photons, and the presence of an electron pair with invariant mass $80 < m_{ee} < 100\,\text{GeV}$ is required. Otherwise the event selection is the same as the analysis selection for diphoton events, except for the additional requirement that the leading electron has $p_T > 40\,\text{GeV}$. This requirement is necessary to ensure that no bias is introduced by the electron trigger, which has selection thresholds at $p_T = 32\,\text{GeV}$ (35 GeV) for 2016 (2017) data.

Figure 6.2 shows the output score of the diphoton BDT for data and simulation. The effect of two systematic uncertainties which affect the diphoton BDT are included; these are the shift in the photon identification BDT score, and the uncertainty on the photon energy resolution (see Chapter 7 for details). Good agreement is observed between the data and simulation in both years, with all discrepancies covered by the statistical and systematic uncertainties.

6.3.4 Category definitions

Once the diphoton BDT has been constructed, a category optimisation procedure can be performed for each Stage 1 bin independently. The reconstructed number of jets and $p_T^{\gamma\gamma}$ define the category type into which a given event falls. Then, independently for each category type, an optimisation procedure is performed to define diphoton BDT boundaries for a given number of subcategories. The Approximate Mean Significance (AMS) is used to define the figure of merit for the optimisation. The value of the AMS metric corresponds to the expected significance of a signal from the likelihood ratio statistic for a simple counting experiment, neglecting the impact of systematic uncertainties. Its derivation is outlined in Ref. [76]. The formula for the AMS for a given analysis category is given by

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

where S is 68% (corresponding to $\pm 1\sigma_{eff}$) of the number of signal events from the desired truth bin (not the total number of signal events), and B is the background. The value of B is calculated by performing an exponential fit to the background, and then integrating the number of events between $125\,\text{GeV} - \sigma_{eff} < m_H < 125\,\text{GeV} + \sigma_{eff}$. This formula reduces to $S/\sqrt{S+B}$ in the limit of small S/B. Results are found to be robust to the choice of metric; only in bins with a very low number of events (e.g. the BSM bins) does the AMS metric return different optimal diphoton BDT boundaries from the $S/\sqrt{S+B}$ metric, and then the differences are relatively small. The total

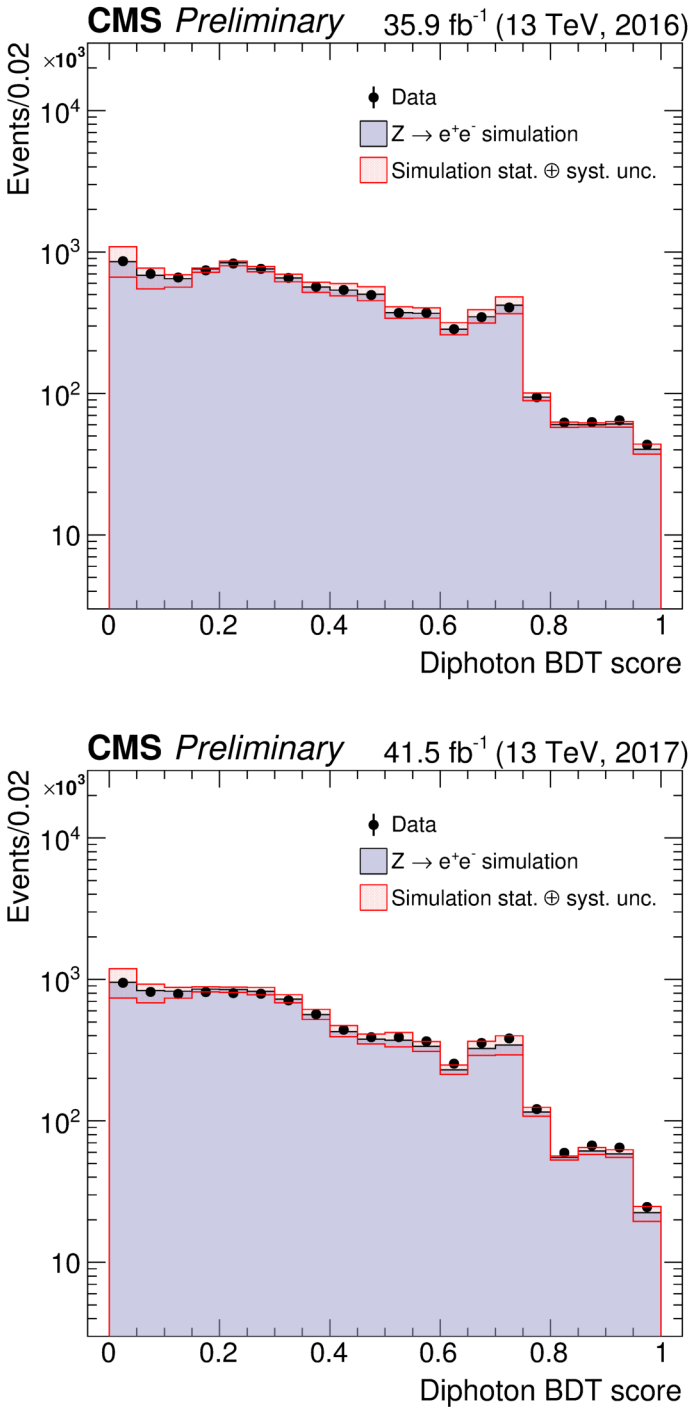


Figure 6.2: Score of the diphoton BDT in $Z \to e^+e^-$ events where the electrons are reconstructed as photons. The points show the score for data, the histogram shows the score for simulated Drell–Yan events, including statistical and systematic uncertainties (pink band). The top plot shows 2016 data and MC, with 2017 data and MC on the bottom. Figures first shown in Ref. [4].

AMS significance is computed by summing the values for each analysis category in quadrature.

The optimisation procedure itself is performed using a random search. This is found to be more computationally efficient than a grid search. After the diphoton BDT boundaries have been established, a cross-check is performed to ensure that a sensible minimum has been found. This is done by evaluating the AMS score for diphoton BDT values close to the nominal optimal point individually for each category type and each boundary. If a slightly better performing point is found in the vicinity, it replaces the value returned by the random search. The cross-check is designed to check that the random search has found a stable minimum; this is determined by examining the shape of the scan around the optimal point. In general, the results of the random search are found to be robust.

This optimisation process is repeated for different numbers of categories. No improvement beyond two categories per target STXS bin is observed, except for the 0J bin, which requires three categories. The boundaries chosen and the expected number of signal and background events, together with the expected significance, are shown in Table 6.2 and Table 6.3 for 2016 and 2017 simulation and data respectively. These tables illustrate the tendency for events with higher p_T^H to have higher diphoton BDT scores.

Category	Diphoton BDT value	S	B	AMS
0J Tag 0	> 0.851	145	1000	4.48
0J Tag 1	$0.796 - 0.851$	201	2973	3.64
0J Tag 2	$0.586 - 0.796$	238	9263	2.46
1J low Tag 0	> 0.832	36	462	1.67
1J low Tag 1	$0.718 - 0.832$	45	1635	1.11
1J med Tag 0	> 0.866	17	1042	1.60
1J med Tag 1	$0.749 - 0.866$	39	7755	1.38
1J high Tag 0	> 0.908	5.5	18	1.14
1J high Tag 1	$0.810 - 0.908$	6.6	112	0.61
1J BSM Tag 0	> 0.917	3.0	7.4	0.89
GE2J low Tag 0	> 0.833	4.8	142	0.39
GE2J low Tag 1	$0.709 - 0.833$	6.7	571	0.28
GE2J med Tag 0	> 0.869	8.3	65	0.99
GE2J med Tag 1	$0.757 - 0.869$	17	462	0.80
GE2J high Tag 0	> 0.910	9.1	33	1.45
GE2J high Tag 1	$0.811 - 0.910$	10	158	0.81
GE2J BSM Tag 0	> 0.938	9.7	9.4	2.48
GE2J BSM Tag 1	$0.865 - 0.938$	4.9	27	0.86

Table 6.2: The chosen diphoton BDT boundaries, the expected number of signal (S) and background (B) events, and the expected significance (defined by the AMS metric) of each category in the ggH phase space for 2016 data and simulation, assuming an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$. The categories targeting each stage 1 bin are ordered such that "Tag 0" has the highest S/B value, "Tag 1" the next highest, and so on.

Category	Diphoton BDT value	S	B	AMS
0J Tag 0	> 0.840	217	2042	4.72
0J Tag 1	$0.769 - 0.840$	250	5063	3.49
0J Tag 2	$0.615 - 0.769$	201	9669	2.03
1J low Tag 0	> 0.817	41	676	1.58
1J low Tag 1	$0.652 - 0.817$	42	2466	0.84
1J med Tag 0	> 0.881	8.5	45	1.18
1J med Tag 1	$0.760 - 0.881$	43	872	1.46
1J high Tag 0	> 0.910	5.5	20	1.10
1J high Tag 1	$0.824 - 0.910$	7.6	109	0.71
1J BSM Tag 0	0.775	4.9	1.2	2.02
GE2J low Tag 0	> 0.861	1.9	44	0.27
GE2J low Tag 1	$0.709 - 0.861$	8.0	649	0.31
GE2J med Tag 0	> 0.835	11	177	0.87
GE2J med Tag 1	$0.786 - 0.835$	6.9	10	1.74
GE2J high Tag 0	> 0.916	6.3	24	1.17
GE2J high Tag 1	$0.815 - 0.916$	10	163	0.78
GE2J BSM Tag 0	> 0.901	11	23	2.15
GE2J BSM Tag 1	$0.596 - 0.901$	3.0	1.8	1.25

Table 6.3: The chosen diphoton BDT boundaries, the expected number of signal (S) and background (B) events, and the expected significance (defined by the AMS metric) of each category in the ggH phase space for 2017 simulation and data, assuming an integrated luminosity of $41.5\,\mathrm{fb}^{-1}$. The categories targeting each stage 1 bin are ordered such that "Tag 0" has the highest S/B value, "Tag 1" the next highest, and so on.

Region	Definition	Fraction		Cross
		VBF	VH had	section (pb)
BSM	Leading jet $p_T > 200$ GeV	4.6%	5.4%	0.23
2J-like	$\geq$ two jets, $\lvert\Delta\eta\rvert > 2.8$, $m_{jj} > 400$ GeV, $p_T^{Hjj} < 25$ GeV	25.8%	0.4%	0.91
3J-like	$\geq$ two jets, $\lvert\Delta\eta\rvert > 2.8$, $m_{jj} > 400$ GeV, $p_T^{Hjj} > 25$ GeV	9.0%	1.7%	0.34
VH-like	$\geq$ two jets, $60 < m_{jj} < 120$ GeV	2.3%	34.5%	0.55
Rest	All other VBF events	59.2%	57.9%	2.86

Table 6.4: The particle level definition of each VBF stage 1 bin and the corresponding fractional and absolute cross sections. The fractions reported are normalised relative to inclusive VBF or VH hadronic production, whilst the cross sections are the sum of the VBF and VH hadronic values. The fractions are estimated from simulated VBF and hadronic VH H $\rightarrow \gamma\gamma$ events within the region $\lvert y_H\rvert < 2.5$. Details of the simulated samples can be found in Section 5. Each bin is exclusive; all bins except the BSM bin are required to have the leading jet $p_T < 200$ GeV.

6.4 Vector boson fusion categorisation

6.4.1 Signal bin definitions

The VBF process is divided into five particle level bins at stage 1 of the STXS framework. Events where the Higgs boson is produced in association with a vector boson (VH, where V = W or Z) and the vector boson decays hadronically are included together with VBF. In this region, there are two bins defined analogously to the VBF-like bins for ggH production, requiring a dijet with $m_{jj} > 400\,\text{GeV}$ and $\Delta\eta > 2.8$, split by a 25 GeV boundary in p_T^{Hjj}. In addition, there is a "VH-like" bin, which requires the presence of a dijet with $60 < m_{jj} < 120\,\text{GeV}$. A "BSM-like" bin is also defined, where the p_T of the leading jet is greater than 200 GeV. All remaining events fall into the "VBF rest" bin. Each bin is exclusive; all bins, except for the BSM bin, are required to have the leading jet $p_T < 200\,\text{GeV}$. Table 6.4 shows a summary of the definition of each bin and the corresponding fraction of the total VBF cross section. The inclusive VBF cross section is 3.779 pb at $m_H = 125.09\,\text{GeV}$ [20], computed at approximately next-to-next-to-leading (NNLO) order in QCD and NLO in EW, meaning some but not all NNLO diagrams are included in the calculation. Of this approximately 3.52 pb is within $\lvert y_H\rvert < 2.5$. For hadronic VH production, the cross section at $m_H = 125.09\,\text{GeV}$ of 1.54 pb [20] is computed at NNLO in QCD and NLO in EW, with around 1.37 pb within $\lvert y_H\rvert < 2.5$.

6.4.2 Categorisation strategy

Different categorisation scenarios are considered for the VBF process. Ideally categories would be constructed to target each stage 1 bin. However, the experimental sensitivity to the BSM-like, VH-like, and VBF rest bins is limited. Care is therefore taken not reduce the sensitivity to inclusive VBF production when designing the categories for each of these scenarios.

The bins which can be most precisely measured are the 2J-like and 3J-like VBF bins. In addition to migrations between the categories targeting each bin, there is substantial contamination from events with two jets arising from ggH production. For this reason the dijet BDT is trained to discriminate between ggH and VBF events. The categorisation is then performed using the detector level equivalents of the quantities used to define the bins: m_{jj}, p_T^{Hjj}, and leading jet p_T. The dijet and diphoton BDTs are subsequently used to reduce the respective number of ggH and background events entering the categories, thus increasing their sensitivity.

6.4.3 Dijet BDT

The dijet BDT is trained to discriminate VBF events from both background and ggH events. This is made possible due to the distinctive signature of VBF events, which typically have two jets with large separation in pseudorapidity and high invariant mass. Therefore the input variables to the dijet BDT consist primarily of jet-related variables. In addition, the jets in VBF events originate from quarks, whereas ggH events typically originate from gluons. This results in subtle differences in the internal structure of the jet objects, as discussed in Chapter 4. Adding the jet shape variables used there as inputs does not significantly improve the performance of the dijet BDT; however it has been shown that more sophisticated techniques using neural network classifiers and detailed jet inputs can improve the performance [92]. The full set of input variables for the dijet BDT are:

- the transverse momentum of the two leading photons, divided by the diphoton mass, $p_T^{1,2}/m_{\gamma\gamma}$;
- the transverse momentum of the two leading jets, $p_T^{j1,j2}$;
- the invariant mass of the dijet, m_{jj};
- the magnitude of the difference in pseudorapidity of the two leading jets, $|\Delta\eta|$;
- the magnitude of the difference in azimuthal angle between the two leading jets, $|\Delta\phi_{jj}|$;

- the magnitude of the difference in azimuthal angle between the diphoton and dijet, $|\Delta\phi_{\gamma\gamma,jj}|$;
- the minimum angular separation between either of the two leading photons and either of the two leading jets, $\Delta R_{\min}(\gamma, j)$;
- the centrality variable, which is given by:

$$C_{\gamma\gamma} = \exp\left(-\frac{4}{(\eta_{j1} - \eta_{j2})^2}\left(\eta_{\gamma\gamma} - \frac{\eta_{j1} + \eta_{j2}}{2}\right)^2\right) \tag{6.4}$$

  where η_{j1} and η_{j2} are the pseudorapidities of the leading and subleading jets respectively.

In previous versions of the analysis [1], both the signal and background for the dijet BDT training have been taken directly from simulation. However there are a limited number of events in the prompt-fake and fake-fake background samples which pass the analysis preselection, and this problem is exacerbated by applying additional requirements on the dijet system. Consequently, there are few events with which to train the dijet BDT, and those events which are present have extremely high weights. This results in multiple issues that reduce the effectiveness of the dijet BDT. If these high-weight events are included in the training, the classifier assigns too much importance to specific instances and does not successfully learn generalised features of the input datasets. Attempts have been made to mitigate this, the first of which is to reduce the weight of the QCD events by a factor of 25. In addition, the VBF preselection is loosened in order to include more events in the training. The standard VBF preselection consists of the analysis preselection with the additional requirements of $p_T > 40$ (30) GeV for the leading (subleading) jet, $m_{jj} > 250\,\text{GeV}$ and photon identification BDT score of greater than -0.2 for both photons. The loosened version reduces the p_T thresholds by 10 GeV each, the m_{jj} threshold to 100 GeV, and removes the tighter photon identification requirement. This slightly improves the training performance; however the background rejection is still suboptimal due to the change in phase space and reduction in event weights. Furthermore, this makes the simulation very difficult to validate against data since the distributions are highly discontinuous. An illustration of the effect is shown in Figure 6.3, which shows the discontinuous m_{jj} distribution after the VBF preselection is applied.

An alternative solution to this problem is to use events from a data control region to replace the MC samples with low numbers of events. In this way, the more abundant data events are used to replace the simulated prompt-fake and fake-fake events used to train the dijet BDT. There are a sufficient number of prompt-prompt events in

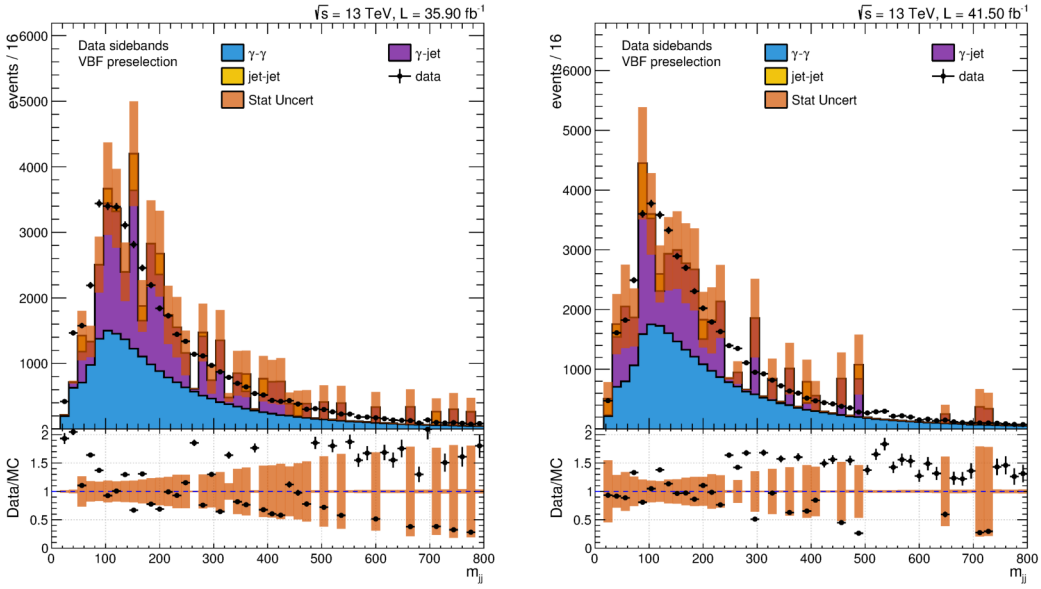


Figure 6.3: Distribution of the dijet invariant mass in background (stacked histogram) and data (black points) events. The statistical uncertainty is shown by the orange band. The left plot shows 2016 data and MC, with 2017 data and MC on the right.

MC, which do not need replacement and continue to be taken from simulation for the training. This method is used for the first time in this analysis; the outline of the procedure is as follows:

- define a control region by inverting the photon identification BDT requirement for events entering the VBF signal region;
- use simulation to compute the relative fraction of events in the signal region and control region, for each photon, as a function of the photon kinematics;
- apply these factors as event weights to the data in the control region and use these events to replace the simulation of prompt-fake and fake-fake events in the training.

There are four distinct regions considered in the data-driven replacement method. Events entering the analysis comprise the signal region, where both photons pass the photon identification BDT requirements; the total number of events is N_{pp}. There are two control regions, where one photon passes the requirement and one fails it. The number of events in each are denoted by N_{pf} (where the leading photon passes and the subleading photon fails) and N_{fp} (where the leading photon fails and the subleading

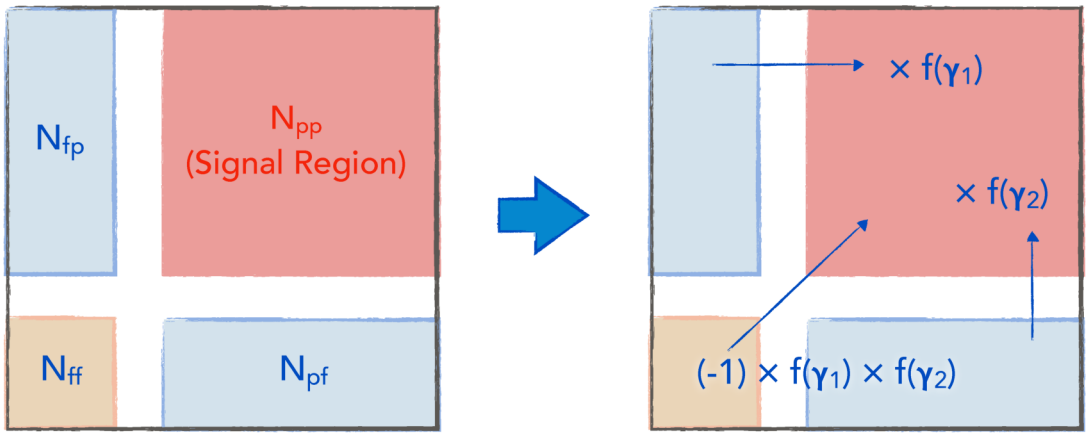


Figure 6.4: A schematic illustrating the data-driven method for replacing the simulated prompt-fake and fake-fake events with reweighted data events from control regions defined by the photon identification BDT. The ratio of events which pass the photon identification requirement to events that fail is taken from simulation in bins of p_T and η, and used to calculate so-called "fake-factors". The fake-factor for the leading photon is denoted by $f(\gamma_1)$ and the fake-factor for the subleading photon by $f(\gamma_2)$. These fake factors are then applied to events in data where one or more photons fail the photon identification requirement, and these reweighted events replace the simulation in the training of the dijet BDT.

photon passes). Lastly there is the region where both photons fail the requirement, with N_{ff} events. Figure 6.4 illustrates how the method works.

Once these regions have been defined, it is possible to define so-called "fake-factors" to extrapolate from the number of events in the control regions to the expected number of events in the signal region. These factors are calculated from simulation, in bins of p_T and η; this ensures that the distributions of p_T and η are well-modelled. Kinematic variables other than p_T and η are assumed to be similar across the control and signal regions. This is checked by comparing the distributions constructed using the data-driven replacement method with data in the $m_{\gamma\gamma}$ sideband. In order to minimise the statistical error on the fake factors, the loosened VBF preselection is used. The full VBF preselection is then applied before training, once the events have been weighted by the fake factors.

The expression for the fake factor consists of the factor which extrapolates from the control region to the signal region, w_{fake}, multiplied by the fraction of events in the

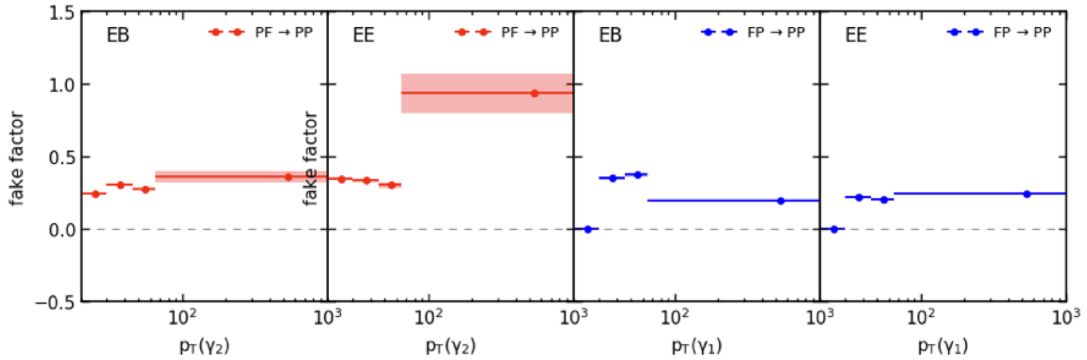


Figure 6.5: Fake factor values in each of the four p_T and two η bins. The left two plots show the values where the subleading photon is a fake, whilst in the right two plots the leading photon is fake. The values are taken from simulation.

data which are either prompt-fake or fake-fake, w_{QCD}. The second term is required since we wish to extract only the prompt-fake and fake-fake components from the data; the prompt-prompt component is still taken from simulation. Each term is applied to the photon which fails the identification requirement, and depends on the p_T and η of this photon. This can be written as:

$$w_{\text{fake}} = w_{\text{fake}}(\eta, p_T) = \left(\frac{N^{SR}}{N^{CR}}\right)_{MC},$$

$$w_{\text{QCD}} = w_{\text{QCD}}(\eta, p_T) = \left(\frac{N^{CR}_{\text{pf}} + N^{CR}_{\text{ff}}}{N^{CR}_{\text{pp}} + N^{CR}_{\text{pf}} + N^{CR}_{\text{ff}}}\right)_{MC},$$

$$f = f(\eta, p_T) = w_{\text{fake}} \times w_{\text{QCD}},$$

where N^{SR} and N^{CR} are the number of events in the signal and control regions respectively; each is a function of the photon p_T and η. The number of prompt-prompt, prompt-fake, and fake-fake events in the control regions are also a function of photon p_T and η, and are denoted by N^{CR}_{pp}, N^{CR}_{pf}, and N^{CR}_{ff} respectively. The fake factors are calculated in four bins of p_T and two bins of η. Figure 6.5 shows the values of the factors in each of these bins.

Validation of the data-driven method is performed to confirm that its output is a suitable replacement for the simulation. The new training inputs, comprising the simulated prompt-prompt events and data-driven prompt-fake and fake-fake events, are compared with data from sidebands in the diphoton mass distribution. The usual signal region selection is applied to the sideband data, but with the requirement that the diphoton mass does not lie in the region $115\,\text{GeV} < m_{\gamma\gamma} < 135\,\text{GeV}$. These events cannot be used for training the dijet BDT since their reuse in the final fits could then

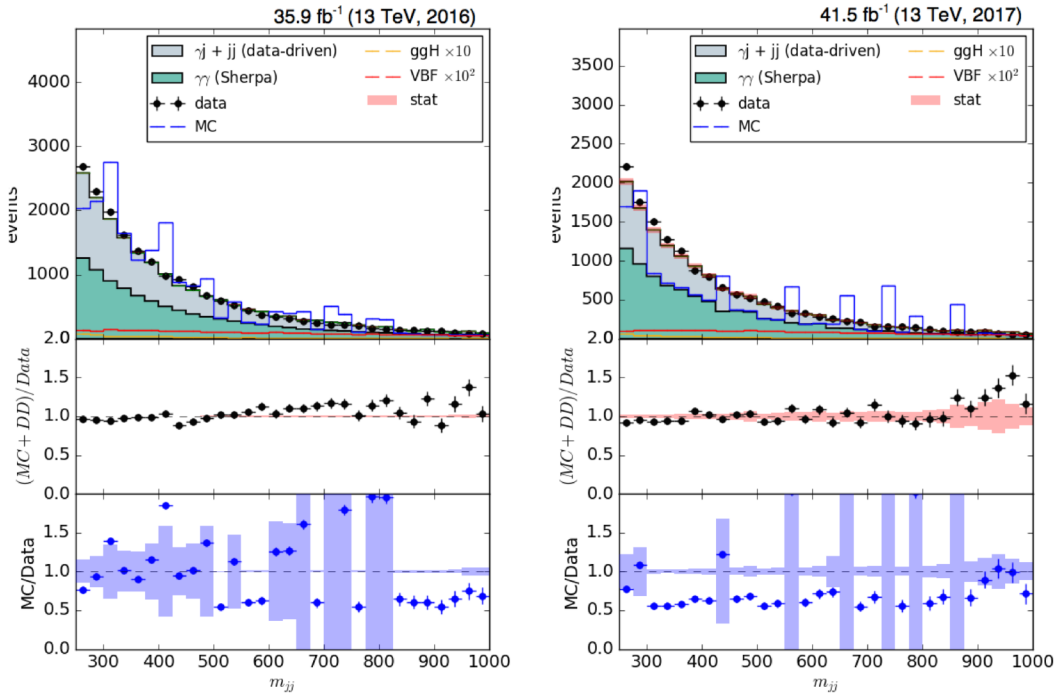


Figure 6.6: Validation of the data-driven method. The upper parts of the plots show the simulated background (blue histogram), the background obtained with simulation for the prompt-prompt component and utilising the data-driven method for the prompt-prompt and prompt-fake components (green and grey stacked histograms), and mass sideband data (black points). The distributions of the ggH (yellow line) and VBF (red line) are also shown. Two ratio plots are also included, with one comparing the data-driven method to the sideband data, and the other comparing the simulation with the sideband data. Data and simulation from 2016 are shown on the left, with 2017 on the right.

induce bias; however they are useful for validating the method, since their kinematic properties should be essentially identical to the data-driven output. Figure 6.6 shows the good agreement between the data-driven output and the sideband data. The figure also shows how the method solves the problem of the high-weight events present in the MC; this confirms that it fulfils its purpose.

The dijet BDT is then trained with the data-driven inputs. For the 2016 dataset, the area under the ROC curve is 0.87. This constitutes a significant improvement over the score when the BDT is trained only with simulation, which is 0.84. As an additional cross-check, the BDT is trained with events from the data sideband. The area of this ROC curve is also 0.87, which confirms that the observed improvement is reasonable. The performance is 2017 is very similar, with the data-driven method improving the

ROC score from 0.83 to 0.86. After the category boundaries have been optimised, this results in an increase in the final value of the AMS score of approximately 10%.

The agreement between data and simulation for the dijet BDT score is validated in a similar way to the diphoton BDT. Events from the $Z \to e^+e^-$ control region are required to pass the VBF preselection. Figure 6.7 shows the output score of the dijet BDT for data and simulation. The systematic uncertainties included in the plot are those affecting the jet energy scale and the jet energy resolution corrections. Good agreement is observed between the data and simulation in both years, with all discrepancies covered by the statistical and systematic uncertainties.

6.4.4 Category definitions

In previous versions of the analysis, the categories targeting VBF production were defined using the so-called "combined BDT". The combined BDT is designed to incorporate both the dijet and diphoton information in order to optimally construct the VBF categories. Its input variables are:

- the diphoton BDT score
- the dijet BDT score
- the transverse momentum of the two leading photons, divided by the diphoton mass, $p_T^{1,2}/m_{\gamma\gamma}$;

The combined BDT is trained with VBF as signal against the three sources of SM background, using simulated events in each case. Gluon fusion is not included in the training because it is found to worsen the performance of the combined BDT to reject non-Higgs boson background. Since these backgrounds have a greater impact on the final sensitivity than the contamination from other signal processes, this is considered a higher priority. However, it does mean that the ggH rejection of the combined BDT can be sub-optimal. Here the efficacy of the combined BDT is compared with the more direct approach of defining the VBF categories by imposing thresholds on the diphoton and dijet BDT scores directly.

In addition to comparing the two different BDT approaches, different possible categorisation scenarios are considered. In Ref. [1], three inclusive categories are defined by placing thresholds on the output score of the combined BDT. For this analysis, further splitting using kinematic variables is required if the different VBF stage 1 bins are to be measured individually. When comparing different categorisation scenarios, the figure of merit considered is the total VBF significance, computed using the AMS

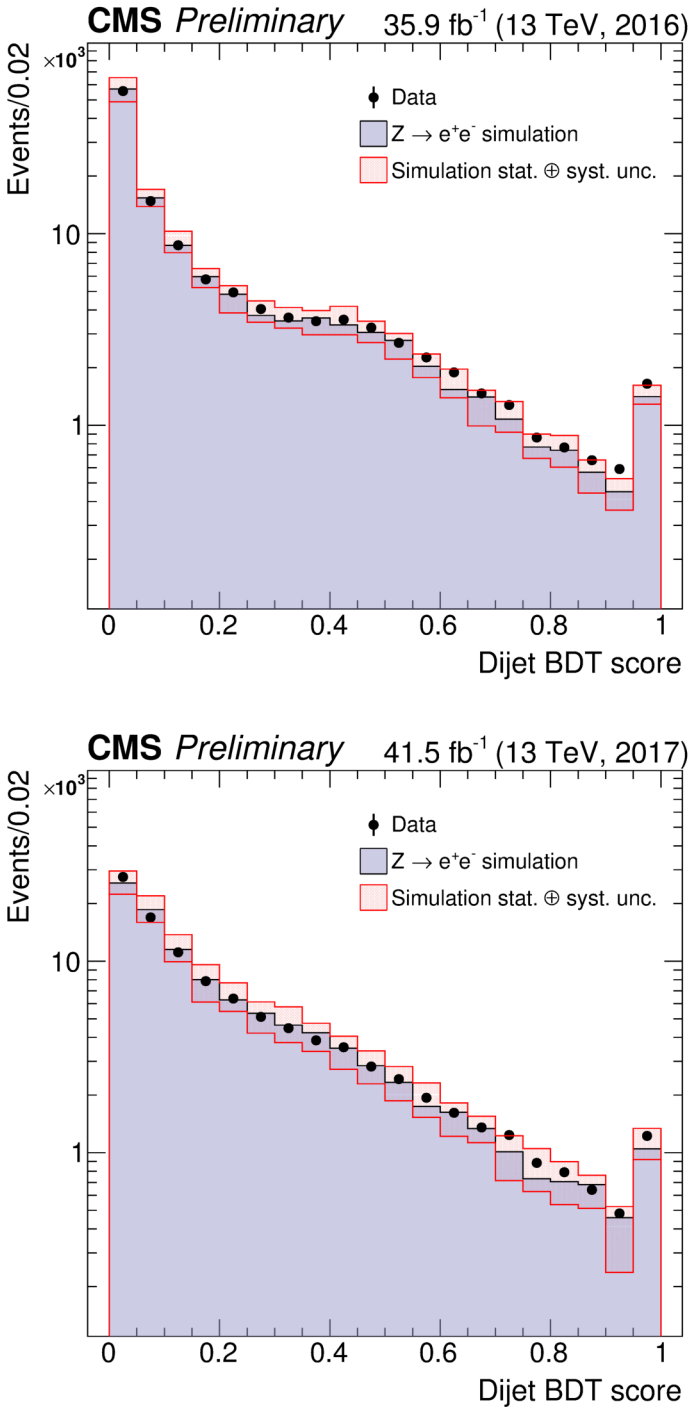


Figure 6.7: Score of the dijet BDT in $Z \to e^+e^-$ events where the electrons are reconstructed as photons. The points show the score for data, the histogram shows the score for simulated Drell–Yan events, including statistical and systematic uncertainties (pink band). The top plot shows 2016 data and MC, with 2017 data and MC on the bottom. Figures first shown in Ref. [4].

metric with all VBF bins considered as signal. Scenarios with additional splits following the stage 1 bin definitions will then be considered preferable, provided they do not substantially reduce the overall VBF significance, since they enable individual stage 1 bins to be measured in addition to the overall VBF cross section. In each scenario, there is a single dedicated VBF BSM category which requires that the p_T of the leading jet is greater than 200 GeV. The list of scenarios considered is as follows:

- **Inclusive two category:** Two categories are considered, with no additional kinematic selection criteria applied. Including the VBF BSM category, this gives three categories in total.
- **Inclusive three category:** Three categories are considered, with no additional kinematic selection criteria applied. It is almost equivalent to the approach in Ref. [1], aside from the additional VBF BSM category. This gives a total of four categories.
- **Split by p_T^{Hjj}:** Two categories are considered for each of the 2J-like and 3J-like VBF bins, using a boundary at 25 GeV on the reconstructed value of p_T^{Hjj}. This gives a total of five categories.
- **Split by p_T^{Hjj} and m_{jj}:** Two categories are considered for each of the 2J-like and 3J-like VBF bins, using a boundary at 25 GeV on the reconstructed value of p_T^{Hjj}. Additionally, the requirement that m_{jj} is greater than 400 GeV is applied. A fifth category targeting principally the "VBF rest" bin, with $250 < m_{jj} < 400$, is therefore included. This gives a total of six categories.

A boundary optimisation procedure is followed for each scenario, using the same methodology as for the ggH categories. For the approach which uses the diphoton and dijet BDT output scores directly, the values of the thresholds on the two BDTs are optimised simultaneously for each category. The resulting significance values of each scenario for the 2016 and 2017 datasets are shown in Tables 6.5 and 6.6 respectively. The performance of the combined BDT approach is compared with the approach using the diphoton and dijet BDT scores directly in each case. For both years, the combined BDT performs slightly worse than the direct use of the diphoton and dijet BDTs. This can be attributed to the fact that the combined BDT is trained only on simulation; it is not possible to use the data-driven method for its training since the photon information is used as inputs. Furthermore, the results demonstrate that there is no substantial reduction in the overall VBF significance with the most granular categorisation scenario. It is therefore considered the best scenario for facilitating measurements of individual stage 1 bins, and is chosen as the VBF categorisation scheme for the analysis.

Scenario	Dijet + Diphoton BDT	Combined BDT
Inclusive 2 category	2.43σ	2.26σ
Inclusive 3 category	2.55σ	2.31σ
Split by p_T^{Hjj}	2.38σ	2.28σ
Split by p_T^{Hjj} and m_{jj}	2.53σ	2.43σ

Table 6.5: The total VBF significance (defined by the AMS metric) for different categorisation scenarios using 2016 data and simulation, assuming an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$. Classification using the combined BDT is compared with setting boundaries with the diphoton and dijet BDTs directly.

Scenario	Dijet + Diphoton BDT	Combined BDT
Inclusive 2 category	2.16σ	2.02σ
Inclusive 3 category	2.18σ	2.09σ
Split by p_T^{Hjj}	2.14σ	2.03σ
Split by p_T^{Hjj} and m_{jj}	2.18σ	2.08σ

Table 6.6: The total VBF significance (defined by the AMS metric) for different categorisation scenarios using 2017 data and simulation, assuming an integrated luminosity of $41.5\,\mathrm{fb}^{-1}$. Classification using the combined BDT is compared with setting boundaries with the diphoton and dijet BDTs directly.

Tables 6.7 and 6.8 show the final thresholds for the diphoton and dijet BDTs in each category, together with the expected number of signal and background events and the per-category significance.

Category	Dijet BDT value	Diphoton BDT value	S	B	AMS
2J-like Tag 0	> 0.12	> 0.62	8.2	7.0	2.04
2J-like Tag 1	$-0.89 - 0.12$	> 0.72	3.1	14.9	0.68
3J-like Tag 0	> 0.48	> 0.61	4.7	8.8	1.07
3J-like Tag 1	$-0.84 - 0.48$	> 0.74	3.2	35.7	0.46
VBF BSM Tag	> -0.41	> 0.73	2.2	7.7	0.53
VBF Rest Tag	> -0.74	> 0.77	2.5	34.3	0.36

Table 6.7: The chosen diphoton and dijet BDT boundaries, the expected number of signal (S) and background (B) events, and the expected significance (defined by the AMS metric) of each category in the VBF phase space for 2016 data and simulation, assuming an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$. The categories targeting each stage 1 bin are ordered such that "Tag 0" has the highest S/B value, "Tag 1" the next highest, and so on.

Category	Dijet BDT value	Diphoton BDT value	S	B	AMS
2J-like Tag 0	> -0.39	> 0.75	7.5	9.2	1.73
2J-like Tag 1	$-0.89 - -0.39$	> 0.68	1.5	14.5	0.33
3J-like Tag 0	> 0.03	> 0.77	4.7	12.0	1.00
3J-like Tag 1	$-0.69 - 0.03$	> 0.57	2.1	36.9	0.31
VBF BSM Tag	> -0.27	> 0.78	2.3	6.3	0.59
VBF Rest Tag	> -0.62	> 0.73	2.8	38.0	0.38

Table 6.8: The chosen diphoton and dijet BDT boundaries, the expected number of signal (S) and background (B) events, and the expected significance (defined by the AMS metric) of each category in the VBF phase space for 2017 data and simulation, assuming an integrated luminosity of $41.5\,\mathrm{fb}^{-1}$. The categories targeting each stage 1 bin are ordered such that "Tag 0" has the highest S/B value, "Tag 1" the next highest, and so on.

6.5 Summary

A set of H $\rightarrow \gamma\gamma$ analysis categories are defined in order to maximise the overall sensitivity of the analysis and to enable the measurement of different signal bins. Categories targeting nine different subdivisions of the ggH production mode are constructed, using the reconstructed $p_T^{\gamma\gamma}$ and number of jets to infer the most likely signal bin. The diphoton BDT is then used to further split these events into categories of differing S/B, which increases the precision of the measurement of each bin. Furthermore, categories targeting the different VBF signal bins are constructed. A dijet BDT is trained using a novel data-driven approach, and is used to discriminate against both ggH production and background processes. It is used together with the diphoton BDT and the reconstructed values of m_{jj}, p_T^{Hjj} and the number of jets to define the final VBF analysis categories.

Chapter 7

Signal and background modelling

7.1 Introduction

The final results of this analysis are extracted by performing a maximum likelihood fit of the signal and background models to the diphoton invariant mass distribution observed in data. In this fit, the value of twice the negative log-likelihood (2NLL) is minimised for each value of the parameter to be measured. The resulting 2NLL curve can be used to extract the best-fit value and uncertainty of the parameter of interest. In addition, systematic uncertainties are included in the fit to account for imperfect knowledge of both theoretical and experimental inputs to the analysis. These uncertainties are implemented in the form of nuisance parameters. A nuisance parameter is any parameter of the model for which an uncertainty interval is not constructed. In the final fit, nuisance parameters are "profiled". This means their values are allowed to vary, either freely or with an additional constraint applied which penalises deviation from the nominal expectation. This has the effect of widening the 2NLL curve and thereby increasing the overall uncertainty on a given measurement. Full details of the statistical procedure used to extract the results are given in Chapter 8.

In order to perform the fits, models of the signal and background $m_{\gamma\gamma}$ distributions in each analysis category are required as inputs. The signal model is derived from simulation, with a model constructed for each particle level stage 1 cross section bin in each reconstructed analysis category. Both the shape and normalisation of the model are parameterised as a function of m_H. The data-driven background model considers a range of different functional forms to represent the smoothly falling background spectrum, following the approach described in Ref. [93]. This methodology follows that used in the previous CMS H $\rightarrow$ $\gamma\gamma$ analysis, which is described in Ref. [1]. The construction of both the signal and background models is described in detail in this chapter. In addition, the treatment of systematic uncertainties affecting the two

models is discussed.

7.2 Signal modelling

The signal model is a parametric function of m_H which describes the shape of the $m_{\gamma\gamma}$ distribution, together with the expected normalisation of this shape. An independent model is constructed for each stage 1 bin in each analysis category Additionally, since the $m_{\gamma\gamma}$ shape depends on whether the right vertex (RV) or wrong vertex (WV) has been chosen, the model for each of these cases is constructed separately.

Each model consists of a sum of up to five Gaussian functions. The number of Gaussian functions required depends on the shape of the $m_{\gamma\gamma}$ spectrum and the available quantity of simulated events. Alternative functional parameterisations have been studied, such as the sum of a Gaussian function to represent the core of the $m_{\gamma\gamma}$ distribution and an exponential to model each tail [94]. This alternative parameterisation yields very similar signal models to the sum of Gaussian functions, and the final results are unaffected. The Gaussian-based model is retained due to its simplicity and computational efficiency.

The parameters of the Gaussian functions are determined by performing a fit to the simulated $m_{\gamma\gamma}$ distribution for each model. In order to account for the fact that the mass of the Higgs boson is not known exactly, the model constructed is a continuous parametric function of m_H. The dependence on m_H is determined by simultaneously fitting events simulated with three values of m_H: 120, 125, and 130 GeV. Each parameter of each Gaussian function is represented as a linear function of m_H; the shape of each model then consists of $2\,(3N_{\text{Gaus}} - 1)$ parameters, where N_{Gaus} is the chosen number of Gaussian functions. The values of these parameters are established by the simultaneous fit across mass points. An example of the evolution of signal model shapes as a function of m_H, for the ggH 0J bin the 0J Tag 0 category, is shown in Figure 7.1 By construction, the shape and normalisation of the models vary smoothly with m_H.

In some cases, there is an insufficient number of simulated events available to accurately model the shape for a given combination of stage 1 bin, analysis category and vertex scenario. The shape is then replaced by that of the stage 1 bin which has the highest expected yield in the category under consideration. This replacement procedure is motivated by the fact that events subject to the same selection tend to have similar values of the diphoton mass resolution.

After each model has been constructed, the shapes from the RV and WV scenarios are combined. The fraction of events in which the correct vertex is chosen is also

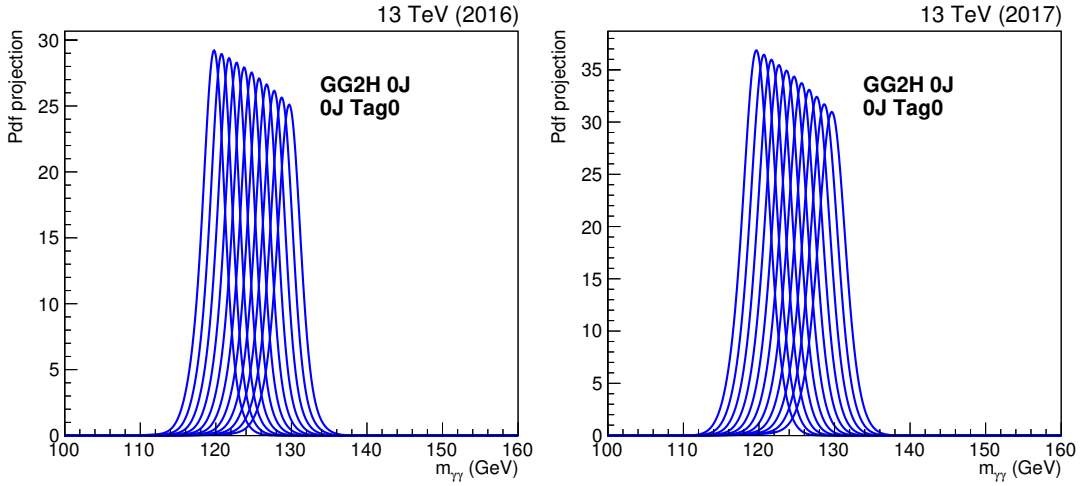


Figure 7.1: The evolution of the parametrised signal model shape as a function of m_H. The plots show the signal model for the ggH 0J bin in the 0J Tag 0 category. The left plot shows 2016 simulation, with 2017 simulation on the right.

described by a linear function of m_H; once determined, this is used to assign the correct normalisations for the RV and WV models. An example of the signal model for the ggH 0J bin in the 0J Tag 0 category, after the models for the RV and WV scenarios have been summed, is shown in Figure 7.2.

For each category, the models corresponding to the contributions from each stage 1 bin are then summed. To normalise the contribution from each stage 1 bin correctly, the total number of expected events for each stage 0 process is obtained using the cross sections and H $\rightarrow \gamma\gamma$ branching ratio from Ref. [20][1], and the measured integrated luminosity in data. The fraction of each stage 1 bin is then taken from the simulated events for each stage 0 process. Finally, the product of the detector efficiency and analysis acceptance is modelled as a linear function of m_H, determined by the ratio of the total number of expected events to the number of events entering each analysis category. Together these allow the total signal model for each category to be computed. An example of the signal model for the 0J Tag 0 category, after the models for all the stage 1 bins have been summed, is shown in Figure 7.3.

[1]The exception is ggH, for which the latest calculations at N3LO are used [89, 90].

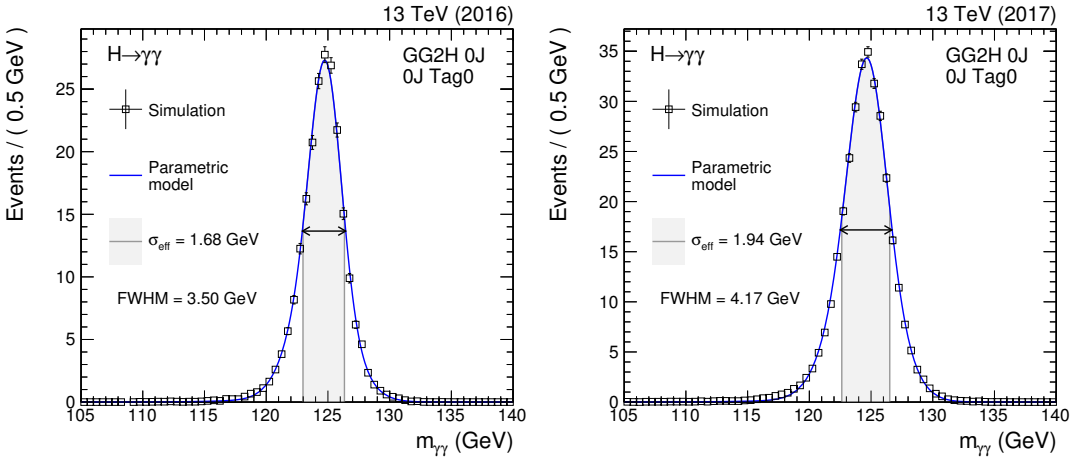


Figure 7.2: The parametrised signal shape for the ggH 0J bin in the 0J Tag 0 category, after the models for the RV and WV scenarios have been summed. The open squares represent weighted simulation events and the blue line the corresponding model. Also shown is the σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) and the full width at half of the maximum (FWHM). The left plot shows 2016 simulation, with 2017 simulation on the right.

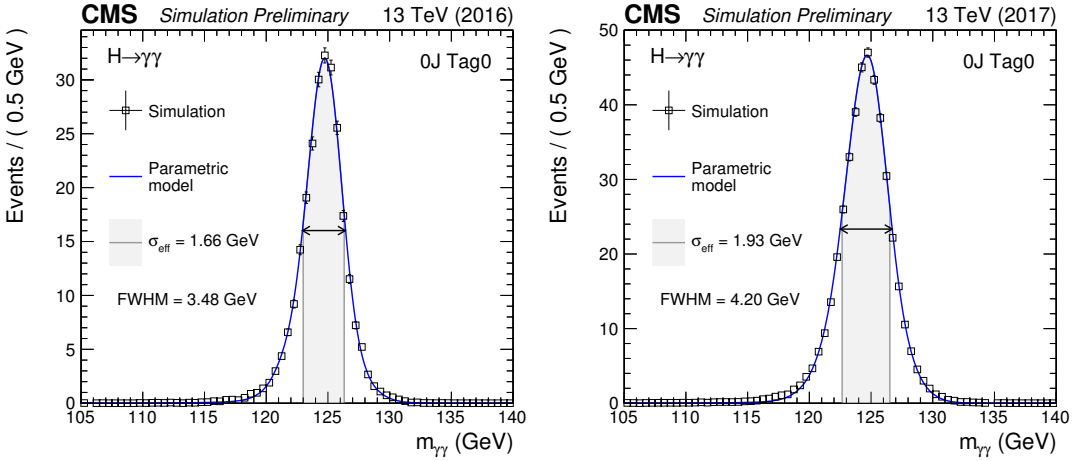


Figure 7.3: The parametrised signal shape for the 0J Tag 0 category, after the models for all the stage 1 bins have been summed. The open squares represent weighted simulation events and the blue line the corresponding model. Also shown is the σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) and the full width at half of the maximum (FWHM). The left plot shows 2016 simulation, with 2017 simulation on the right. Figures first shown in Ref. [4].

7.3 Background modelling

The background model represents the smoothly falling spectrum in the $m_{\gamma\gamma}$ distribution that results from processes other than Higgs boson production. The shape of this falling distribution is not known a priori; therefore different functional forms must be considered when constructing the model. Each choice of function results in a different number of estimated events under the signal peak produced by the Higgs boson, and hence affects the measured value of parameters representing the size of the signal contribution. The uncertainty in the measurement associated with this choice must be accounted for in the final results. In this analysis, the discrete profiling method is used, as first described in Ref. [93]. The model is constructed independently for each analysis category.

The discrete profiling method incorporates the uncertainty on the background into the measurement by treating the choice of function used to model the background distribution as a discrete nuisance parameter. Ordinarily, a nuisance parameter is a continuous variable which affects the measured value but is not in itself of any interest. The choice of background function can be treated in the same way as any other nuisance parameter, with the only difference being that its value is discrete rather than continuous. To construct the final 2NLL curve, the so-called "envelope" of all the possible choices of background function is taken. This procedure is illustrated in Figure 7.4, which shows how the 2NLL curve for the fully profiled fit can be approximated by taking the envelope of the curves generated with a nuisance parameter fixed to different values. The discrete profiling method accounts for the uncertainty in the background function analogously; a different curve is generated by each choice of background function, and the final curve represents the envelope of each of these individual curves. This envelope is necessarily wider than any of the curves corresponding to an individual function, and therefore returns a greater uncertainty, reflecting the imperfect knowledge of the shape of the background distribution.

In principle, all functions which provide a sufficiently good fit to the observed $m_{\gamma\gamma}$ distribution in data could be included in the discrete profiling method. To apply the method in practice, four different families of functions representing smoothly falling distributions are considered.

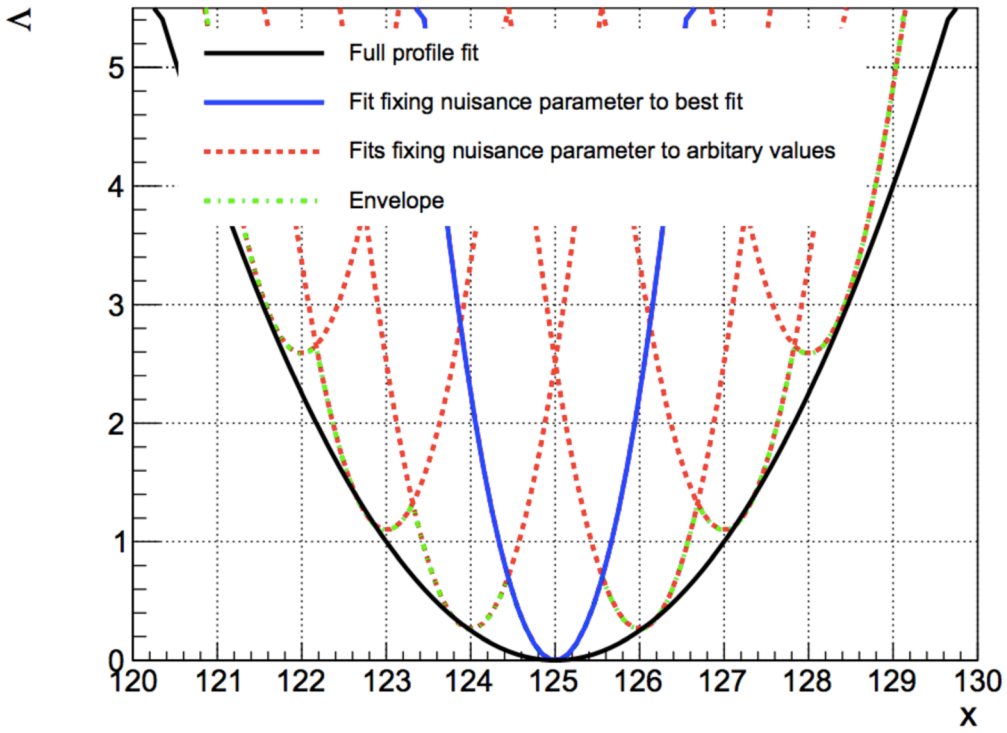


Figure 7.4: An illustration of the discrete profiling, or envelope, method. The envelope of all the curves corresponding to fixed values of a given nuisance parameter approximates the full curve obtained when the nuisance parameter is profiled. Figure taken from Ref. [93].

These groups, for an N-parameter function with parameters $p_0, p_1, ..., p_N$ to be determined, are:

- Sum of exponential functions, where

$$f_N(x) = \sum_{i=0}^{N} p_{2i} \exp\left(p_{2i+1} x\right)$$

- Sum of power law functions, where

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

- Bernstein polynomials, where

$$f_N(x) = \sum_{i=0}^{N} p_i \binom{N}{i} x^i \left(1-x\right)^{N-i}$$

- Laurent series, where

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

with

$$L(i) = \sum_{j=0}^{i} (-1)^j j.$$

Only a subset of functions from each family are considered; this is necessary to maintain a reasonable computational efficiency for the method. For each family of functions, multiple orders can be included in the final set of candidate functions. A likelihood fit is performed for each order, and then the procedure to decide which orders to include works as follows. A penalty equal to the number of parameters in the function is added to the value of the 2NLL; this prevents functions of arbitrarily high order being chosen. First, the order is increased until a minimum goodness-of-fit threshold is reached; there is no need to include functions which do not fit the data well. For each subsequent order, an F-test [95] is performed to gauge the size of the improvement in the fit quality which the increase in function complexity brings. This test computes a p-value by assuming the difference in 2NLL values is distributed as a χ^2, with degrees of freedom equal to the difference in number of parameters between the two orders. If the p-value is below a certain threshold, the higher order function is deemed to constitute a worthwhile improvement, is added to the set of functions

considered, and the process continues. Otherwise, the higher order function is not included and the procedure is complete. The different functions chosen for the 1J high Tag 0 category are shown in Figure 7.5; it can be seen that different choices of function lead to a different number of events when integrating under the signal peak. In this case, the range in the integrated number of events around the Higgs boson mass amongst the candidate functions shown is of the order ten.

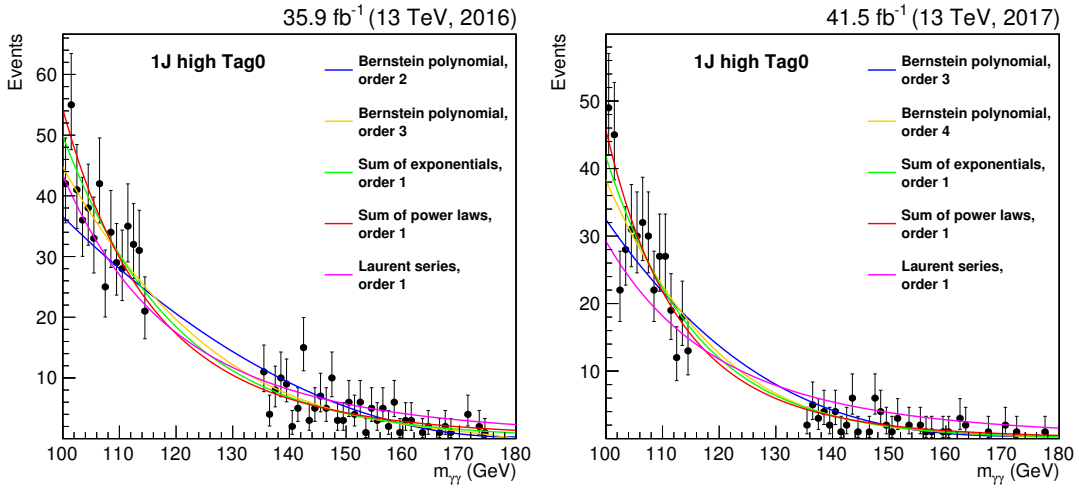


Figure 7.5: The functions chosen for consideration in the final fit for the 1J high Tag 0 category, with the data (black points) blinded, meaning the points with $115 < m_{\gamma\gamma} < 135\,\text{GeV}$ are not shown. The left plot shows 2016 data, with 2017 data on the right.

In the final fit, the background function is chosen from the set of candidates as described above. All parameters of each function are free to vary and determined in the fit. As for the F-test above, a regularisation term is added to the 2NLL for each parameter in the chosen function. This serves to penalise unnecessary complexity. Further details of the discrete profiling method, including studies of the bias and coverage, are contained within Ref. [93].

7.4 Systematic uncertainties

In this analysis, the systematic uncertainty associated with the data-driven background estimation is handled using the discrete profiling method, as described above. There are many systematic uncertainties which affect the signal model; these are handled in one of two ways. Uncertainties which modify the shape of the $m_{\gamma\gamma}$ distribution are incorporated into the signal model as nuisance parameters, where the mean and width of each Gaussian function can be affected. These uncertainties are typically

experimental uncertainties relating to the energy of the individual photons. Conversely if the shape of the $m_{\gamma\gamma}$ distribution is unaffected, the uncertainty is treated as a log-normal variation in the event yield. These uncertainties include theory sources and experimental uncertainties such as those affecting the BDTs used for categorising events. The magnitude of each uncertainty's impact is determined individually for each stage 1 bin in each analysis category.

7.4.1 Theoretical uncertainties

The sources of theoretical uncertainties considered in this analysis are as follows:

- *QCD scale uncertainty*: the uncertainty arising from variations of the renormalisation and factorisation scales used when computing the expected SM cross section and event kinematics. These account for the missing higher order terms in perturbative calculations. The recommendations provided in Ref. [20] are followed. This involves estimating the uncertainty using three sources: varying the renormalisation by a factor of two, varying the factorisation scale by a factor of two, and varying both in the same direction simultaneously. Depending on the production process, the size of the uncertainty varies from around 0.5% to 9%.

- *Uncertainty on the SM ggH production*: for ggH production, additional sources are included which account for the uncertainty in the modelling of the p_T^H distributions, the number of jets in the event, and the ggH contamination in the VBF categories. Two sources reflect the migration uncertainty around the p_T^H bin boundaries, at 60 and 120 GeV respectively; their magnitude depends on the number of jets and the p_T^H in the event. An additional source covers the uncertainty on p_T^H arising from the treatment of the top quark mass in the ggH loop. Two further sources account for the migration between the zero, one and two or more jet bins. The uncertainty on the ggH production of events with a VBF-like dijet system is covered by two sources corresponding to the prediction in the two-jet-like and three-jet-like bins. The total magnitude of these uncertainties vary from around 5% to 30%, with events that have one or more jets and high values of p_T^H typically having the greatest associated uncertainty.

- *PDF (parton density functions) uncertainties*: these account for the uncertainty due to imperfect knowledge of the composition of the proton, which affects which partons are most likely to initiate high energy events. The overall normalisation uncertainties are computed following the PDF4LHC prescription [40, 96], while the bin-to-bin migrations are calculated from the NNPDF3.0 [97] PDF set using

the MC2HESSIAN procedure [98]. The overall normalisation uncertainties are between 1-5%, with the migrations significantly smaller, usually less than 1%.

- *Uncertainty in the strong force coupling constant*: the uncertainty in the value of the strong force coupling constant α_s is included in the treatment of the PDF uncertainties, following the PDF4LHC prescription.

- *Uncertainty in the* $H \to \gamma\gamma$ *branching fraction*: the probability of the Higgs boson decaying to two photons is required to calculate the SM expected cross section, but this branching fraction is not known exactly. The uncertainty is currently estimated to be 2% [20].

- *Underlying event and parton shower uncertainties*: these uncertainties are obtained using dedicated simulated samples where the choice and specific tune of the event generator have been modified. The impact of the uncertainty ranges from 1-16% depending on the process and category.

The QCD scale and PDF uncertainties impact the theoretical modelling in two ways. Both the overall cross section prediction for a given STXS bin, and the distributions of kinematic variables used in the event selection and categorisation, can be affected. The effects on the overall process cross section predictions are omitted when measurements of simplified template cross sections are performed, and are instead considered as uncertainties on the SM prediction. However the uncertainties on the event kinematics, which affect the efficiency and acceptance of the analysis, are still taken into account. All uncertainties are included when measurements of signal strength modifiers are performed. Other uncertainties which only affect the overall SM predicted yields, such as the $H \to \gamma\gamma$ branching fraction, are also omitted from measurements of simplified template cross sections.

7.4.2 Experimental uncertainties

The uncertainties which affect the shape of the signal $m_{\gamma\gamma}$ distribution are:

- *Photon energy scale and resolution*: the uncertainties associated with the corrections applied to the photon energy scale in data and the resolution in simulation are evaluated using $Z \to e^+e^-$ events. The estimate is computed by varying the regression training scheme, the R_9 variable, and the electron selection criteria.

- *Non-linearity of the photon energy scale*: a further source of uncertainty covers possible remaining differences in the linearity of the photon energy scale between data and simulation. The uncertainty is estimated using boosted $Z \to e^+e^-$ events.

- *Modelling of electromagnetic showers*: this uncertainty accounts for the imperfect knowledge of electromagnetic showering processes. The size of the effect is estimated by changing the model used to generate the bremsstrahlung energy spectrum, which can modify the photon and electron energy scales.

- *Shower shape corrections*: an uncertainty on the shower shape corrections accounts for the imperfect modelling of shower shapes in simulation. The impact is estimated by comparing the energy scale before and after the corrections to shower shape variables (which improve the agreement between data and simulation) are applied.

- *Light collection non-uniformity*: an uncertainty is associated with the modelling of the light collection as a function of emission depth within a given ECAL crystal. It is estimated by comparing simulation with analytical estimates of longitudinal shower profiles.

- *Modelling of material in front of the ECAL*: the amount of material through which objects pass before reaching the ECAL affects the behaviour of the electromagnetic showers, and may not be perfectly modelled in simulation. Dedicated samples with variations in the amount of upstream material are used to estimate the impact on the photon energy scale.

- *Vertex fraction*: the largest contribution to the right and wrong vertex fraction uncertainty comes from the modelling of the underlying event, in addition to the uncertainty on the ratio of data and simulation obtained using $Z \to \mu^+\mu^-$ events. It is handled as an additional nuisance parameter built into the signal model which allows the fraction of events in the right vertex/wrong vertex scenario to change.

The uncertainties which only modify the event yield include:

- *Integrated luminosity*: estimated to be 2.5% and 2.3% for the 2016 and 2017 datasets respectively [99, 100].

- *Photon identification BDT score*: the uncertainty arising from the photon identification BDT score is estimated by requiring the systematic variation to cover the observed discrepancies between data and simulation. This is described further in Chapter 5. The uncertainty on the signal yields is estimated by propagating this uncertainty through the full category selection procedure. The impact in the most sensitive analysis categories is around 5%.

- *Jet energy scale and smearing corrections*: The energy scale of jets is measured using the p_T balance of jets with Z bosons and photons in $Z \to e^+e^-$, $Z \to \mu^+\mu^-$ and γ+jets events, as well as the p_T balance between jets in dijet and multijet events [101]. The uncertainty in the jet energy scale is a few per-cent and depends on p_T and η. The impact of jet energy scale uncertainties on event yields is evaluated by varying the jet energy corrections within their uncertainties and propagating the effect to the final result. The resulting uncertainty in the final analysis categories can be as high as 21% for the scale, but is less than around 4% for the resolution.

- *Per-photon energy resolution estimate*: the uncertainty on the per-photon resolution is parametrised as a rescaling of the resolution by $\pm 5\%$ about its nominal value. This is designed to cover all differences between data and simulation in the distribution, which is an output of the energy regression.

- *Trigger efficiency*: the efficiency of the trigger selection is measured with $Z \to e^+e^-$ events using the tag-and-probe technique. The size of its uncertainty is less than 1%.

- *Photon preselection*: the uncertainty on the preselection efficiency is computed as the ratio between the efficiency measured in data and in simulation. Its magnitude is less than 1%.

- *Missing transverse energy*: this uncertainty is computed by shifting the reconstructed p_T of the particle candidates entering the missing transverse energy computation, within the momentum scale and resolution uncertainties appropriate to each type of reconstructed object, as described in Ref. [101]. In this analysis, the size of the uncertainty is very small.

- *Pileup jet identification*: the uncertainty on the pileup jet classification output score is estimated by comparing the score of jets in events with a Z boson and one balanced jet in data and simulation. The magnitude is of the order 1%.

- *Lepton isolation and identification*: this uncertainty affecting electrons and muons is computed by varying the ratio of the efficiencies and using the tag and probe technique in $Z \to e^+e^-$ events. The resulting impact on the categories selecting leptons is up to around 3%.

- *Efficiency of b-tagging*: uncertainties on the b-tagging efficiency are evaluated by comparing data and simulated distributions for the b-tag discriminator. The

uncertainties include the statistical component on the estimate of the fraction of heavy and light flavour jets in data and simulation. Its magnitude is around 1%.

7.4.3 Correlation of uncertainties

Since the analysis inputs and category definitions are constructed independently for the 2016 and 2017 datasets, the uncertainties affecting each year can either be chosen to be correlated or uncorrelated. If an uncertainty is deemed to be correlated, then there is only one nuisance parameter in the final fit whose value will affect both the 2016 and 2017 categories. Otherwise, if the uncertainties are uncorrelated, there is a separate nuisance parameter for each year and the impact on the categories from each year is independent.

In this analysis, all the sources of theory systematic uncertainties are taken to be correlated between years. In most cases this is clearly the correct treatment, since the underlying theoretical predictions are the same for each year and independent of the data-taking conditions. For the experimental uncertainties, most uncertainties are left uncorrelated. This is motivated by the fact that both the data-taking conditions and the reconstruction vary between the two years. An exception is the uncertainty on the photon identification BDT output, which is kept correlated because the uncertainty prescription is identical for both years. The final results were also computed with all the experimental uncertainties taken to be correlated, and it was confirmed that the difference between the two correlation scenarios is minimal. This is in accordance with expectation, since the predominant source of uncertainty in all the stage 1 measurements is statistical in origin.

7.5 Summary

The inputs to the final fits of this analysis are the simulated signal model and the data-driven background model. The uncertainty on the background model is accounted for using the discrete profiling method, where the choice of background function is treated as a nuisance parameter. The signal model is parametric in m_H, with the contribution from each stage 1 bin in each analysis category modelled separately. There are numerous uncertainties included in the signal model, which either affect the shape of the signal mass distribution or the event yield. An illustration of the impact of the systematic uncertainties on different signal strength measurements, from the previous CMS H $\rightarrow \gamma\gamma$ analysis documented in Ref. [1], is shown in Figure 7.6. The systematic uncertainties with the greatest impact on the inclusive Higgs boson signal strength are the photon identification uncertainty and theoretical sources relating to ggH produc-

tion. In contrast, the leading systematic uncertainties affecting VBF signal strength measurement are those on the jet energy scale and resolution corrections. The same sources of systematic uncertainties are included in this analysis. The precision of the measurements of stage 1 simplified template cross sections presented here is limited by statistical uncertainties. However this will change as more data are collected, and once the full dataset from Run 2 is analysed the systematic uncertainties will become more important.

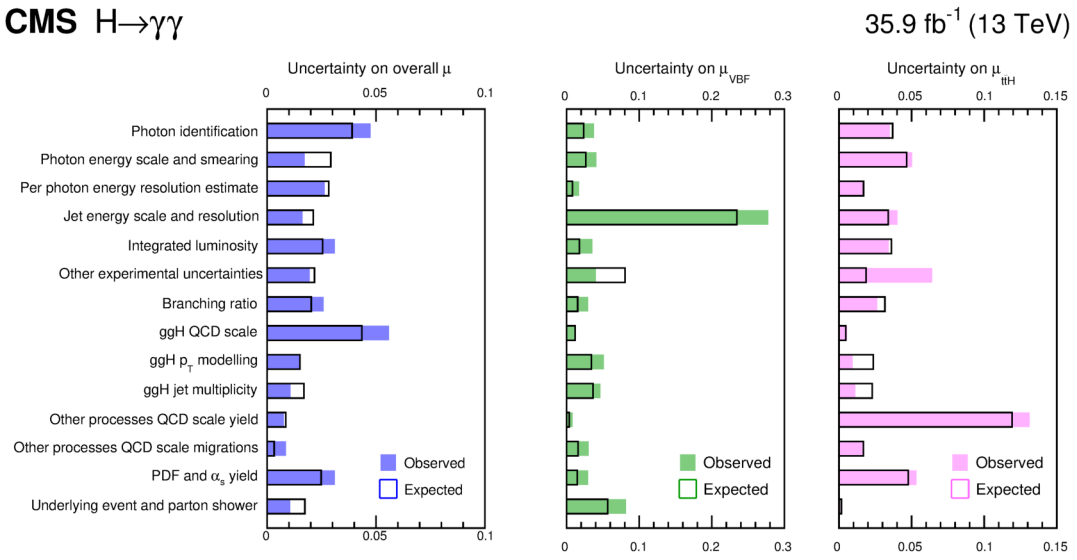


Figure 7.6: The impact of the different systematic uncertainties on the inclusive, VBF, and ttH signal strength modifiers in the previous CMS H $\rightarrow \gamma\gamma$ analysis [1]. The observed (expected) results are shown by the solid (empty) bars. The expected uncertainties are constructed using the Asimov dataset [**Cowan**]. The observed uncertainties are those arising in the fit to data. This figure was first shown in Ref. [1], and is based on the results described therein.

Chapter 8

Results

8.1 Introduction

The principal aim of this analysis is to measure Higgs boson simplified template cross sections, at both stage 0 and stage 1, and their associated uncertainties. This is achieved by performing a simultaneous fit of the signal and background models to the observed $m_{\gamma\gamma}$ distribution in each category. A binned maximum likelihood fit is performed in the range $100 < m_{\gamma\gamma} < 180\,\text{GeV}$, with a bin size of 250 MeV; this is sufficiently small relative to the diphoton mass resolution that a negligible amount of information is lost.

The likelihood function in each category, $\mathcal{L}_c$, is expressed as:

$$\mathcal{L}_c(\text{data}\,|\,\vec{\sigma}, m_H, \vec{\theta}) = \prod_{i=1}^{N_b} \text{Poisson}\left(d_i \,|\, s_i(\vec{\sigma}, m_H, \vec{\theta}) + b_i(\vec{\theta})\right) \times C(\vec{\theta}), \tag{8.1}$$

where $\vec{\sigma}$ is the set of parameters of interest (POIs), which in this analysis are always a set of one or more cross section parameters; $\vec{\theta}$ is the set of nuisance parameters which affect the measurements but are not themselves of interest; N_b is the number of bins used in the category's $m_{\gamma\gamma}$ distribution; Poisson indicates a Poisson function evaluated with the observed number of events in the i^{th} bin d_i and expected number of events given by the sum of the signal expectation s_i and the background expectation b_i; $C(\vec{\theta})$ is the constraint term which penalises deviations from the expected values of the signal nuisance parameters, and applies a penalisation term according to the total number of degrees of freedom in the background model. The expected number of signal events in each bin depends on the POIs, m_H, and the nuisance parameters, whilst the expected number of background events depends only on unconstrained nuisance parameters.

The total likelihood $\mathcal{L}$ is then given by the product of the likelihoods over all

categories:

$$\mathcal{L}(\text{data}\,|\,\vec{\sigma}, m_H, \vec{\theta}) = \prod_{c=1}^{N_c} \mathcal{L}_c(\text{data}\,|\,\vec{\sigma}, m_H, \vec{\theta}), \tag{8.2}$$

where N_c is the total number of analysis categories. The fit is then performed by minimising the value of the negative log-likelihood, 2NLL, where

$$2\text{NLL} = -2\ln\mathcal{L}(\text{data}\,|\,\vec{\sigma}, m_H, \vec{\theta}). \tag{8.3}$$

The free parameters in the fit are the parameters of interest, mH, and the background nuisance parameters; the signal nuisance parameters can vary but are constrained by the $C(\theta)$ term. The 2NLL is constructed and minimised numerically within the RooFit [102] software package for statistical data analysis. The values of the parameters which give the minimum value of the 2NLL are then described as the "best-fit" values.

A frequentist approach is followed in order to extract the uncertainties on the POIs, in addition to their best-fit values. The likelihood ratio test statistic, 2ΔNLL, is constructed for a range of POI values:

$$2\Delta\text{NLL} = -2\ln\frac{\mathcal{L}(\text{data}\,|\,\vec{\sigma}, \hat{\hat{m}}_H, \vec{\hat{\hat{\theta}}})}{\mathcal{L}(\text{data}\,|\,\vec{\hat{\sigma}}, \hat{m}_H, \vec{\hat{\theta}})}, \tag{8.4}$$

where $\hat{\hat{m}}_H$ and $\vec{\hat{\hat{\theta}}}$ are the best-fit values of the Higgs boson mass and nuisance parameters at the POI values $\vec{\sigma}$; $\vec{\hat{\sigma}}$, $\hat{m}_H$, and $\vec{\hat{\theta}}$ are the global best-fit values of the POIs, Higgs boson mass, and nuisance parameters respectively. The distribution of the likelihood ratio test statistic can then be used to infer the approximate uncertainties on the measurements. For a sufficient number of events, the distribution tends to that of a χ^2 [76], where the number of degrees of freedom is equal to the number of POIs being measured. In this case, the 68% confidence level (CL) intervals are given approximately by the corresponding region for a χ^2 distribution, which depends on the number of degrees of freedom. For a single POI, the region is defined by $2\Delta\text{NLL} < 1$. The interpretation of the 68% CL intervals within the frequentist paradigm is that in an ensemble of identical pseudo-experiments, the observed interval should contain the true value of the POI in 68% of cases. The crossing points of the 2ΔNLL at ± 1 are therefore quoted as the 68% CL uncertainties on the POI in question.

In this analysis, the POIs considered are Higgs boson simplified template cross sections, which are defined at various levels of granularity and denoted by the symbol σ. The stage 0 cross sections are equivalent to the sum of the individual stage 1

cross sections. This makes clear that parameters can be defined as sums of different STXS bins, not just individual bins themselves. Measuring a wider set of STXS bins provides more information, but the uncertainties are correspondingly larger than if fewer parameters are measured. Therefore in this section results are reported under various scenarios, with between one and thirteen POIs in total.

In each case, the fit is performed with cross sections as the parameters of interest. After each cross section has been measured, the value is divided by the SM prediction. This procedure differs from that used to measure a signal strength μ, as defined in Chapter 2, where the parameter in the fit is the ratio of the observed cross section to the SM prediction. The key difference between the two is that in the signal strength measurements, the uncertainty on the SM prediction must be considered in the fit because it enters directly in the denominator of the parameter of interest. In contrast, the STXS measurements do not include these uncertainties on the SM yield; the measured cross sections do not directly depend on the SM prediction. This ensures the measurements are as independent as possible of the current SM predictions and their uncertainties, which means they remain useful if theoretical advances are made in the future.

In the following section, the observed diphoton mass distribution and the composition of the analysis categories are presented. The remainder of chapter describes the results of stage 0 and stage 1 measurements within the STXS framework. All results were first reported in Ref. [4].

8.2 Observed diphoton mass distributions

The observed diphoton mass distribution is displayed together with the result of a signal plus background fit in Figure 8.1. The fit contains one signal parameter, which includes all signal events where $|y_H| < 2.5$. The fit is performed simultaneously to all analysis categories. In the plot, each category is summed with a weight corresponding to the ratio of signal events to background events. The uncertainty on the background prediction is also shown. The signal peak due to Higgs boson production is clearly visible.

The result of the same single parameter fit is also shown in Figure 8.2. In this case, only certain subsets of categories are included in the sum. The $m_{\gamma\gamma}$ distributions for the weighted sum of the categories targeting ggH 0J, ggH 1J, ggH 2J, and VBF production are shown. The plots indicate the total number of events and approximate signal to background ratio for the different processes targeted in this analysis.

The full set of unweighted diphoton mass distributions for each category considered

in the analysis are contained in Appendix A.

8.3 Composition of analysis categories

All analysis categories are contaminated, to varying extents, by background events and other signal processes which are not being targeted. The level of contamination then affects the sensitivity of the analysis when the final fits are performed. Tables 8.1 and 8.2 show the expected number of signal events for the 2016 and 2017 datasets respectively. The relative contribution to each category from each of the individual stage 0 bins is shown, together with the σ_{eff} and σ_{HM} (the FWHM divided by 2.35) for the category's signal model. Also reported is the expected number of background events per GeV in a $\pm 1\sigma_{\text{eff}}$ window around 125 GeV, calculated using the best-fit background function. The table illustrates how the ratio of expected ratio of signal to signal plus background events (S/S+B) is highest in the "Tag 0" categories, with lower S/S+B values but a greater number of events overall as the tag number increases.

The signal composition of the analysis categories in terms of the stage 1 bins being targeted is shown in Figures 8.3 and 8.4. The contribution of each bin to the total number of expected signal events in a category is displayed, meaning the values in each row sum to 100%. In general the migration between categories due to mis-measurement of $p_T^{\gamma\gamma}$ is very low, whilst there are significantly higher migrations arising from jet counting.

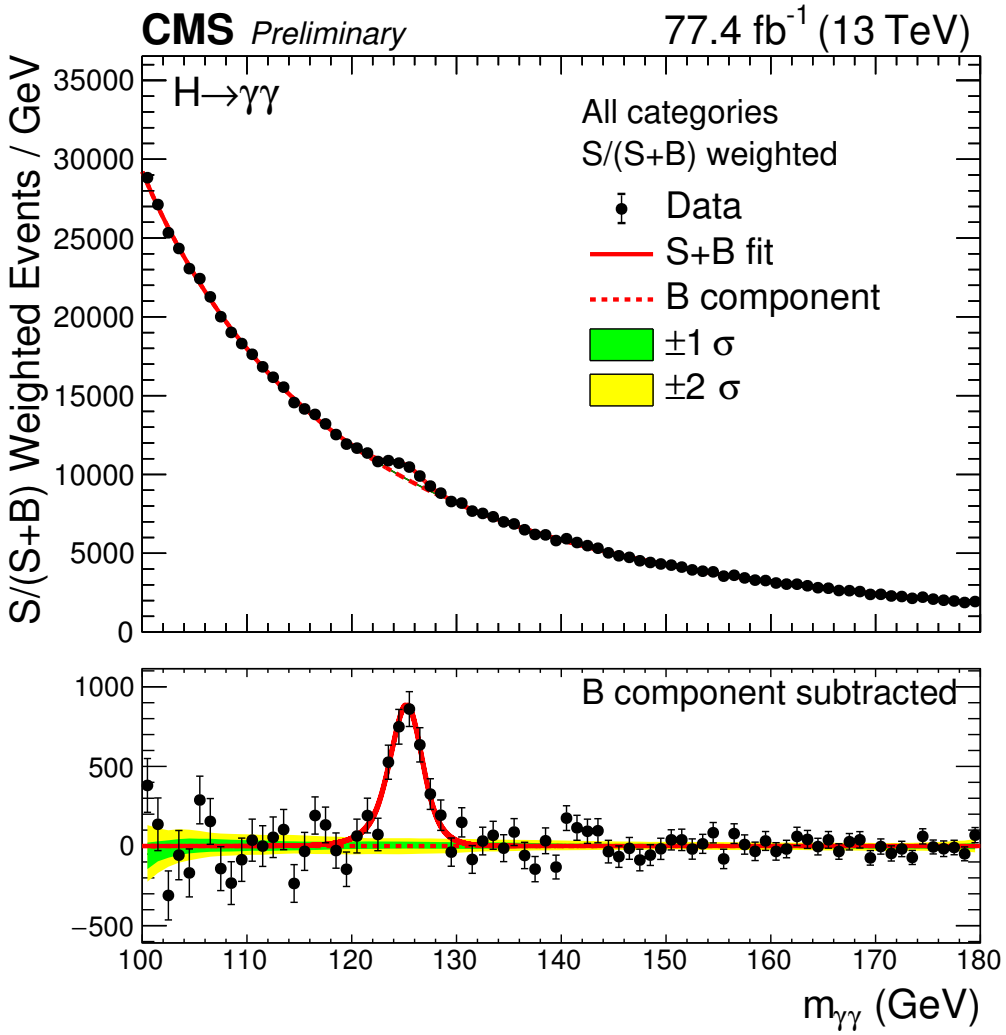


Figure 8.1: Data points (black) and signal plus background model fit for the sum of all analysis categories is shown. Each category is weighted by S/(S + B), where S and B are the numbers of expected signal and background events, respectively, in a $\pm 1\sigma_{\text{eff}}$ mass window centred on m_H. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component. Figure first shown in Ref. [4].

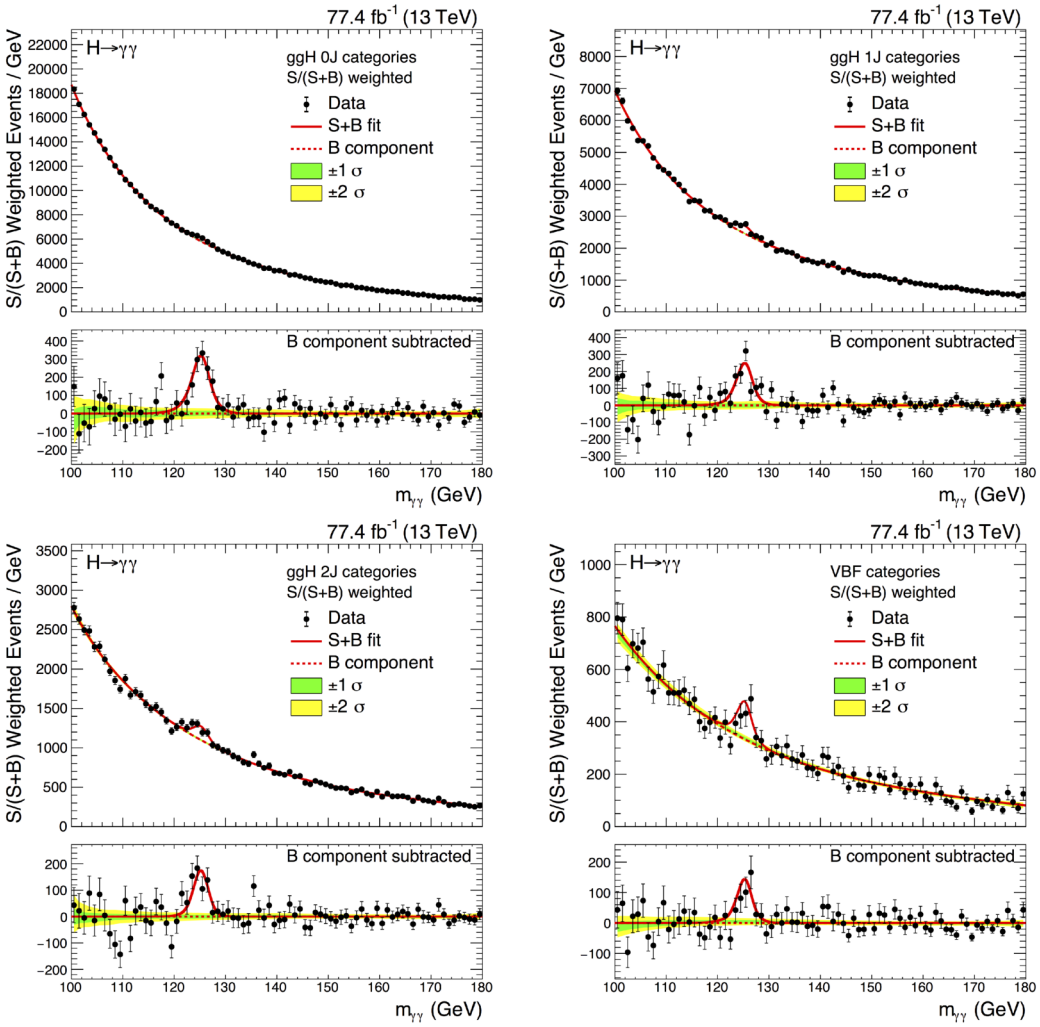


Figure 8.2: Data points (black) and signal plus background model fit for the ggH 0J, ggH 1J, ggH 2J, and VBF categories is shown. Each category is weighted by S/(S + B), where S and B are the numbers of expected signal and background events, respectively, in a $\pm 1\sigma_{\text{eff}}$ mass window centred on m_H. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

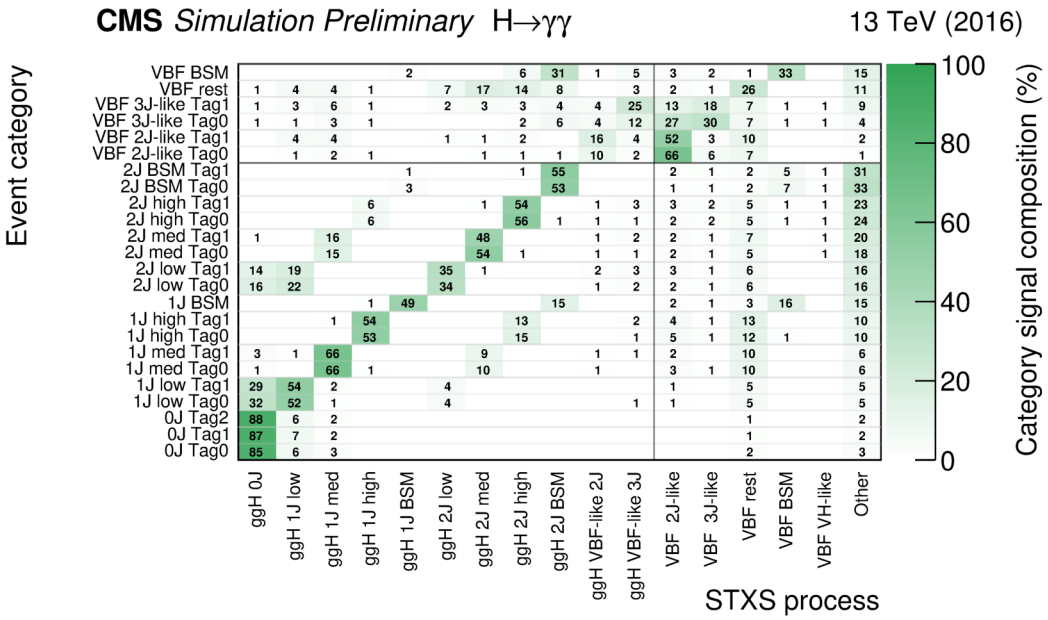


Figure 8.3: The composition of each analysis category in terms of stage 1 bins is shown. The colour scale corresponds to the fraction of each category (rows) accounted for by each stage 1 process (columns). Each row therefore sums to 100%. Entries with values less than 0.5% are not shown. Simulation corresponding to 2016 conditions is shown. Figure first shown in Ref. [4].

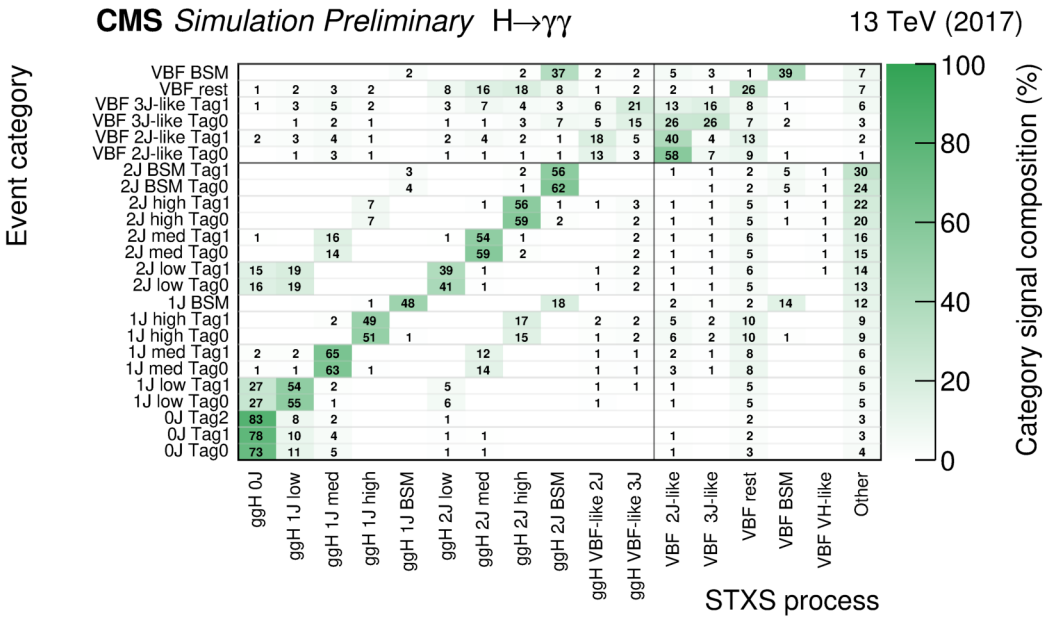


Figure 8.4: The composition of each analysis category in terms of stage 1 bins is shown. The colour scale corresponds to the fraction of each category (rows) accounted for by each stage 1 process (columns). Each row therefore sums to 100%. Entries with values less than 0.5% are not shown. Simulation corresponding to 2017 conditions is shown. Figure first shown in Ref. [4].

Event Categories	SM 125 GeV Higgs boson expected signal														Bkg	S/(S+B)
	Total	ggH	VBF	ttH	tHq	tHW	bbH	ggZH	WH lep	WH had	ZH lep	ZH had	σ_{eff}	σ_{HM}	(GeV^{-1})	
0J Tag 0	257.1	95.0 %	1.9 %	<0.05 %	<0.05 %	<0.05 %	0.9 %	0.1 %	0.9 %	0.3 %	0.7 %	0.2 %	1.66	1.48	522.7	0.09
0J Tag 1	356.4	96.0 %	1.6 %	<0.05 %	<0.05 %	<0.05 %	0.9 %	<0.05 %	0.6 %	0.4 %	0.4 %	0.2 %	2.10	1.74	1182.2	0.05
0J Tag 2	417.2	96.5 %	1.4 %	<0.05 %	<0.05 %	<0.05 %	0.8 %	<0.05 %	0.6 %	0.3 %	0.3 %	0.2 %	2.38	1.97	3229.8	0.02
1J Low Tag 0	115.1	88.9 %	6.5 %	0.1 %	<0.05 %	<0.05 %	1.3 %	<0.05 %	0.6 %	1.4 %	0.2 %	0.8 %	1.61	1.37	269.6	0.08
1J Low Tag 1	145.5	89.2 %	6.2 %	0.1 %	<0.05 %	<0.05 %	1.2 %	<0.05 %	0.6 %	1.6 %	0.3 %	0.9 %	2.13	1.82	722.3	0.03
1J Medium Tag 0	48.7	79.9 %	13.7 %	0.1 %	0.1 %	<0.05 %	0.8 %	0.3 %	1.1 %	2.2 %	0.4 %	1.4 %	1.54	1.40	61.4	0.15
1J Medium Tag 1	109.1	81.1 %	12.6 %	0.1 %	0.1 %	<0.05 %	0.9 %	0.1 %	1.0 %	2.2 %	0.5 %	1.4 %	1.86	1.61	383.9	0.05
1J High Tag 0	17.6	70.3 %	19.8 %	0.2 %	0.1 %	<0.05 %	0.5 %	0.7 %	2.7 %	2.9 %	1.0 %	1.7 %	1.47	1.34	15.6	0.21
1J High Tag 1	21.2	70.8 %	19.5 %	0.3 %	0.1 %	<0.05 %	0.4 %	0.9 %	2.5 %	2.7 %	1.1 %	1.7 %	1.74	1.64	61.3	0.06
1J BSM	8.6	63.9 %	21.6 %	0.3 %	0.2 %	0.1 %	0.3 %	1.6 %	4.9 %	3.4 %	2.0 %	1.8 %	1.40	1.35	6.0	0.26
2J Low Tag 0	28.8	75.7 %	8.2 %	4.2 %	0.5 %	<0.05 %	3.0 %	0.1 %	0.9 %	4.1 %	0.5 %	2.7 %	1.61	1.21	86.2	0.07
2J Low Tag 1	38.5	73.3 %	10.3 %	4.2 %	0.5 %	<0.05 %	2.7 %	0.3 %	0.9 %	4.5 %	0.5 %	2.8 %	1.98	1.70	254.7	0.03
2J Medium Tag 0	24.8	72.1 %	9.6 %	5.0 %	0.6 %	0.1 %	1.7 %	0.6 %	0.8 %	5.7 %	0.5 %	3.3 %	1.50	1.36	38.7	0.13
2J Medium Tag 1	50.5	68.8 %	11.2 %	6.0 %	0.6 %	0.1 %	2.1 %	0.7 %	0.9 %	5.7 %	0.5 %	3.3 %	1.85	1.54	213.2	0.04
2J High Tag 0	22.6	65.3 %	11.2 %	7.3 %	1.0 %	0.3 %	0.9 %	1.4 %	1.4 %	6.9 %	0.5 %	3.8 %	1.52	1.41	21.7	0.19
2J High Tag 1	28.4	65.0 %	11.9 %	7.8 %	0.9 %	0.2 %	1.0 %	1.7 %	1.1 %	6.4 %	0.5 %	3.6 %	1.78	1.72	79.8	0.06
2J BSM Tag 0	14.6	56.3 %	11.1 %	11.5 %	1.9 %	1.2 %	0.3 %	2.6 %	1.5 %	8.1 %	0.7 %	4.9 %	1.40	1.33	6.5	0.35
2J BSM Tag 1	9.7	57.8 %	11.5 %	12.1 %	1.6 %	0.9 %	0.4 %	1.4 %	1.6 %	7.7 %	0.7 %	4.2 %	1.64	1.53	17.3	0.10
VBF 2J-like Tag 0	12.9	19.5 %	79.8 %	0.1 %	0.1 %	<0.05 %	0.3 %	0.1 %	0.1 %	0.1 %	<0.05 %	<0.05 %	1.70	1.41	4.9	0.35
VBF 2J-like Tag 1	6.2	32.4 %	65.5 %	0.3 %	0.2 %	<0.05 %	0.7 %	0.1 %	0.2 %	0.5 %	<0.05 %	0.1 %	1.85	1.52	8.4	0.12
VBF 3J-like Tag 0	12.0	30.6 %	65.6 %	1.3 %	0.7 %	0.1 %	0.8 %	0.4 %	0.1 %	0.3 %	<0.05 %	0.2 %	1.59	1.34	6.0	0.30
VBF 3J-like Tag 1	13.6	52.5 %	38.9 %	3.2 %	1.2 %	0.1 %	1.1 %	0.4 %	0.5 %	1.1 %	0.3 %	0.7 %	1.71	1.51	19.5	0.12
VBF Rest	13.0	59.5 %	29.4 %	3.5 %	1.0 %	0.2 %	1.4 %	0.7 %	1.0 %	2.0 %	0.3 %	1.0 %	1.50	1.29	21.2	0.12
VBF BSM	7.2	45.5 %	39.9 %	6.5 %	1.0 %	0.8 %	0.9 %	1.1 %	1.0 %	2.2 %	0.1 %	1.0 %	1.48	1.29	6.0	0.22
Total	1779.2	87.5 %	6.7 %	0.9 %	0.1 %	<0.05 %	1.0 %	0.2 %	0.8 %	1.4 %	0.4 %	0.8 %	1.96	1.67	7238.9	0.04

Table 8.1: The expected number of signal events per category and the percentage breakdown per production mode in that category. The σ_{eff}, computed as the smallest interval containing 68.3% of the invariant mass distribution, and σ_{HM}, computed as the FWHM divided by 2.35, are also shown as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events per GeV around 125 GeV is listed. The expected ratio of signal to signal plus background events, S/(S + B), is also shown, where S and B are the numbers of expected signal and background events, respectively, in a $\pm 1\sigma_{eff}$ mass window centred on m_H. Data and simulation from 2016 are shown.

Event Categories	SM 125 GeV Higgs boson expected signal														Bkg	S/(S+B)
	Total	ggH	VBF	ttH	tHq	tHW	bbH	ggZH	WH lep	WH had	ZH lep	ZH had	σ_{eff}	σ_{HM}	(GeV^{-1})	
0J Tag 0	401.1	91.8 %	4.4 %	<0.05 %	<0.05 %	<0.05 %	1.4 %	0.1 %	1.0 %	0.4 %	0.6 %	0.2 %	1.94	1.79	870.3	0.07
0J Tag 1	552.3	93.7 %	3.1 %	<0.05 %	<0.05 %	<0.05 %	1.3 %	<0.05 %	0.7 %	0.4 %	0.4 %	0.2 %	2.42	2.06	2121.9	0.04
0J Tag 2	347.3	95.0 %	2.2 %	<0.05 %	<0.05 %	<0.05 %	1.3 %	<0.05 %	0.5 %	0.4 %	0.3 %	0.2 %	2.72	2.41	3035.8	0.01
1J Low Tag 0	130.8	89.5 %	5.9 %	0.1 %	<0.05 %	<0.05 %	1.1 %	<0.05 %	0.5 %	1.7 %	0.2 %	0.9 %	1.91	1.71	360.2	0.06
1J Low Tag 1	111.5	89.2 %	6.1 %	0.1 %	<0.05 %	<0.05 %	1.1 %	<0.05 %	0.5 %	1.8 %	0.2 %	1.0 %	2.47	2.22	689.4	0.02
1J Medium Tag 0	71.4	81.5 %	12.4 %	0.2 %	0.1 %	<0.05 %	0.5 %	0.2 %	0.9 %	2.5 %	0.4 %	1.3 %	1.85	1.67	110.8	0.11
1J Medium Tag 1	91.1	82.7 %	11.4 %	0.2 %	0.1 %	<0.05 %	0.5 %	0.2 %	0.8 %	2.3 %	0.4 %	1.4 %	2.13	1.91	342.2	0.04
1J High Tag 0	14.7	71.7 %	19.4 %	0.3 %	0.2 %	<0.05 %	0.3 %	1.0 %	2.3 %	2.5 %	1.0 %	1.5 %	1.54	1.51	8.7	0.27
1J High Tag 1	28.2	72.4 %	18.4 %	0.4 %	0.2 %	<0.05 %	0.3 %	0.8 %	2.2 %	2.8 %	0.9 %	1.7 %	1.76	1.77	47.7	0.10
1J BSM	15.5	66.9 %	20.9 %	0.4 %	0.3 %	0.1 %	0.1 %	1.0 %	4.0 %	3.0 %	1.6 %	1.8 %	1.76	1.71	17.5	0.15
2J Low Tag 0	10.9	80.2 %	7.0 %	1.7 %	0.4 %	<0.05 %	1.0 %	0.3 %	0.7 %	4.8 %	0.3 %	3.4 %	1.55	1.52	35.1	0.06
2J Low Tag 1	40.8	77.6 %	8.1 %	3.0 %	0.5 %	<0.05 %	0.8 %	0.3 %	0.7 %	5.4 %	0.3 %	3.1 %	2.06	1.94	249.0	0.03
2J Medium Tag 0	16.8	76.6 %	8.1 %	1.9 %	0.5 %	0.1 %	0.3 %	1.0 %	0.7 %	7.0 %	0.4 %	3.4 %	1.60	1.46	28.9	0.11
2J Medium Tag 1	49.7	74.6 %	9.1 %	3.4 %	0.6 %	0.1 %	0.4 %	0.8 %	0.9 %	6.1 %	0.4 %	3.6 %	2.12	1.86	228.8	0.03
2J High Tag 0	14.0	71.1 %	9.2 %	1.7 %	0.6 %	0.1 %	0.2 %	2.7 %	1.0 %	8.2 %	0.7 %	4.6 %	1.54	1.52	14.2	0.18
2J High Tag 1	24.4	69.1 %	9.4 %	3.7 %	0.8 %	0.2 %	0.2 %	2.3 %	1.1 %	8.2 %	0.5 %	4.7 %	1.42	1.31	64.4	0.08
2J BSM Tag 0	15.8	66.4 %	9.4 %	2.6 %	0.9 %	0.4 %	0.1 %	2.7 %	1.9 %	9.3 %	0.9 %	5.4 %	1.67	1.63	11.1	0.22
2J BSM Tag 1	5.7	60.4 %	9.5 %	9.2 %	1.4 %	0.7 %	0.1 %	2.7 %	1.4 %	9.0 %	1.0 %	4.7 %	1.89	1.82	24.3	0.04
VBF 2J-like Tag 0	13.5	24.8 %	74.4 %	0.1 %	0.1 %	<0.05 %	0.1 %	0.1 %	<0.05 %	0.2 %	<0.05 %	0.2 %	1.90	1.73	5.7	0.30
VBF 2J-like Tag 1	4.8	41.7 %	56.5 %	0.2 %	0.2 %	<0.05 %	0.2 %	0.2 %	0.2 %	0.5 %	<0.05 %	0.3 %	2.28	1.94	9.3	0.07
VBF 3J-like Tag 0	12.7	36.8 %	60.6 %	0.4 %	0.5 %	<0.05 %	0.1 %	0.4 %	0.2 %	0.5 %	0.1 %	0.2 %	1.90	1.69	7.8	0.23
VBF 3J-like Tag 1	7.6	56.0 %	37.8 %	0.8 %	0.9 %	<0.05 %	0.2 %	0.8 %	0.5 %	1.6 %	0.2 %	1.0 %	1.86	1.79	11.1	0.11
VBF Rest	12.9	63.4 %	29.9 %	1.0 %	0.6 %	0.1 %	0.4 %	0.8 %	0.6 %	2.0 %	0.3 %	1.1 %	1.80	1.71	21.3	0.10
VBF BSM	6.5	44.7 %	47.8 %	1.0 %	0.5 %	0.3 %	0.1 %	1.4 %	0.7 %	2.1 %	0.4 %	1.0 %	1.75	1.45	4.5	0.22
Total	1999.8	88.2 %	6.7 %	0.4 %	0.1 %	<0.05 %	1.1 %	0.2 %	0.8 %	1.4 %	0.4 %	0.8 %	2.22	1.98	8320.2	0.04

Table 8.2: The expected number of signal events per category and the percentage breakdown per production mode in that category. The σ_{eff}, computed as the smallest interval containing 68.3% of the invariant mass distribution, and σ_{HM}, computed as the FWHM divided by 2.35, are also shown as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events per GeV around 125 GeV is listed. The expected ratio of signal to signal plus background events, S/(S + B), is also shown, where S and B are the numbers of expected signal and background events, respectively, in a $\pm 1\sigma_{eff}$ mass window centred on m_H. Data and simulation from 2017 are shown.

8.4 Results in the STXS framework

Results in the STXS framework are presented with three different parameterisations; for each result the underlying signal bins are grouped into different parameters which are free to vary in the fit. The recommendations contained in Ref. [20] concerning how to treat sub-dominant processes are followed in each case. The ggH parameters include bbH events. The ggZH process is grouped together with leptonic VH production if the Z boson decays leptonically, and with ggH otherwise. The hadronic VH processes are grouped with VBF production to form the qqH parameters. In each fit, the ttH, tH, and VH leptonic parameters are constrained to their SM prediction. This is necessary since there are no categories targeting these production modes, and therefore the parameters would be almost unconstrained and cause increased uncertainties in the other parameters of interest. In all fits the mass of the Higgs boson is profiled.

8.4.1 Stage 0 cross sections

Measurements of stage 0 STXS bins are performed in a fit with two parameters, ggH and qqH. The resulting cross sections, relative to the SM prediction, are found to be $\sigma_{ggH}/\sigma_{ggH}^{\mathrm{SM}} = 1.15 \pm 0.15$ and $\sigma_{qqH}/\sigma_{qqH}^{\mathrm{SM}} = 0.83^{+0.37}_{-0.31}$. The individual likelihood scans are shown in Figures 8.5 and 8.6. Two scans are shown, one corresponding to the full fit and one corresponding to the fit without systematic uncertainties. The systematic component of the uncertainty is then determined by subtracting the statistical component from the total uncertainty. In both measurements the statistical component of the uncertainty is greater than the systematic component. However for the ggH cross section, the magnitude of each is comparable. With the full Run 2 dataset, which will increase the available integrated luminosity to around $137\,\mathrm{fb}^{-1}$, the ggH measurement is likely to become systematics-dominated.

8.4.2 Stage 1 cross sections

Two different measurements are performed at stage 1 of the STXS framework. In both cases, some stage 1 bins are merged in order to improve the statistical sensitivity of the measurement. In the first fit, the definition of parameters is motivated by merging as few bins as possible whilst maintaining the uncertainty on each parameter at less than 100% of the SM predicted value. This results in a total of seven signal parameters. There are six ggH parameters, of which four correspond to a single stage 1 bin; these are the zero-jet (0J), one-jet low (1J low), medium (1J med), and high (1J high) p_T^H bins. The two-jet or greater parameter (GE2J) groups together five individual stage 1 bins, comprising the low, medium and high p_T^H bins as well as the two VBF-like

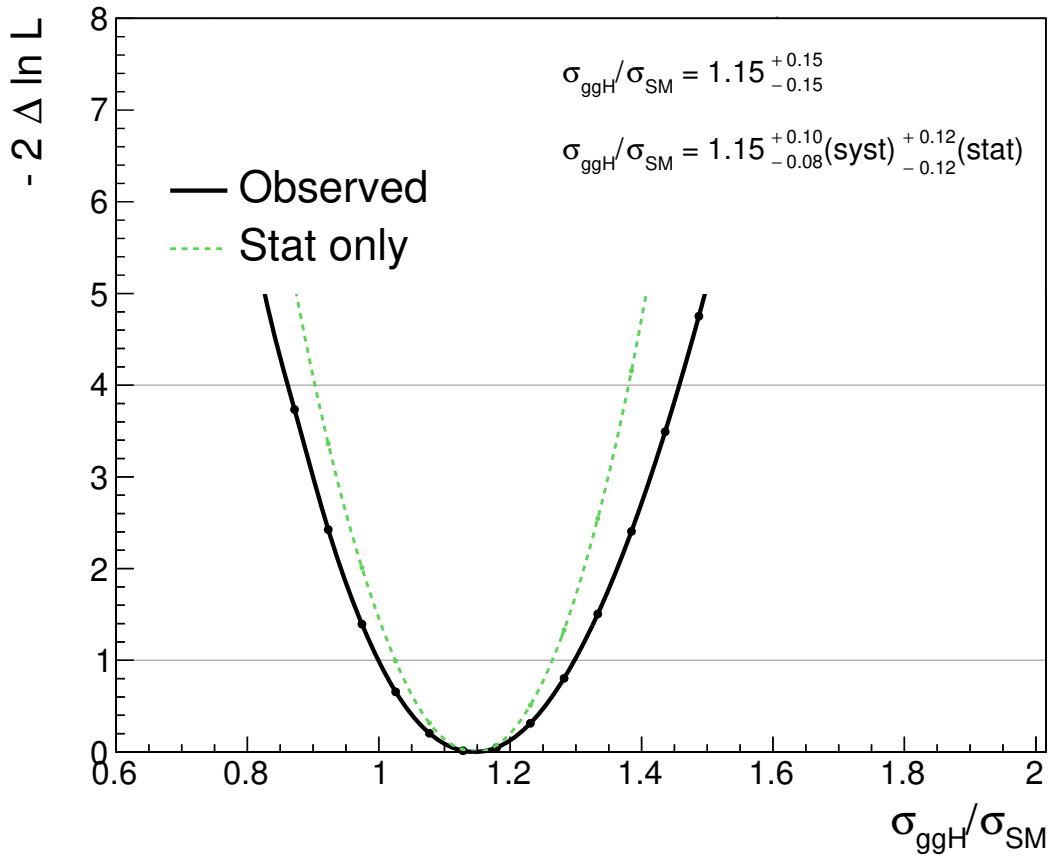


Figure 8.5: The results of a two-parameter fit in the STXS framework, showing the scan of the profiled likelihood ratio in the ggH cross section. All ggH bins are grouped together in the fit to form one parameter, with all VBF bins comprising the second parameter. The ggH parameter includes bbH components, while the qqH parameter includes the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. The solid black line shows the full scan, whilst the dashed green line shows the scan without any systematic uncertainties included.

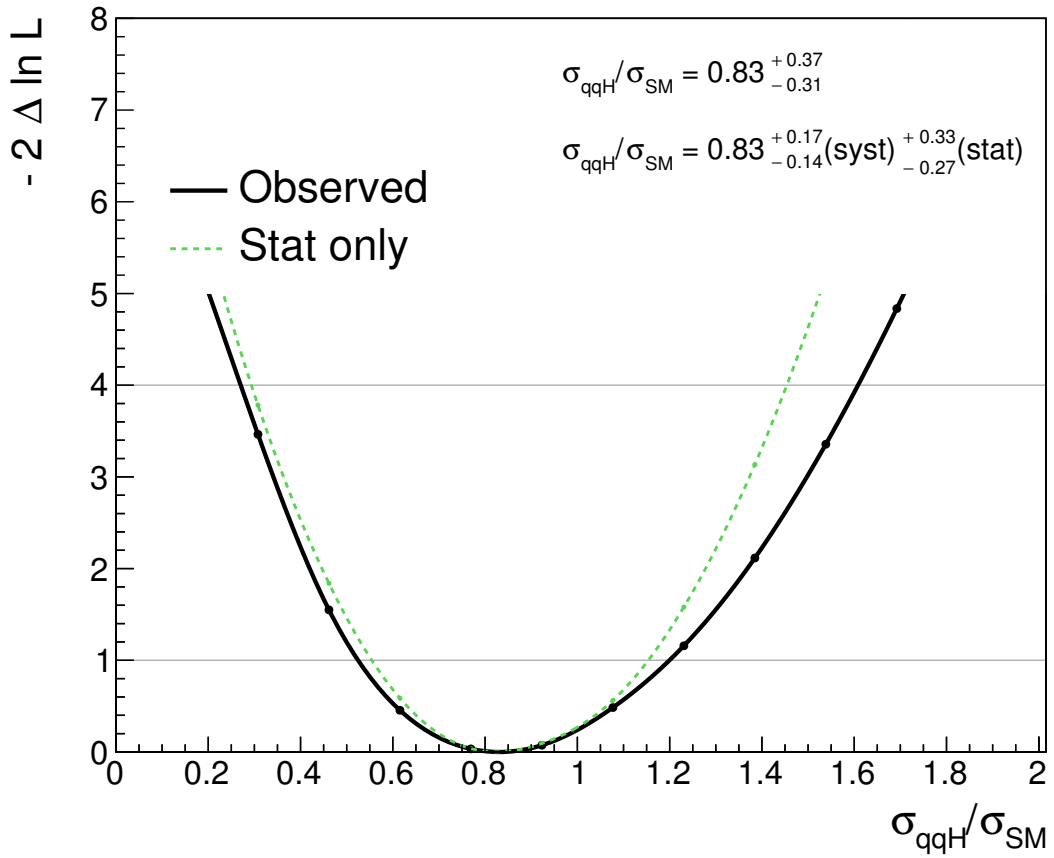


Figure 8.6: The results of a two-parameter fit in the STXS framework, showing the scan of the profiled likelihood ratio in the qqH cross section. All ggH bins are grouped together in the fit to form one parameter, with all VBF bins comprising the second parameter. The ggH parameter includes bbH components, while the qqH parameter includes the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. The solid black line shows the full scan, whilst the dashed green line shows the scan without any systematic uncertainties included.

bins. The ggH BSM parameter is the sum of the one-jet and two-jet BSM bins where $p_T^H > 200\,\text{GeV}$. Finally, the qqH parameter is unchanged from stage 0; all five bins are grouped together. The results of this seven-parameter fit are shown in Figure 8.7. The observed 68% CL intervals for each parameter are compared to the SM predictions and their associated uncertainties. There is very good agreement with the SM; the p-value with respect to the SM hypothesis is approximately 64%. Furthermore, for several parameters, the experimental precision is less than a factor of three greater than the theoretical uncertainty on the SM prediction.

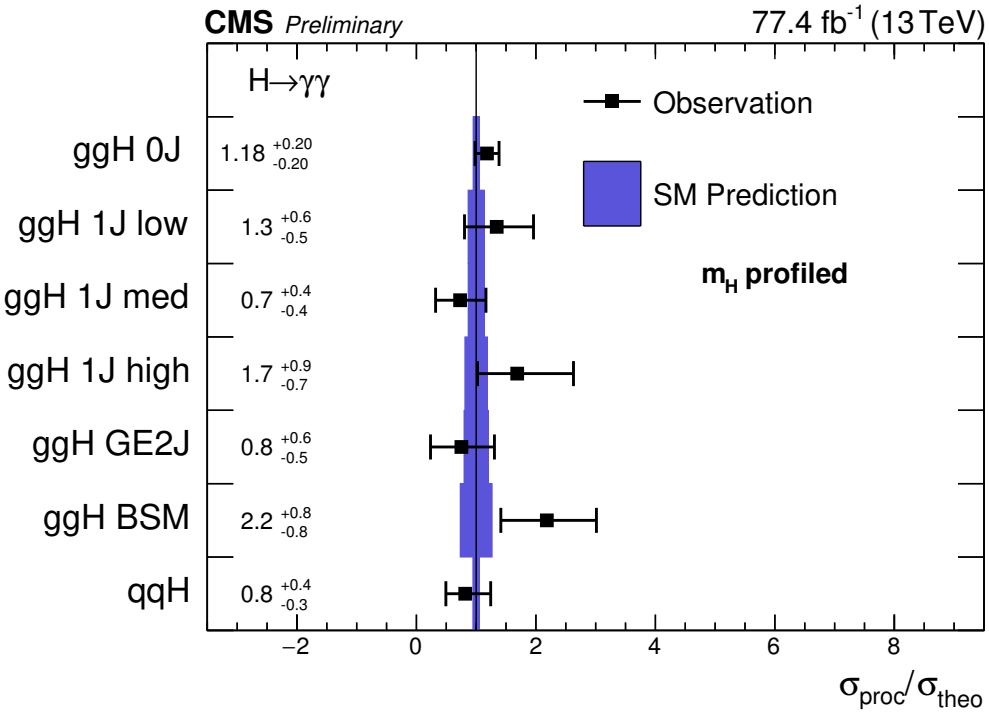


Figure 8.7: The results of a seven-parameter fit in the STXS framework. The ggH 1J and 2J BSM bins are grouped together in the fit; the remaining five ggH bins with two or more jets are also grouped. All five VBF bins are grouped together. The ggH parameters include bbH components, while the qqH parameter includes the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. Cross section ratios are shown with approximate 68% CL intervals (black points), and compared to the SM expectations and their uncertainties (blue bands). The compatibility of this fit with the SM prediction, expressed as a p-value with respect to the SM, is approximately 64%. Figure first shown in Ref. [4].

In the second measurement at stage 1, the fit contains thirteen signal parameters. This choice represents the minimal possible merging of bins whilst retaining a reasonable sensitivity of less than around 200% of the SM prediction. Nine of the parameters

correspond to individual ggH stage 1 bins; the only merged ggH parameter is the so-called ggH VBF-like parameter, where the 2J-like and 3J-like ggH VBF-like bins are grouped together. For qqH, the 2J-like and 3J-like parameters represent individual bins, while the "qqH other" parameter is composed of the VH-like, Rest and BSM bins. The resulting cross sections relative to the corresponding SM predictions are shown in Figure 8.8. In the fit all parameters are constrained to be non-negative – this is necessary to ensure the fit converges. The parameters whose best-fit values are constrained to be zero are known to have 68% CL intervals which slightly under-cover. This is checked and confirmed using an ensemble of pseudo-experiments. Therefore the compatibility of those parameters with the SM prediction is higher than the quoted 68% CL intervals would imply. The compatibility of the thirteen-parameter fit with the SM prediction, expressed as a p-value with respect to the SM hypothesis, is approximately 18%.

In addition to the best-fit values and 68% CL intervals of each fit, the correlation between parameters is reported. The correlation matrices are essential for reinterpretation of the measurements. The observed correlations between the signal parameters are therefore shown for the seven-parameter and thirteen-parameter scenarios in Figures 8.9 and 8.10 respectively. In the seven-parameter fit, the magnitudes of the correlations are generally small. The contamination from ggH 0J events in the 1J low categories results in a slightly higher anti-correlation, whilst the difficulty in distinguishing ggH 2J production from VBF is illustrated in the high anti-correlation between the ggH 2J and qqH parameters. The thirteen-parameter fit displays higher correlation values, partly due to the low sensitivity to the qqH other parameter.

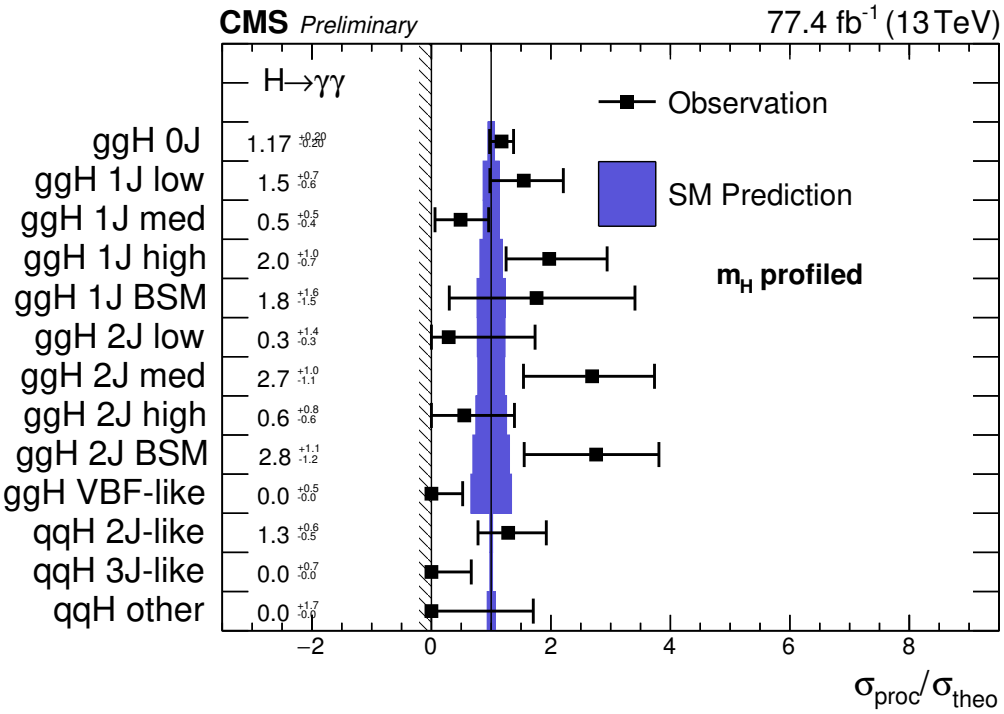


Figure 8.8: The results of a thirteen-parameter fit in the STXS framework. The two VBF-like ggH bins are grouped to form one parameter, as are the VBF BSM-like, VH-like and Rest bins. No further merging is performed. The ggH parameters include bbH components, while the qqH parameters include the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. Cross section ratios are shown with approximate 68% CL intervals (black points) and compared to the SM expectations and their uncertainties (blue bands). The cross section ratios are constrained to be non-negative, as indicated by the vertical line and hashed pattern. The parameters whose best-fit values are at zero are known to have 68% CL intervals which slightly under-cover; this is checked using pseudo-experiments. The compatibility of this fit with the SM prediction, expressed as a p-value with respect to the SM, is approximately 18%. Figure first shown in Ref. [4].

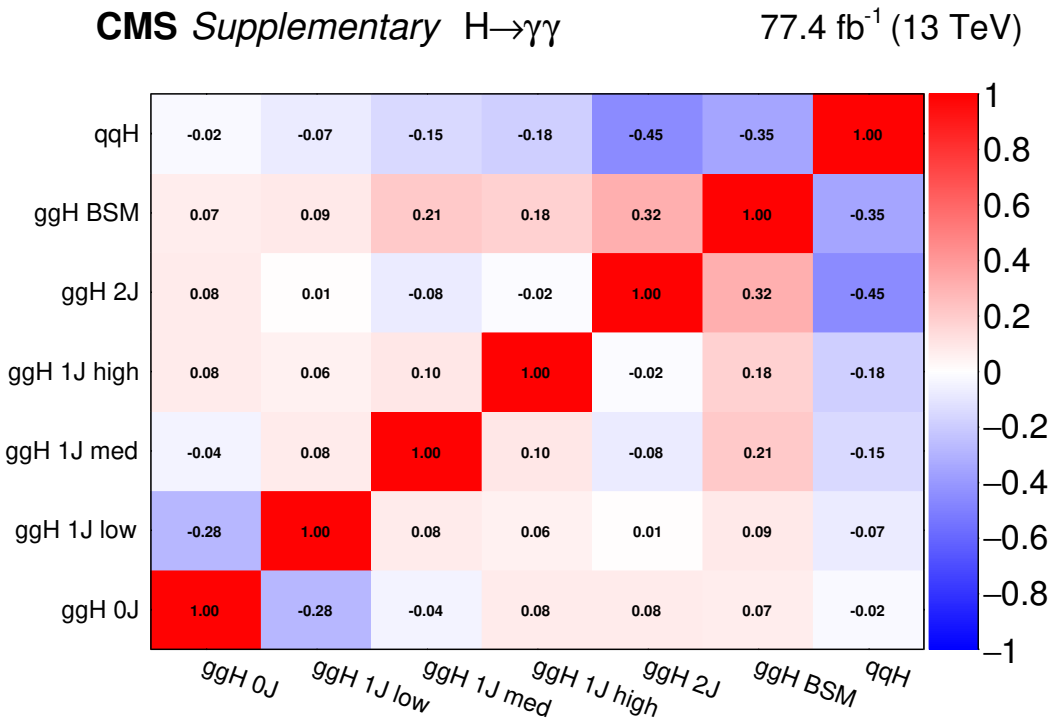


Figure 8.9: Observed correlations in a seven-parameter fit in the STXS framework. The ggH 1J and 2J BSM bins are grouped together in the fit; the remaining five ggH bins with two or more jets are also grouped. All five VBF bins are grouped together. The ggH parameters include bbH components, while the qqH parameter includes the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. The size of the correlation is indicated by the colour scale. Figure first shown in Ref. [4].

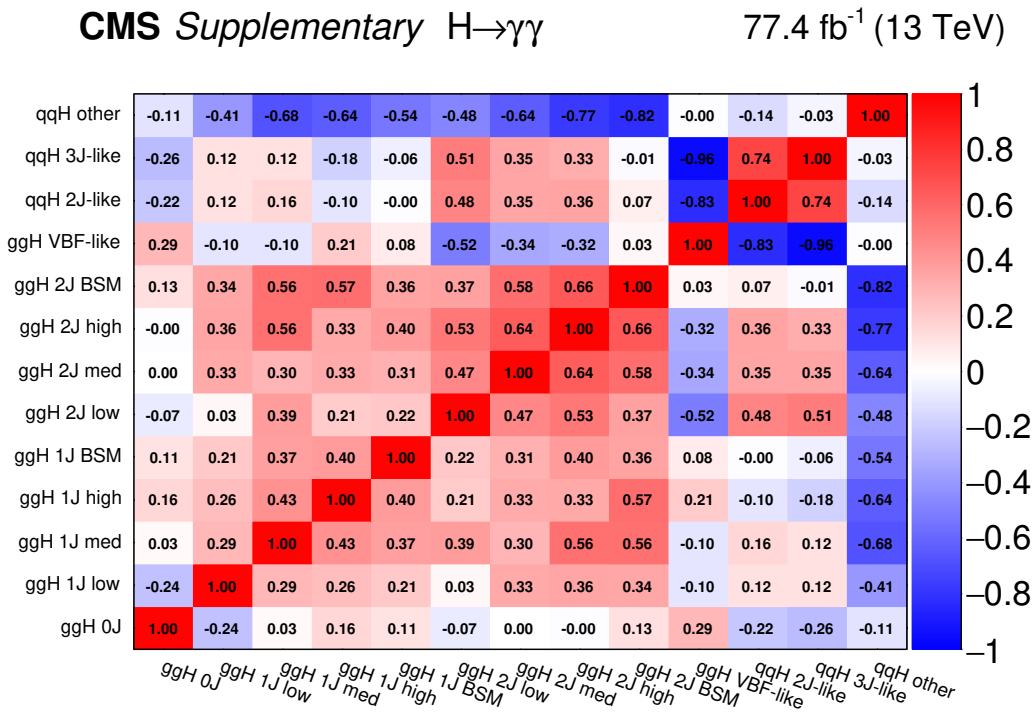


Figure 8.10: Observed correlations in a thirteen-parameter fit in the STXS framework. The two VBF-like ggH bins are grouped to form one parameter, as are the VBF BSM-like, VH-like and Rest bins. No further merging is performed. The ggH parameters include bbH components, while the qqH parameters include the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. The size of the correlation is indicated by the colour scale. Figure first shown in Ref. [4].

8.5 Summary

Measurements of cross sections at various levels of granularity within the STXS framework have been presented. At stage 0, two parameters corresponding to ggH production and electroweak qqH production are measured, with the observed values $\sigma_{ggH}/\sigma_{ggH}^{\mathrm{SM}} = 1.15 \pm 0.15$ and $\sigma_{qqH}/\sigma_{qqH}^{\mathrm{SM}} = 0.83^{+0.37}_{-0.31}$ highly consistent with the SM prediction. At stage 1, two different fits with seven and thirteen signal parameters respectively are performed. Both are consistent with the SM; the compatibility in terms of p-values with respect to the SM hypothesis are 64% and 18% respectively.

The results at stage 1 of the STXS framework are summarised in Tables 8.3 and 8.4, where the measured cross section of each parameter is shown together with the SM prediction. A breakdown of the uncertainties is also provided, with the total uncertainty split into its statistical, experimental, and theoretical components.

In this analysis, based upon the datasets collected by CMS in 2016 and 2017, no significant deviations from the SM predictions are observed. The size of the uncertainties on stage 1 cross sections is still relatively large, and there is the scope for these to be substantially reduced in the near future. This will be achieved by analysing the data collected during 2018, and then combining the results of multiple decay channels. It is likely that the precision of some combined stage 1 cross section measurements will be comparable to the theoretical uncertainty on the SM prediction, which necessitates an effort to improve the theoretical predictions. Furthermore, the statistical component of the uncertainties will decrease; in some cases, this will lead to systematic uncertainties limiting the measurement sensitivity. Therefore an important aspect of future analyses will be in understanding and minimising the impact of these systematic uncertainties. This will become crucial in the longer term, with Run 3 of the LHC due to commence in 2021, and the HL-LHC due to eventually provide an integrated luminosity of $3000\,\mathrm{fb}^{-1}$ by 2040. With these datasets, stringent tests of the SM at the level of a few per-cent are expected to be feasible [64].

Signal parameter	Cross section (fb)		$\sigma/\sigma_{\mathrm{SM}}$	Uncertainty on $\sigma/\sigma_{\mathrm{SM}}$			
	SM pred.	Measured		Total	Stat.	Exp.	Theo.
ggH 0J	61 ± 3	72 ± 12	1.18	$^{+0.20}_{-0.20}$	$^{+0.18}_{-0.18}$	$^{+0.10}_{-0.08}$	$^{+0.06}_{-0.05}$
ggH 1J low	15 ± 2	21^{+9}_{-8}	1.3	$^{+0.6}_{-0.5}$	$^{+0.6}_{-0.5}$	$^{+0.2}_{-0.2}$	$^{+0.2}_{-0.1}$
ggH 1J med	10 ± 1	$7.6^{+4.3}_{-4.1}$	0.7	$^{+0.4}_{-0.4}$	$^{+0.4}_{-0.4}$	$^{+0.1}_{-0.1}$	$^{+0.1}_{-0.0}$
ggH 1J high	1.7 ± 0.3	$2.9^{+1.6}_{-1.1}$	1.7	$^{+0.9}_{-0.7}$	$^{+0.8}_{-0.6}$	$^{+0.3}_{-0.2}$	$^{+0.2}_{-0.1}$
ggH 2J	11 ± 2	$8.4^{+6.1}_{-5.7}$	0.8	$^{+0.6}_{-0.5}$	$^{+0.5}_{-0.5}$	$^{+0.1}_{-0.1}$	$^{+0.3}_{-0.1}$
ggH BSM	1.3 ± 0.4	$2.9^{+1.1}_{-1.0}$	2.2	$^{+0.8}_{-0.8}$	$^{+0.6}_{-0.6}$	$^{+0.4}_{-0.3}$	$^{+0.3}_{-0.2}$
qqH	11 ± 1	$9.1^{+4.7}_{-3.0}$	0.8	$^{+0.4}_{-0.3}$	$^{+0.4}_{-0.3}$	$^{+0.2}_{-0.1}$	$^{+0.1}_{-0.0}$

Table 8.3: The results of a seven-parameter fit in the STXS framework. The ggH 1J and 2J BSM bins are grouped together in the fit; the remaining five ggH bins with two or more jets are also grouped. All five VBF bins are grouped together. The ggH parameters include bbH components, while the qqH parameter includes the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. Both the measured value and the standard model prediction for the product of the cross section and branching ratio are shown. The ratio of the measured cross section to the SM prediction is also shown, together with its uncertainty. In addition, the statistical, experimental, and theoretical components of the uncertainty on each parameter are reported. Table first shown in Ref. [4].

Signal parameter	Cross section (fb)		$\sigma/\sigma_{\text{SM}}$	Uncertainty on $\sigma/\sigma_{\text{SM}}$			
	SM pred.	Measured		Total	Stat.	Exp.	Theo.
ggH 0J	61 ± 3	72 ± 12	1.17	$^{+0.20}_{-0.20}$	$^{+0.18}_{-0.18}$	$^{+0.08}_{-0.07}$	$^{+0.06}_{-0.04}$
ggH 1J low	15 ± 2	24^{+10}_{-9}	1.5	$^{+0.7}_{-0.6}$	$^{+0.6}_{-0.5}$	$^{+0.2}_{-0.1}$	$^{+0.2}_{-0.1}$
ggH 1J med	10 ± 1	$5.1^{+4.7}_{-4.3}$	0.5	$^{+0.5}_{-0.4}$	$^{+0.4}_{-0.4}$	$^{+0.1}_{-0.1}$	$^{+0.1}_{-0.0}$
ggH 1J high	1.7 ± 0.3	$3.4^{+1.6}_{-1.2}$	2.0	$^{+1.0}_{-0.7}$	$^{+0.8}_{-0.7}$	$^{+0.3}_{-0.1}$	$^{+0.4}_{-0.2}$
ggH 1J BSM	0.4 ± 0.1	$0.6^{+0.6}_{-0.5}$	1.8	$^{+1.7}_{-1.5}$	$^{+1.5}_{-1.4}$	$^{+0.3}_{-0.2}$	$^{+0.4}_{-0.1}$
ggH 2J low	2.9 ± 0.7	$0.8^{+4.2}_{-0.8}$	0.3	$^{+1.5}_{-0.3}$	$^{+1.4}_{-0.3}$	$^{+0.3}_{-0.1}$	$^{+0.3}_{-0.0}$
ggH 2J med	4.6 ± 1.0	12 ± 5	2.6	$^{+1.1}_{-1.1}$	$^{+1.0}_{-1.0}$	$^{+0.3}_{-0.2}$	$^{+0.4}_{-0.3}$
ggH 2J high	2.3 ± 0.6	$1.3^{+1.9}_{-1.3}$	0.6	$^{+0.8}_{-0.6}$	$^{+0.7}_{-0.6}$	$^{+0.2}_{-0.1}$	$^{+0.3}_{-0.0}$
ggH 2J BSM	1.0 ± 0.3	$2.7^{+1.1}_{-1.2}$	2.8	$^{+1.1}_{-1.2}$	$^{+0.8}_{-1.0}$	$^{+0.3}_{-0.3}$	$^{+0.5}_{-0.4}$
ggH VBF-like	1.5 ± 0.5	$0.0^{+0.8}_{-0}$	0.0	$^{+0.5}_{-0.0}$	$^{+0.5}_{-0.0}$	$^{+0.2}_{-0.0}$	$^{+0.1}_{-0.0}$
qqH 2J-like	2.1 ± 0.1	$2.6^{+1.3}_{-0.8}$	1.3	$^{+0.6}_{-0.5}$	$^{+0.4}_{-0.4}$	$^{+0.4}_{-0.3}$	$^{+0.1}_{-0.1}$
qqH 3J-like	0.8 ± 0.03	$0.0^{+0.5}_{-0}$	0.0	$^{+0.7}_{-0.0}$	$^{+0.6}_{-0.0}$	$^{+0.2}_{-0.0}$	$^{+0.0}_{-0.0}$
qqH other	8.2 ± 0.6	0^{+14}_{-0}	0.0	$^{+1.7}_{-0.0}$	$^{+1.6}_{-0.0}$	$^{+0.6}_{-0.0}$	$^{+0.2}_{-0.0}$

Table 8.4: The results of a thirteen-parameter fit in the STXS framework. The two VBF-like ggH bins are grouped to form one parameter, as are the VBF BSM-like, VH-like and Rest bins. No further merging is performed. The ggH parameters include bbH components, while the qqH parameters include the hadronic VH contribution. The ttH, tH and VH leptonic processes are constrained to the SM prediction. Both the measured value and the standard model prediction for the product of the cross section and branching ratio are shown. The ratio of the measured cross section to the SM prediction is also shown, together with its uncertainty. In addition, the statistical, experimental, and theoretical components of the uncertainty on each parameter are reported. Table first shown in Ref. [4].

Chapter 9

Conclusions

Measurements of Higgs boson production cross sections are performed using $77.4\,\text{fb}^{-1}$ of data at $\sqrt{s} = 13\,\text{TeV}$ collected by the CMS experiment at the LHC. Events with two photons consistent with the decay of the SM Higgs boson are selected and subsequently categorised using kinematic variables, the diphoton BDT, and the dijet BDT. Models of the signal and background contributions to the diphoton invariant mass distribution in each analysis category are constructed as inputs to the final fits from which the results of the analysis are extracted. The results are presented as measurements of cross sections within the STXS framework; the best-fit values and their uncertainties are estimated with an approach based on the profile likelihood ratio test statistic.

The observed measurements are all consistent with the hypothesis of a SM Higgs boson. Cross sections for gluon fusion and vector boson fusion production, relative to the corresponding standard model predictions, are measured to be 1.15 ± 0.15 and $0.83^{+0.37}_{-0.31}$ respectively [4]. Two additional measurements are performed with a mixture of merged and unmerged stage 1 STXS bins as the parameters of interest. In the first case, seven signal parameters are defined, each with a measured uncertainty of less than 100% of the cross section predicted by the SM. The second fit has thirteen signal parameters, all of which are measured to a precision of better than 200% of the SM prediction. The results of both fits are compatible with the SM; the p-values with respect to the SM hypothesis are found to be approximately 64% and 18% respectively.

Despite the excellent agreement with the SM, many of the measured uncertainties are still relatively large. This is particularly true for the stage 1 cross section measurements, for which the uncertainties are currently dominated by the statistical component. Plausible BSM theories often predict that Higgs coupling parameters deviate from the SM at the per-cent level [103]. The STXS framework is well-suited to systematically characterising any possible deviations from the SM, for example within effective field theories [104]. Furthermore, in the near future combinations of analyses

using the full LHC Run 2 dataset will be performed. Inputs will include results from the various decay channels and from different experiments, resulting in significant improvements in the precision of stage 1 cross section measurements. In many cases, it is likely that the magnitude of the uncertainties will be comparable to both the errors on the SM theoretical predictions and the experimental systematic uncertainties on the measurements. Further progress will require advancement in experimental techniques and understanding of collected data, as well as dedicated efforts to improve the accuracy of SM predictions.

In the longer term, Run 3 of the LHC and subsequently the Phase 2 upgrade to the HL-LHC will provide unprecedented amounts of data. There are significant experimental challenges to be met in order for these data to remain of the same quality as those collected during Run 2. One aspect of the CMS Phase 2 upgrade is the HGCAL, which brings exciting possibilities for novel reconstruction techniques that have only begun to be explored. In addition to facilitating improved precision on existing measurements in the Higgs sector [64], the HL-LHC will enable entirely new measurements to be made. For example, the nature of the Higgs potential has not yet been confirmed experimentally. This can be measured directly by searching for events in which two Higgs bosons are produced, but can also be accessed indirectly, for example by precisely measuring differential distributions of single Higgs boson production [105]. A multitude of other further measurements will also be possible, including observing rare Higgs boson decay modes.

In summary, there is huge potential for further progress in Higgs boson measurements at the LHC. Performing optimal measurements will require a sustained and dedicated experimental effort to understand the data and reduce systematic uncertainties, as well as adopting new and innovative analysis techniques. Progress on the theoretical side will also be necessary to keep the uncertainties on the SM predictions below the experimental precision. In doing so, the SM will be tested as thoroughly as possible. Hopefully, insights into how to address its shortcomings will be found, and our fundamental understanding of the universe improved.

Appendix A

Observed diphoton mass distributions

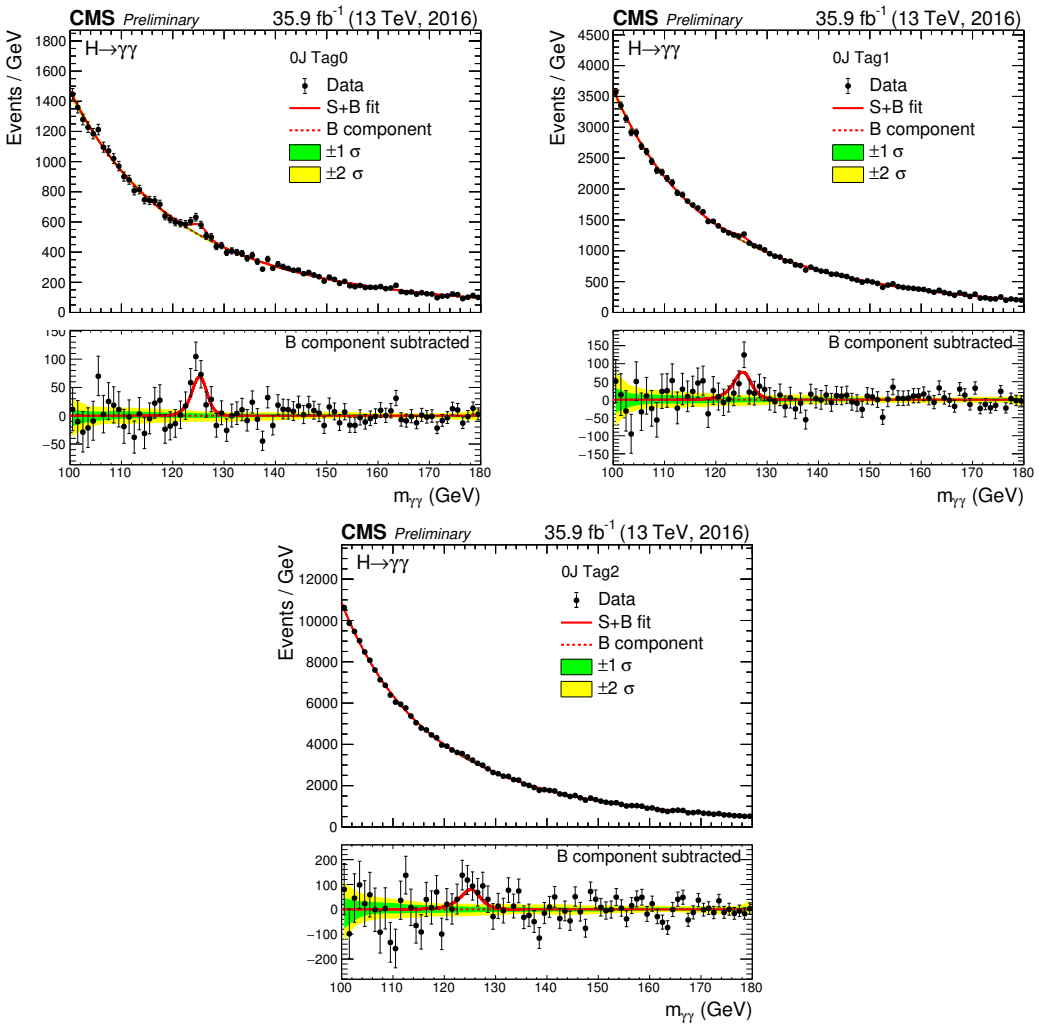


Figure A.1: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

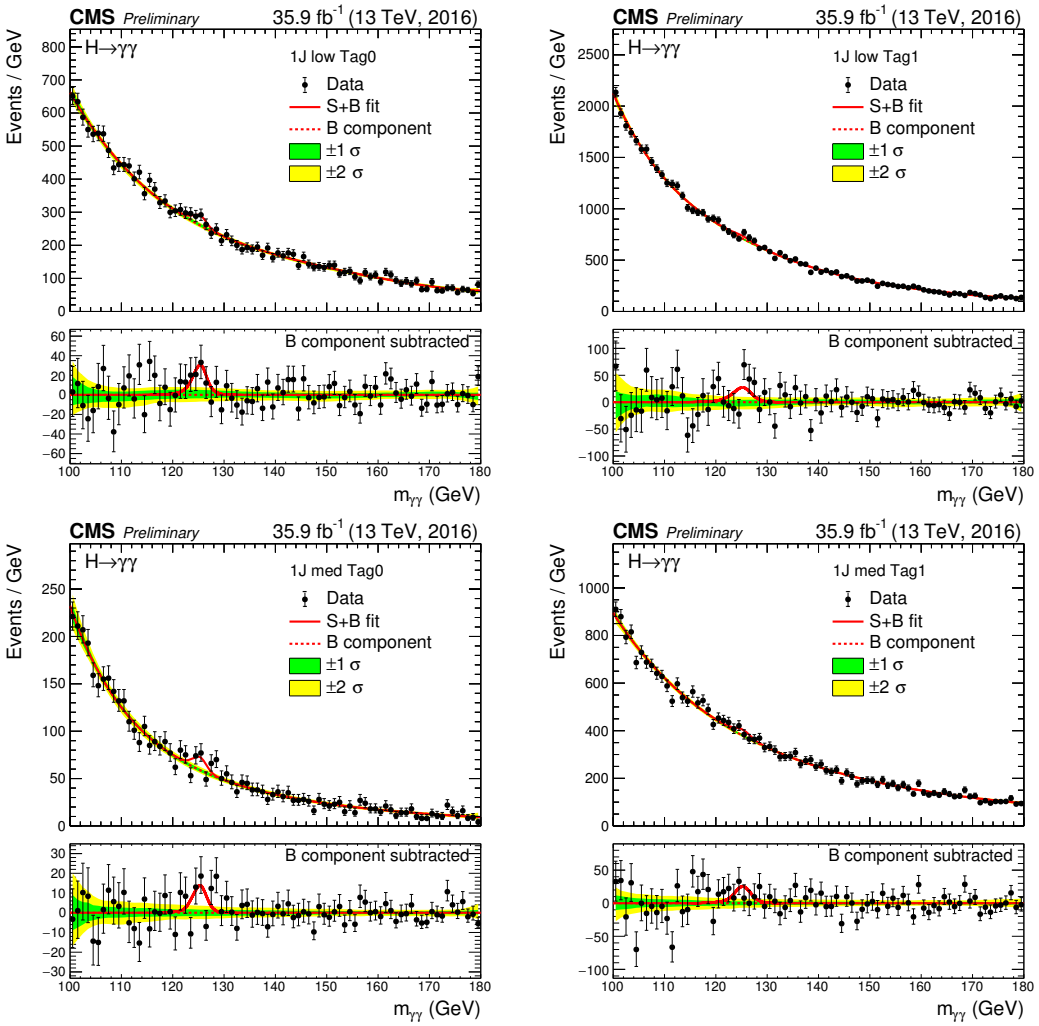


Figure A.2: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

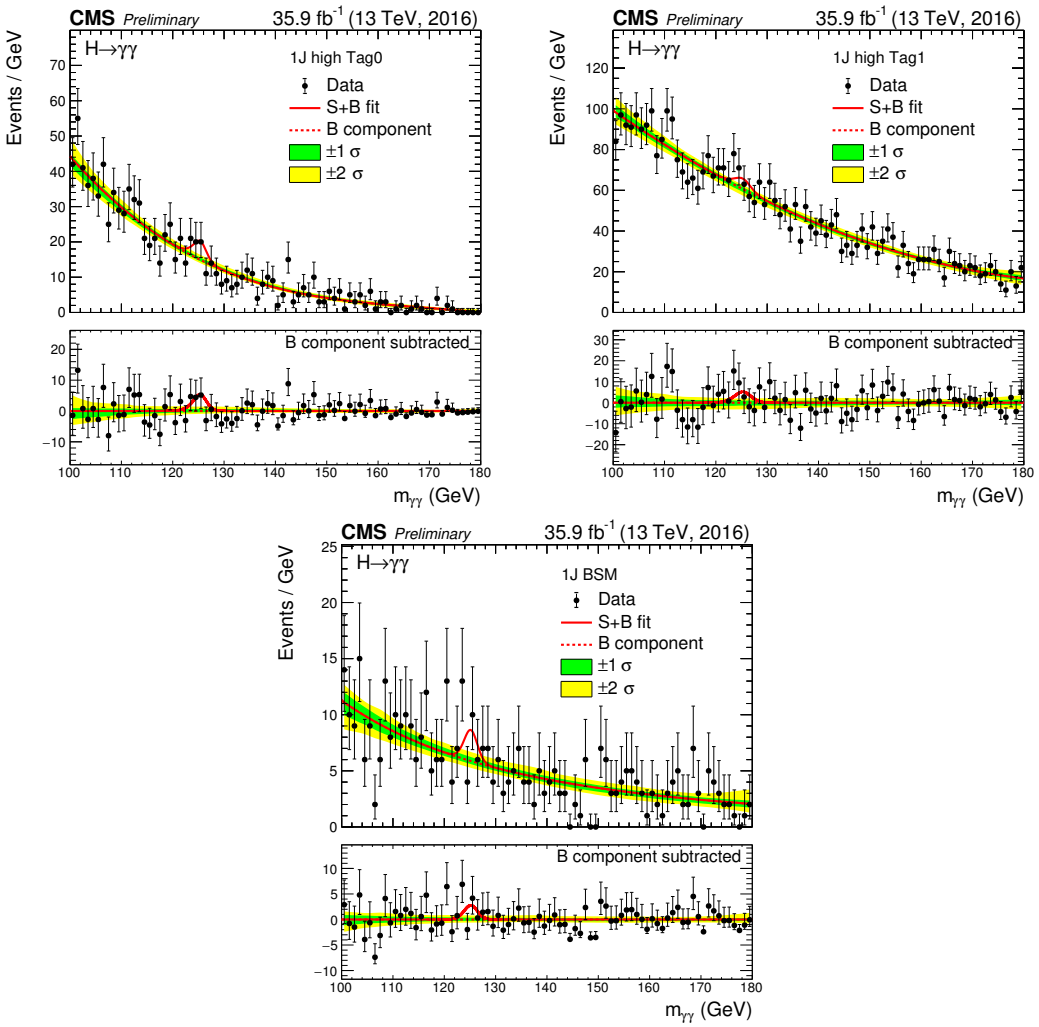


Figure A.3: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

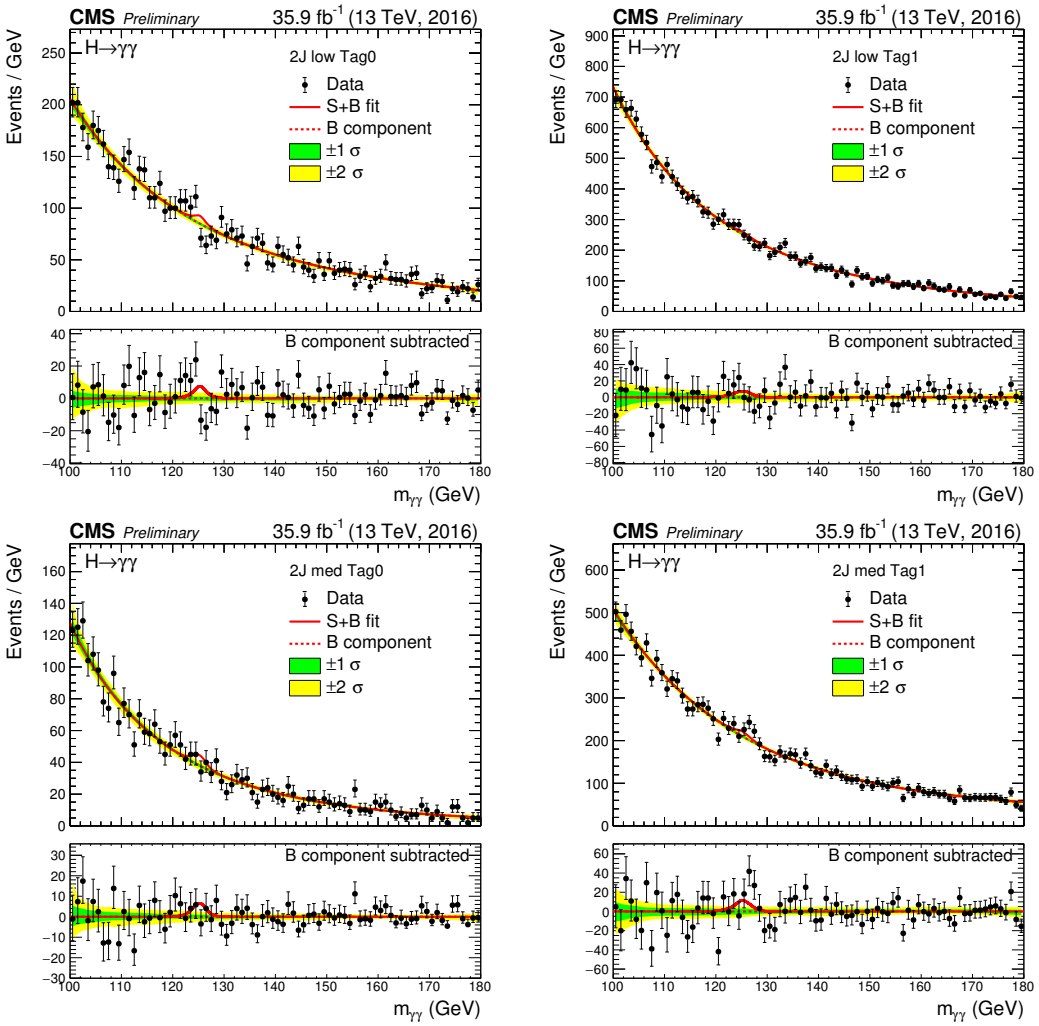


Figure A.4: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

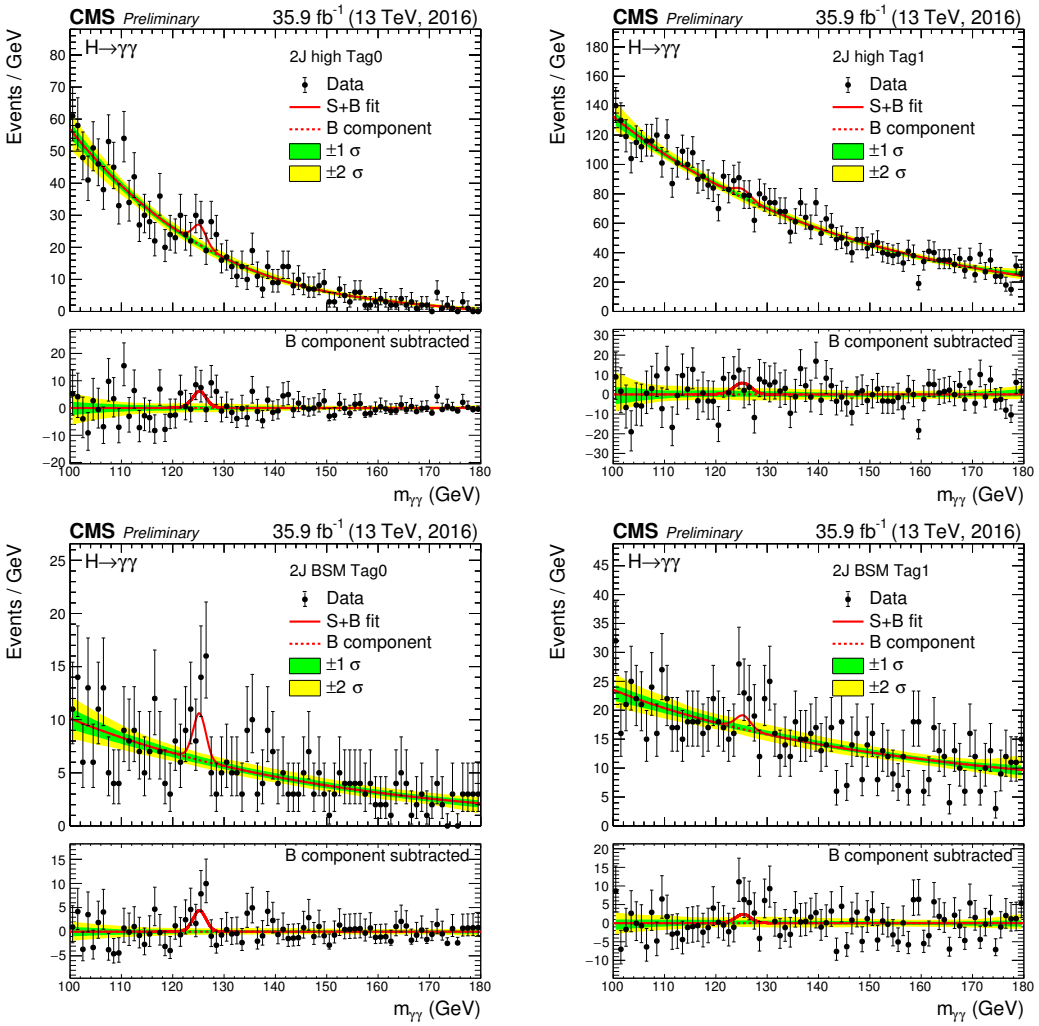


Figure A.5: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

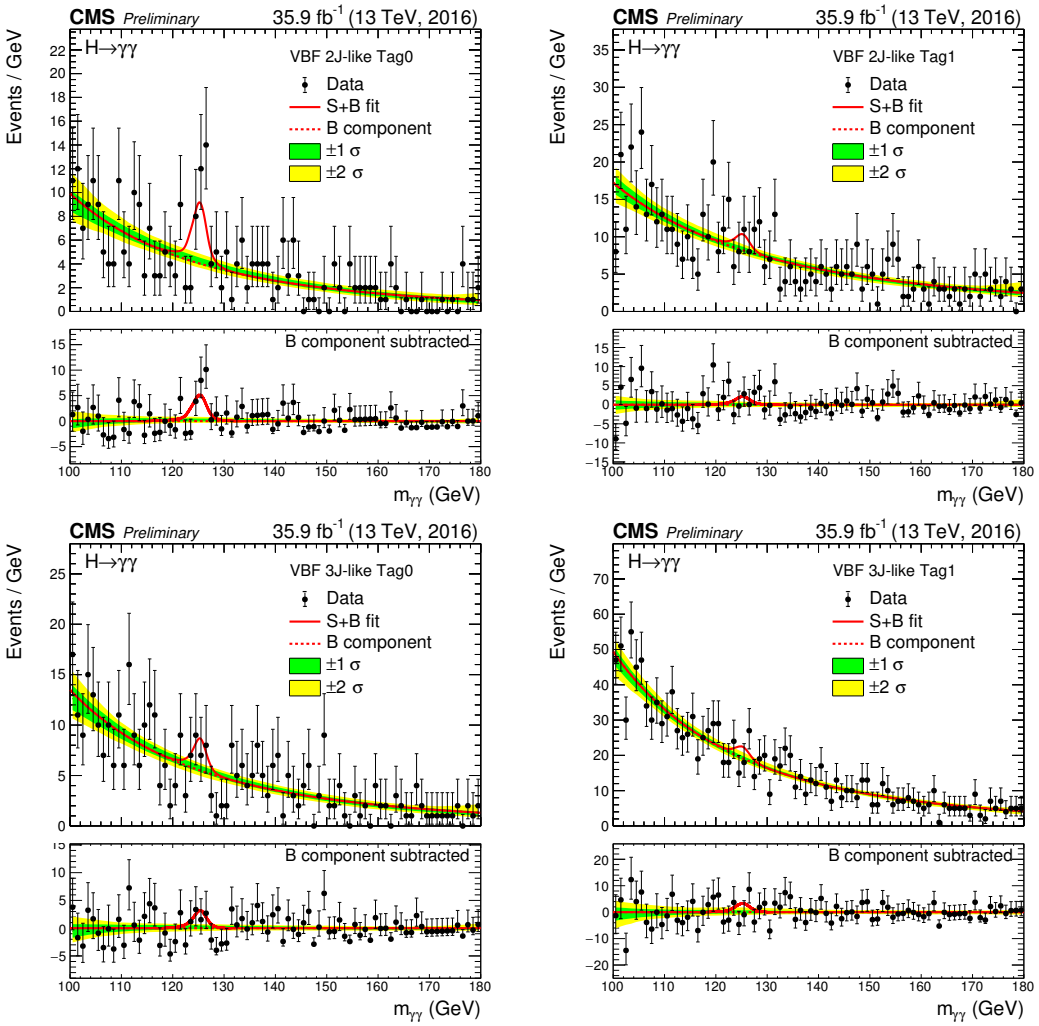


Figure A.6: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

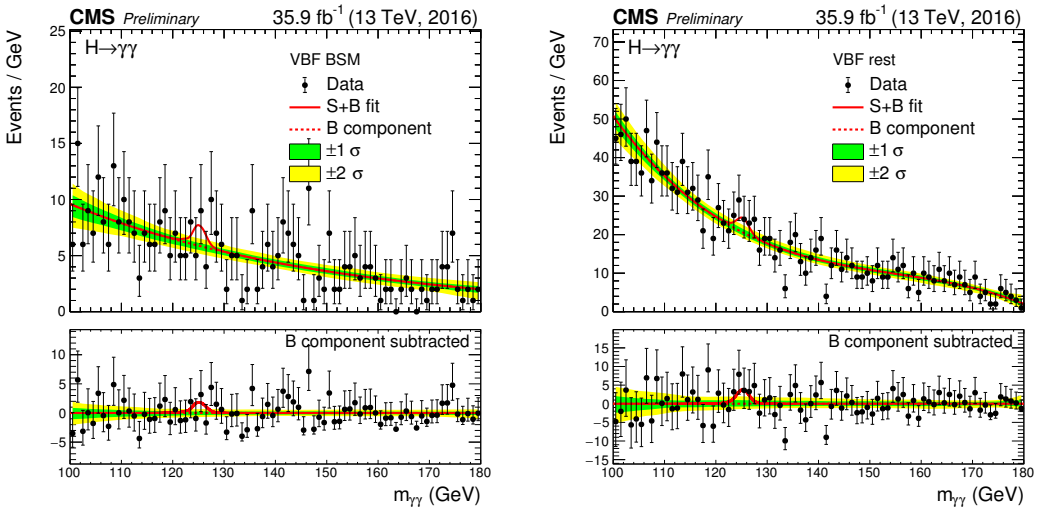


Figure A.7: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

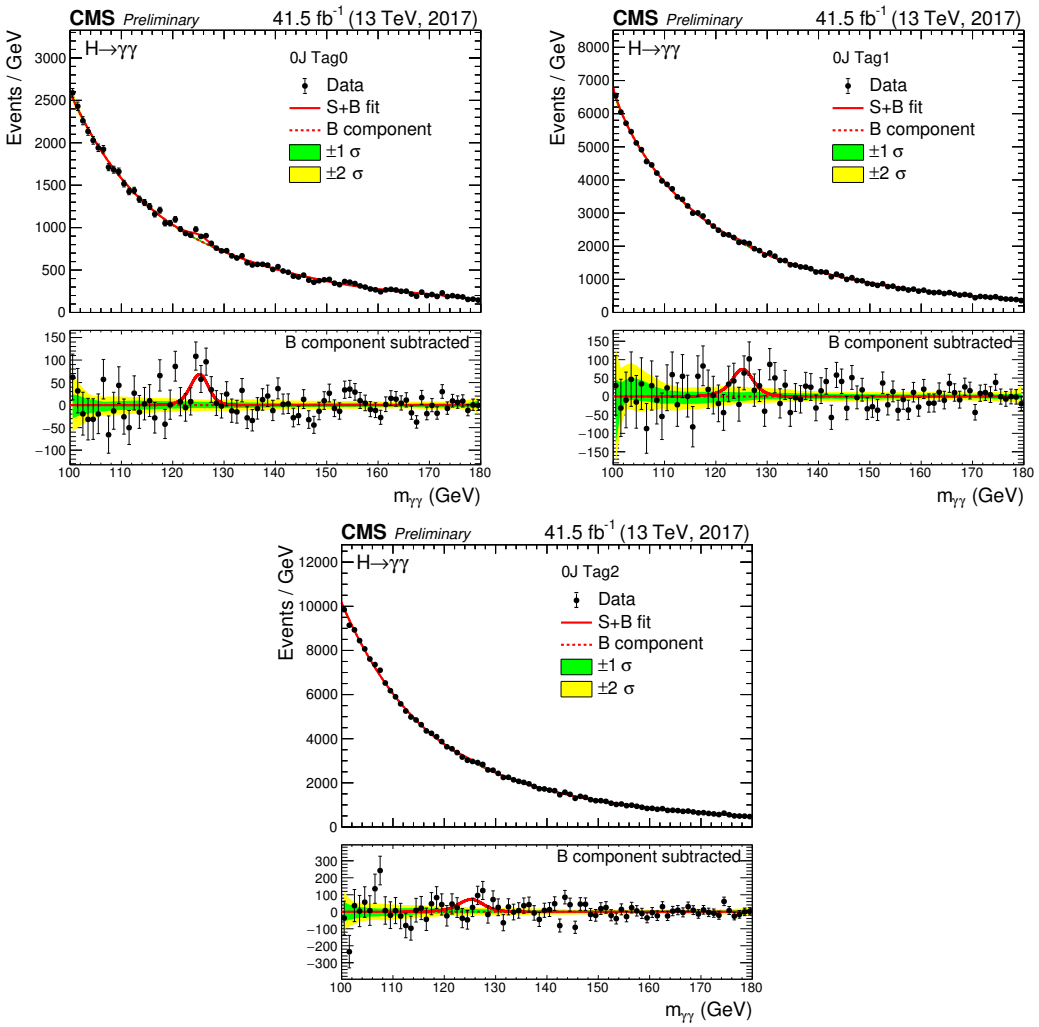


Figure A.8: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

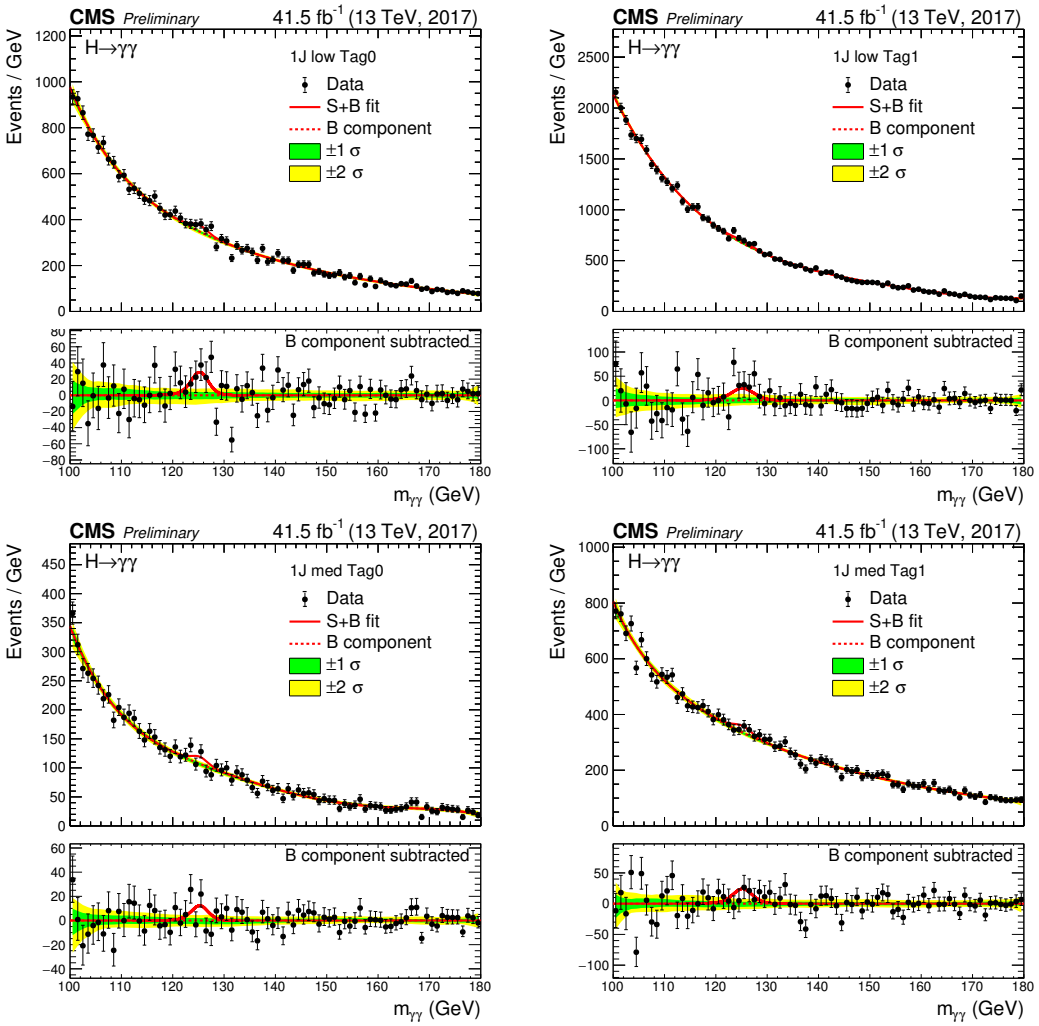


Figure A.9: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

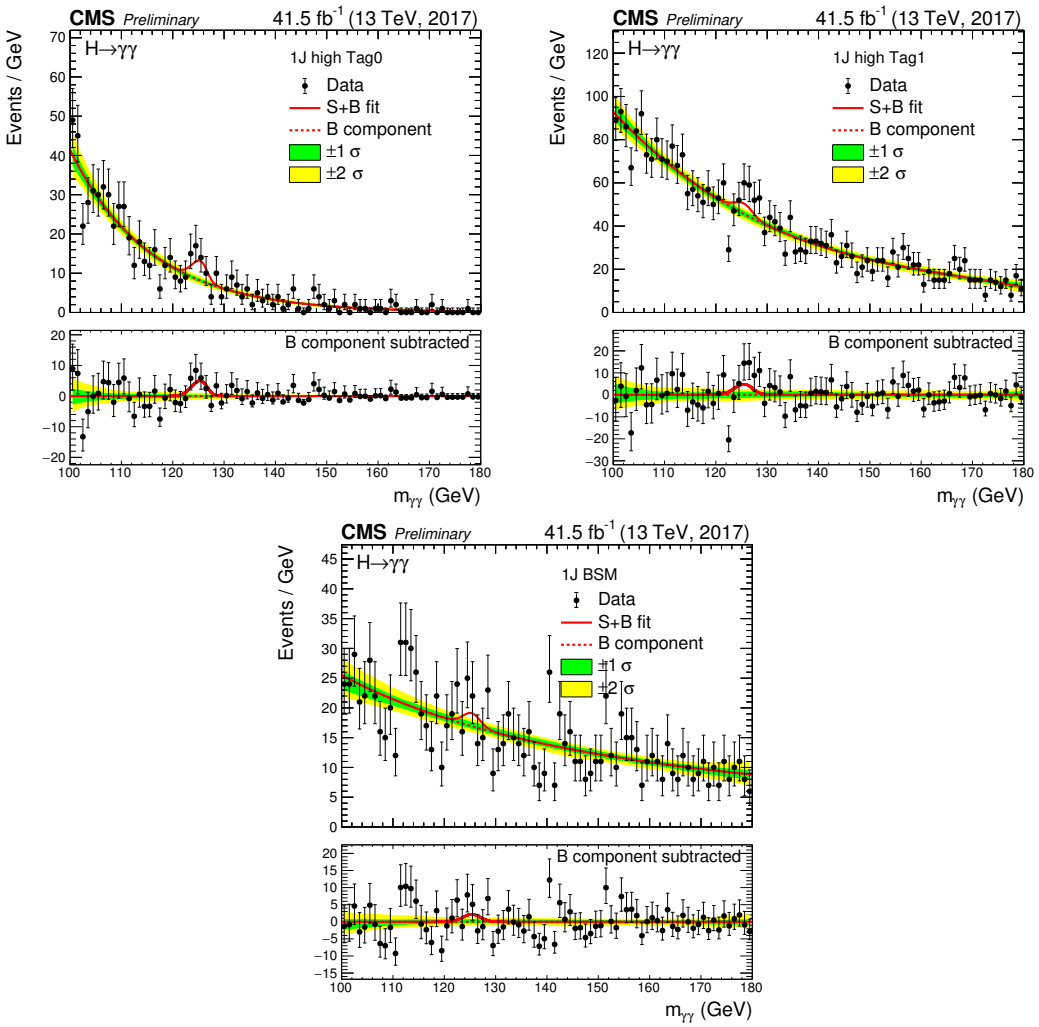


Figure A.10: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

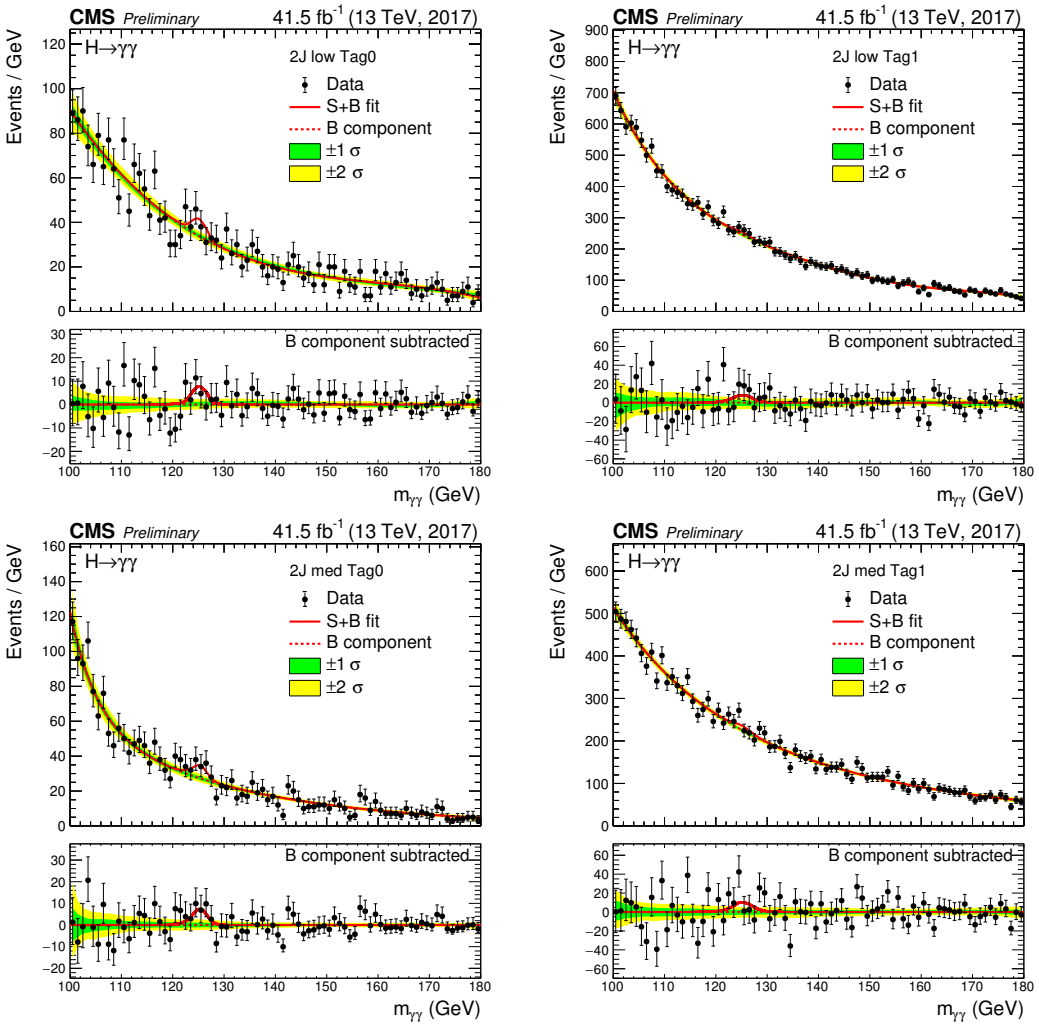


Figure A.11: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

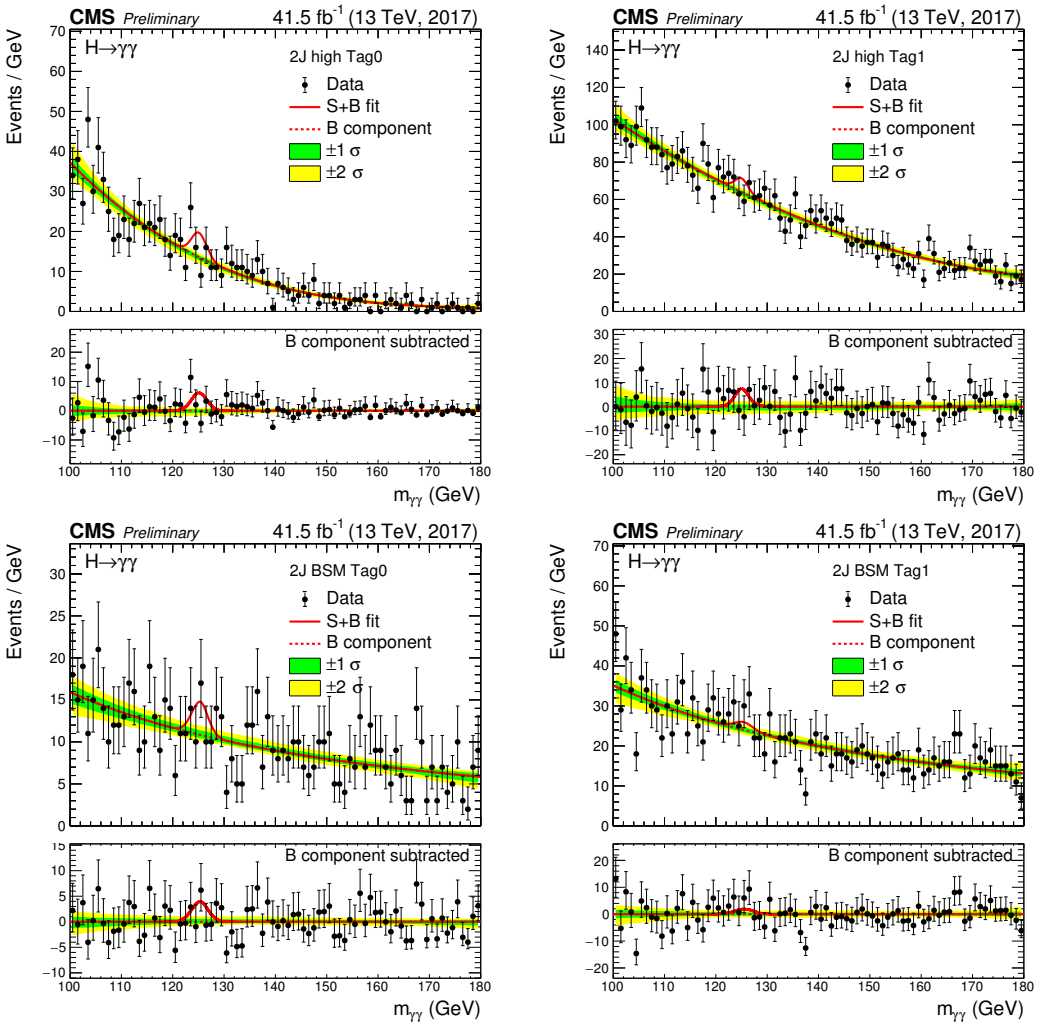


Figure A.12: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

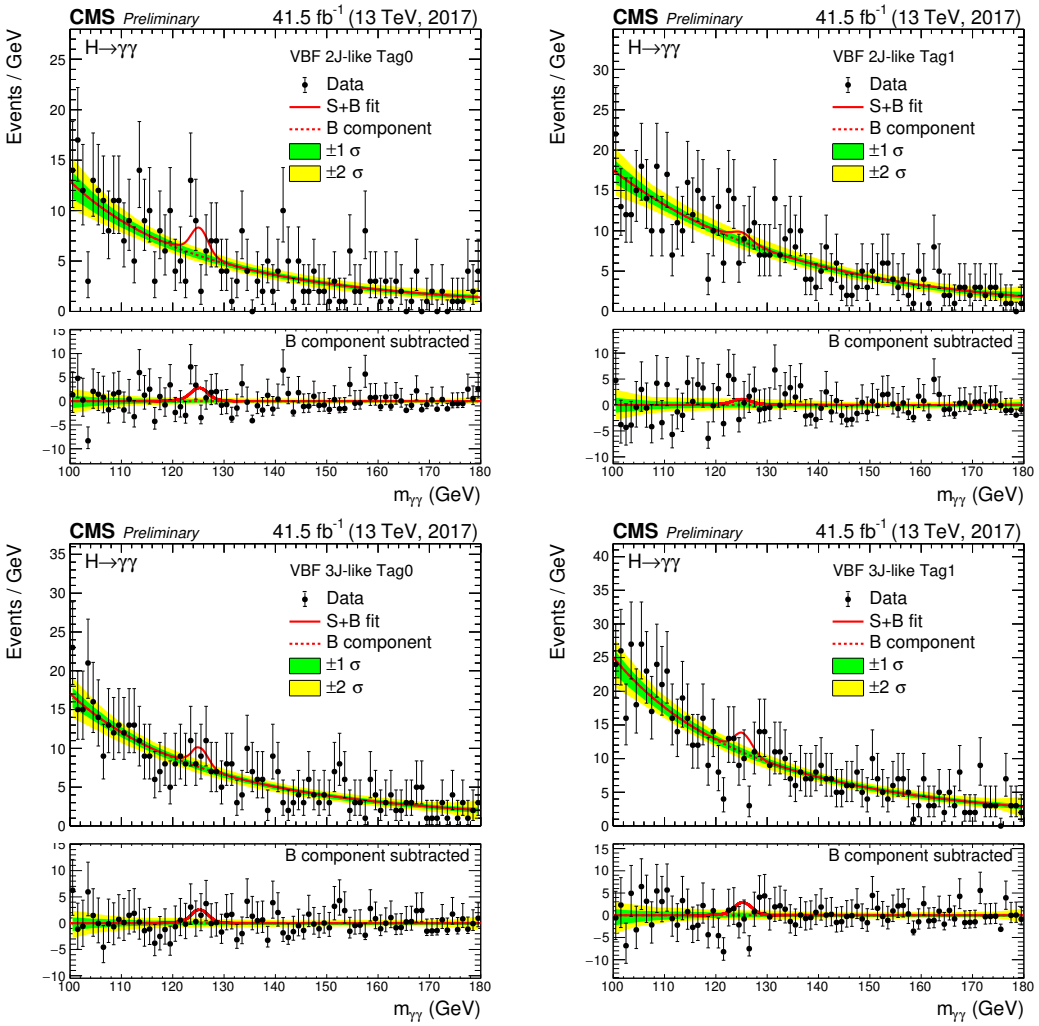


Figure A.13: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

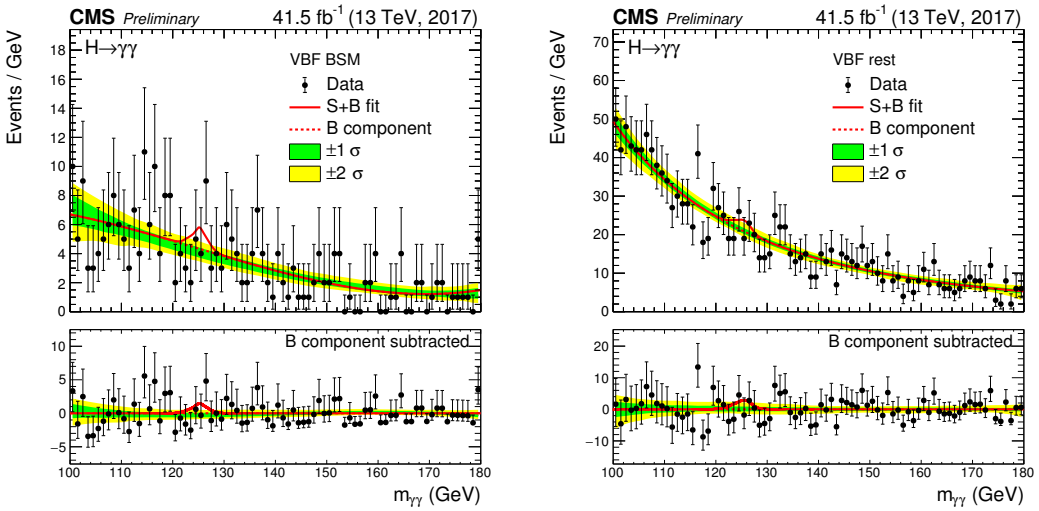


Figure A.14: Data points (black) and signal plus background model fit are shown. The one standard deviation (green) and two standard deviation (yellow) bands include the uncertainties in the background component of the fit. The solid red line shows the contribution from the total signal, plus the background contribution. The dashed red line shows the contribution from the background component of the fit. The bottom plot shows the residuals after subtraction of this background component.

Bibliography

[1] CMS Collaboration. "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 11 (2018), p. 185. DOI: 10.1007/JHEP11(2018)185. arXiv: 1804.02716 [hep-ex].

[2] Ziheng Chen et al. "Offline Reconstruction Algorithms for the CMS High Granularity Calorimeter for HL-LHC". In: *Proceedings, 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference and 24th international Symposium on Room-Temperature Semiconductor X-Ray & Gamma-Ray Detectors (NSS/MIC 2017): Atlanta, Georgia, USA, October 21-28, 2017*. 2018, p. 8532605. DOI: 10.1109/NSSMIC.2017.8532605.

[3] CMS Collaboration. *The Phase-2 Upgrade of the CMS Endcap Calorimeter*. Tech. rep. CERN-LHCC-2017-023. CMS-TDR-019. Technical Design Report of the endcap calorimeter for the Phase-2 upgrade of the CMS experiment, in view of the HL-LHC run. Geneva: CERN, Nov. 2017. URL: https://cds.cern.ch/record/2293646.

[4] CMS Collaboration. *Measurements of Higgs boson production via gluon fusion and vector boson fusion in the diphoton decay channel at $\sqrt{s} = 13$ TeV*. Tech. rep. CMS-PAS-HIG-18-029. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2667225.

[5] S. Glashow. "The renormalizability of vector meson interactions". In: *Nucl. Phys.* 10 (1959), pp. 107–117. DOI: 10.1016/0029-5582(59)90196-8.

[6] S. Weinberg. "A Model of Leptons". In: *Physical Review Letters* 19 (Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[7] A. Salam and J. Ward. "Weak and electromagnetic interactions". In: *Il Nuovo Cimento* 11 (Feb. 1959), pp. 568–577. DOI: 10.1007/BF02726525.

[8] P. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13 (Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[9] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Physical Review Letters* 13 (Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[10] G. Guralnik, C. Hagen, and T. Kibble. "Global Conservation Laws and Massless Particles". In: *Physical Review Letters* 13 (Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[11] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[12] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[13] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[14] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett.* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[15] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001.

[16] D. Clowe et al. "A direct empirical proof of the existence of dark matter". In: *Astrophys. J.* 648 (2006), pp. L109–L113. DOI: 10.1086/508162. arXiv: astro-ph/0608407 [astro-ph].

[17] N. Aghanim et al. "Planck 2018 results. VI. Cosmological parameters". In: (2018). arXiv: 1807.06209 [astro-ph.CO].

[18] Y. Fukuda et al. "Evidence for oscillation of atmospheric neutrinos". In: *Phys. Rev. Lett.* 81 (1998), pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562. arXiv: hep-ex/9807003 [hep-ex].

[19] S. Martin. "A Supersymmetry primer". In: (1997). [Adv. Ser. Direct. High Energy Phys.18,1(1998)], pp. 1–98. DOI: 10.1142/9789812839657_0001, 10.1142/9789814307505_0001. arXiv: hep-ph/9709356 [hep-ph].

[20] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[21] G. Schneider et al. "Double-trap measurement of the proton magnetic moment at 0.3 parts per billion precision". In: *Science* 358.6366 (2017), p. 1081. ISSN: 0036-8075. DOI: 10.1126/science.aan0207.

[22] M. Peskin and D. Schroeder. *An Introduction to quantum field theory*. Addison-Wesley, 1995. ISBN: 9780201503975, 0201503972.

[23] E. Nöther. "Invariant variation problems". In: *Transport Theory and Statistical Physics* 1.3 (1918), p. 186.

[24] P. Dirac. "The quantum theory of the electron". In: *Proc. Roy. Soc. Lond.* A117 (1928), p. 610. DOI: 10.1098/rspa.1928.0023.

[25] D. Griffiths. *Introduction to Elementary Particles*. Wiley, 2009. ISBN: 9783527406012.

[26] M. Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. ISBN: 9781107034266.

[27] ATLAS Collaboration. "Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC". In: *Phys. Lett.* B726 (2013). [Erratum: Phys. Lett.B734,406(2014)], pp. 88–119. DOI: 10.1016/j.physletb.2014.05.011,10.1016/j.physletb.2013.08.010. arXiv: 1307.1427 [hep-ex].

[28] CMS Collaboration. "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV". In: *Eur. Phys. J.* C75.5 (2015), p. 212. DOI: 10.1140/epjc/s10052-015-3351-7. arXiv: 1412.8662 [hep-ex].

[29] ATLAS and CMS Collaborations. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV". In: *JHEP* 08 (2016), p. 045. DOI: 10.1007/JHEP08(2016)045. arXiv: 1606.02266 [hep-ex].

[30] ATLAS Collaboration. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". In: *Phys. Lett.* B784 (2018), pp. 173–191. DOI: 10.1016/j.physletb.2018.07.035. arXiv: 1806.00425 [hep-ex].

[31] ATLAS Collaboration. "Observation of $t\bar{t}H$ production". In: *Phys. Rev. Lett.* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex].

[32] ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector". In: *Phys. Lett.* B786 (2018), pp. 59–86. DOI: 10.1016/j.physletb.2018.09.013. arXiv: 1808.08238 [hep-ex].

[33] CMS Collaboration. "Observation of Higgs boson decay to bottom quarks". In: *Phys. Rev. Lett.* 121.12 (2018), p. 121801. DOI: 10.1103/PhysRevLett.121.121801. arXiv: 1808.08242 [hep-ex].

[34] ATLAS Collaboration. "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev.* D99 (2019), p. 072001. DOI: 10.1103/PhysRevD.99.072001. arXiv: 1811.08856 [hep-ex].

[35] CMS Collaboration. "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector". In: *Phys. Lett.* B779 (2018), pp. 283–316. DOI: 10.1016/j.physletb.2018.02.004. arXiv: 1708.00373 [hep-ex].

[36] ATLAS Collaboration. "Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector". In: *Eur. Phys. J.* C75.10 (2015). [Erratum: Eur. Phys. J.C76,no.3,152(2016)], p. 476. DOI: 10.1140/epjc/s10052-015-3685-1,10.1140/epjc/s10052-016-3934-y. arXiv: 1506.05669 [hep-ex].

[37] CMS Collaboration. "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV". In: *Phys. Rev.* D92.1 (2015), p. 012004. DOI: 10.1103/PhysRevD.92.012004. arXiv: 1411.3441 [hep-ex].

[38] ATLAS and CMS Collaborations. "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments". In: *Phys. Rev. Lett.* 114 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. arXiv: 1503.07589 [hep-ex].

[39] CMS Collaboration. "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 11 (2017), p. 047. DOI: 10.1007/JHEP11(2017)047. arXiv: 1706.09936 [hep-ex].

[40] S. Heinemeyer et al. "Handbook of LHC Higgs cross sections: 3. Higgs Properties". In: (2013). DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347. URL: http://cds.cern.ch/record/1559921.

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

[42] CMS Collaboration. "Combined measurements of Higgs boson couplings in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Submitted to: Eur. Phys. J. C* (2018). arXiv: 1809.10733 [hep-ex].

[43] ATLAS Collaboration. "Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 98 (2018), p. 052005. DOI: 10.1103/PhysRevD.98.052005. arXiv: 1802.04146 [hep-ex].

[44] ATLAS Collaboration. *Measurements of Higgs boson properties in the diphoton decay channel using 80 fb^{-1} of pp collision data at $\sqrt{s}$ = 13 TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2018-028. 2018. URL: https://cds.cern.ch/record/2628771.

[45] ATLAS Collaboration. *Measurements of the Higgs boson production, fiducial and differential cross sections in the 4l decay channel at $\sqrt{s}$ =13 TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2018-018. 2018. URL: http://cds.cern.ch/record/2621479.

[46] L. Evans and P. Bryant. "LHC Machine". In: *JINST* 3.08 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[47] CERN. *LEP Design Report*. Geneva: CERN, 1984. URL: http://cds.cern.ch/record/102083.

[48] C. Lefèvre. *The CERN accelerator complex*. Tech. rep. CERN-DI-0812015. Geneva: CERN, Dec. 2008. URL: https://cds.cern.ch/record/1260465.

[49] LHCb Collaboration. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (2008), S08005. URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08005.

[50] ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08002. URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08002.

[51] B. Acharya et al. "The Physics Programme Of The MoEDAL Experiment At The LHC". In: *Int. J. Mod. Phys.* A29 (2014), p. 1430050. DOI: 10.1142/S0217751X14300506. arXiv: 1405.7662 [hep-ph].

[52] The TOTEM Collaboration. "The TOTEM Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (2008), S08007. DOI: 10.1088/1748-0221/3/08/s08007. URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08007.

[53] E. Berti et al. “The LHCf experiment: present status and physics results”. In: 2017. arXiv: 1710.03991 [hep-ex].

[54] CMS Collaboration. *CMS Luminosity - Public Results*. http://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[55] T. Sakuma and T. McCauley. *Detector and Event Visualization with SketchUp at the CMS Experiment*. 2014. URL: http://stacks.iop.org/1742-6596/513/i=2/a=022032.

[56] CMS Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[57] CMS Collaboration. *CMS Technical Design Report for the Pixel Detector Upgrade*. Tech. rep. CERN-LHCC-2012-016. CMS-TDR-11. Sept. 2012. URL: https://cds.cern.ch/record/1481838.

[58] A. Benaglia. “The CMS ECAL performance with examples”. In: *Journal of Instrumentation* 9.02 (2014), p. C02008. URL: http://stacks.iop.org/1748-0221/9/i=02/a=C02008.

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

[60] CMS Collaboration. “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det].

[61] CMS Collaboration. “The CMS trigger system”. In: *JINST* 12.01 (2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020. arXiv: 1609.02366 [physics.ins-det].

[62] D. Barney. “CMS Detector Slice”. CMS Collection. Jan. 2016. URL: https://cds.cern.ch/record/2120661.

[63] G. Apollinari et al. “High Luminosity Large Hadron Collider HL-LHC”. In: *CERN Yellow Report* 5 (2015), pp. 1–19. DOI: 10.5170/CERN-2015-005.1. arXiv: 1705.08830 [physics.acc-ph].

[64] ATLAS and CMS Collaborations. “Report on the Physics at the HL-LHC and Perspectives for the HE-LHC”. In: *HL/HE-LHC Physics Workshop: final jamboree Geneva, CERN, March 1, 2019*. 2019. arXiv: 1902.10229 [hep-ex].

[65] D Contardo et al. *Technical Proposal for the Phase-II Upgrade of the CMS Detector*. Tech. rep. CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02. Geneva, June 2015. URL: https://cds.cern.ch/record/2020886.

[66] CMS Collaboration. *Technical proposal for a MIP timing detector in the CMS experiment Phase 2 upgrade*. Tech. rep. CERN-LHCC-2017-027. LHCC-P-009. Geneva: CERN, Dec. 2017. URL: https://cds.cern.ch/record/2296612.

[67] CMS Collaboration. *The Phase-2 Upgrade of the CMS Tracker*. Tech. rep. CERN-LHCC-2017-009. CMS-TDR-014. Geneva: CERN, June 2017. URL: https://cds.cern.ch/record/2272264.

[68] CMS Collaboration. *The Phase-2 Upgrade of the CMS Barrel Calorimeters*. Tech. rep. CERN-LHCC-2017-011. CMS-TDR-015. Geneva: CERN, Sept. 2017. URL: https://cds.cern.ch/record/2283187.

[69] CMS Collaboration. *The Phase-2 Upgrade of the CMS Muon Detectors*. Tech. rep. CERN-LHCC-2017-012. CMS-TDR-016. Geneva: CERN, Sept. 2017. URL: https://cds.cern.ch/record/2283189.

[70] CMS Collaboration. *The Phase-2 Upgrade of the CMS L1 Trigger Interim Technical Design Report*. Tech. rep. CERN-LHCC-2017-013. CMS-TDR-017. Geneva: CERN, Sept. 2017. URL: https://cds.cern.ch/record/2283192.

[71] CMS Collaboration. *The Phase-2 Upgrade of the CMS DAQ Interim Technical Design Report*. Tech. rep. CERN-LHCC-2017-014. CMS-TDR-018. Geneva: CERN, Sept. 2017. URL: https://cds.cern.ch/record/2283193.

[72] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[73] A. Rodriguez and A. Laio. "Clustering by fast search and find of density peaks". In: *Science* 344.6191 (2014), pp. 1492–1496. ISSN: 0036-8075. DOI: 10.1126/science.1242072. URL: http://science.sciencemag.org/content/344/6191/1492.

[74] CMS Collaboration. "Performance of Photon Reconstruction and Identification with the CMS Detector in Proton-Proton Collisions at sqrt(s) = 8 TeV". In: *JINST* 10.08 (2015), P08010. DOI: 10.1088/1748-0221/10/08/P08010. arXiv: 1502.02702 [physics.ins-det].

[75] ATLAS Collaboration. "Light-quark and gluon jet discrimination in *pp* collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector". In: *Eur. Phys. J.* C74.8 (2014), p. 3023. DOI: 10.1140/epjc/s10052-014-3023-z. arXiv: 1405.6583 [hep-ex].

[76] G. Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J.* C71 (2011). [Erratum: Eur. Phys. J.C73,2501(2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z. arXiv: 1007.1727 [physics.data-an].

[77] CALICE Collaboration. "Tests of a particle flow algorithm with CALICE test beam data". In: *JINST* 6 (2011), P07005. DOI: 10.1088/1748-0221/6/07/P07005. arXiv: 1105.3417 [physics.ins-det].

[78] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[79] T. Sjostrand, S. Mrenna, and P. Z. Skands. "A Brief Introduction to PYTHIA 8.1". In: *Comput. Phys. Commun.* 178 (2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[80] K. Hamilton et al. "NNLOPS simulation of Higgs boson production". In: *JHEP* 10 (2013), p. 222. DOI: 10.1007/JHEP10(2013)222. arXiv: 1309.0017 [hep-ph].

[81] C. Oleari. "The POWHEG-BOX". In: *Nucl. Phys. Proc. Suppl.* 205-206 (2010), pp. 36–41. DOI: 10.1016/j.nuclphysbps.2010.08.016. arXiv: 1007.3893 [hep-ph].

[82] T. Gleisberg et al. "Event generation with SHERPA 1.1". In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

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

[84] M. Cacciari, G. Salam, and G. Soyez. "The anti- k t jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (2008), p. 063. URL: http://stacks.iop.org/1126-6708/2008/i=04/a=063.

[85] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[86] CMS Collaboration. "Measurement of the associated production of a Higgs boson and a pair of top-antitop quarks with the Higgs boson decaying to two photons in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: (2018).

[87] CMS Collaboration. "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV". In: *JINST* 10.06 (2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det].

[88] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. ISBN: 9780387848587.

[89] C. Anastasiou et al. "Higgs boson gluon-fusion production in QCD at three loops". In: *Phys. Rev. Lett.* 114 (2015), p. 212001. DOI: 10.1103/PhysRevLett.114.212001. arXiv: 1503.06056 [hep-ph].

[90] C. Anastasiou et al. "High precision determination of the gluon fusion Higgs boson cross-section at the LHC". In: *JHEP* 05 (2016), p. 058. DOI: 10.1007/JHEP05(2016)058. arXiv: 1602.00695 [hep-ph].

[91] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *CoRR* abs/1603.02754 (2016). arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754.

[92] Jack Wright. "Study of Higgs boson production through vector boson fusion at the CMS experiment using a dense convolutional neural network". PhD thesis. Imperial College London, 2018.

[93] P. Dauncey et al. "Handling uncertainties in background shapes: the discrete profiling method". In: *JINST* 10.04 (2015), P04015. DOI: 10.1088/1748-0221/10/04/P04015. arXiv: 1408.6865 [physics.data-an].

[94] Louie Corpe. "Study of Higgs boson production through its decay to two photons using data collected at a centre-of-mass energy of 13 TeV with the CMS detector". PhD thesis. Imperial College London, 2017.

[95] R. Fisher. "On the interpretation of χ^2 from contingency tables, and the calculation of P". In: *Journal of the Royal Statistical Society* 85.1 (1922), pp. 87–94.

[96] J. Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys.* G43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph].

[97] R. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[98] S. Carrazza et al. "An Unbiased Hessian Representation for Monte Carlo PDFs". In: *Eur. Phys. J.* C75.8 (2015), p. 369. DOI: 10.1140/epjc/s10052-015-3590-7. arXiv: 1505.06736 [hep-ph].

[99] CMS Collaboration. *CMS Luminosity Measurements for the 2016 Data Taking Period*. Tech. rep. CMS-PAS-LUM-17-001. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2257069.

[100] CMS Collaboration. *CMS luminosity measurement for the 2017 data-taking period at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-LUM-17-004. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2621960.

[101] CMS Collaboration. *Jet algorithms performance in 13 TeV data*. Tech. rep. CMS-PAS-JME-16-003. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2256875.

[102] W. Verkerke and D. Kirkby. "The RooFit toolkit for data modeling". In: *eConf* C0303241 (2003). arXiv: physics/0306116 [physics].

[103] S. Dawson et al. "Working Group Report: Higgs Boson". In: *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013*. 2013. arXiv: 1310.8361 [hep-ex]. URL: http://www.slac.stanford.edu/econf/C1307292/docs/EnergyFrontier/Higgs-18.pdf.

[104] C. Hays, V. Sanz Gonzalez, and G. Zemaityte. "Constraining EFT parameters using simplified template cross sections". In: (Oct. 2017). URL: https://cds.cern.ch/record/2290628.

[105] CMS Collaboration. *Constraints on the Higgs boson self-coupling from ttH+tH, H to gamma gamma differential measurements at the HL-LHC*. Tech. rep. CMS-PAS-FTR-18-020. 2018. URL: http://cds.cern.ch/record/2647986.