# Observation of a New Particle in the Search for the Standard Model Higgs Boson at the CMS Detector

Nicholas Wardle
Imperial College London

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

# Abstract

The discovery of the Standard Model (SM) Higgs boson is one of the primary physics objectives of the Large Hadron Collider at CERN. This thesis describes a search carried out for the SM Higgs boson on data collected during the 2011 and 2012 proton-proton (pp) collision runs with the CMS detector corresponding to integrated luminosities of $5.1 fb^{-1}$ and $5.3 fb^{-1}$ respectively. A detailed description of the search for the SM Higgs boson decaying to two photons from the full dataset collected at CMS during the 2011 pp collision run is provided. In particular, the development of signal and background modelling techniques used for statistical interpretations of the data are highlighted. Results of the search using these techniques from the 2011 dataset are presented. In addition, an update to the analysis including data taken during 2012 is described and the results from the combined 2011 and 2012 analyses given. Results from the combination of several Higgs decay channels at CMS are reported, including those presented in the International Conference on High Energy Physics in July 2012 at which the announcement of discovery was made. Ongoing studies to ascertain the properties of the new particle are discussed and some preliminary results from the combined 7 and 8 TeV datasets (corresponding to $5.1 fb^{-1}$ and $12.2 fb^{-1}$ respectively) are presented.

# Declaration

I, the author of this thesis, hereby declare the work contained in this document to be my own. Studies conducted and results produced by the author are indicated in the main body of text. All figures labelled "CMS" are sourced directly from CMS publications, including those produced by the author and have, been referenced as such in the figure caption. Where the figure is sourced from a CMS document which is unpublished or from a preliminary public document (marked "CMS Preliminary"), a reference to that document is included. All figures and studies taken from external sources are referenced appropriately throughout this document.

Nicholas Wardle

# Acknowledgements

I would like to thank foremost my parents, Pat and David, who have provided me every opportunity to pursue research in Physics. Their unwaivering support and encouragement has been an endless source of determination throughout my education. In addition, I would like to thank my friends and colleagues (they know who they are) who provided much needed distraction from study and helping me appreciate other aspects of life in Geneva and London. Secondly, I thank my supervisors Jonathan Hays and Gavin Davies for guiding me through my PhD research. The mix of enthusiasm for the subtleties of data analysis techniques and expertise in maintaining the bigger picture have provided many hours of educational and entertaining discussion. I'd like to thank the Imperial College $H \to \gamma\gamma$ group and the CMS $H \to \gamma\gamma$ group for providing a platform to discuss ideas and results in an open and often welcoming manner. Finally, I would like to thank the STFC for providing the funding for my research and in particular allowing for the time spent in Geneva.

# Contents

*"Un bon mot ne prouve rien."*
— François-Marie Arouet (Voltaire)

# Chapter 1.

# Introduction

The discovery of a new particle was announced by the ATLAS and CMS Collaborations on the 4th of July 2012. The long-awaited discovery followed decades of experimental endeavours in the search for the Higgs boson, the missing piece of the Standard Model (SM) of particle physics. If further measurements of the properties of the new particle fit the SM predictions, the discovery will serve as compelling evidence for the mechanism by which spontaneous symmetry breaking in the SM occurs, giving rise to the masses of the fundamental fermions and bosons.

In Chapter 2, an introduction to the fundamental constituents of matter and the interactions between them is given. The mechanism by which the fundamental fermions and bosons acquire mass in the SM, spontaneous symmetry breaking, is outlined, serving as a motivation for the search for the SM Higgs boson. Previous searches and indirect constraints are discussed with the chapter concluding in the search strategies employed at the LHC.

Chapter 3 describes the experimental apparatus required to undertake such a search, in particular the CMS detector which was used to collect the data upon which the majority of the author's research was conducted. This chapter includes a section describing a set of jet energy calibrations derived by the author which were subsequently used in the Level-1 trigger system at CMS.

The main analysis conducted by the author is detailed in Chapter 4. This chapter contains a description of the search for the Standard Model Higgs boson in the two photon decay channel carried out on proton-proton collision data collected at CMS during 2011. The focus of the chapter is on the background modelling technique developed by the author used for statistical interpretations of the data. This method was one of two

developed at CMS, which served as a cross-check of the background model used for the published result. The template signal modelling technique developed for this analysis is also used regularly by the $H \to \gamma\gamma$ working group at CMS for fast production of results and analysis development in a common analysis framework. The chapter concludes with the updates for the 2012 analysis including data collected at a centre of mass energy of 8 TeV.

Finally, in Chapter 6, the statistical tools employed and developed at CMS for the purposes of combined Higgs boson searches are detailed. The chapter includes the results presented at the July 2012 International Conference of High Energy Physics during which the announcement of the discovery of the new particle was made by the ATLAS and CMS Collaborations. The section concludes with a discussion of the ongoing research at CMS intended to ascertain the properties of the newly discovered particle and includes results produced by the author for the Hadron Collider Physics (HCP) symposium in November 2012.

In addition to the work contained in this thesis, the author contributed towards early studies in electroweak physics at CMS. The studies undertaken involved the development of a robust signal extraction technique used to measure the production cross-section of $W$ bosons, via their decay to electrons, in proton-proton collisions at 7 TeV. The technique utilised control samples in data to subtract backgrounds from QCD, exploiting the kinematic signature of the decay $W \to e\nu$. Re-establishing well measured Standard Model processes, such as $pp \to W \to e\nu$, was one of the first major goals of CMS, ensuring a high level of understanding of the detector components and their calibration. The analysis was performed on the first $36 fb^{-1}$ of data collected at CMS during 2010 and contributed towards the publication containing the $W$ cross-section measurement from that dataset [1, 2].

# Chapter 2.

# Theory and Motivations

The goal of particle physics is to identify the most elemental constituents of matter and understand the nature of the fundamental forces acting between them. In this chapter, a brief summary of the components of the Standard Model will be given along with the motivation for the search for the Standard Model Higgs boson. Section 2.1 introduces the mechanism by which mass is generated in the Standard Model and its relation to the SM Higgs boson is highlighted. In Section 2.2, searches for, and indirect constraints on, the SM Higgs boson before the start up of the Large Hadron Collider (LHC) are discussed. The section concludes with how the Higgs boson can be produced and observed in proton-proton collisions at the LHC.

## 2.1. The Standard Model of Particle Physics

The Standard Model (SM) is a well tested, precision model of particle physics. Within the confines of quantum field theory (QFT), the SM provides a description of the electromagnetic, weak-nuclear and strong nuclear interactions, incorporating both relativistic and quantum mechanical effects.

### 2.1.1. Fundamental Matter Particles

All of the known fundamental constituents of matter are spin-$\frac{1}{2}$ fermions. The equation of motion for a spin-$\frac{1}{2}$ particle with mass $m$, given in Equation 2.1, was provided by

Dirac.

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \tag{2.1}$$

The matrices $\gamma^\mu$, $\mu \in 0, 1, 2, 3$, are defined by the anti commutator relation $\gamma^\mu\gamma^\nu + \gamma^\mu\gamma^\nu = 2\eta^{\mu\nu} I_4$ where $\eta^{\mu\nu}$ is the flat space-time metric $(+, -, -, -)$ and $I_4$ is the $4 \times 4$ identity matrix. The solutions, $\psi$, to Equation 2.1 yield the particle and anti-particle states which satisfy the relativistic expression, $E^2 = \mathbf{p} \cdot \mathbf{p} + m^2$, for a massive particle with momentum $\mathbf{p}$ and energy $E$.

The fundamental fermions are separated into those which do (quarks) and do not (leptons) interact with the strong nuclear force. Quarks and leptons are grouped into three generations which share the same properties but increase in mass. Unlike the leptons, quarks are not seen as free particles in nature, but rather are confined to exist within baryons composed of three quarks and quark-anti-quark pairs known as mesons. A summary of the known fundamental fermions in their three generations is given in Table 2.1.

| | **I** | | **II** | | **III** | | **Charge** |
|---|---|---|---|---|---|---|---|
| Leptons | electron | $e$ | muon | $\mu$ | tau | $\tau$ | -1 |
| | electron neutrino | $\nu_e$ | muon neutrino | $\nu_\mu$ | tau neutrino | $\nu_\tau$ | 0 |
| Quarks | up | $u$ | charm | $c$ | top | $t$ | $+\frac{2}{3}$ |
| | down | $d$ | strange | $s$ | bottom | $b$ | $-\frac{1}{3}$ |

**Table 2.1.:** Fundamental fermions in the Standard Model. All of the fundamental fermions are spin-$\frac{1}{2}$ particles. The anti-fermion counterparts are not listed here.

### 2.1.2. Fundamental Forces

The fundamental forces of nature are mediated by the exchange of gauge bosons. They are all spin-1 particles which arise from consideration of the symmetries which the relevant theory possesses (See Section 2.1.3). The quantum field theories of electromagnetism, Quantum Electro-dynamics (QED), and the strong nuclear force, Quantum Chromo-dynamics (QCD), yield massless mediator bosons, the photon and the gluons, which are a direct consequence of the gauge invariance of those theories. Despite this, the typical ranges over which the two interactions occur are dramatically different; strong

interaction effects are only apparent on a scale of around $10^{-15}$m whereas the range of electromagnetic interactions are effectively infinite.

The mediators of the weak nuclear and electromagnetic forces arise through the unification of the theories of weak and electromagnetic interactions and the mixing of the associated gauge fields. The weak gauge bosons, $W^{\pm}$ and $Z$, unlike the photon and gluons, have a finite mass which has been measured experimentally [3, 4]. A summary of the fundamental gauge bosons of the Standard Model is given in Table 2.2. A quantum description of gravity is not included in the Standard Model. This is a reasonable approximation as the strength of this interaction is much smaller than the other three, thereby having no impact on the predictive power of the model.

| | **Mediator Particle** | | **Charge** | **Mass (GeV)** |
|---|---|---|---|---|
| Electromagnetism | photon | $\gamma$ | 0 | 0 |
| Strong Nuclear | gluon | $g_j$, $j \in \{1, \cdots 8\}$ | 0 | 0 |
| Weak Nuclear | | $W^+$ | +1 | 80.39 |
| | | $W^-$ | -1 | 80.39 |
| | | $Z$ | 0 | 91.19 |

**Table 2.2.:** Fundamental gauge bosons in the Standard Model. All of the gauge-bosons are spin-1 particles. The masses of the $W^{\pm}$ and $Z$ bosons are taken from References [3] and [4] respectively.

### 2.1.3. Electroweak Gauge Symmetry

Symmetries in nature are often found to relate to some underlying physical principle or fundamental law. It was first shown by Emmy Noether that for any physical system which can be described in the Lagrangian formalism, any symmetry of the Lagrangian has an associated conserved quantity [5]. In the context of dynamical quantum theories, the particular characteristics of particle interactions can be used to constrain the appropriate Lagrangian by identifying a particular group of transformations under which the Lagrangian should be symmetric (invariant).

One of the major achievements of the twentieth century in the development of the Standard Model was the unification of the electromagnetic and weak interactions [6, 7, 8]. The original proposal, by Glashow in 1961, was to construct a theory which incorporates

the characteristics of the weak and electromagnetic interactions by associating them with a particular symmetry group [6]. The physical nature of electroweak interactions is encoded into a Lagrangian which is invariant under transformations of the group $SU(2)_L \times U(1)_Y$. This group has three generators for $SU(2)_L$, $T_i = \frac{1}{2}\tau_i$ where $\tau_i,\ i \in \{1,2,3\}$ are the $2 \times 2$ Pauli-spin matrices, and one additional generator for $U(1)_Y$, $Y$. The quantum numbers associated with the $SU(2)_L$ group, weak isospin $t_{1,2,3}$, and $U(1)_Y$ group, hypercharge $y$, are related to the electric charge $Q$ as,

$$Q = t_3 + \frac{y}{2}, \tag{2.2}$$

where the factor of $\frac{1}{2}$ is chosen by convention. The associated gauge fields are $\hat{\mathbf{W}}_\mu = \left(\hat{W}^1_\mu, \hat{W}^2_\mu, \hat{W}^3_\mu\right)$ and $\hat{B}_\mu$. An example Lagrangian for interactions within the first leptonic generation of fermions, $\hat{\mathscr{L}_G}$, is given in Equation 2.3,

$$\begin{aligned} \hat{\mathscr{L}_G} &= \bar{\chi}_L \gamma^\mu \left[ i\partial_\mu - g\frac{1}{2}\boldsymbol{\tau}\cdot\hat{\mathbf{W}}_\mu - g'\left(-\frac{1}{2}\right)\hat{B}_\mu \right]\chi_L \\ &\quad + \bar{e}_R \gamma^\mu \left[ i\partial_\mu - g'(-1)\hat{B}_\mu \right] e_R - \frac{1}{4}\hat{\mathbf{W}}_{\mu\nu}\cdot\hat{\mathbf{W}}^{\mu\nu} - \frac{1}{4}\hat{B}_{\mu\nu}\hat{B}^{\mu\nu} \end{aligned} \tag{2.3}$$

where the bar notation denotes the adjoint of the field, $\bar{\psi} = \psi^\dagger\gamma^0$ and $\chi_L$ is the left handed component of the leptonic fermion doublet. The field tensors, $\hat{\mathbf{W}}_{\mu\nu}$ and $\hat{B}_{\mu\nu}$ given in Equations 2.4 and 2.5, describe the kinematics of the gauge fields.

$$\hat{\mathbf{W}}_{\mu\nu} = \partial_\mu\hat{\mathbf{W}}_\nu - \partial_\nu\hat{\mathbf{W}}_\mu - g\hat{\mathbf{W}}_\mu \wedge \hat{\mathbf{W}}_\nu \tag{2.4}$$

$$\hat{B}_{\mu\nu} = \partial_\mu\hat{B}_\nu - \partial_\nu\hat{B}_\mu. \tag{2.5}$$

Experimentally, it has been verified that the weak nuclear force explicitly violates parity, that is transformations under spatial inversions $x \to -x$ [9]. A fermionic field, $\psi$, can be projected into its left and right handed components, $\psi_L$ and $\psi_R$, using the operators $\frac{1}{2}(1 \mp \gamma^5)$ respectively, where $\gamma^5 = \gamma^0\gamma^1\gamma^2\gamma^3$. As the weak nuclear force only interacts with left-handed fermions, right-handed components of the fermion fields are invariant under $SU(2)_L$ transformations. The right-handed component of the neutrino field therefore does not appear in the Lagrangian, $\hat{\mathscr{L}_G}$, since it interacts with neither the electromagnetic nor the weak interactions. Under the $SU(2)_L \times U(1)_Y$ group, the left handed components of the leptonic fermion fields, $\chi_L$ of Equation 2.3, transform as a

doublet

$$\chi_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \longrightarrow \exp(-i\boldsymbol{\alpha}\cdot\frac{\boldsymbol{\tau}}{2} - i\alpha) \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \tag{2.6}$$

whereas the right-handed component of the electron field transforms as a singlet.

$$e_R \longrightarrow \exp(-2i\alpha)e_R. \tag{2.7}$$

The transformations are "local" in the sense that the coefficients $\boldsymbol{\alpha}$ and $\alpha$ are functions of space-time. To maintain the symmetry under local transformations of this type, the gauge fields transform as follows,

$$\hat{\mathbf{W}}_\mu \longrightarrow \hat{\mathbf{W}}_\mu - \frac{1}{g}\partial_\mu\boldsymbol{\alpha} - \boldsymbol{\alpha}\wedge\hat{\mathbf{W}}_\mu \tag{2.8}$$

$$\hat{B}_\mu \longrightarrow \hat{B}_\mu - \frac{1}{g'}\partial_\mu\alpha \tag{2.9}$$

The Lagrangian of Equation 2.3 contains no explicit terms which relate to the mass of the electron ($m_e$). Including the electron's mass directly would require the addition of the term,

$$\begin{aligned} -m_e\bar{e}e &= -m_e\bar{e}\left[\frac{1}{2}\left(1-\gamma^5\right)+\frac{1}{2}\left(1+\gamma^5\right)\right]e \\ &= -m_e\left(\bar{e}_R e_L + \bar{e}_L e_R\right). \end{aligned} \tag{2.10}$$

As $e_L$ transforms as a member of a doublet and $e_R$ as a singlet, the addition of this term to Equation 2.3 would break the symmetry of the Lagrangian which motivated its construction, namely transformations under the $SU(2)_L$ group [10].

The physical electroweak boson fields, $\hat{W}^\pm_\mu$, $\hat{Z}_\mu$ and photon field, $\hat{A}_\mu$, are obtained through a mixture of the electroweak gauge fields as,

$$\begin{aligned} \hat{W}^\pm_\mu &= \sqrt{\frac{1}{2}}\left(\hat{W}^1_\mu \mp i\hat{W}^2_\mu\right) \\ \hat{Z}_\mu &= \cos\theta_w\hat{W}^3_\mu - \sin\theta_w\hat{B}_\mu \\ \hat{A}_\mu &= \sin\theta_w\hat{W}^3_\mu + \cos\theta_w\hat{B}_\mu, \end{aligned} \tag{2.11}$$

where the mixing angle, $\theta_w = \tan^{-1}\frac{g'}{g}$, relates the couplings of the weak neutral and electromagnetic interactions. As expected, there is no term which corresponds to the mass of the photon, however, the same is true for the $W$ and $Z$ bosons. The masses of the $W$ and $Z$ bosons, given in Table 2.2, have been measured experimentally and found to be non-zero. The inclusion of mass terms for these bosons in Equation 2.3 would also break the symmetry of the Lagrangian. Furthermore, it has been shown that the inclusion of these mass terms results in a loss of re-normalizability of the theory, making it less powerful for predicting observables such as cross-sections and decay rates [11]. Instead, these masses can be generated via a spontaneous, rather than explicit, breaking of the symmetry.

### 2.1.4. Spontaneous Symmetry Breaking: The Higgs Mechanism

In quantum field theory, a symmetry is "spontaneously" broken when the Lagrangian itself remains invariant while the vacuum state, for which the Hamiltonian of the theory attains its minimum, does not [10]. In the context of the electroweak theory, spontaneous symmetry breaking is achieved through the introduction of a complex scalar field which attains a non-zero vacuum expectation value (VEV) [12, 13, 14, 15, 16]. This field is an $SU(2)$ doublet,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{2.12}$$

The Lagrangian, $\hat{\mathscr{L}_G}$, of Equation 2.3 is modified to include an additional term which is $SU(2)_L \times U(1)_Y$ invariant, $\hat{\mathscr{L}_\phi}$ given by,

$$\hat{\mathscr{L}_\phi} = (\hat{D}_\mu\phi)^\dagger(\hat{D}^\mu\phi) + \mu^2\phi^\dagger\phi - \frac{\lambda}{4}(\phi^\dagger\phi)^2, \tag{2.13}$$

where the covariant derivative $\hat{D}^\mu$ which acts on $\phi$ is given by,

$$\hat{D}^\mu = \partial^\mu + ig\frac{1}{2}\boldsymbol{\tau}\cdot\hat{\mathbf{W}}^\mu + ig'\frac{1}{2}\hat{B}^\mu. \tag{2.14}$$

The second two terms in Equation 2.13 correspond to the Higgs potential. In order to generate masses for the gauge bosons, the parameters, $\mu$ and $\lambda$, must satisfy $\mu^2 > 0$ and

$\lambda > 0$. The choice of non-zero VEV must then be made so that only the $W$ and $Z$ bosons acquire mass, while the symmetry associated with electromagnetism remains unbroken, leaving the photon massless. The choice suggested by Weinberg in 1967 [7] was,

$$\text{VEV} = \langle 0|\phi|0\rangle = \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix}, \tag{2.15}$$

where $v = \frac{2\mu}{\sqrt{\lambda}}$. In order to obtain the physical particle spectrum, perturbations around the vacuum state are considered. If $\hat{\boldsymbol{\theta}}$ and $\hat{H}$ represent small variations in the four degrees of freedom of the field $\phi$ then,

$$\phi = \exp(-i\hat{\boldsymbol{\theta}} \cdot \frac{1}{2v}\boldsymbol{\tau}) \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + \hat{H}) \end{pmatrix}. \tag{2.16}$$

This can be simplified by choosing the phase fields $\hat{\boldsymbol{\theta}}$ to be zero. The Lagrangian obtained by inserting $\phi$ with this form into Equation 2.13 and adding it to the Lagrangian of Equation 2.3 is,

$$\begin{aligned} \hat{\mathscr{L}_\phi} + \hat{\mathscr{L}_G} &= \frac{1}{2}\partial_\mu \hat{H} \partial^\mu \hat{H} - \mu^2 \hat{H}^2 + \frac{1}{8} g^2 v^2 \hat{W}_{1\mu} \hat{W}_1^\mu + \frac{1}{8} g^2 v^2 \hat{W}_{2\mu} \hat{W}_2^\mu \\ &\quad - \frac{v^2}{8}\left(g^2 + g'^2\right) \hat{Z}_\mu \hat{Z}^\mu + KB, \end{aligned} \tag{2.17}$$

where only terms which are at most second order in the fields are kept, illustrating the physical particle spectrum, and the fermion fields are dropped altogether. The relation between the $\hat{W}_3^\mu$ and $\hat{B}^\mu$ fields from Equation 2.11 has been used to obtain the physical photon, $\hat{A}^\mu$, and $\hat{Z}^\mu$ fields. From this form of the Lagrangian, it is clear that the $\hat{W}_1^\mu$, $\hat{W}_2^\mu$ and $\hat{Z}$ fields acquire mass. As the $\hat{W}_1^\mu$ and $\hat{W}_2^\mu$ fields mix to form the physical $\hat{W}^\pm$ fields, the $W^\pm$ bosons acquire a mass of $m_W = \frac{gv}{2}$. The mass of the $Z$ boson is given by $m_Z = \frac{v}{2}\sqrt{g^2 + g'^2}$, while there is no term associated with the mass of the photon. An additional scalar field, $\hat{H}$ (the Higgs boson), remains in the Lagrangian with mass $\sqrt{2}\mu$. The term $KB$ in Equation 2.17 denotes additional kinetic terms for the $\hat{W}_1^\mu$, $\hat{W}_2^\mu$, $\hat{Z}^\mu$ and $\hat{A}^\mu$ fields. The masses of the fermions are generated by adding Yukawa coupling terms,

$$-\lambda_f \bar{\chi}_L \phi \psi_R + \lambda_{f'} \bar{\psi}_R (-i\tau_2 \phi^*) \chi_L, \tag{2.18}$$

to $\hat{\mathscr{L}_\phi}$. The couplings $\lambda_f$, $f = u, d, e, \mu \cdots$, are directly related to the mass of the fermions, specifically $\lambda_f \propto m_f$ such that the heavier fermions have stronger coupling to the Higgs boson. Although the SM does not predict the values of these couplings, the masses of the fermions are experimentally measurable allowing access to, and providing constraints on, the properties of the Higgs boson.

## 2.2. The SM Higgs Boson

The introduction of a complex scalar field into the Standard Model to generate masses for the SM particles results in the prediction of a new massive scalar boson, the Higgs boson [12, 13, 14, 15, 16]. The discovery of such a particle would give strong evidence as to the nature of electroweak symmetry breaking and hence many searches for it have been launched since its existence was first proposed.

The mass of the SM Higgs boson, $m_H$, is not a predicted quantity in the SM but is rather a function of the self-coupling parameter, $\lambda$, and $v$. The latter is determined experimentally to be $v = 246$ GeV by precisely measuring the rate of muon decay [17]. However, since $\lambda$ is unconstrained, a large range in $m_H$ remains theoretically acceptable for the Higgs boson mass.

### 2.2.1. Constraints and Previous Searches

Several theoretical considerations constrain the mass of the SM Higgs boson [18]. The desire to avoid the need for non-pertubative calculations for electroweak processes at high energies constrains the SM Higgs boson mass to be less than around 770 GeV [19]. Conversely, if $m_H$ is too small, then the Higgs potential of Equation 2.13 contains a global minimum at large values of the scalar field $\phi$. Additional physics, beyond that of the SM, would be required so that this global minimum corresponds to the observed vacuum with $v = 246$ GeV. This places a loose lower bound on the SM Higgs boson mass of about 115 GeV [20].

#### Direct Searches

The first direct constraints on the Higgs boson at higher masses were provided by the four experiments operating at the Large Electron-Positron (LEP) collider. By steadily

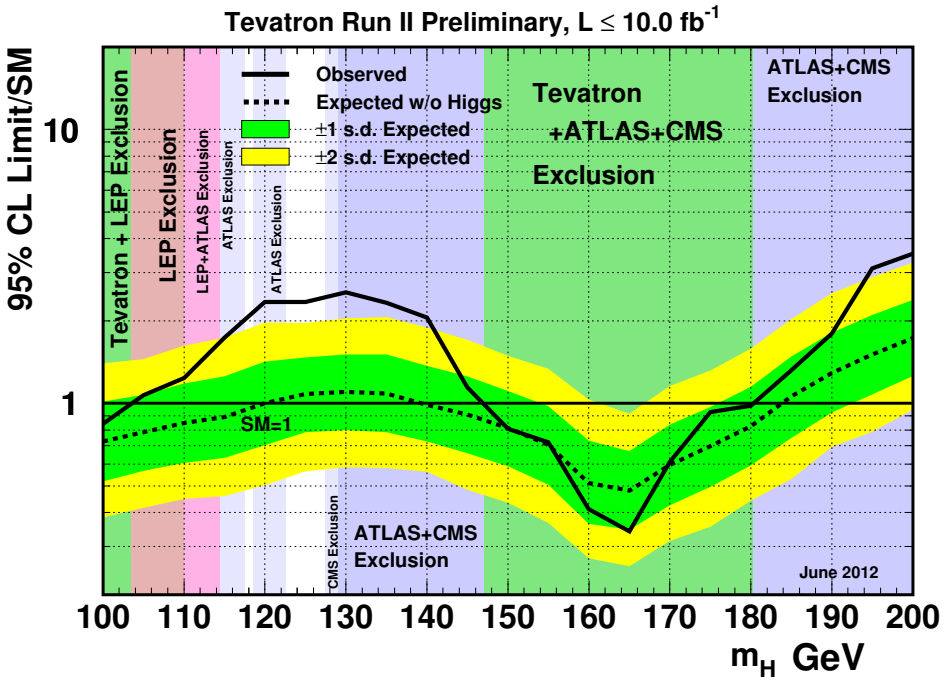


**Figure 2.1.:** The 95% confidence upper limits on the ratio of Higgs boson production to the SM prediction as a function of $m_H$. The dotted line indicates the median expected exclusion assuming no SM Higgs boson exists while the solid line indicates the observed exclusion obtained from the data. Where this line falls below 1, a SM Higgs boson with that mass is excluded at the 95% confidence level as indicated by the green bands. The other coloured bands indicate exclusion limits resulting from direct searches for the SM Higgs boson conducted by other Collaborations before June 2012. The figure has been altered from its original source [22].

increasing the centre of mass energy of the collisions, LEP was able to exclude masses of $m_H < 114.4$ GeV at the 95% confidence level [21]. Prior to the LHC turn on, the CDF and D0 experiments at the Tevatron collider provided additional limits on the mass of the Higgs boson through direct searches in proton anti-proton collisions. The centre of mass energy available in these collisions, $\sqrt{s} = 1.96$ TeV, provided sensitivity to Higgs boson masses between 90 and 190 GeV. Priority at the Tevatron experiments was given to the $H \to WW$ channel at high mass and $H \to bb$, with associated production of a $W$ or $Z$ boson, at low mass. By the shutdown of the Tevatron in 2011, the two experiments had collected combined datasets corresponding to a total integrated luminosity of $10 fb^{-1}$. Figure 2.1 shows the 95% confidence upper limits on the ratio of the excluded Higgs boson production cross-section to that predicted by the Standard Model as a function of $m_H$ obtained from this dataset. Mass hypotheses in the ranges $100 \leq m_H \leq 119$ GeV and $141 \leq m_H \leq 184$ GeV are excluded at the 95% confidence level [22].

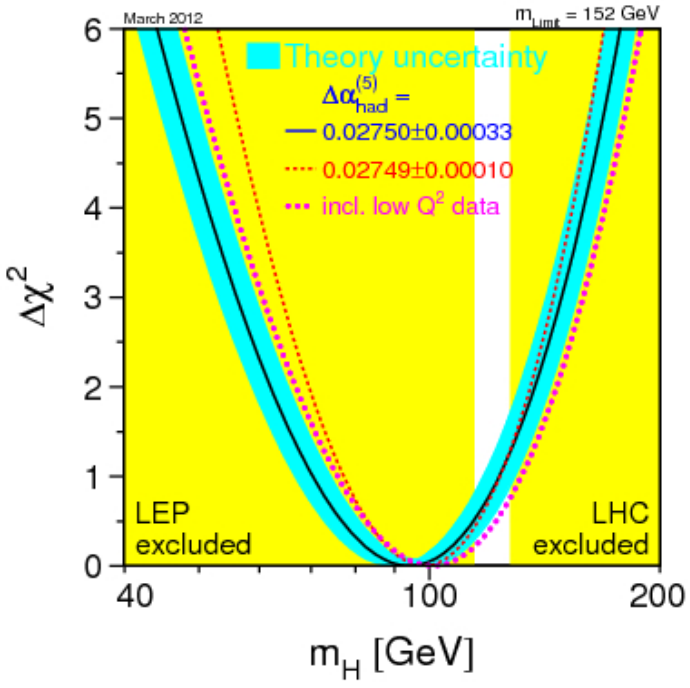


**Figure 2.2.:** Delta chi-squared from global fit to combined data from CDF, D0, SLD and the LEP Collaborations as a function of $m_H$ [23]. The solid line is the nominal fit with theoretical uncertainties indicated in blue while the dashed lines indicate alternative theoretical prescriptions. The yellow bands indicate the regions excluded at the 95% confidence level from direct searches for the SM Higgs boson conducted at LEP and the LHC before March 2012.

### Precision Measurements

Collision data taken at the Tevatron are combined with precision measurements of electroweak observables performed at LEP and by the SLD Collaboration based at SLAC to constrain the mass of the Higgs boson. Figure 2.2 shows the relative chi-squared from a fit to these data as a function of $m_H$. The minimum of the curve is at 94 GeV with an experimental uncertainty of +29 and -24 GeV. The theoretical uncertainty is indicated by the blue band. The yellow bands indicate the excluded regions in $m_H$ provided by direct searches for the SM Higgs boson conducted at LEP and the LHC by March 2012.

### 2.2.2. Higgs Boson Production and Decay at the LHC

At the LHC, protons are accelerated to higher energies than previously available at the Tevatron. The increased centre-of-mass energy enhances the rate at which Higgs boson production occurs and improves the sensitivity to higher masses. The four main mechanisms by which a Higgs boson can be produced are shown at leading order in Figure 2.3. The dominant production mechanism is through gluon-gluon fusion ($ggH$).

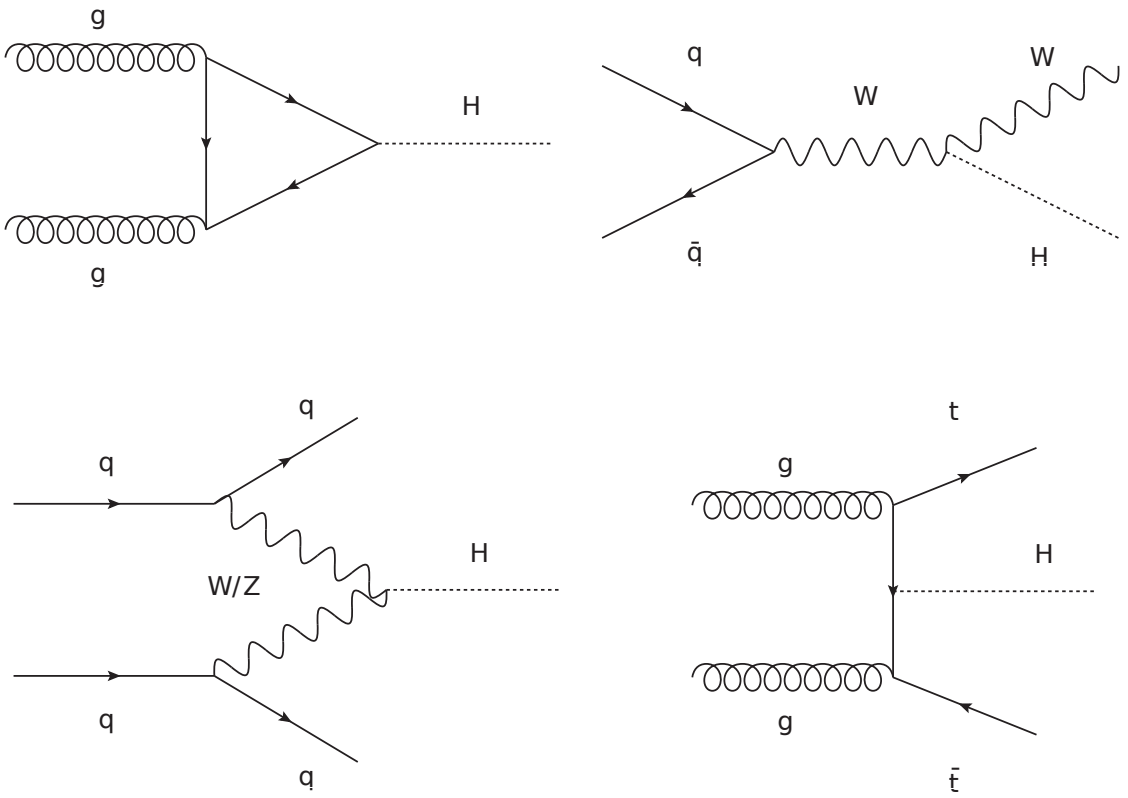


**Figure 2.3.:** Dominant SM Higgs boson production mechanisms: Gluon-gluon fusion (top left), vector-boson fusion (bottom left), associated production with vector boson (top right) and top anti-top quark pair (bottom right).

As the gluons are massless particles, the gluons couple to the Higgs boson via a quark-loop. The three other production mechanisms which dominate Higgs boson production are vector boson fusion ($qqH$) and production in association with a $W$ or $Z$ boson ($VH$) or top anti-top quark pair ($ttH$). Although these modes are at least an order of magnitude smaller in cross-section than gluon-gluon fusion, their specific topologies can be exploited experimentally to enhance the signal over background processes (see Chapter 6). Figure 2.4 shows the production cross-sections and their theoretical errors for the four main production modes of the SM Higgs boson in p-p collisions at the

LHC [24, 25]. The Higgs boson is an unstable particle so will be observable directly at

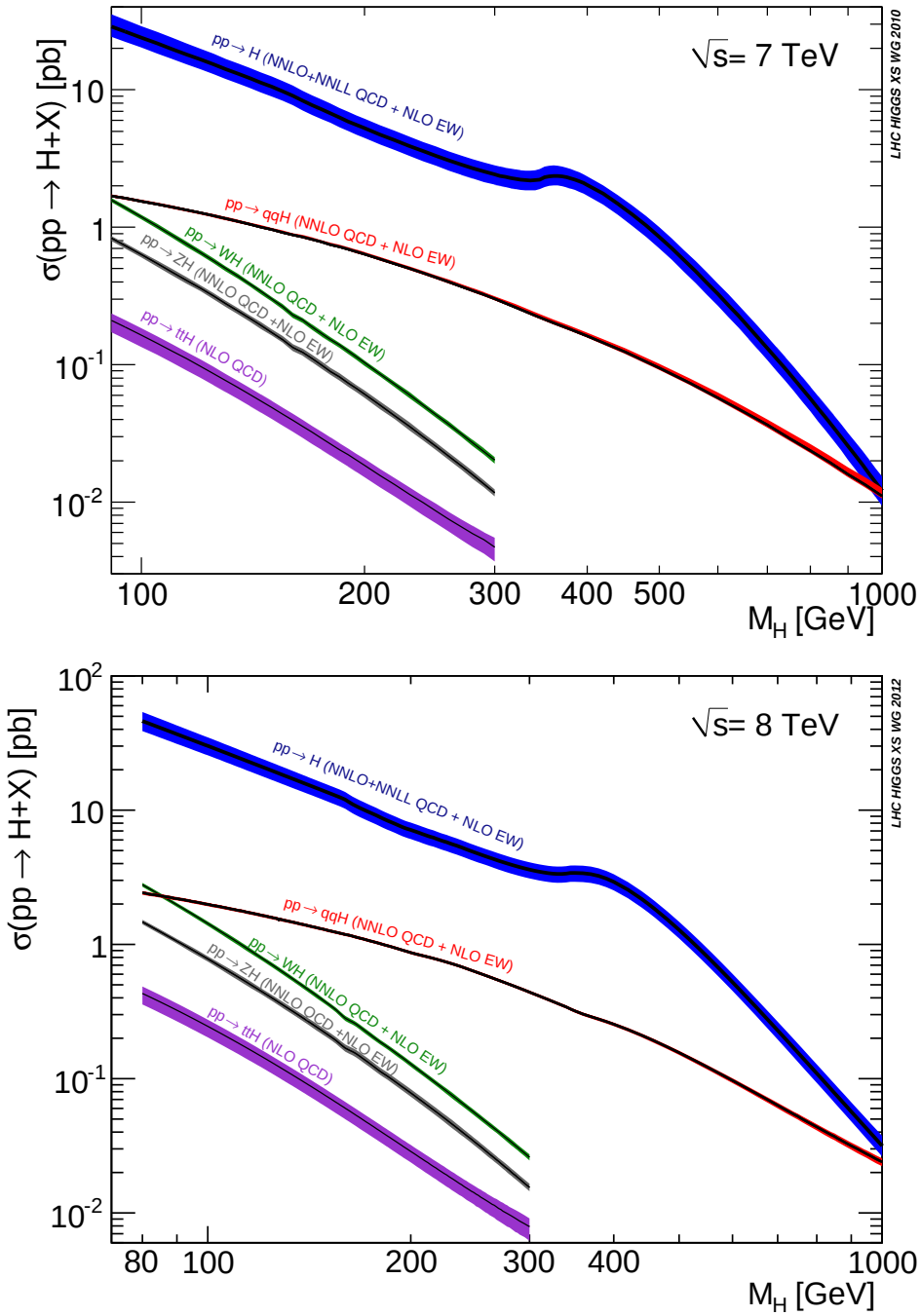


**Figure 2.4.:** SM Higgs boson production cross-sections at $\sqrt{s} = 7$ TeV (top) and 8 TeV (bottom) of the four main production mechanisms, $pp \to H + X$, along with their theoretical uncertainties as a function of $m_H$ [24, 25]. The coloured bands indicate the theoretical uncertainties.

the LHC only through its decay products. The relative decay rates (branching ratios)

to different SM particles vary as a function of the Higgs boson mass. At low mass, $m_H < 135$ GeV, Higgs boson decay to a $b$ anti-$b$ quark pair dominates. In proton-proton collisions, pairs of $b$-quarks are produced frequently making the background levels too high to compete with for an experimental search. For higher masses, $m_H > 180$ GeV, the Higgs boson is heavy enough to facilitate production of real $W$ and $Z$ bosons which dominate its decay. As the gluon and photon are massless, they do not directly couple to the Higgs boson hence these decays are mediated by virtual loops of massive particles. The branching ratios of the Higgs boson to SM particles are shown as a function of $m_H$ in Figure 2.5 (left).

For small $m_H$, the natural width of the SM Higgs boson, $\Gamma_H$, is several orders of magnitude smaller than its mass. Figure 2.5 (right) shows the value of the SM Higgs boson total width as a function of its mass. This means that for decays in which the products are fully reconstructible in particle detectors, the width of the invariant mass spectrum of the decay products will depend almost entirely on the experimental resolution. In particular the ATLAS and CMS detectors provide excellent energy and momentum resolution for electrons, muons and photons. Despite having lower branching ratios, the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ channels are therefore of particular importance for direct detection of the SM Higgs boson at the LHC.

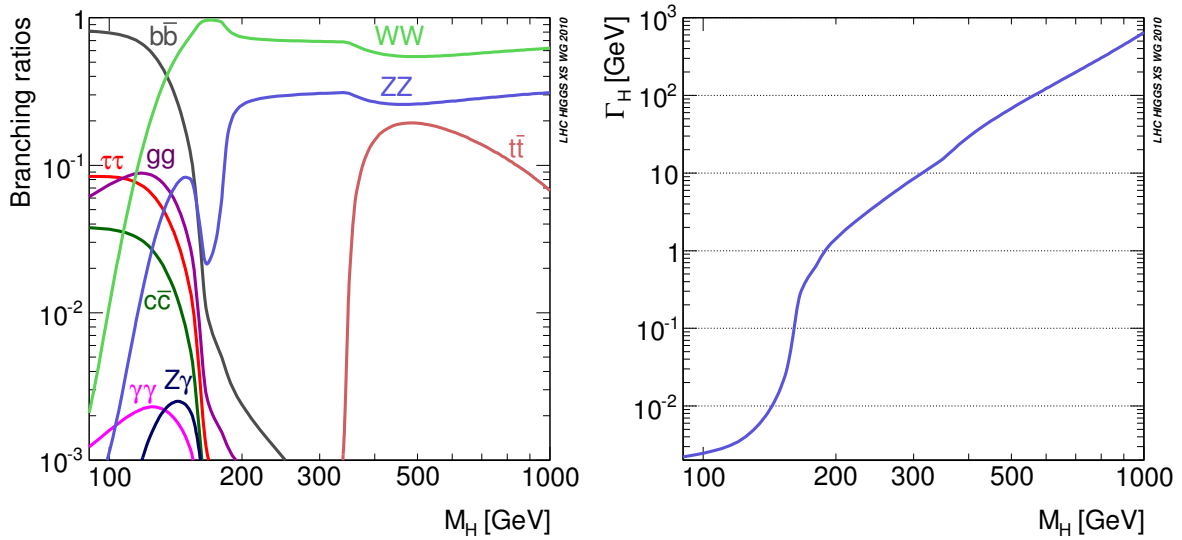


**Figure 2.5.:** Left: SM Higgs boson production branching ratios for the dominant decays as a function of $m_H$. Right: SM Higgs boson total width, $\Gamma_H$, as a function of $m_H$ [24].

# Chapter 3.

# The LHC and the CMS Detector

One of the many physics goals of the LHC is the establishment of the mechanism by which the fundamental fermions and bosons acquire mass in the SM. The discovery of the new particle announced by the ATLAS and CMS Collaborations in July 2012, if found to be the Higgs boson, will have provided a major step towards this goal. This of course could not have been achieved without the use of the particle detectors designed by those Collaborations. This chapter is intended to serve as an introduction to the CMS detector, being the experiment the author worked on. Section 3.2 provides an overview of the components of the CMS detector, paying attention in particular to those which are the most relevant for the search for the Higgs boson in the two photon decay mode. Section 3.3 will cover work performed by the author, as service to the CMS Collaboration, on improving the jet resolution in the GCT component of the L1 trigger. A set of calibrations (derived by the author) to be used online during CMS data-taking are described.

## 3.1. The LHC

The Large Hadron Collider (LHC) at CERN is the only collider, currently in operation, designed to study physics at the TeV scale. The collider is an octagonal ring, 27 km in circumference, hosted in the former LEP tunnel in France/Switzerland. Both proton-proton (pp) and heavy ion (PbPb) collisions are studied as part of the LHC physics programme with the former used for direct searches for new physics. Proton beams are formed inside the Proton Synchrotron (PS) from bunches of protons 50 ns apart with an energy of 26 GeV. The protons are then accelerated in the Super

Proton Synchrotron(SPS) to 450 GeV before being injected into the LHC. Around 1200 superconducting dipole magnets maintain two beams of protons accelerating around the ring in opposite directions before being collided at one of the sites of the four major experiments; ALICE [26], ATLAS [27], CMS [28] and LHCb [29]. Figure 3.1 is a cartoon of the accelerator indicating the sites of the four experiments.

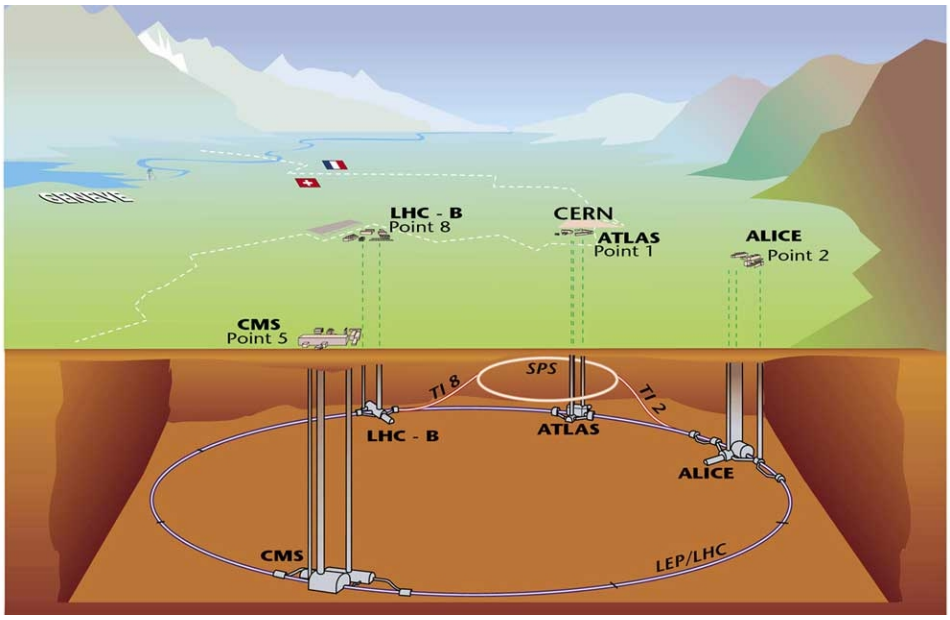


**Figure 3.1.:** LHC accelerator ring. The relative locations of the four main experiments are indicated along with their points of access to the beam.

The first major physics run began in May 2010 with a centre of mass energy $\sqrt{s} = 7$ TeV and continued until November providing a dataset of $44pb^{-1}$. The LHC resumed collisions in April 2011 delivering a further $6fb^{-1}$ by the end of October. The centre of mass energy was increased to $\sqrt{s} = 8$ TeV for the 2012 pp collision run, improving the sensitivity of searches for new physics. A total of $6fb^{-1}$ of 8 TeV data were taken by July 2012 which were combined with earlier data resulting in the discovery of the new boson reported by ATLAS and CMS Collaborations at the ICHEP conference that year.

## 3.2. The CMS Detector

The Compact Muon Solenoid (CMS) detector is one of two general purpose detectors at the LHC designed to search for new physics. Among the wide range of physics programmes at CMS, the search for the SM Higgs boson has a high priority. The decay rates of the SM Higgs boson in different channels vary dramatically as a function of its

mass ($m_H$). A key feature of the experiment's design was, therefore, the necessity to maintain a high sensitivity to the SM Higgs for a wide range of masses in as many decay channels as possible. To achieve this, several detector components are layered around the beam axis to reconstruct many types of particle produced at the interaction point. Each component consists of a cylindrical barrel section and two endcaps to provide an almost hermetic coverage of the outgoing particle flux.

The tracker, providing measurements of the momentum of charged particles and the location of primary and secondary vertices (from decays of heavy flavour mesons), is the first layer of detection. This is followed by the electromagnetic calorimeter which is used to measure energy deposited in electromagnetic showers from particles such as electrons and photons. The hadronic calorimeter (HCAL) complements this by providing energy measurements of sprays of hadrons, known as jets, which deposit energy through nuclear interactions. The HCAL is a sampling calorimeter in that the active material (plastic scintillators) are sandwiched between dense absorbing material to increase the depth of the calorimeter to around 11 radiation lengths. The addition of the forward calorimeter (HF) extends the HCAL coverage in the forward regions. The tracker and calorimeters are situated within a 4T axial magnetic field provided by the superconducting magnet surrounding them. The magnetic flux return is implemented within the muon detector systems which lie outside the superconducting coil and form the outermost detection layers. Muons deposit very little energy throughout the detector and can carry on into the surrounding cavern. The barrel muon system is constructed from layers of drift-tubes (DT) interleaved with resistive plate chambers. The combination of the two provides high resolution timing and hit positions which are used to determine the trajectory of muons both from p-p collisions and cosmic sources for calibration. For the endcaps, the DTs are replaced with cathode strip chambers as the higher flux of particles along the beam line requires the use of components which can operate under high levels of radiation.

CMS uses a right-handed Cartesian coordinate system with the origin at the interaction point and the $z$-axis pointing along the beam axis. The $x$-axis points towards the centre of the LHC ring and the $y$-axis points vertically upwards. The azimuthal angle, $\phi \, \epsilon \, [-\pi, \pi]$, is defined with respect to the $x$-axis in the transverse $(x - y)$ plane. The polar angle $\theta$ is measured from the $z$-axis. Commonly, the direction of an outgoing particle is defined by $\phi$ and its pseudo-rapidity $\eta$ defined as

$$\eta = -\log \tan \left( \frac{\theta}{2} \right). \tag{3.1}$$

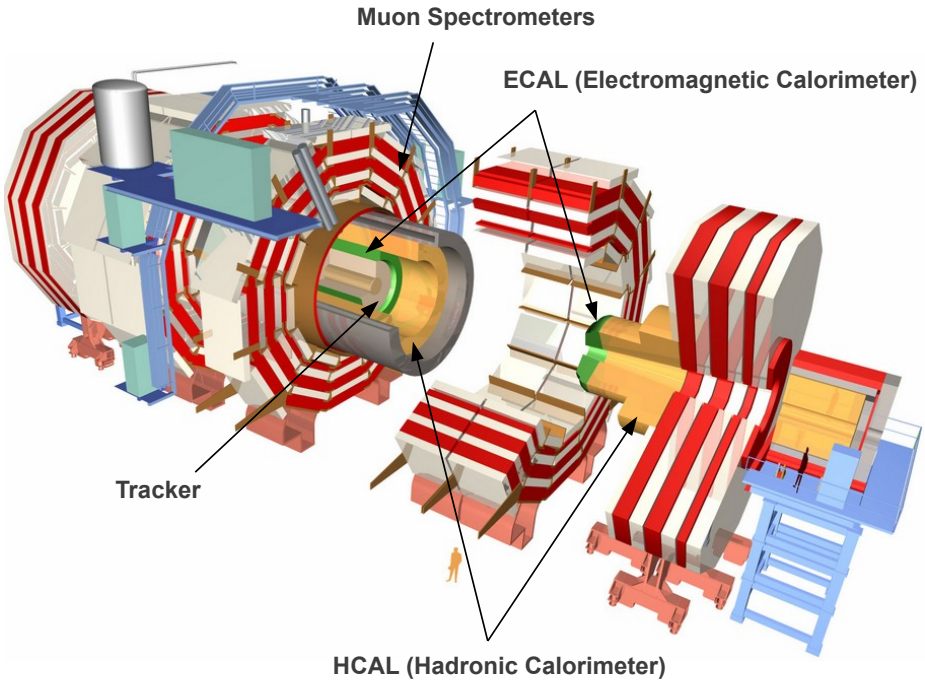


**Figure 3.2.:** Diagram of the CMS Detector. The arrows indicate the main detector elements. The figure has been altered from its original source [30].

As hard collisions produce high momentum particles travelling perpendicular to the beam line, particles are often characterised by the magnitude of the projection of their momenta onto the transverse plane, $p_T = \sqrt{p_x^2 + p_y^2}$. Similarly, the transverse energy is defined as $E_T = E \sin\theta$. Figure 3.2 shows the geometry of the CMS detector and its major components.

### 3.2.1. Tracker

The CMS tracker is designed to reconstruct charged particle tracks which are ubiquitous in high energy p-p collisions. The tracker provides precise measurements of observables such as the momentum of charged particles and the location of the vertex at which they are produced. In addition to the high level of granularity required to make such measurements, the high rate of interaction at LHC requires a fast response from the tracking elements. The tracker is formed of a pixel detector component encased by layers of silicon strip detectors. The pixel detector is the closest tracking element to the interaction point. It is a composite of 66 million individual silicon pixels, $100\mu\text{m} \times 150\mu\text{m}$ in size, forming three cylindrical layers around the beam line and two forward disks.

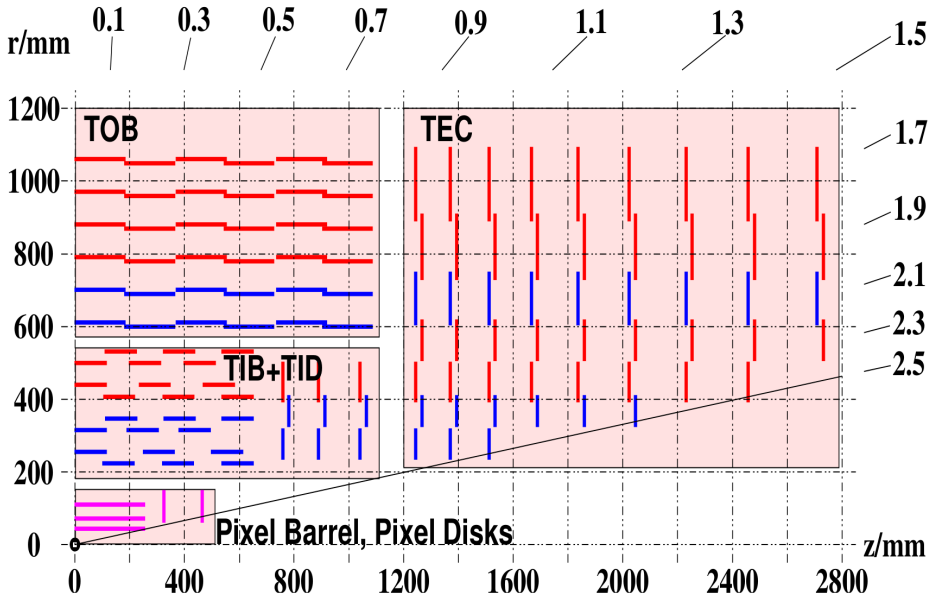


**Figure 3.3.:** Cross-section of the pixel and silicon strip detector components of the CMS tracker [32].

The resolution of the pixel detector is around 10 $\mu$m in the $\hat{r}$ and $\hat{\phi}$ direction and 17 $\mu$m in $\hat{z}$ [31]. Outside the pixel detector, ten cylindrical layers of silicon strip detectors (TIB/TOB) and twelve discs (TID/TEC) extend the tracking system out to a radius of 120cm from the beam line. The tracker geometry, as shown in Figure 3.3, covers a pseudo-rapidity range $|\eta| < 2.5$.

By making multiple precise measurements throughout the tracker system, the trajectories (tracks) of charged particles can be reconstructed. Tracks are associated to a common point of origin (primary vertex) by grouping those which are separated by less than 1cm in the $z$ coordinate of the point of closest approach to the beam line. The vertex resolution is dependent both on the number of tracks associated to the vertex and their average transverse momenta ($\bar{p}_T$). The resolution was measured in early data from 2010 by splitting tracks associated to a vertex randomly into two groups with equal kinematic distributions. The difference between the vertex locations calculated from the two groups was used to provide an estimate of the resolution [33]. Figure 3.4 shows the resolution in $z$ as a function of the track multiplicity measured in data and simulation. The simulation provides a good description of both the trend with number of associated tracks and the improvement in resolution with $\bar{p_T}$ in the data.

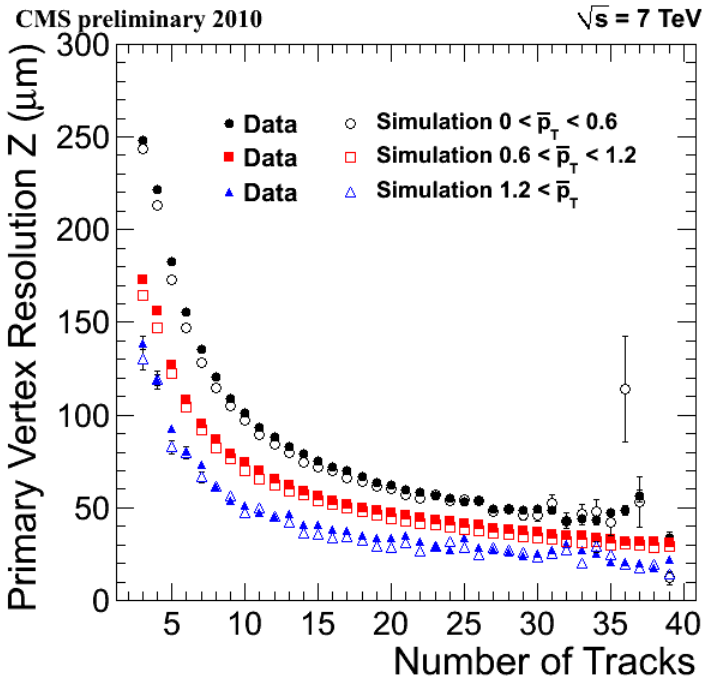


**Figure 3.4.:** Resolution of vertex $z$-position as a function of the number of tracks associated to the vertex measured in simulation and 2010 data [33]. The resolution is given for three different average track momenta.

### 3.2.2. Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is used to reconstruct energy deposits in electromagnetic showers from particles such as electrons and photons. It is constructed from high density lead tungstate ($PbWO_4$) crystals which form a barrel section (EB) and two endcaps (EE) outside the tracker. Two lead plates in front of a fine grained silicon strip detector are situated just before the endcaps forming the ECAL pre-shower (PS). Photons travelling at high $\eta$ will convert in the lead and the resulting electron-positron pair will produce tracks which can be used to pinpoint the position of the incoming photon. The additional information obtained using the two layers can be used to distinguish prompt photons from those produced in neutral pion decays.

The ECAL is designed to cover a pseudo-rapidity range of $|\eta| < 3$. The crystals are arranged to form modules which surround the beam line in a non-projective geometry: the gaps between crystal modules are offset by 3°, beyond the interaction point, with respect to the trajectories of particles produced at the centre of the interaction point. Electrons and photons deposit most of their energy within the crystals as the depth of the crystals is equivalent to 25.8 radiation lengths [34]. Electromagnetic showers

produced by the interaction of electrons and photons in the ECAL crystals produce scintillation light which is collected to measure the energy of the particle. The scintillation output of the crystals is, however, low and temperature dependent ($\sim 2.1\%$/K at the ECAL operating temperature of 291 K). Avalanche photo-diodes (APDs) and vacuum photo-triodes (VPTs) are used to collect the scintillation light and amplify the signal in the calorimeter barrel and endcaps respectively. These technologies are chosen to withstand the high magnetic field inside CMS. For the endcaps, VPTs are used as they are less sensitive to the high radiation conditions in the forward regions. Around 4.5 photo-electrons per MeV are produced in both APDs and VPTs.

The energy resolution of the ECAL can be parametrised as the combination of three uncorrelated sources as given in equation 3.2. The parameters $a$, $b$ and $c$ are the stochastic, noise and constant contributions respectively. These constants have been derived from test-beam data [35]. The stochastic term ($a = 2.83 \pm 0.3\%$) is very low for lead tungstate since the shower can be mostly contained within the crystals. As the noise term ($b$ =124 MeV) is determined by the electronics, it is mostly the constant term ($c = 0.26 \pm 0.04\%$) which will limit the ECAL accuracy at high energies. Maintaining a high resolution over the long term running of the LHC will allow accurate reconstruction of high energy photons, such as those produced by $H \to \gamma\gamma$ decays.

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{a}{\sqrt{E}}\right)^2 + \left(\frac{b}{E}\right)^2 + c^2 \tag{3.2}$$

### Electron and Photon Reconstruction

Electron and photon candidates are formed by clustering deposits of energy caused by electromagnetic showers in the ECAL. For unconverted photons, these clusters will likely be well localised in $\eta$ and $\phi$ around the incident photon. However, for photons which convert in the material in front of the calorimeter, the resulting electron-positron pair will deposit energy across several regions of the calorimeter. In the presence of the axial magnetic field, electrons radiate bremsstrahlung photons causing deposits which are spread over a wide range in $\phi$ while being fairly narrow in $\eta$. This characteristic is exploited by the "Hybrid" clustering algorithm used to reconstruct high energy electrons and photons in the ECAL barrel [36]. The values of the particular thresholds used for seeding clusters were tuned providing an efficiency for electrons with $p_T > 7$ GeV greater than 99% [37]. Figure 3.5 is an illustration of the Hybrid clustering algorithm. The algorithm proceeds as follows;

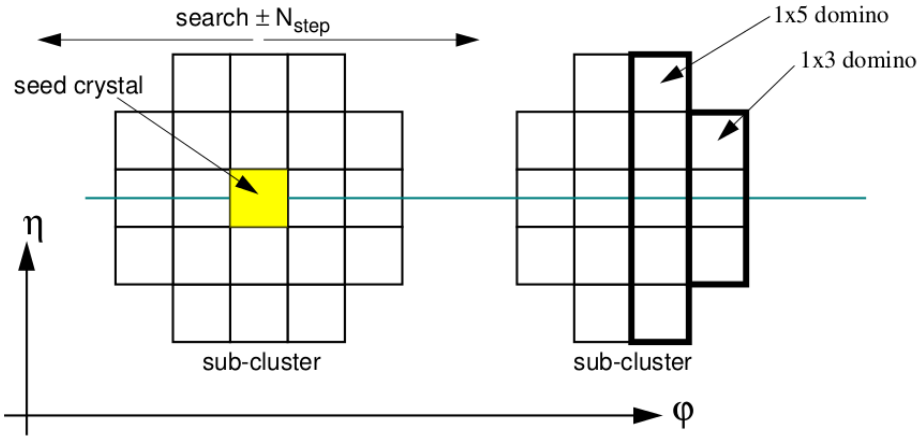


**Figure 3.5.:** Sub-cluster construction of the Hybrid algorithm used to reconstruct photons and electrons in the ECAL barrel.

- Step 1: A seed crystal is determined to be a single crystal in the barrel with the highest $E_T$ satisfying $E_T > 1$ GeV.
- Step 2: $1 \times 3$ ($\phi \times \eta$) crystal dominoes are formed with their central crystal aligned with the seed crystal in $\eta$. If the energy contained in the $1 \times 3$ domino is larger than 1 GeV, the domino is extended by two crystals in $\eta$. A maximum of 10 dominoes are added in each direction in $\phi$ starting from the seed crystal forming a sub-cluster.
- Step 3: Dominoes containing less than 100 MeV are removed and the remaining dominoes are grouped into sub-clusters providing each seeding domino for a sub-cluster contains more than 350 MeV. The final group of sub-clusters form a supercluster for the electromagnetic object.

In the ECAL endcaps, superclusters are built using the "Multi5×5" algorithm which connects overlapping 5×5 grids of crystals whose positions lie within 0.3 radians in $\phi$ [38]. Additional information is used from the PS to enhance the energy reconstruction in the endcaps.

Superclusters are associated to electron candidates where a compatible track can be reconstructed from compatible hits in the tracker using a Gaussian sum filter algorithm [39]. This provides an additional measure of the electron's momentum which is used to improve the resolution of the electron energy. Apart from this, the reconstruction of photons and electrons is identical which is an important feature allowing for data driven calibrations and validations of photons using electrons such as those described in Chapter 4.

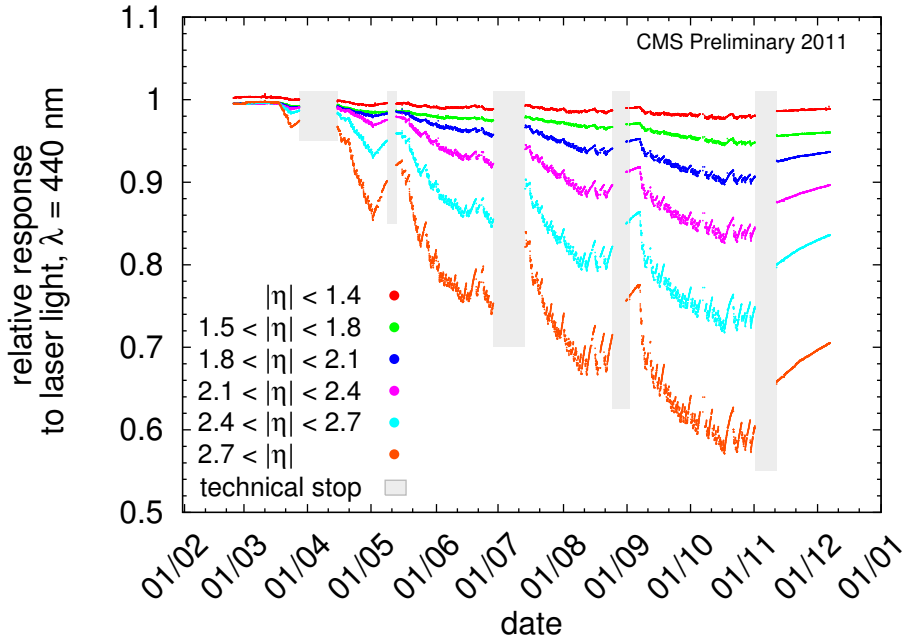


**Figure 3.6.:** Relative ECAL crystal response to blue laser light (440 nm) in bins of pseudo-rapidity, for the 2011 data taking period. The grey bands indicate periods during which there was no beam.

### Laser Calibration

ECAL crystals suffer from loss of optical transmission when irradiated through the formation of crystal-lattice defects which absorb some of the scintillation light. Annealing acts to recover from damage due to radiation which results in an equilibrium optical transmission which is dose-dependent [34]. At the LHC, the dose varies during each run. This requires that the time varying optical transmission of the ECAL crystals be monitored to assess the impact on energy measurements. The crystal transparency is monitored by comparing the relative transmission in blue laser light (440 nm), which is close to the scintillation emission peak, to infra-red (796 nm), which is far from the peak and relatively unaffected by the radiation damage. Figure 3.6 shows the relative response to the blue laser of the monitoring system averaged over all the crystals in bins of $|\eta|$ throughout the 2011 data taking runs [40]. The time dependence of the response is stronger at higher values of $|\eta|$ due to the larger flux of particles along the beam axis.

The response of the crystals measured using the laser monitoring system is used to calibrate the energy reconstruction of the ECAL. These calibrations are validated in $W \to e\nu$ data events by comparing the electron energy ($E$) as measured by the ECAL to the momentum ($p$) of the electron measured in the tracker [40]. Figure 3.7 shows the

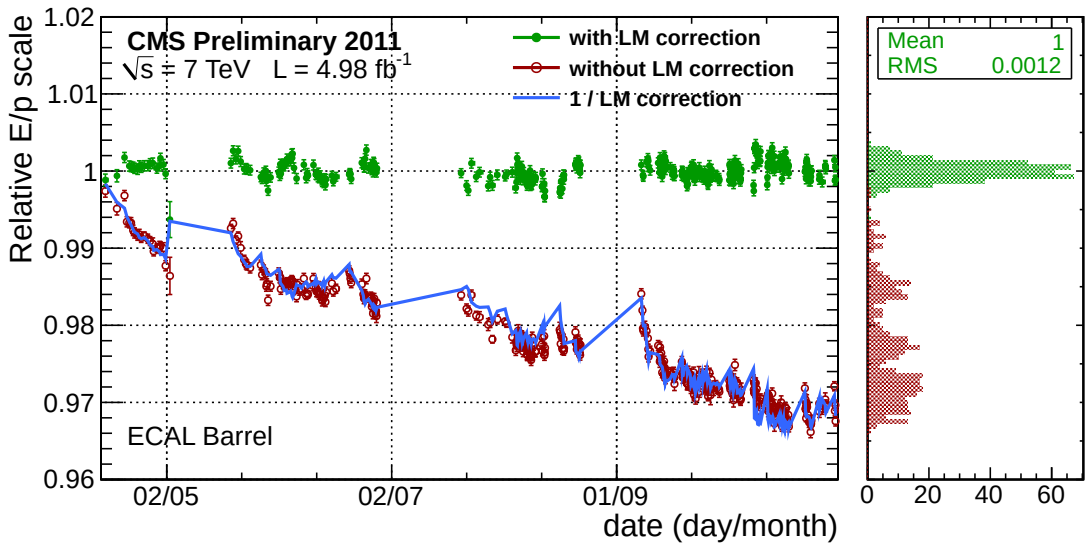


**Figure 3.7.:** Ratio $E/p$ in electrons reconstructed in the ECAL Barrel from $W \to e\nu$ events in 2011 data as a function of time before and after applying transparency corrections from the laser monitoring (LM) system. The blue line indicates the correction applied per point averaged over all crystals used in the electron energy measurement.

relative variation in the ratio $E/p$ as a function of time throughout 2011. A stable energy scale is achieved through application of the laser calibrations.

### 3.2.3. Shower-shape and Isolation

In addition to providing a measurement of the energy of incoming electromagnetic particles, the ECAL's fine granularity provides additional information which can be used to characterise the supercluster and distinguish prompt electrons and photons from fakes. The shape of the electromagnetic shower can be described by the ratio of the energy contained in the central $3 \times 3$ cluster surrounding the seed crystal to the total energy of the supercluster ($r_9$). Superclusters associated with real unconverted photons will typically have a larger value of $r_9$ than those which are in reality due to narrow $\pi^0$ decays. Another common variable used for identification is the energy weighted crystal width of the sub-cluster used to seed the supercluster $\sigma_{i\eta i\eta}$,

$$\sigma_{i\eta i\eta} = \frac{\sum_i w_i(\eta_i - \eta_{sc})^2 \Delta\eta_{xtal}^2}{\sum_i w_i}, \tag{3.3}$$

where $\eta_{sc}$ is the pseudo-rapidity of the seed crystal, $\Delta\eta_{xtal}$ is the crystal width and $w_i$ is the crystals weight determined as $w_i = \max\{0, 4.7 + \log(E_i/E_{tot})\}$. Prompt photons will tend to have a more localised cluster leading to lower values of $\sigma_{i\eta i\eta}$. The distributions of $r_9$ and $\sigma_{i\eta i\eta}$ are shown for a simulated sample of superclusters identified as photons from real and fake sources in Figure 3.8. The two distinct peaks in the $\sigma_{i\eta i\eta}$ distribution are due to the different superclustering algorithms used in the barrel and endcaps.

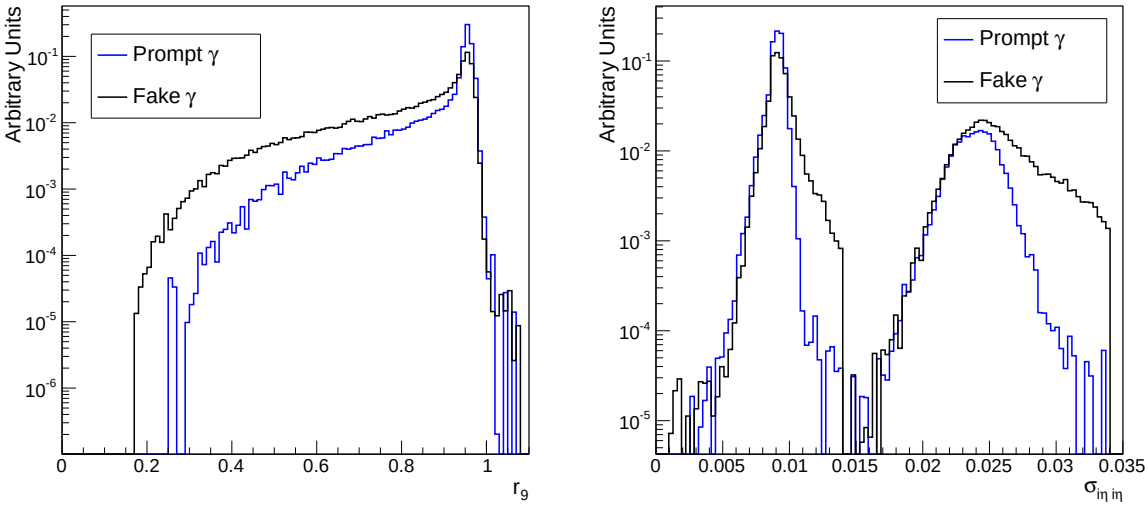


**Figure 3.8.:** Shower shape variable $r_9$ (left) and $\sigma_{i\eta i\eta}$ (right) distributions for superclusters associated with simulated real and fake photons. The real photon is taken from simulated $H \to \gamma\gamma$ events while the fake photon is taken from a $\gamma + jet$ sample where the photon candidate is matched to a generated quark leg. In the right hand plot, two distributions can be distinguished. The narrower is from photons in the barrel and the wider from photons in the endcaps.

Hard interaction processes tend to produce electromagnetic particles which are well isolated in the detector. A cone is defined around the candidate with radius $\Delta R$ defined as

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} \tag{3.4}$$

where $\Delta\phi$ and $\Delta\eta$ are defined as the $\phi$ and $\eta$ co-ordinates relative to the weighted centre of the supercluster.

The sum of $E_T$ for each crystal inside the cone, after removing those associated to the supercluster itself, quantifies the isolation of the electron or photon candidate. Similar isolation variables are defined for the HCAL and tracker by summing over the $E_T$ and $p_T$ of HCAL deposits and tracks respectively.

## 3.3. Level-1 Trigger

In order to cope with the high collision rate, a two-tier trigger system is implemented at CMS. The trigger is able to use limited information from each event to decide whether or not to record the event. This allows a large reduction in the rate of data-taking while maintaining a high efficiency to select events producing interesting physics objects. The first level, the Level-1 (L1) trigger, uses custom-built electronics in order to reduce the output rate from 40 MHz to 100 kHz [41]. Events which satisfy some relatively loose set of criteria are passed to the second level, the high-level trigger (HLT), where more sophisticated algorithms, much closer to those used in the offline reconstruction, are used to decide whether or not to store an event [42].

The L1 calorimeter trigger is able to use coarse measurements of the energy deposited in the ECAL and HCAL to form candidate physics objects such as electrons, photons, tau leptons decaying hadronically and hadronic jets. With the exception of electrons and photons, all of the L1 algorithms run in the Global Calorimeter Trigger (GCT). The following section is a description of a set of calibrations for the GCT designed to improve the resolution of the L1 jets.

### 3.3.1. Jet Energy Calibration

The response of the hadronic calorimeter varies considerably across its barrel, endcap and forward sections. The energies of jets are corrected offline to account for these effects; however, if left uncalibrated at L1, this can lead to inefficiencies in the trigger system. The response is measured in QCD Monte Carlo (MC) simulation by comparing the $p_T$ of L1 jet candidates to generated jets. The generated jets are reconstructed using an anti-$k_T$ jet finding algorithm [43]. L1 jets are matched to generator jets by determining the minimum separation, $\Delta R$, between each generator jet and any L1 jet candidate and requiring it be less than 0.7. This is much looser than typical matching requirements applied offline due to the coarser spatial resolution of the L1 jets. The generator and the closest of these L1 jets is defined as a matched pair and the response is calculated as $E_T^{L1}/E_T^{Gen}$ for that pair. Figure 3.9 shows the response as a function of the pseudo-rapidity of the generated jet $|\eta^{Gen}|$.

The response is measured in 11 $|\eta|$ bins which correspond to the 11 GCT regions. Corrections for each region are derived as a function of $E_T^{L1}$ by determining the average

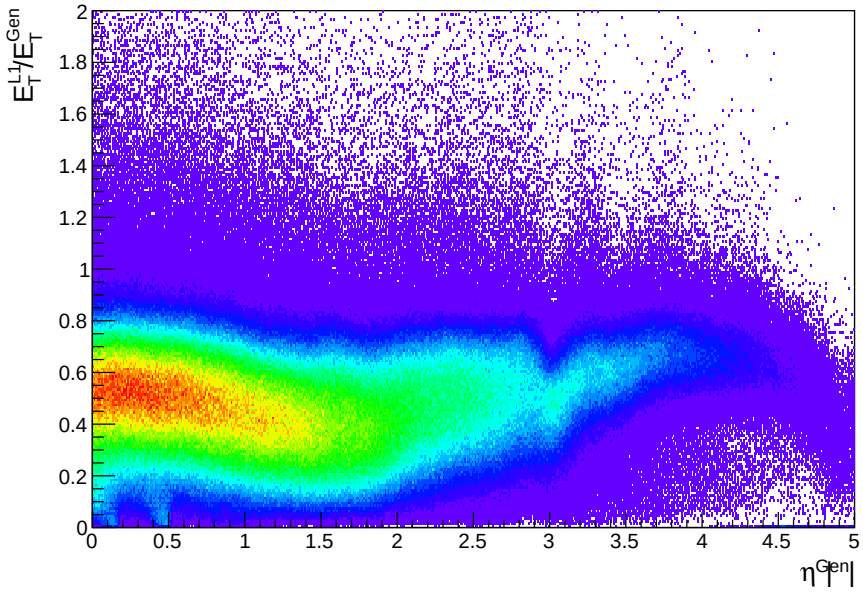


**Figure 3.9.:** Response measured from matched generator-L1 jet pairs in MC as a function of the generator jet pseudo-rapidity $|\eta^{Gen}|$.

response, $\langle E_T^{L1}/E_T^{Gen}\rangle$, and $\langle E_T^{L1}\rangle$ in 4 GeV bins of $E_T^{Gen}$ between 14 GeV and 200 GeV. Below 14 GeV, the resolution in $E_T^{L1}$ restricts a proper measurement of the response while above 200 GeV, the response approaches unity. The average response is taken from the mean of a Gaussian fit to the distribution of $E_T^{L1}/E_T^{Gen}$ while $< E_T^{L1} >$ is taken as the mean average of the $E_T^{L1}$ distribution. For low values of $E_T^{Gen}$, the response becomes very non-Gaussian due to the limited resolution of the L1 trigger, so in this case, the average response is taken as the mean of the $E_T^{L1}/E_T^{Gen}$ distribution. The response is inverted to provide a corrective scale factor in each region as a function of $E_T^{L1}$. This is then parameterised by performing a chi-squared fit of the functional form given in Equation 3.5.

$$\langle E_T^{L1}/E_T^{Gen}\rangle^{-1} = E_T^{L1} \cdot \left(p_0 + \frac{p_1}{(\log E_T^{L1})^2 + p_2} + p_3 \exp(-p_4(\log E_T^{L1} - p_5)^2)\right) \quad (3.5)$$

The functional form chosen provides a good description of the shape at low $E_T^{L1}$ in the high $|\eta|$ regions and is the same as that used for offline jet calibration at CMS [44]. The parameterisation provides a multiplicative correction to be applied to L1 jets online.

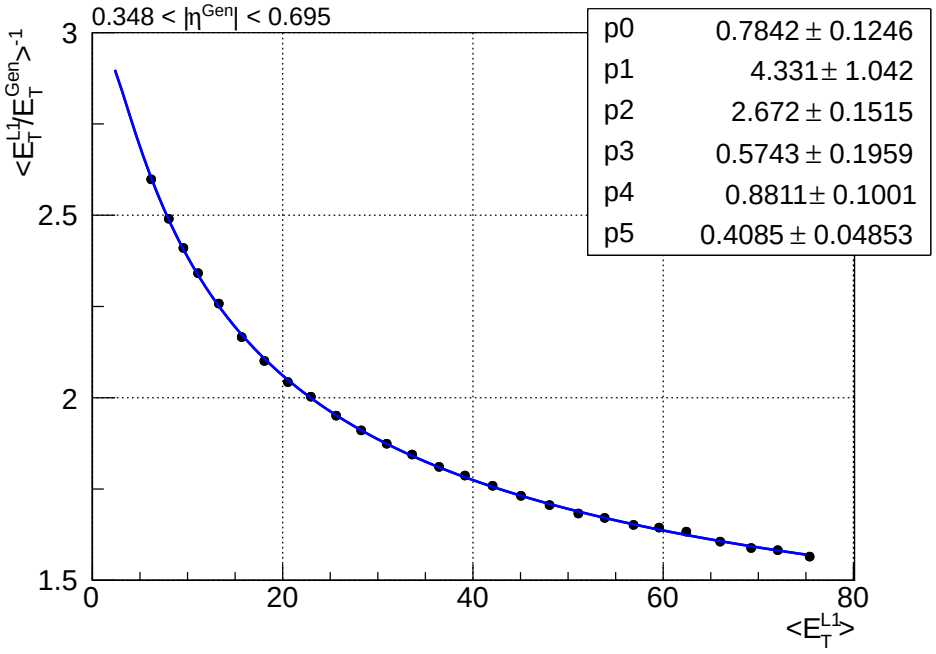


**Figure 3.10.:** Correction function for the $0.348 < |\eta^{Gen}| < 0.695$. The points represent the average quantities as measured in MC. The blue line is a parametric fit to the points using a chi-squared minimisation. The error bars, estimated from the number of MC events, are too small to be visible in this plot.

Figure 3.10 is an example of the fit in the $0.348 < |\eta^{Gen}| < 0.695$ bin. The full set of fits in each of the 11 $E_T^{Gen}$ bins can be found in Appendix A.1.

### 3.3.2. Calibration Performance

The calibrations derived were applied using the GCT emulation software to the same MC sample used to derive them to provide a closure test of their performance. The response is shown in Figure 3.11 as a function of $E_T^{L1}$ and $\eta^{Gen}$. The points in each figure are calculated from a Gaussian fit to the distribution of $E_T^{L1}/E_T^{Gen}$ in bins of $E_T^{L1}$ and $\eta^{Gen}$ respectively. The results show that the procedure closes to a precision of between 5% and 10%. The improvement in L1 jet resolution expected from MC is demonstrated in Figure 3.12. The resolution, calculated by fitting a Gaussian to the distribution of the difference in $E_T$ measured at L1 to that of the generator jet in bins of $E_T^{L1}$ (see Appendix A.2), for L1 jets is shown before and after applying the corrections.

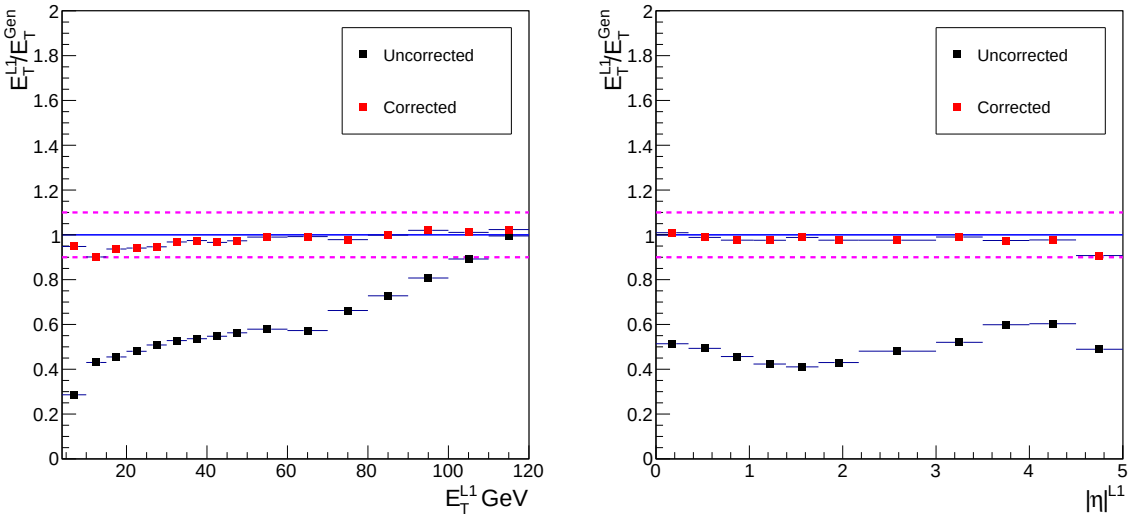


**Figure 3.11.:** Closure tests performed in MC as a function of $E_T^{L1}$ (left) and $\eta^{Gen}$ (right). The test shows that after applying the corrections, the response is within 10% (dashed lines) of unity. The error bars are too small to be visible in these plots.

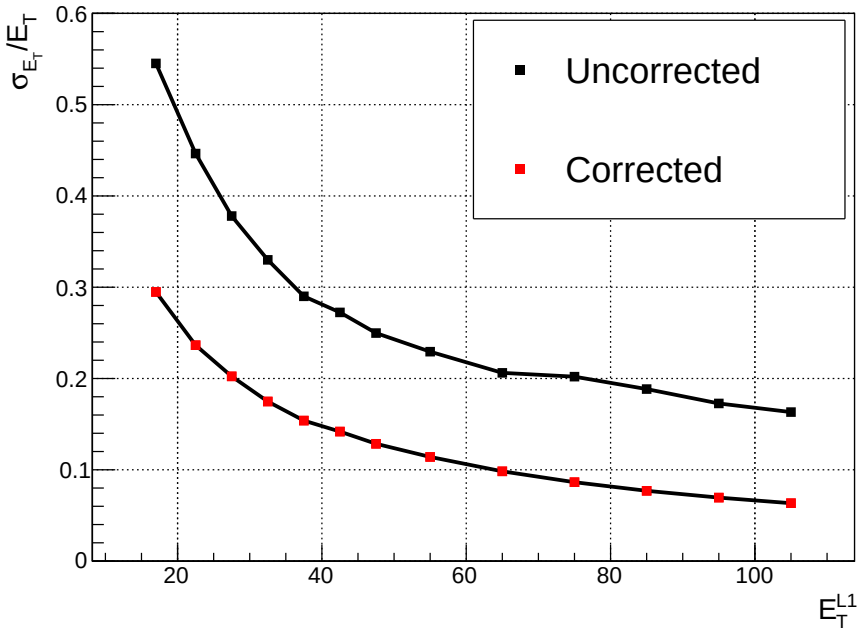


**Figure 3.12.:** Jet energy resolution at L1 as a function of $E_T^{L1}$ before and after application of the derived calibrations. The error bars are too small to be visible in these plots.

### Performance in Data

The corrections derived in MC were applied to data online during and since run 2011B. The resolution as a function of $E_T^{L1}$ was measured using events in data from that run period using the following method. First, the fraction of L1 jets above some threshold in $E_T^{L1}$ is determined as a function of the fully reconstructed jet $E_T$ in data. This is then fit with an error function of the form,

$$\mathrm{erf}(x) \propto \int_0^{\frac{x-\mu}{\sigma}} e^{-t^2} dt, \tag{3.6}$$

to provide a measure of the average energy in the calorimeters for jets which just pass the threshold at L1 ($\mu$) and the resolution of those jets ($\sigma$). As the full energy reconstruction for jets at CMS is much more accurate than the value reconstructed at the L1 trigger, the effects of the jet energy resolution after applying the full jet reconstruction are negligible. This is repeated for different thresholds in $E_T^{L1}$. Figure 3.13 shows the resolution as a function of $E_T$, where the value of $E_T$ is taken from the $\mu$ parameter of each fit. The uncertainties on each point represent the statistical uncertainty from the error function fits. The points are fit with the parameterisation given in Equation 3.2 to extract the parameters which describe the resolution of the calorimeter. The energy resolution of the L1 jets is improved after applying these calibrations in the GCT [45].

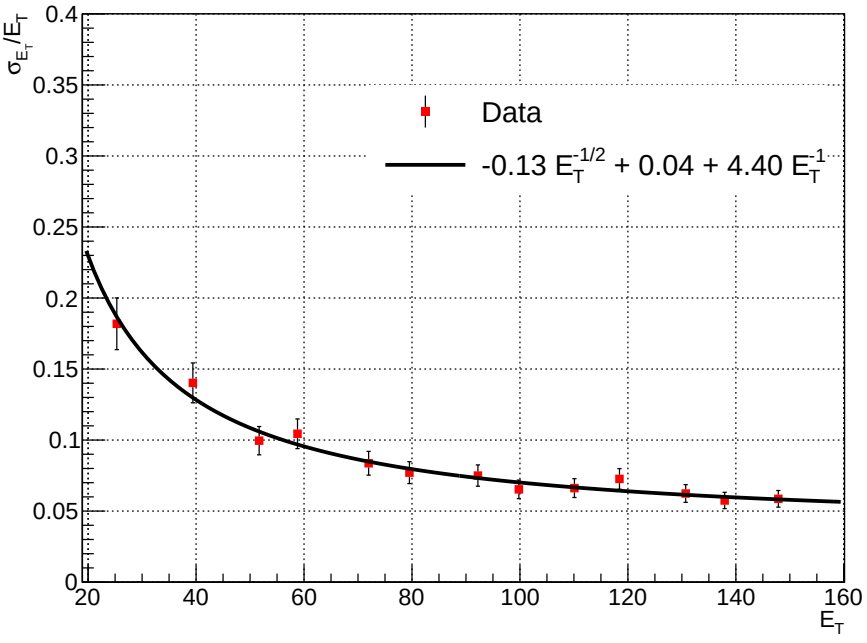


**Figure 3.13.:** Energy resolution, $\sigma_E$, of L1 jets as a function of transverse energy deposited in the calorimeter, $E_T$. The coefficients of the functional form shown are the result of a fit to the points.

# Chapter 4.

# Higgs Boson Decay to Two Photons

The two photon channel is one of the most promising decay modes in the search for the SM Higgs boson at the LHC. Despite having a relatively small branching ratio, the decay $H \to \gamma\gamma$ provides a very clean final state in which the kinematics of the Higgs boson are fully reconstructed. It is, therefore, one of the most sensitive channels at low $m_H$. The dominant source of background is from real, prompt diphoton events from QCD processes, $pp \to \gamma\gamma$ (prompt-prompt). In addition, there are contributions from $pp \to \gamma + jet$ (prompt-fake) and $pp \to jet + jet$ (fake-fake) in which jets are misidentified as photons. In high-priority analyses such as the search for the Higgs boson, cross-checking of analysis techniques is essential to ensure a robust result. At CMS, it is the policy to design at least two different techniques for the same analysis ideally implemented in independent code bases. In particular, as the signal rate in the $H \to \gamma\gamma$ decay mode is small compared to the background rates, the sensitivity of the search is heavily influenced by how well the backgrounds are understood. For these reasons, two data-driven techniques for extracting the signal were developed. The first (method A) uses a fully parametric description of the background which is fitted to the data. The type of parameterisation and the number of parameters are chosen so that the systematic uncertainty introduced by potentially choosing the wrong functional form is less than $\frac{1}{5}$ of the statistical uncertainty in the fit [46]. The second (method B) uses a binned model constructed from sidebands in the diphoton invariant mass spectrum. Method B was developed by the author and serves as an independent cross-check of method A. In particular, this is achieved by allowing direct inclusion of the background modelling systematic uncertainties in the signal extraction, thereby building confidence in the understanding of the background. This chapter describes a search for a Higgs boson decaying to two photons which was performed on the full 2011 dataset corresponding to $5.1 fb^{-1}$of proton-proton collisions recorded at CMS at a center of mass energy of

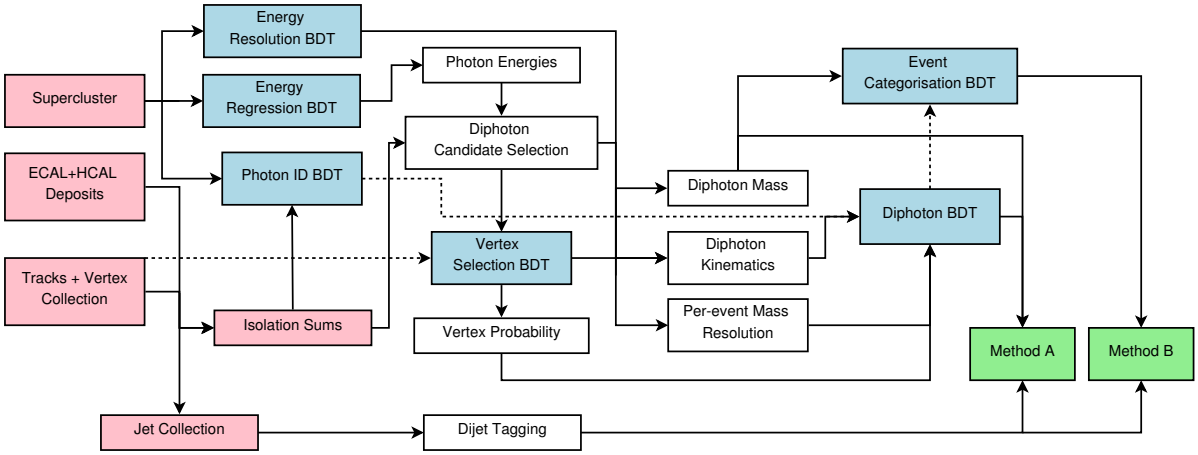


**Figure 4.1.:** Flow chart of the $H \to \gamma\gamma$ analysis performed on the 2011 dataset. The blue boxes indicate stages which involve the use of a boosted decision tree (BDT). The red boxes indicate inputs from the common CMS reconstruction and are not detailed in this chapter. The two methods for signal extraction, labelled A and B, are indicated by the green boxes.

7 TeV. In Sections 4.1 to 4.3, the reconstruction and selection of events used for this analysis is detailed. Section 4.4 then describes method B for modelling the background for the purposes of extracting a potential signal and statistical interpretations of the data. Results from the 2011 dataset are given in Section 5.2 and the update for the ICHEP conference in July 2012 at which the discovery announcement of the new boson was made is given in 5.2.1.

## 4.1. Data Samples

The dataset used for this analysis is taken from a combination of the 2011A (March-August) and 2011B (September-October) proton-proton collision runs. The selection of the dataset is based around dedicated diphoton triggers which select events satisfying one of two sets of criteria. The first set requires two HLT photon candidates, one with $p_T > 26$ GeV and the other with $p_T > 18$ GeV, which are well isolated in the calorimeter [47]. The second has a lower threshold on the first photon, $p_T > 22$ GeV but requires that both photons have localised showers in the ECAL ($r_9 > 0.8$ in 2011A and $r_9 > 0.9$ in 2011B). Additionally, the invariant mass of the two trigger objects is required to be greater than 60 (70) GeV in the 2011A(B) datasets. Events which would pass the full offline selection but fail to trigger at the HLT lead to an inefficiency, reducing the number of signal events with respect to that expected from an integrated luminosity of $5.1 fb^{-1}$.

However, the thresholds applied offline are chosen to be much tighter than those of the trigger; the trigger efficiency is >99% with respect to the analysis selection [47].

Signal Monte Carlo (MC) events are generated for a Higgs boson decaying to two photons via the four main production processes, gluon-gluon fusion ($ggH$), vector boson fusion ($qqH$) and associated $W/Z$ ($VH$) and $t\bar{t}$ ($ttH$) production. The gluon-gluon fusion and vector boson fusion processes were generated with `POWHEG` [48] with next-to leading order (NLO) contributions whereas the two associated production processes were generated to leading order (LO) only. The $p_T$ spectrum of the Higgs boson ($p_T^H$) from gluon-gluon fusion was calculated at next-to-next-to leading plus next-to leading log resummed order (NNLO+NLL) using the `HqT` program [49]. The production cross-sections and branching ratios are taken from the LHC Cross-section Working Group [50].

Background processes were generated at LO using `POWHEG` interfaced with `PYTHIA` [51]. The QCD dijet and $\gamma + jet$ samples were filtered by requiring the generated photons, electrons and neutral mesons with $p_T > 15$ GeV have at most one charged particle in a cone, $\Delta R < 0.2$, to increase the production efficiency with respect to the tracker isolation requirements of the full selection. The background samples considered for this analysis are summarized in Table 4.1. A full simulation of the CMS detector is provided in `GEANT4` which is used for all signal and background MC samples [52]. The MC includes a simulation of additional interaction vertices expected in data from pileup. This is where multiple pp collisions occur within a single bunch crossing which results in several primary vertices being reconstructed. The distribution in the number of reconstructed vertices in MC is corrected to match that observed in data.

## 4.2. Object Reconstruction and Identification

The reconstruction of all objects used for this analysis, in both data and MC, is based on the standard CMS reconstruction software `CMSSW_4_2_X` [53]. Additional sensitivity can be gained by refining the object selection and reconstruction specifically to the search for $H \to \gamma\gamma$.

| **Process** | | **Cross-section** $(pb)$ | **Luminosity** $(pb^{-1})$ |
|---|---|---|---|
| DiPhotonJets | | 154.7 | 7400 |
| DiPhoton Box | $\hat{p_T}$ $25-250$ | 12.37 | 41900 |
| QCD Dijet | $\hat{p_T}$ $30-40$ | 10870 | 560 |
| | $\hat{p_T}$ $40-\infty$ | 43571 | 920 |
| Gamma+Jet | $\hat{p_T}$ $20-\infty$ | 493.44 | 2400 |
| DrellYan+Jets to $ll$ | $\hat{p_T}$ $50-\infty$ | 2475 | 14000 |

**Table 4.1.:** Background MC used throughout the analysis with production cross-sections and corresponding equivalent integrated luminosity. The prompt-prompt ($\gamma\gamma$) sample comprises events from the DiphotonJets and Diphoton Box samples. Both the QCD dijet and Gamma+Jet contain prompt-fake ($\gamma j$) events. The samples are filtered to avoid double counting of this background. Fake-fake ($jj$) events are taken from the QCD Dijet sample.

### 4.2.1. Boosted Decision Trees

Multivariate analysis (MVA) techniques are often used in high energy physics analyses which suffer from the presence of large background rates. These techniques provide greater distinction between signal and background than traditional selection techniques. This is due to the fact that the full information from each event can be utilised by describing the event with a set of measurable quantities. The distributions of these quantities and correlations between them can be exploited to provide the maximum separation between signal and background. Of the many MVA techniques available, Boosted Decision Trees (BDT) are commonly used in high energy physics as they are robust against the inclusion of non-informative variables; the performance will not degrade due to the addition of information which offers little or no separation power [54].

BDTs are used in the $H \to \gamma\gamma$ search described in this chapter in order to achieve the maximum sensitivity to a potential signal. For most of the BDTs used in this analysis, the BDT is used to provide separation between signal and background processes. The first step in producing a BDT of this type is to identify a list of "input" variables which describe the event objects and provide discrimination between signal and background. These can be variables such as the $r_9$ and $\sigma_{i\eta i\eta}$ of the photon superclusters or kinematic variables of the event such as the transverse momentum of the diphoton system ($p_T^{\gamma\gamma}$). Additional variables, such as the positions of the photons within the detector, are included so that the BDT can account for correlations between the input variables due to detector effects.

A decision tree (DT) is then trained on a MC simulation sample of signal and background events. The DT splits the events into two sub-samples by applying a series of cuts on the input variables. The purity, $p$, of each sub-sample is defined as the fraction of the events which are signal,

$$p = \frac{N_s}{N_s + N_b}, \tag{4.1}$$

where $N_s$ and $N_b$ are the sum of weights for the signal and background samples. A separation criterion is defined to decide whether or not to further sub-divide that subset. A number of criterion definitions exist, though commonly the Gini index [54], $p(1-p)$ is used. Values which are close to one indicate the sample is polluted by background, so above some configurable threshold, the sub-sample is further split and the process continues. This continues until either all sub-samples are below the threshold or the user-defined maximum number of splitting levels (tree depth) is reached. Each MC event is then assigned a value of -1 or +1 depending on whether it is or is not in a sub-sample with $p > 0.5$. The DT is trained by varying the cuts applied in order to maximise the purity in each sub-sample. Several events will be incorrectly classified by simply taking the output of the DT. This is mitigated by training a set of DTs on the same sample and modifying the contribution from each DT through a processes known as "boosting". A weight is applied simultaneously to all of the DTs by minimising the binomial log-likelihood loss function for the sample of $K$ events,

$$L(F, y) = \sum_{k=1}^{K} \ln\left(1 + e^{-2Fy}\right), \tag{4.2}$$

where $y = -1$ for a background event and $y = +1$ for a signal event. Each DT can be considered as a member of a family of functions $f(\mathbf{x}; \mathbf{b}_n)$ for a particular set of cut values $\mathbf{b} = \mathbf{b}_n$ [55]. The function $F$ is the weighted average over the individual DTs given by,

$$F(\mathbf{x}; \mathbf{a}, \mathbf{b}_n) = \sum_{n=0}^{N} a_n f(\mathbf{x}; \mathbf{b}_n);\ \mathbf{a} = (a_1, a_2 \cdots a_N). \tag{4.3}$$

Although other weighting schemes can be used, for the BDTs trained in this analysis, the scheme described, known as gradient boosting, was found to give the best performance (see Section 4.4). The resulting set of trees is known as a boosted decision tree. All of the BDTs used for this analysis were trained using the `TMVA` toolkit [54] available within the `ROOT` framework.

### 4.2.2. Supercluster Energy Correction

As the natural width of the Higgs boson is around 100 MeV at low $m_H$, the width of a reconstructed mass peak from a $H \to \gamma\gamma$ decay is driven by the experimental energy resolution of the photons. This resolution can be improved dramatically by correcting the raw energy of the supercluster at the per-photon level. These corrections are derived using a multivariate technique in which a regression BDT is trained on prompt photons in the gamma+jet MC sample using the ratio of the generated photon energy to the raw energy of the reconstructed supercluster [56]. As this ratio can vary across different regions of the detector, the input variables include both the $\eta$ and $\phi$ positions of the supercluster. In addition, several variables are included which describe the shower shape: $r_9$, the energy weighted widths in $\eta$ and $\phi$ of the supercluster, the energy weighted crystal width ($\sigma_{i\eta i\eta}$) and the ratio of hadronic energy behind the supercluster to the energy of the supercluster itself ($H/E$). In the endcaps, there is additional information available from the pre-shower measurement so the ratio of the energy in the pre-shower to the raw supercluster energy is included. Figure 4.2 shows the improvement in resolution after applying the regression corrections (Raw + Regression) compared to the standard calibrations used at CMS (Standard). In parallel, a similar set of corrections were derived by fitting an analytical expression (Standard + IC Residual) to the residual energy difference between the generated and reconstructed photon energy as a function of supercluster energy, position and $r_9$ [57]. The regression technique reduces the effective resolution of the Higgs mass peak ($\sigma_{\text{eff}}$) resolution by around 30% over using the raw supercluster energy compared to the analytic fit which improves the resolution by 15%.

An estimate of the per-photon energy resolution, $\sigma_E$, is obtained by training a second regression BDT targeting the absolute deviation between the correction estimated by the first BDT and the true correction to generator level. This second BDT is trained on an independent set of events to the first. The per-photon resolution is used to calculate an estimate of the per-event mass resolution, $\sigma_{m_{\gamma\gamma}}$, which is used during the event selection (Section 4.3).

#### Energy Scale Measured in Data

Despite correcting the energy of the photons using the regression technique, discrepancies between data and MC are still observed. This is due to additional detector effects which may not be simulated, such as the time dependence of the ECAL crystal transparency [40].

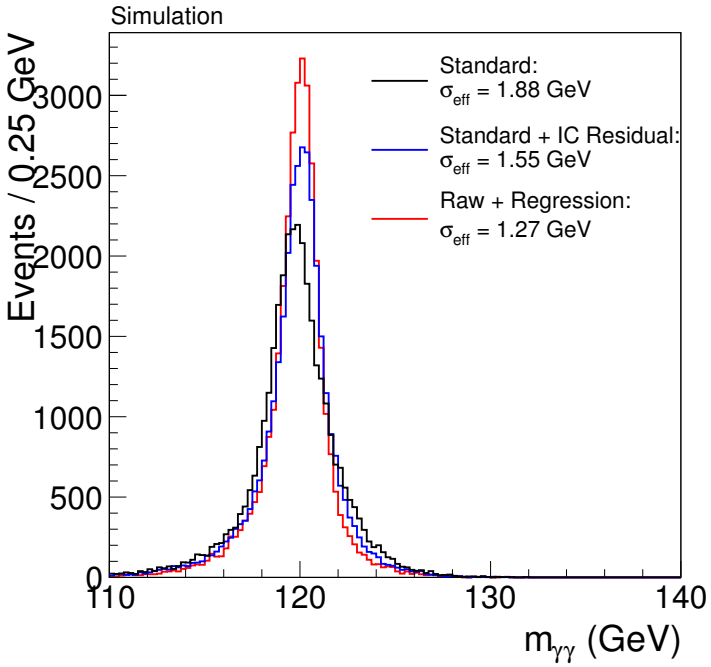


**Figure 4.2.:** Comparison of the diphoton mass peak in Higgs MC with a mass of 120 GeV using different measurements of the photon energy. The black line is from using the raw energy of the supercluster, the blue is from using the analytic fit method (Standard + IC Residual) and the red from using the regression method (Raw + Regression). The quantity $\sigma_{eff}$, the narrowest range in $m_{\gamma\gamma}$ which contains 68% of the distribution, is given for each peak [47].

Further corrections are derived based on $Z \to e^+e^-$ events which provide an invariant mass peak (with almost no background) constructed from electromagnetic objects reconstructed using a similar procedure to photons. An additional regression BDT is trained on $Z \to e^+e^-$ MC which is used to compare the supercluster energy scale in data and MC [47]. The energy scale of the superclusters is measured by matching the electron invariant mass peak in data to that in MC. This is achieved using an analytic fit to the $Z \to e^+e^-$ peak in data and MC separately. The natural peak of the $Z$ is described using a Breit-Wigner distribution whose parameters are fixed to those given by the Particle Data Group, $m_Z = 91.188$ GeV, $\Gamma_Z = 2.495$ GeV [4]. This is then convoluted with a Crystal Ball (CB) function which describes the resolution effects of the calorimeter and energy losses from bremsstrahlung before the ECAL [58]. The CB parameter $\Delta m$ is a free parameter of the fit giving the offset of the peak position from the $Z$ pole.

The values of these fitted parameters vary with the position of the supercluster ($|\eta|$). Moreover the variation in data is strongly dependent on the run during which the data were taken. The scale is extracted in six run ranges and four $|\eta|$ regions to account for this effect, providing a first set of corrections. The difference between MC and data is dependent on whether the electron showered in the material before the calorimeter or not, which is characterised by the $r_9$ of the supercluster. The data-MC difference in each $|\eta|$ region is therefore measured a second time after applying the first set of corrections to the data and obtaining the residual difference for electrons with $r_9 < 0.94$ and $r_9 > 0.94$ separately. This dependency, unlike that with $|\eta|$, is found to be constant with time. The final energy scale correction is then defined as the product of the two corrections. The relative correction,

$$1 - \Delta P = 1 - \frac{\Delta m_{data} - \Delta m_{MC}}{m_Z}, \tag{4.4}$$

is applied to the photons in data. The values for the scale in each category, $\Delta P$, are given in Tables B.2 and B.3 of Appendix B.1. The uncertainties on these measurements are primarily due to the difference in the $r_9$ distribution of electrons and photons. In addition, smaller systematics are included due to the variation of the measurements when changing the electron selection and between using the electron-trained and photon-trained regression corrections. These uncertainties are incorporated into the signal model for the purposes of signal extraction as described in Section 4.4.5.

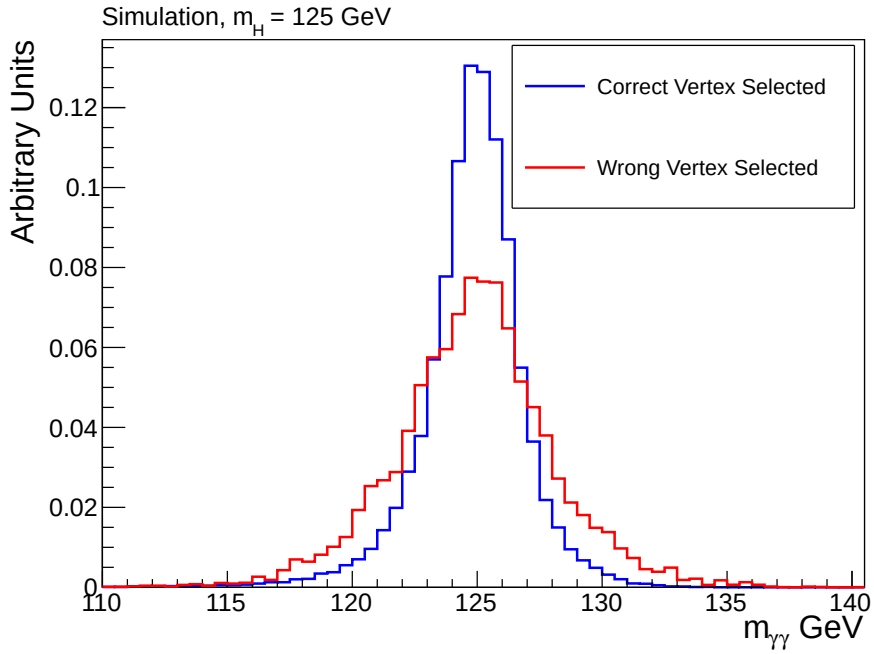


**Figure 4.3.:** Invariant mass peak in $H \to \gamma\gamma$ MC with $m_H = 125$ GeV. The blue histogram is from events in which the generated vertex is within 10mm of the vertex assigned to the diphoton pair. The red histogram is from events in which the incorrect vertex is assigned. Both distributions are normalised to unit area for ease of comparison.

### 4.2.3. Vertex Selection

The assignment of the correct vertex to the diphoton pair is an important step in the reconstruction of its invariant mass. Figure 4.3 shows the invariant mass distributions from a SM Higgs boson for events in which the vertex selected is within 10mm of the generated vertex compared to those in which an incorrect vertex is assigned. Since photons do not leave tracks, computing the angle between the two photons depends strongly on determining the vertex at which they were produced.

A BDT was trained to rank the standard collection of reconstructed vertices. The input variables are chosen to exploit the correlation between the diphoton system and the recoiling tracks. These are the $p_T$-balance,

$$-\sum_{alltracks}\left(\mathbf{p}_T^{track}\cdot\frac{\mathbf{p}_T^{\gamma\gamma}}{p_T^{\gamma\gamma}}\right), \tag{4.5}$$

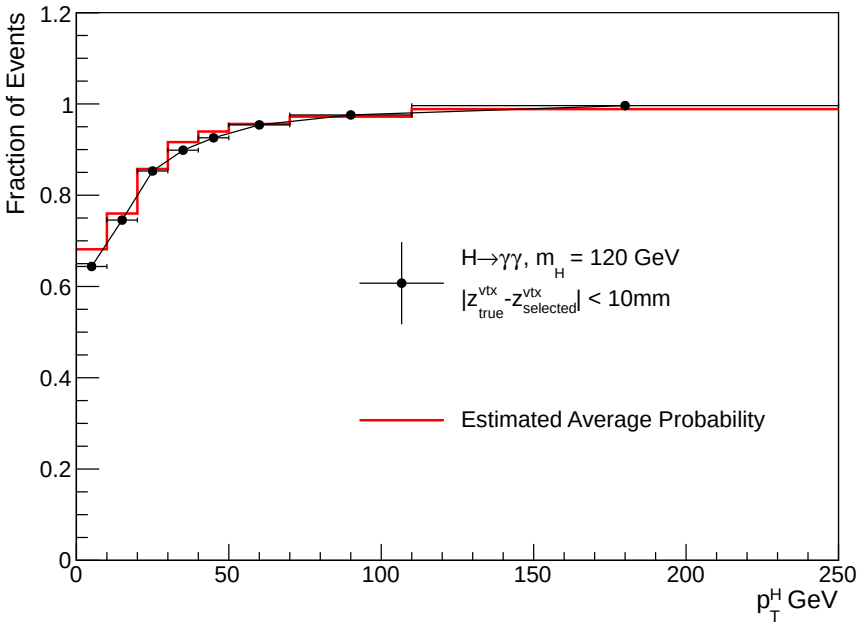


**Figure 4.4.:** Fraction of simulated gluon-gluon fusion events in which the $z$ position of the selected vertex is within 10mm of the true vertex as a function of Higgs boson $p_T$. The red histogram is the average probability to select the correct vertex in each bin estimated from the per-event BDT.

and the $p_T$-asymmetry calculated as,

$$\frac{|\sum_{alltracks} \mathbf{p}_T^{track}| - p_T^{\gamma\gamma}}{|\sum_{alltracks} \mathbf{p}_T^{track}|}. \tag{4.6}$$

In addition, the sum of the squares of the transverse momenta of all the tracks associated to a given vertex is included to preferentially select hard interactions. If at least one of the photons converts to an $e^+e^-$ pair, the difference between the position in $z$ as calculated using the electron-positron pair and that from the standard vertex, relative to the resolution in $z$, is included as an input variable. The BDT was trained on $H \to \gamma\gamma$ MC with a mass of 120 GeV. Figure 4.4 shows the fraction of events in a gluon-gluon MC sample in which the vertex with the highest BDT score is within 10mm of the true vertex as a function of $p_T^H$. The fraction of events in which this occurs in data is measured using $Z \to \mu^+\mu^-$ events as a function of the $p_T$ of the $Z$ boson [47]. Figure 4.5 shows the fraction of events for which the chosen vertex is within 10mm of the true vertex as measured in $Z \to \mu^+\mu^-$ data and MC. The BDT (MVA) selection method described is

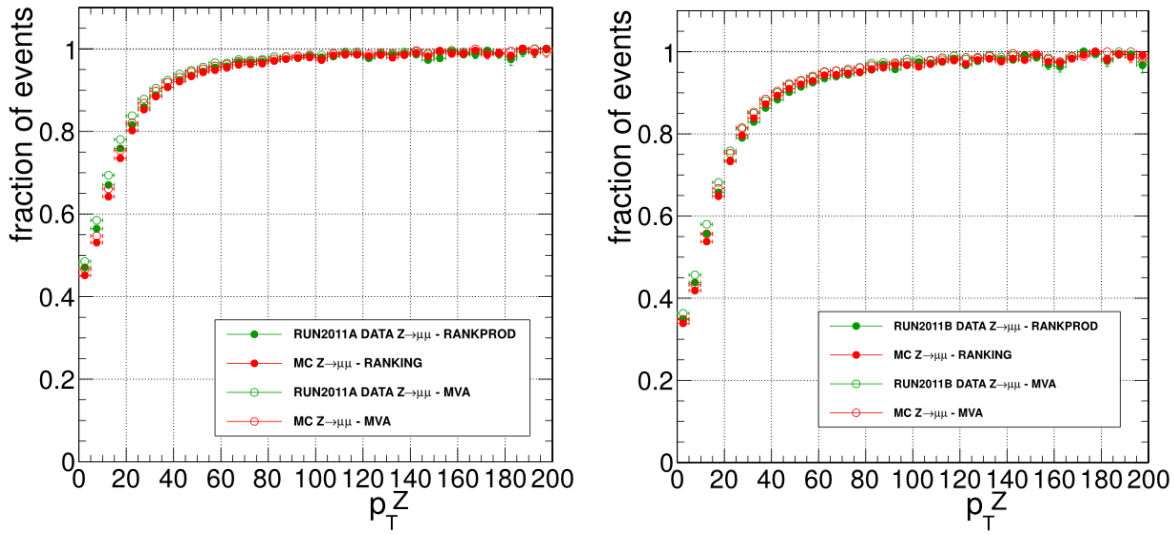


**Figure 4.5.:** Fraction of $Z \to \mu^+\mu^-$ events in which the selected vertex is with 10mm of the true vertex in Run 2011A (left) and Run 2011B (right) data and MC as a function of $p_T^Z$ [47]. The BDT selection, labelled MVA, is shown by the open circles where the ranking method, labelled RANK is shown as points.

compared with the standard CMS vertex ranking algorithm (RANK) [59] which is less efficient for low $p_T^Z$. The measurements in data are used to correct the Higgs signal MC for the purpose of signal modelling.

A second, per-event, BDT is trained using the output of the first to identify under which conditions the correct vertex is selected. The output of this BDT is then used to calculate the probability in a given event that the correct vertex is assigned. The red line in Figure 4.4 shows a comparison of the per-event vertex probability estimated from the second BDT against the fraction of the events in which the selected vertex is located within 10mm from the true vertex. The per-event vertex probability estimate provides a good model of the actual fraction of events in which the correct vertex is selected.

### 4.2.4. Photon Identification

A large portion of the fake background in the $H \to \gamma\gamma$ search is due to high momentum neutral mesons which decay to two photons where both the photons are combined into the same supercluster [46]. Information from the shower shape of the photon supercluster can be used, in addition to the energy isolation within the calorimeter, in order to distinguish these from prompt photons from the primary interaction point. A BDT was trained on MC events to combine the relevant information into a single photon identification

(ID) discriminator. The signal used for the training was taken from simulated $H \to \gamma\gamma$ events with a Higgs boson mass of 121 GeV while the background was taken from non-prompt photons in the gamma+jet (prompt-fake) sample. Before training, events are required to pass a loose pre-selection designed to avoid training where the MC is unable to properly describe the data and to match the variables used in the trigger [47]. In addition, photon candidates are removed if there is a reconstructed electron matched to the photon supercluster with no matching conversion reconstruction. This greatly reduces the contribution from electrons in $Z \to e^+e^-$ faking photons. The same pre-selection is applied to all MC and data for extracting the signal. The efficiency of the pre-selection for signal was measured in $Z \to e^+e^-$ data and MC using a tag-and-probe method [60].

The measurement is made using events which are selected using a dielectron trigger with supercluster transverse energies of at least 20 GeV. The two objects are then randomly assigned as either the "tag" or "probe" candidate. The tags are then required to pass a tight electron selection based on their isolation and supercluster shower shape. Events are then split into those in which the probe candidate passes or fails the pre-selection and the ratio of signal events, after background subtraction, in each class provides the efficiency measurement of the pre-selection. For the pre-selection efficiency measurement, the background subtraction is performed using a likelihood fit to the invariant mass of the tag-probe pair, modelling the signal with a Breit-Wigner convoluted with a Crystal Ball and the background with an exponential function. Both models are multiplied by an error function of the form,

$$\mathrm{erf}(x) \propto \int_0^{\frac{x-\mu}{\sigma}} e^{-t^2} dt, \tag{4.7}$$

with parameters $\mu$ and $\sigma$ freely floating. The measurement is performed in four probe categories depending on the $r_9$ and $|\eta|$ of its supercluster. The results are shown in Table 4.2.

The input variables are chosen to be insensitive to the kinematics of the diphoton system itself including the diphoton invariant mass. The first set of variables describe the shower shape of the supercluster: $H/E$ (the ratio of energy deposited in the HCAL behind the ECAL to that of the supercluster), $\sigma_{i\eta i\eta}$, $r_9$ and the energy weighted widths of the supercluster in $\eta$ and $\phi$ ($\sigma_\eta$, $\sigma_\phi$). The $\eta$ of the supercluster is included as the shower shape is dependent on the position within the calorimeter. The second set of input variables describe the isolation of the photon in the calorimeter and tracker scaled

| Category | Data | MC | Data/MC |
|---|---|---|---|
| EB $r_9 > 0.9$ | 0.927 ± 0.001 | 0.928 ± 0.001 | 0.999 ± 0.001 |
| EB $r_9 < 0.9$ | 0.888 ± 0.002 | 0.903 ± 0.001 | 0.984 ± 0.003 |
| EE $r_9 > 0.9$ | 0.944 ± 0.001 | 0.938 ± 0.001 | 1.006 ± 0.001 |
| EE $r_9 < 0.9$ | 0.864 ± 0.001 | 0.852 ± 0.001 | 1.014 ± 0.001 |

**Table 4.2.:** Signal efficiency for the preselection measured in data and MC using tag-and-probe in $Z \to e^+e^-$ events. The Data/MC ratios are applied as corrections to the signal MC for the purposes of signal modelling. The uncertainties listed here are statistical only.

to account for the additional expected energy density due to pileup, $\rho$ [61]. These are: the sum of the track isolation, calculated relative to the chosen vertex and the vertex giving the maximum track isolation, ECAL and HCAL isolations in cones with $\Delta R < 0.3$ minus $\rho$ times the effective area of the cone [61]; and the absolute ECAL and HCAL isolations within cones of $\Delta R < 0.3$ and $\Delta R < 0.4$ respectively. In addition, the number of reconstructed vertices in the bunch crossing is included to reduce the pileup dependence of the isolation variables.

Separate BDTs are trained for the ECAL barrel and endcaps as the shower shape and isolation variables are rather distinct between the two. A cut is made on the photon ID BDT output to select events used for the signal extraction which keeps practically all ($> 99\%$) of the signal while removing around 22% of background events. The cut is chosen to be loose as the output of the photon ID will be used as an input to the event selection (diphoton BDT) as described in Section 4.3.1.

## 4.3. Event Selection

In addition to passing the pre-selection, the two photons are required to pass mass-dependent transverse momenta cuts, $p_T/m_{\gamma\gamma} > 1/3,\ 1/4$ for the leading and sub-leading photon respectively. Where more than one diphoton pair in an event satisfies these criteria, the pair which has the largest sum of photon transverse momenta is selected as the Higgs boson candidate. The final selection of diphoton candidates used for the signal extraction is based on using as much information in the event as possible to distinguish likely signal candidate events from the background. Although the photon ID BDT is successful at rejecting fake backgrounds, a large portion of the background is due to real

prompt diphotons from QCD processes. In order to distinguish these from a Higgs signal, the specific kinematics and topology of the event are exploited.

### 4.3.1. Diphoton BDT

A BDT was trained to utilise the kinematics of the selected diphoton pair to discriminate prompt photons from QCD background from those produced by the decay $H \to \gamma\gamma$. The BDT was trained using the prompt-prompt, prompt-fake and fake-fake samples for background and Higgs MC with a mass of 123 GeV for signal. As the mass of the Higgs boson is unknown, the search is performed under different mass hypotheses. In order to allow for the application of the same selection to the data under any mass hypothesis, the input variables to the BDT are chosen to be mass-independent. In addition, this allows for a fully data-driven estimation of the background shape as described in Section 4.4.4. The input variables which describe the kinematics are: the relative transverse momenta of the leading and sub-leading photons ($p_T^1/m_{\gamma\gamma},\ p_T^2/m_{\gamma\gamma}$ respectively), their pseudo-rapidities, $\eta^1,\ \eta^2$ and the cosine of the angle between the two photons in the transverse plane $cos(\Delta\phi) = cos(\phi^1 - \phi^2)$. In addition, information regarding the quality of the objects, the two photons and the selected vertex, is included in the form of the output of the photon ID and the vertex probability BDTs. The per-photon resolution estimate, $\sigma_E$ is combined for each photon to produce a per-event mass resolution estimate under the assumption that the correct vertex is selected, $\sigma_{m_{\gamma\gamma}}(\text{right} - \text{vertex})$,

$$\sigma_{m_{\gamma\gamma}}(\text{right} - \text{vtx}) = \frac{m_{\gamma\gamma}}{2}\sqrt{\left(\frac{\sigma_E^1}{E^1}\right)^2 + \left(\frac{\sigma_E^2}{E^2}\right)^2} \tag{4.8}$$

where $E^1$ and $E^2$ are the energies of the two photons.

Since the correct vertex is not always selected, the mass resolution assuming the incorrect vertex is chosen is calculated using the average length of the region in which the two proton beams collide in data, $\sigma_Z = 5.8$cm. In this case, the distance between the selected and true vertex will be distributed as a Gaussian with width $\sqrt{2}\sigma_Z$. The contribution to the resolution, $\sigma_{m_{\gamma\gamma}}^{vtx}$, can be calculated analytically given the positions of the two photons. The mass resolution estimator under the assumption that the incorrect vertex is chosen is given by the sum in quadrature of $\sigma_{m_{\gamma\gamma}}^{vtx}$ with the mass resolution assuming the correct vertex is chosen. Both estimators for the mass resolution relative to the invariant mass, $\sigma_{m_{\gamma\gamma}}/m_{\gamma\gamma}$ right/wrong-vtx, are included as inputs to the diphoton

BDT. Figures 4.6 and 4.7 show the input variables from the final set of selected diphoton candidates in data and MC. The expectation in each plot from a SM Higgs boson with a mass of 125 GeV, scaled by 10, is shown in red. The invariant mass distribution in data and MC for events passing the full selection is given in Figure 4.9. After the application of the full selection, the total background contains around 76% prompt diphoton events. Figure 4.8 shows the diphoton BDT distribution in data and MC. The final events used for the signal extraction are selected as those with a diphoton BDT output greater than 0.05. This cut was chosen following an optimization study to minimize the expected exclusion limit in the absence of signal. Events below this cut value were found to provide negligible improvement in the expected limit [47].

### Diphoton BDT Validation with $Z \to e^+e^-$ Data

By using a BDT for the full event selection, subtle correlations between the input variables are accounted for which improve the separation between the signal and background. Unlike the background model, the signal model is derived from corrected MC. It is important therefore to ensure that the BDT will respond in the same way in data as for the signal MC used for the signal extraction. The MC can be validated using $Z \to e^+e^-$ data-MC comparisons by inverting the electron veto and treating the electrons as though they were photons. This is done by using the supercluster associated to the electron for the electron's energy measurement and ignoring the track information. In this way, the reconstruction of the electrons is the same as that of the photons allowing for validation of the BDT's response to real photons from a resonant decay. Figure 4.10 shows the diphoton BDT distribution in $Z \to e^+e^-$ MC and data after applying the full selection using this technique. The discrepancies between MC and data are well covered by the systematic uncertainties on the photon ID BDT output and $\sigma_E$ which are included in the final signal model.

Both the photon ID and regression BDT rely on a detailed simulation of electromagnetic showering in MC to correctly describe the data. Due to imperfections of this simulation, systematic uncertainties are included in the signal model to cover the residual difference observed between MC and data for high $p_T$ photons. These uncertainties are validated using $Z \to e^+e^-$ data in the same way as the diphoton BDT. Figures 4.11 and 4.12 show the distributions of the per photon energy resolution estimator $\sigma_E$ relative to the photon energy and the output of the photon ID BDT in $Z \to e^+e^-$ MC and data treating the electrons as photons. The red lines show the $\pm 1\sigma$ error envelope

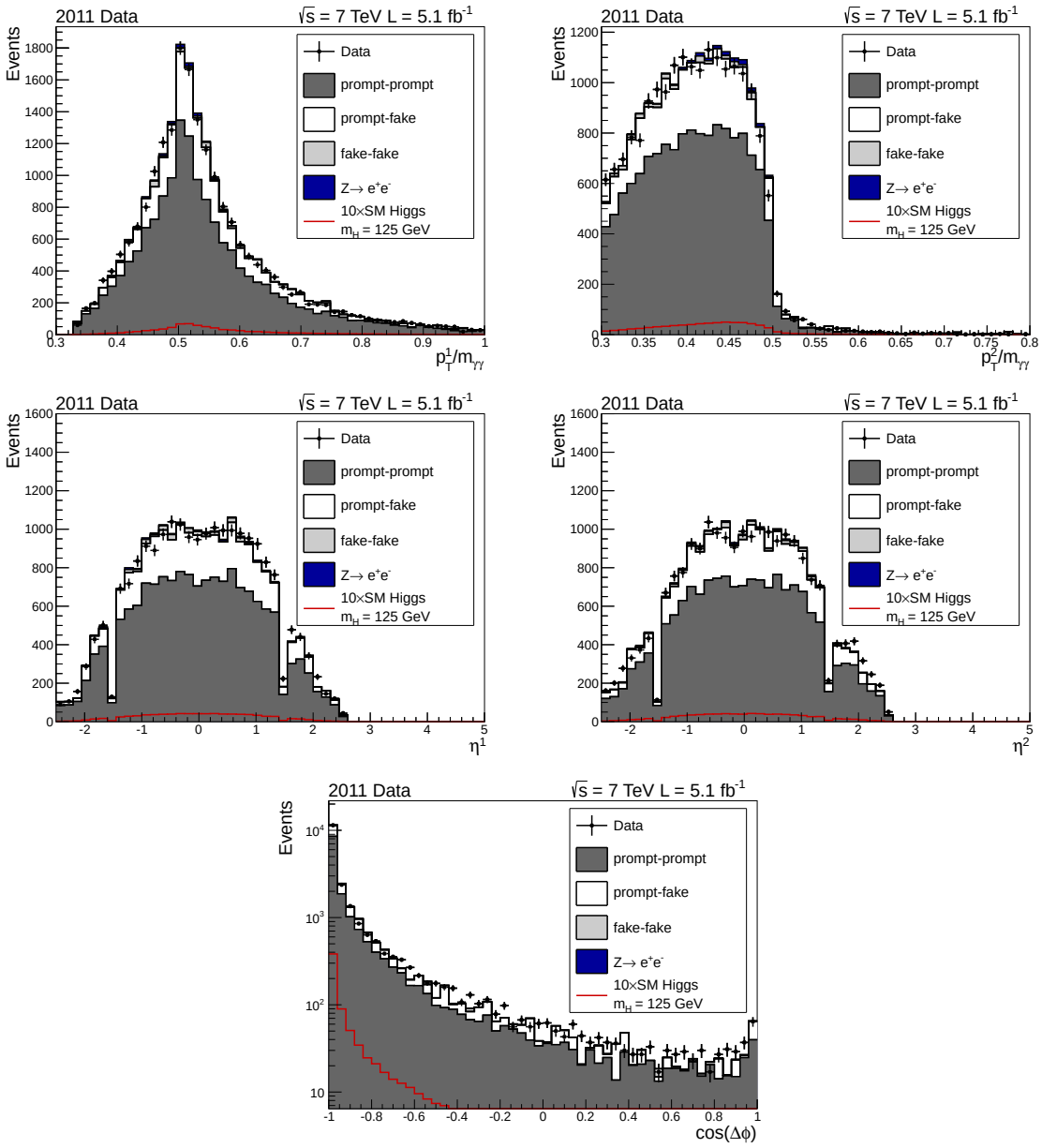


**Figure 4.6.:** Kinematic inputs to the diphoton BDT in data and MC. The distributions are for events which pass the full selection including a cut on the diphoton BDT output of 0.05. The expectation from a SM Higgs boson with 125 GeV is shown in red.

attributed to the systematic uncertainty on the shower simulation. These uncertainties are propagated through the diphoton BDT and included in the signal model as described in Section 4.4.5.

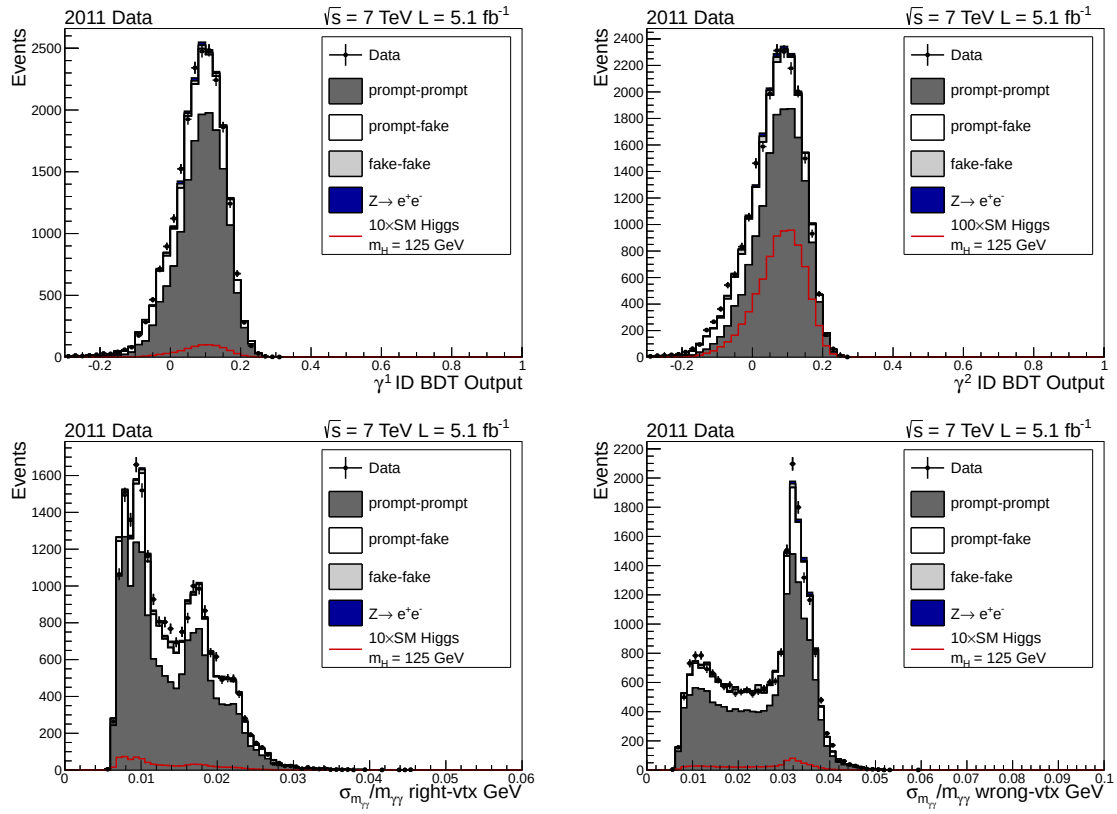


**Figure 4.7.:** Additional input variables to the diphoton BDT in data and MC. The distributions are for events which pass the full selection including a cut on the diphoton BDT output of 0.05. The expectation from a SM Higgs boson with 125 GeV is shown in red.

### 4.3.2. Dijet Tagging

The contribution to Higgs boson production from vector boson fusion is around a factor ten smaller than that of gluon-gluon fusion. However, additional information from the two jets associated with $qqH$ production allows further reduction of the diphoton background [46]. Events containing two jets which pass the full selection and in addition satisfy a series of criteria designed to target the specific dijet topology are tagged as likely to have originated from $qqH$ production. For example, Figure 4.13 shows the separation in $\eta$ between the two jets. Signal events from vector boson fusion production are more likely to have a large separation than those from background processes. The full set of criteria is given in Table 4.3. The dijet tagged events are categorized separately to the remaining events, thereby exploiting their high signal to background ratio for the purpose of signal extraction.

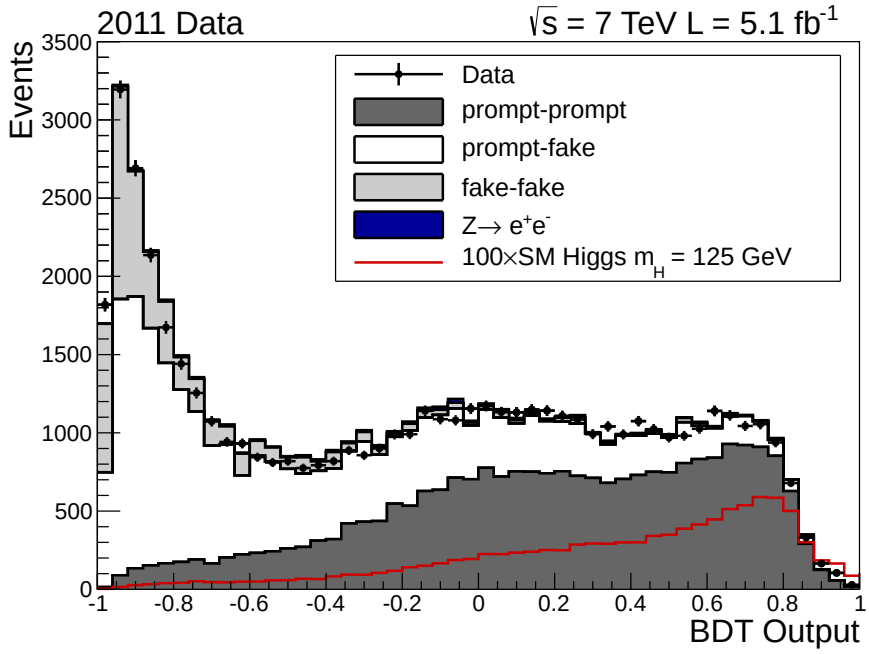


**Figure 4.8.:** Diphoton BDT distribution in data and MC. The contribution expected from a SM Higgs boson with mass 125 GeV, scaled by 100, is shown in red.

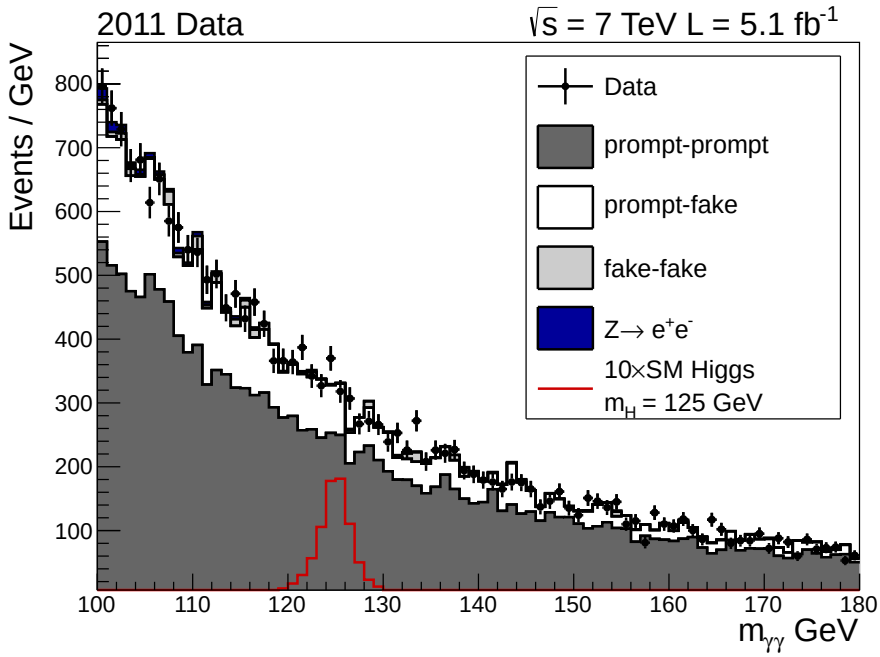


**Figure 4.9.:** Invariant mass distribution in data and MC after applying the full event selection in the range 100 to 180 GeV. The contribution expected from a SM Higgs boson with mass 125 GeV, scaled by 10, is shown in red.

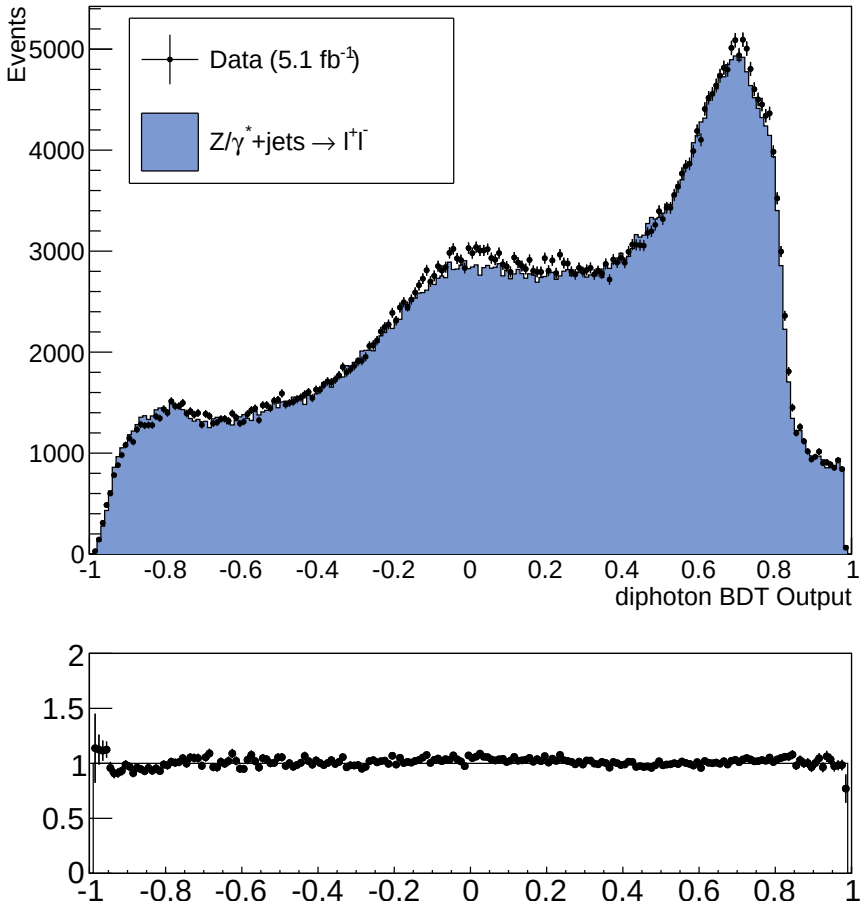


**Figure 4.10.:** Diphoton BDT output distribution in $Z \to e^+e^-$ MC and data after the full selection treating the electrons as photons for the purposes of energy reconstruction. The electron veto is inverted to preferentially select electrons. The lower panel show the data/MC ratio.

| **Variable** | **Cut Value** |
|---|---|
| $E_T^{j^1}$ | $> 30$ GeV |
| $E_T^{j^2}$ | $> 20$ GeV |
| $m_{jj}$ | $> 350$ GeV |
| $\lvert\eta_{j^1} - \eta_{j^2}\rvert$ | $> 3.5$ |
| $\lvert\phi_{jj} - \phi_{\gamma\gamma}\rvert$ | $> 2.6$ |
| $\lvert\frac{1}{2}(\eta_{j^1} + \eta_{j^2}) - \eta_{\gamma\gamma}\rvert$ | $< 2.5$ |

**Table 4.3.:** Dijet selection criteria for the two $qqH$ jets. The leading and sub-leading $E_T$ jets are denoted $j^1$ and $j^2$ respectively.

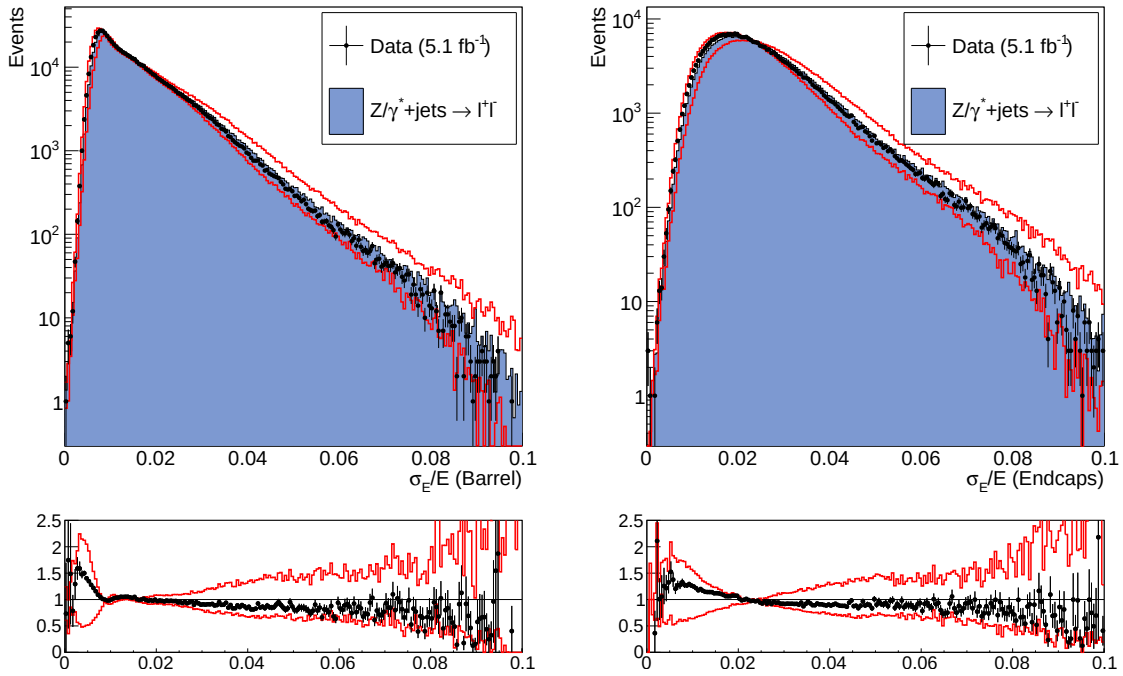


**Figure 4.11.:** Per-photon resolution estimator, $\sigma_E$, relative to the measured energy in $Z \to e^+e^-$ MC and data treating the electrons as photons in the barrel (left) and endcaps (right). The red lines show the $\pm 1\sigma$ systematic error envelope obtained by scaling the value of $\sigma_E$ by $\pm 10\%$. The lower panels show the ratios to the nominal MC distributions.

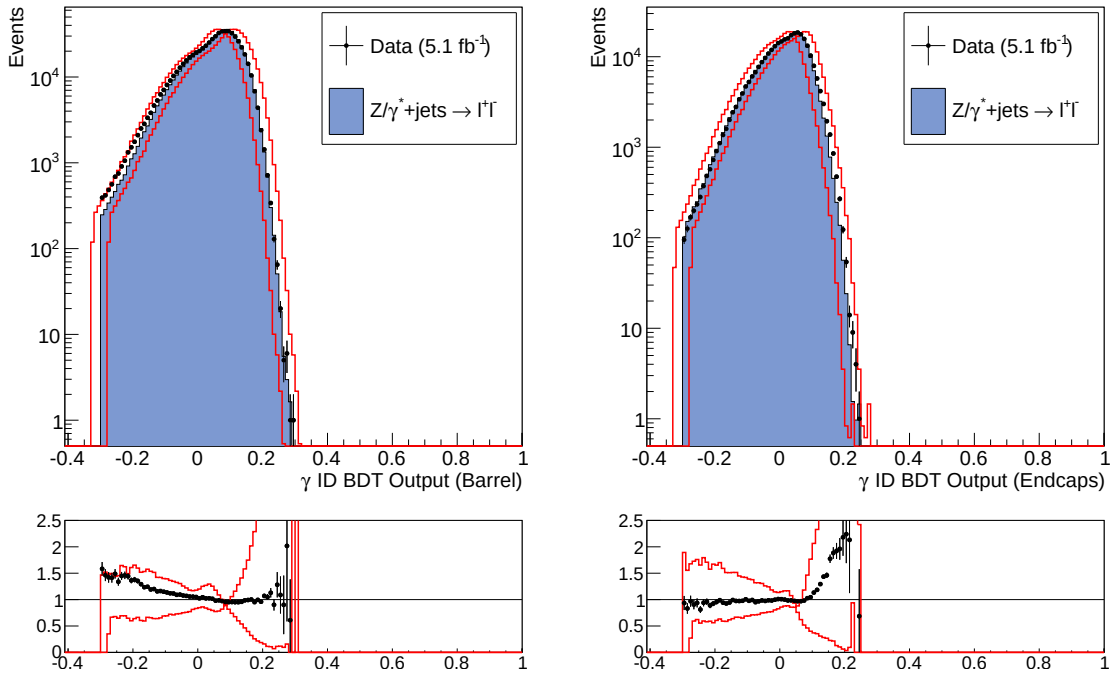


**Figure 4.12.:** Photon ID BDT output in $Z \to e^+e^-$ MC and data treating the electrons as photons in the barrel (left) and endcaps (right). The red lines show the $\pm 1\sigma$ systematic error envelope obtained by shifting the output value by $\pm 0.025\%$. The lower panels show the ratios to the nominal MC distributions.

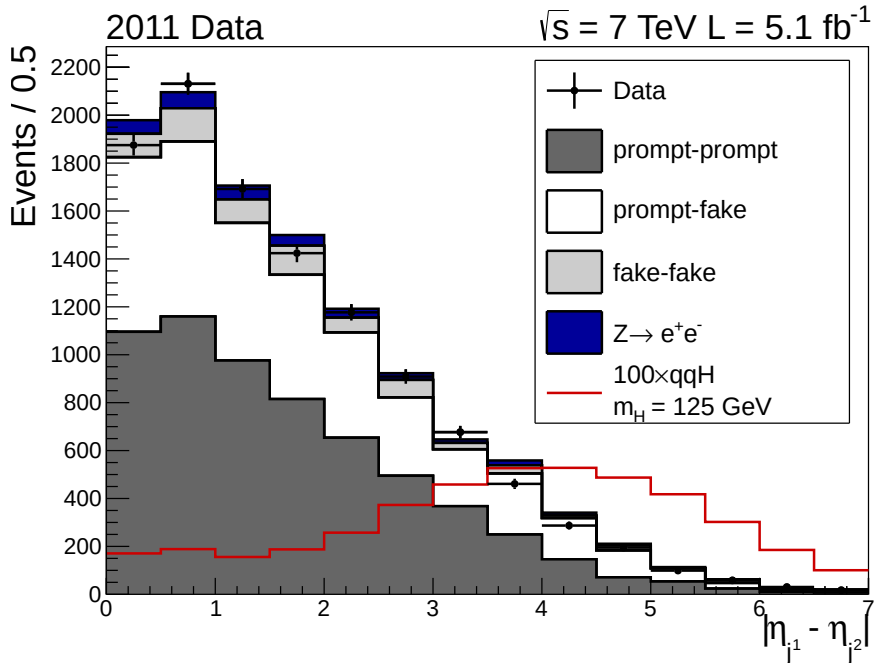


**Figure 4.13.:** Separation in $\eta$ between two identified jets in data and MC. The expectation from a SM Higgs boson produced via vector boson fusion ($qqH$), scaled by 100, is shown in red. All cuts other than the one on $\Delta\eta(Jet1, Jet2)$ are applied to these distributions.

## 4.4. Signal Extraction

The signature for the decay $H \to \gamma\gamma$ is the presence of a narrow peak on a smoothly falling background in the invariant mass spectrum. The signal to background ratio can be dramatically increased by focusing on events falling in a window around the mass of the Higgs boson, $m_H$. Since this mass is not predicted in the Standard Model, the search is performed for a range of mass hypotheses effectively sliding the signal window across the diphoton invariant mass spectrum, $m_{\gamma\gamma}$. As the signal yield for a SM Higgs boson decaying to two photons is small, additional event information from the detector and the kinematics of the diphoton system can be used to increase the sensitivity of the search.

This section describes a multivariate analysis (MVA) based approach to extracting the signal, categorizing events within a sliding signal window based on a single event discriminator (categorisation BDT). The approach allows for use of data in sidebands to determine expected event yields within the signal region, making few assumptions about the specific composition and kinematics of the background. This approach is the second

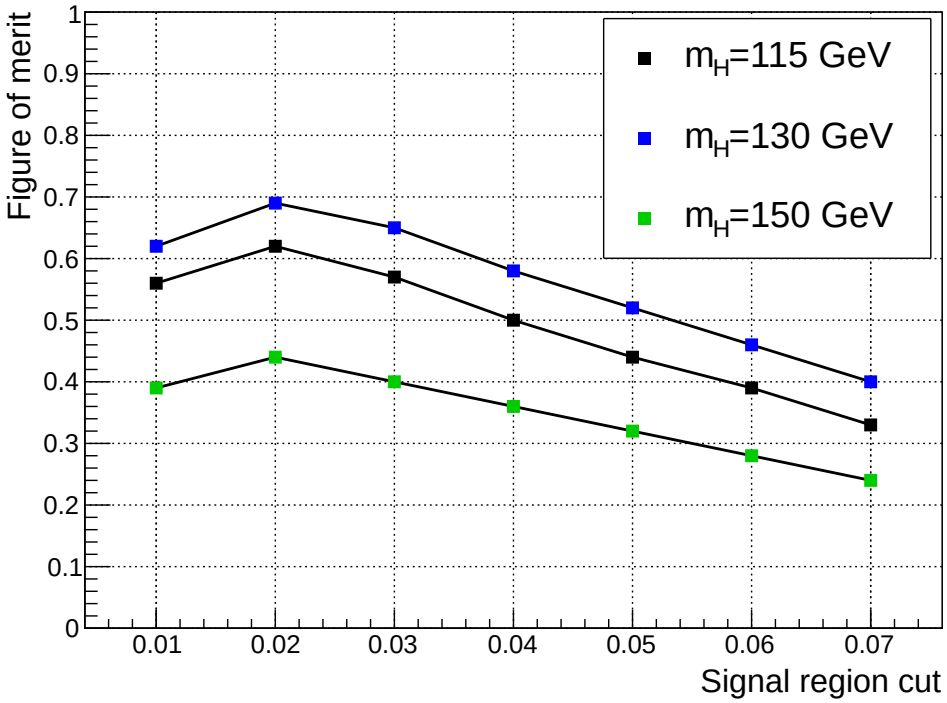


**Figure 4.14.:** Figure of merit for selection of the signal region cut value, $w$. Each colour shows the evaluation under different Higgs boson mass hypotheses.

of the two methods (method B) for extracting the signal, used by the CMS $H \to \gamma\gamma$ group.

### 4.4.1. Definition of the Signal Region

Once the expected resolution of the $H \to \gamma\gamma$ peak is determined, the choice of signal window can be optimized to reduce the uncertainty on the background while selecting as many signal events as possible. The size of the signal window is chosen using a simplified analysis in which the number of signal events from a SM Higgs boson with hypothesised mass $m_H$ expected within the range $|\Delta M/M_H| = |(m_{\gamma\gamma} - m_H)/m_H| < w$ is compared to the uncertainty on the total number of events (from background and signal) in that range. The figure of merit, $N_S/\sigma = N_S/\sqrt{\sigma_S^2 + \sigma_B^2}$, is calculated as a function of signal region cut value, $w$, for a range of mass hypotheses as shown in Figure 4.14. The error on the number of background events, $\sigma_B$, is calculated using the procedure described in Section 4.4.4 whereas the error on the signal is purely statistical. For this analysis, $w = 0.02$ was chosen as the optimal signal region cut value.

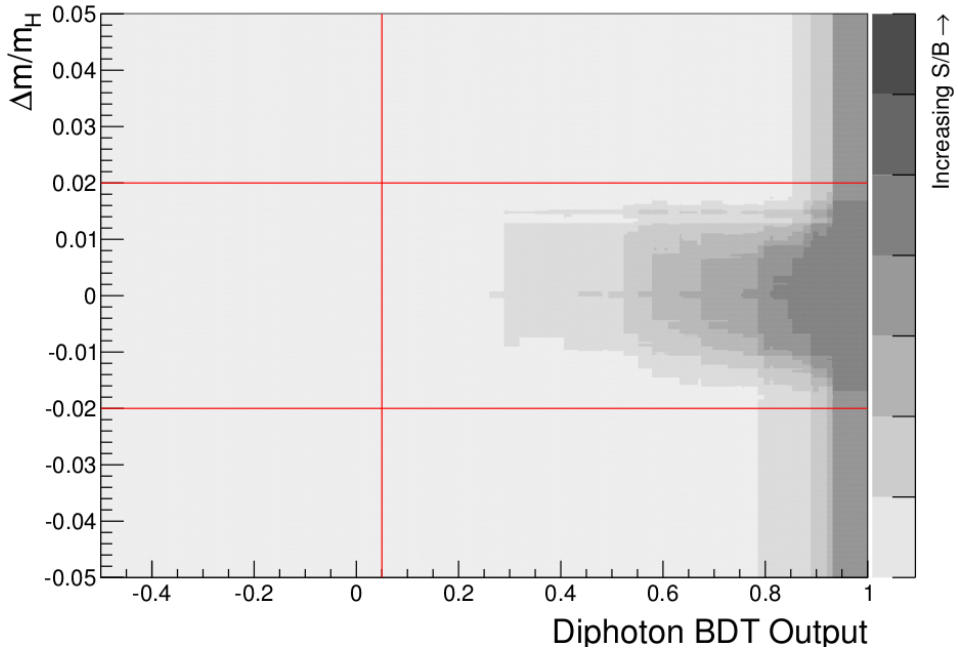


**Figure 4.15.:** Signal to background ratio as a function of diphoton BDT output and $\Delta m/m_H$. The red lines indicate the cuts applied before the training and for applying the event selection. Darker shades indicate a regions with a higher signal to background ratio. The seven shades indicate the region contained in each of the seven BDT bins used for the signal extraction at $m_H = 123$ GeV.

### 4.4.2. Event Categorisation BDT

The inputs to the diphoton BDT contain information from the event kinematics and the quality of the photons and vertex location in the form of the photon ID BDT output and event resolution estimators. The output of the diphoton BDT combined with the invariant mass of the diphoton system therefore provides the necessary information to separate signal from background.

Figure 4.15 shows the variation in the signal to background ratio ($S/B$) across different regions in the two-dimensional plane defined by the output of the diphoton BDT and $\Delta m/m_H$. Events close to the centre of the peak ($\Delta m/m_H = 0$) with a high score in the diphoton BDT are more likely signal events than those far from the high $S/B$ regions.

The two variables are combined to produce a single event discriminator by training a BDT using the diphoton BDT output and $\Delta m/m_H$ as inputs. The BDT is trained with Higgs signal MC with $m_H = 123$ GeV including all four production processes and background MC including prompt-prompt, prompt-fake and fake-fake events. The

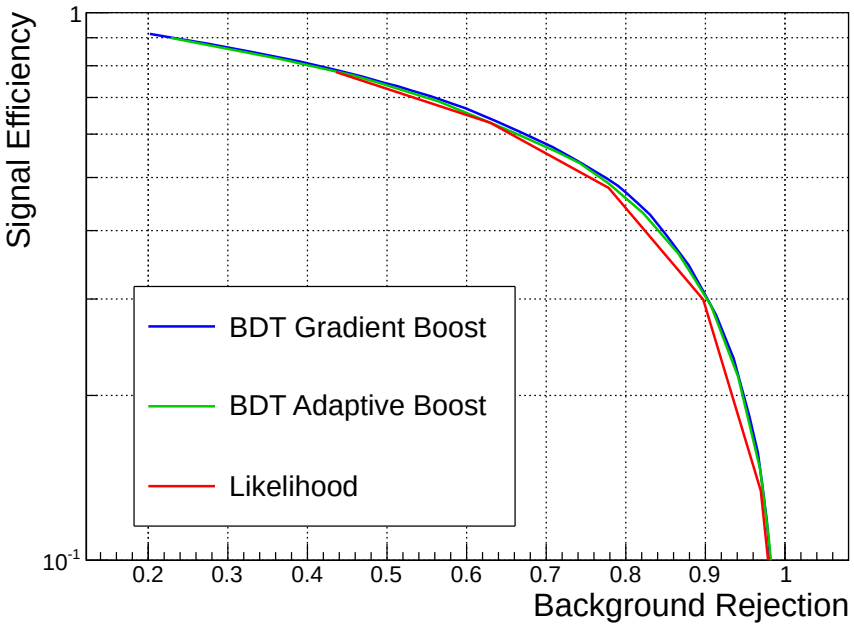


**Figure 4.16.:** Signal efficiency vs background rejection curves for three different MVA techniques used to train the signal-background event discriminator. The curves give the (in)efficiencies for signal (background) after applying sequentially tighter cuts on the discriminator output.

performance of several different training methodologies was compared to find which gave the optimum separation of signal and background. Two different choices of boosting were studied, adaptive and gradient boosting, both of which weight decision trees to optimize the performance in terms of signal-background separation [54]. In addition, these were compared to a simple likelihood which does not account for correlations between the diphoton BDT and $\Delta m/m_H$ as shown in Figure 4.16. The gradient boosting method was found to give the best performance although the variation between methodologies is small.

With finite statistics, a BDT can be over-trained by allowing the training to emphasise statistical fluctuations which are not physical and will not necessarily be representative of the data. To test for this, the MC samples are split into two equal samples, the first of which is used to train the BDT. The distribution of the output values of the BDT from the second set is compared to that of the training sample as shown in Figure 4.17. The comparison is shown using both an arbitrary binning scheme and the final set of bins derived in Section 4.4.3. A $\chi^2$ test was performed on the distributions with the final

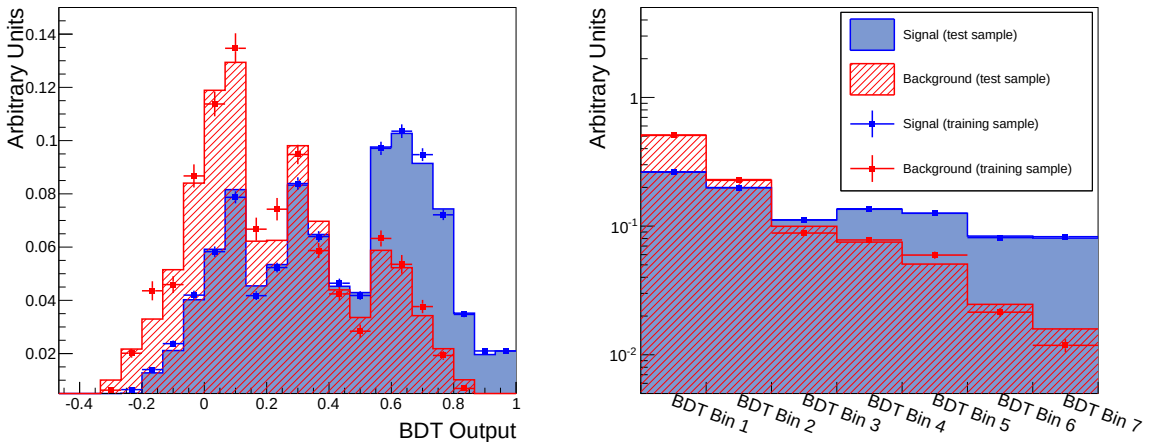


**Figure 4.17.:** Signal and background BDT output distribution with the training sample (points) and testing sample (solid area) superimposed. The comparison is shown using an arbitrary uniform binning (left) and the bins used for extracting the signal (right).

bins giving p-values of 0.06 for the background and 0.95 for the signal indicating that over-training has not occurred.

In this analysis, the background is estimated entirely from data. This means that disagreement between data and background MC will affect the performance of the BDT rather than the validity of the final results. The agreement between the data and MC is shown in Figure 4.18 for the mass hypothesis, $m_H = 125$ GeV. The level of agreement is sufficient so as not to require in-depth study of the BDT output distributions of the background MC.

### 4.4.3. Binning of the BDT Output Distribution

The BDT provides a single variable with which to classify events based on their signal to background ratio, $S/B$, which will have a discrete number of response values based on the number of trees used. The boosting procedure provides a pseudo-continuous distribution which is used to model the signal and background. However, the resulting distribution will still be only pseudo-continuous. In addition, the BDT response does not directly correspond to a physical distribution and it is therefore difficult to motivate any parameterisation of either the signal or background distributions. To overcome these issues, a binning procedure is defined to construct templates which are used as models for the signal and background expectation as a function of BDT response range (BDT bin).

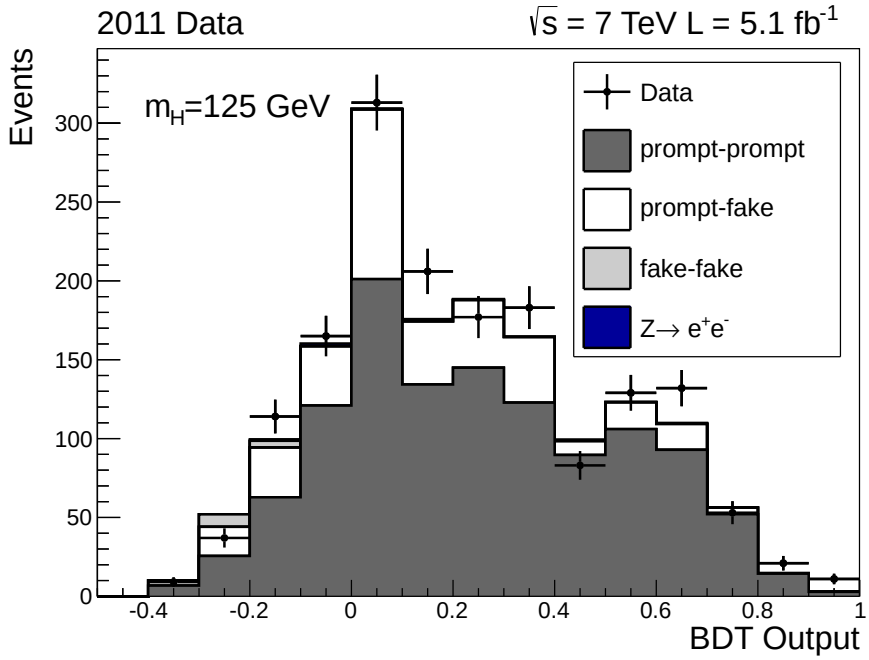


**Figure 4.18.:** Comparison of the distributions of BDT output at $m_H = 125$ GeV for data and background MC. The distributions are arbitrarily binned for the purposes of comparison only.

This procedure is designed firstly to ensure that no bin has zero background expectation and secondly that as few bins as possible are used without reducing the sensitivity of the BDT. These requirements are desirable such that the expected background yield in each bin can be derived using data outside of the signal region as described in Section 4.4.4.

A scan is performed in which the definitions of the bin boundaries are varied in order to find the maximum expected significance in the presence of a SM Higgs signal. For $N$ bins ($N-1$ boundaries) with background and signal expectation yields $b_i$ and $s_i$ respectively, the expected significance, $\sigma_{exp}$, is given by

$$\sigma_{exp} = \left( 2 \sum_{i=1}^{N} (s_i + b_i) \ln \left( \frac{s_i}{b_i} + 1 \right) - s_i \right)^{1/2}, \tag{4.9}$$

using the log-likelihood ratio for Poisson likelihoods. The binning procedure is defined as follows:

1. The distribution of background MC is binned very finely to provide an almost discrete dataset (5000 equally spaced bins are used). The background is re-binned such that there are 20 expected events per bin at a luminosity of $5.1 fb^{-1}$.

2. Smoothed versions of the signal (at each 5 GeV step mass) and background MC templates are produced in order to obtain a stable model of $S/B$ as a function of BDT bin. The smoothing procedure is done via binning a fit (of a 9th order polynomial) to the signal distribution.

3. $N$ bin edges (boundaries), $b_i$, are defined on the remaining bins such that $N+1$ bins are formed with $b_1 < b_2 < \ldots < b_N$. The first bin is defined as $[-1, b_1)$ and the last is defined as $[b_N, 1]$. The $N$ dimensional scan is performed varying these bin edges to find the maximum expected significance in the presence of a SM Higgs signal.

4. An extra boundary is added, the scan is repeated and the maximum expected significance is found for $N+1$ boundaries. If the maximum expected significance is increased by more than 0.1% compared to that of step 3, the new boundary is kept and step 4 is repeated, if not, the procedure terminates.

The scan in step 3 is split into two parts, first using a large step size to find the region where the maximum lies followed by a fine scan in small steps within that region. The ratio of small to large step size is chosen to be that which minimizes the total number of iterations in the scan to reduce the time taken for the procedure. An example of the binning procedure is shown in Figure 4.19. The red histogram is the $S/B$ distribution after step 1, the blue after step 2 and the black vertical lines show the final set of 7 bins chosen for this analysis. Dijet tagged events are treated in the same way as the rest of the events in the analysis by introducing an eighth bin containing events from any BDT output bin inside the range $\Delta m/m_H < w$ which pass the dijet tag.

### 4.4.4. Background Model

The SM background is expected to have a smoothly varying invariant mass spectrum. However, detector effects such as selection, trigger efficiencies and energy resolution shape this distribution in ways which are imperfectly modelled in MC simulation. Moreover, the background contains fakes whose contribution varies as a function of $m_{\gamma\gamma}$. This means the exact composition of the background is needed to model the shape with MC.

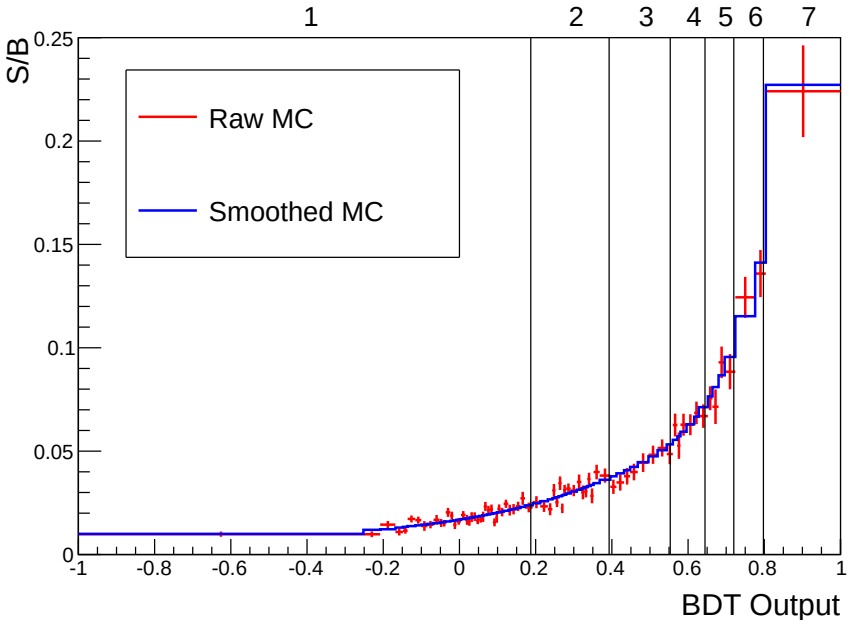


**Figure 4.19.:** Signal to background ratio as a function of BDT output bin. The red and blue histograms show the distribution after applying step 1 of the binning procedure before and after smoothing respectively. The black vertical lines indicate the boundaries of the final binning choice from the full procedure.

In order to remove the impact of systematic uncertainties associated with this, an entirely data-driven approach for modelling the background is used.

For a given mass hypothesis, the shape and normalization of the background model are obtained separately. The shape, meaning the fraction of events in each BDT output bin, is extracted from the BDT output distributions in mass-sidebands, while the overall normalization is obtained from a parametric fit to the mass distribution for all selected events excluding the signal region.

Figure 4.20 shows the invariant mass distribution after event selection in the range $100 < m_{\gamma\gamma} < 180$ GeV for the full 2011 dataset. The red band indicates the signal region for $m_H = 124$ GeV, while the six blue bands indicate the corresponding sidebands used to determine the shape of the background model. The blue line indicates the fit of a sum of two power laws which is used to determine the normalisation of the background in the signal region.

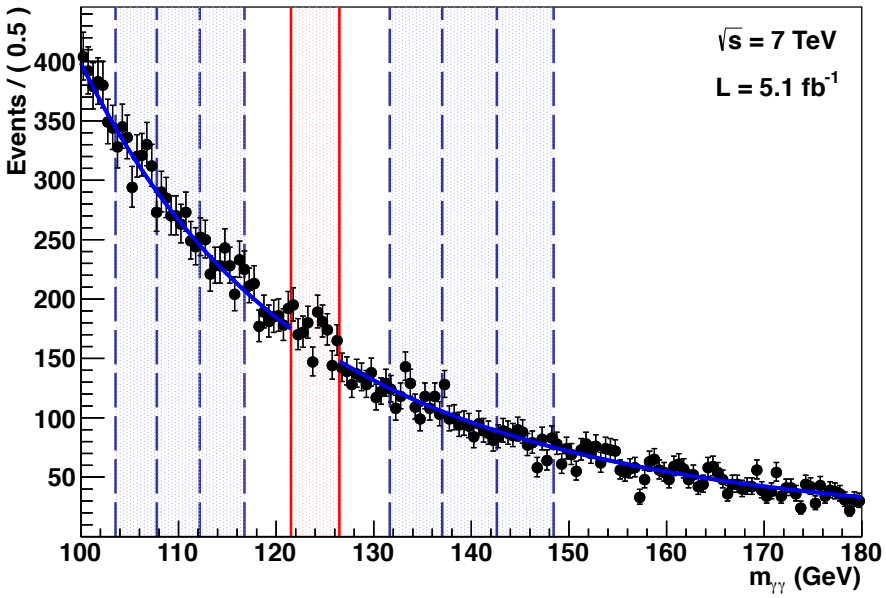


**Figure 4.20.:** Invariant mass distribution of the full 2011 dataset after selection over the mass range used in the analysis (100 to 180 GeV). The ±2% signal region for $m_H = 124$ GeV is indicated in red, while the six corresponding sidebands are indicated as blue bands. The blue line is the double power law fit to the data for the background normalisation for this mass hypothesis.

### Obtaining the Normalisation of the Background

The normalisation of the background model is estimated using an un-binned maximum likelihood fit of a parametric function to the diphoton invariant mass distribution in the range $100 < m_{\gamma\gamma} < 180$ GeV. The normalisation of the background model is given by the integral of the function over the ±2% signal region for each mass hypothesis. The signal region is excluded from the fit to avoid potential bias in the presence of a signal.

The particular parameterization used was chosen following a study of different parametric forms which also provide a good fit to the data. Since the actual functional form is unknown, the choice of parameterization is taken to be that which minimises the total uncertainty when comparing to the other functional forms. Twelve different functional forms were considered, which can be grouped into four general classes: exponentials, power laws, real Laurent polynomials and standard polynomials. Within each of these classes, three functions were used. For the exponentials and power law cases, these were sums of one, two or three exponential or power law ($m_{\gamma\gamma}^{-r}$) terms, while only first, third and fifth order standard polynomials were used. For the Laurent polynomials, the

functions were sums of two, four or six terms, specifically

$$m_{\gamma\gamma}^{-4} + a m_{\gamma\gamma}^{-5},$$
$$m_{\gamma\gamma}^{-4} + a m_{\gamma\gamma}^{-5} + b m_{\gamma\gamma}^{-3} + c m_{\gamma\gamma}^{-6},$$
$$m_{\gamma\gamma}^{-4} + a m_{\gamma\gamma}^{-5} + b m_{\gamma\gamma}^{-3} + c m_{\gamma\gamma}^{-6} + d m_{\gamma\gamma}^{-2} + f m_{\gamma\gamma}^{-7}.$$

For each class therefore, the three functions have one, three or five parameters for the shape.

To assess the bias introduced through choosing one particular parameterisation, pseudo-experiments are generated from each functional form and the invariant mass of those experiments are fit with the other functional forms. The parameters for generation of the pseudo-experiments are fixed by fitting each functional form to the data in the full mass range. In each pseudo-experiment, the integral of a particular fitting function, A, over the signal region is compared to that from a generating function, B. The distribution of the difference between the two values across all of the pseudo-experiments is used to determine the bias introduced from choosing function A when B was the true function. The distributions are then weighted according to the probability of the initial fit to the data and combined so that the total uncertainty from choosing a particular function is computed as the RMS from zero of the weighted summed distributions for all generating functions. Since one of the generating functions can also be the fitting function, the error includes both the statistical uncertainty from the limited data sample and the systematic uncertainty due to an incorrect choice of parameterisation. This study is repeated at 5 GeV intervals in $m_H$ as the overall uncertainty varies as a function of mass hypothesis. Figure 4.21 shows the total error determined for each of the twelve functions at each value of $m_H$ tested. The sum of two power laws was found to give a low total uncertainty while also demonstrating good fit stability in the pseudo-experiments. The total error on the background normalisation is included as a single systematic uncertainty for the purpose of signal extraction (Section 4.4.6).

### Obtaining the Shape of the Background

As the signal yield expected from a SM Higgs boson is small compared to the background, the sensitivity of the search is strongly dependent on how well the relative contribution from the background in each bin is understood. Both inputs to the BDT are designed to be insensitive to the invariant mass of the diphoton system, therefore the BDT

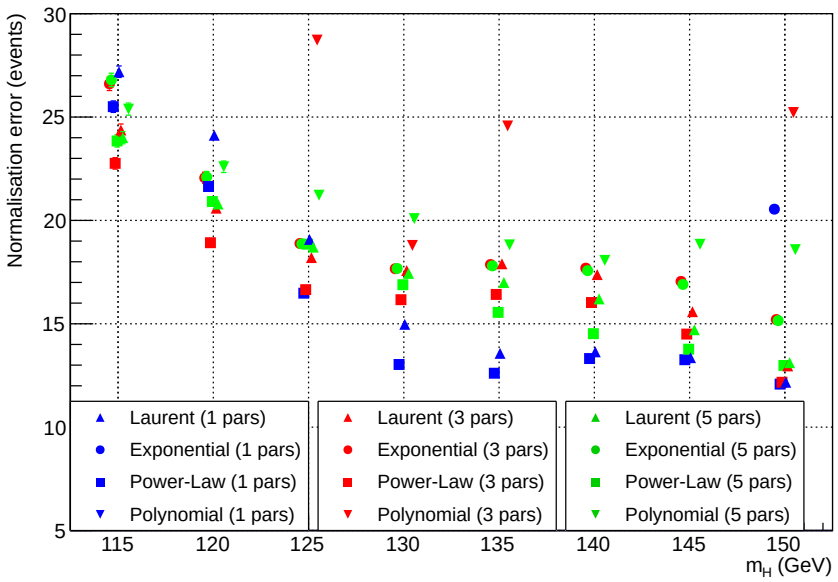


**Figure 4.21.:** Total error on the background normalisation as a function of $m_H$ from different choices of the background shape parameterisation of $m_{\gamma\gamma}$. The total error for the one-parameter exponential and polynomial functions are off the scale of this plot.

output distribution should be the same for any region of the $m_{\gamma\gamma}$ spectrum. Since the background composition remains relatively constant across the range 100 to 180 GeV, data in sidebands of $m_{\gamma\gamma}$, away from the signal, can be defined to determine the BDT distribution of the background inside the signal region. For a particular $m_H$, a contiguous set of lower/upper sidebands are defined to be the ranges $|(m_{\gamma\gamma} - m_{H,i})/m_{H,i}| < w$ centered on $m_{H,i}$ as given in Equation 4.10 where $w = 0.02$.

$$m_{H,i} = m_H \left(\frac{1+w}{1-w}\right)^i \tag{4.10}$$

The two sidebands adjacent to the signal window (corresponding to $i = \pm 1$ in Equation 4.10) are not used in order to avoid signal contamination. Dijet tagged events are treated in the same way as the rest of the events by introducing an eighth bin containing dijet tagged events inside the range $\Delta m/m_H < w$. The distributions for the two input variables, diphoton BDT output and $\Delta m/m_H$, for each of the six sidebands corresponding to $m_H = 125$ are shown in Figure 4.22. Each distribution is normalised to unit area. The resulting BDT output distributions are shown in Figure 4.23. The distributions from each sideband are not distinguishable within the statistical uncertainties.

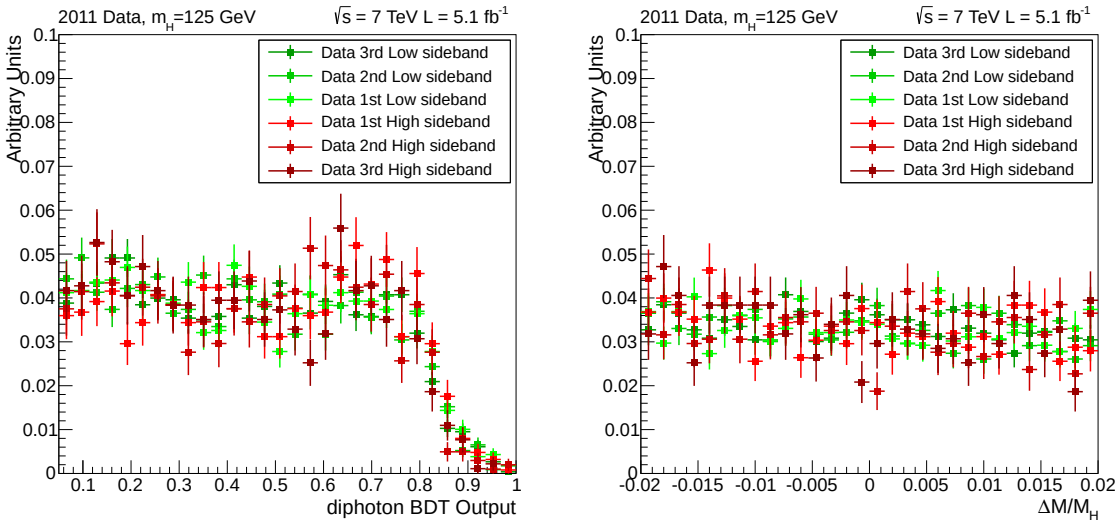


**Figure 4.22.:** Distribution in data from the six sidebands corresponding to $m_H = 125$ GeV of the two BDT input variables, diphoton BDT (left) and $\Delta m/m_H$ (right).

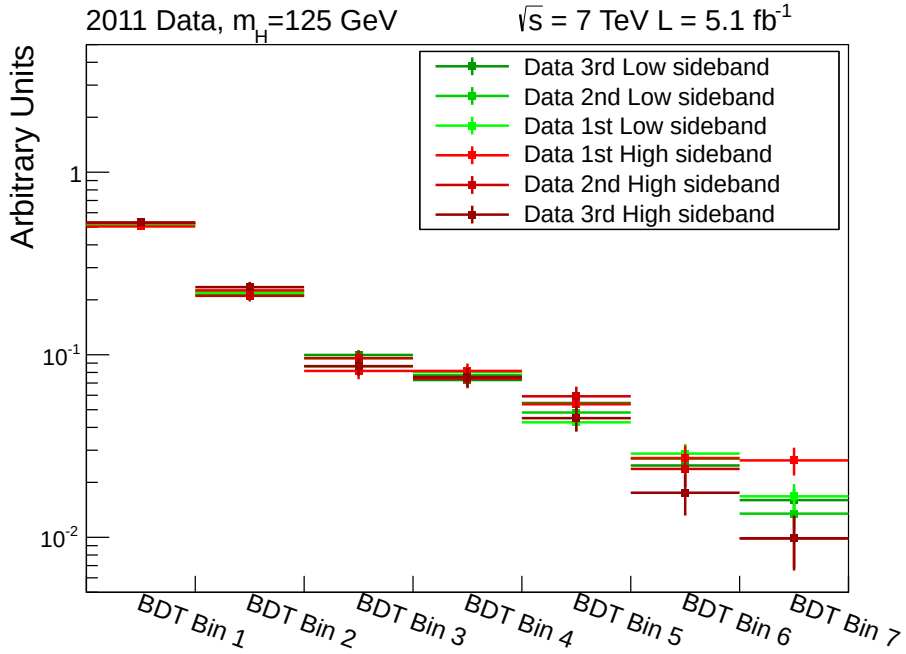


**Figure 4.23.:** Distribution in data from the six sidebands corresponding to $m_H = 125$ GeV of the BDT output binned in the 7 BDT output bins used for signal extraction.

The residual variation in BDT output is due to the small variation in background composition with mass. This is mostly due to the photon ID BDT distribution being sensitive to the fake component which varies with mass. In order to account for this variation, the background model is constructed using a simultaneous linear fit to the BDT output shape in the data sidebands. The expected fraction of events in each bin, $f_j$, for a given mass hypothesis, $m_{H,i}$, is given by Equation 4.11, where $j\epsilon\{1,8\}$ and $i\epsilon\{\cdots,-4,-3,-2,2,3,4\cdots\}$.

$$f_j = p_{0,j} + p_{1,j}(m_{H,i} - m_H) \tag{4.11}$$

Since the normalisation for the background model is determined independently, the sum over all bins is constrained to be one. The expectation value for the background in each bin, $j$, is then determined as $Nf_j$ where $N$ is the normalisation estimated in section 4.4.4. This constraint is imposed for all $m_{H,i}$ by fixing

$$p_{0,1} = 1 - \sum_{i=2}^{8} p_{0,j} \qquad p_{1,1} = -\sum_{j=2}^{8} p_{1,j} \tag{4.12}$$

The coefficients $p_{0,j}, p_{1,j}$ of Equation 4.11 are determined by performing a binned maximum likelihood fit to the observed fractions in the data assuming the contents of each bin in each sideband are Poisson distributed. The results of the fit for $m_H = 124$ GeV are shown in Figure 4.24 and the resulting covariance matrix obtained is shown in Figure 4.25. The fit was performed using `TMinuit` under `ROOT 5.2.0` [62].

There are seven degrees of freedom (eight bins minus one constraint) which are correlated in the simultaneous fit. In order to account for the statistical uncertainty on this fit, a set of seven uncorrelated variables are determined from the covariance matrix using eigenvector decomposition [63]. These variables are treated as seven independent sources of systematic uncertainty on the background shape for the purpose of signal extraction (Section 4.4.6). Figure 4.26 shows the total relative fit error for each bin, at $m_H = 130$ GeV, as the number of sidebands, is varied. Increasing the number of sidebands beyond six, three on each side of the signal region, provides negligible reduction in the statistical uncertainty. In order to avoid contamination from $Z \to e^+e^-$ at the lower mass hypotheses any lower sideband whose lower boundary is less than 100 GeV is removed and an additional higher sideband is introduced. Consequently mass hypotheses in the range $111 \leq m_H < 115.5$ have two lower and four upper sidebands and mass hypotheses in the range $110 \leq m_H < 111$ have one lower and five upper sidebands.

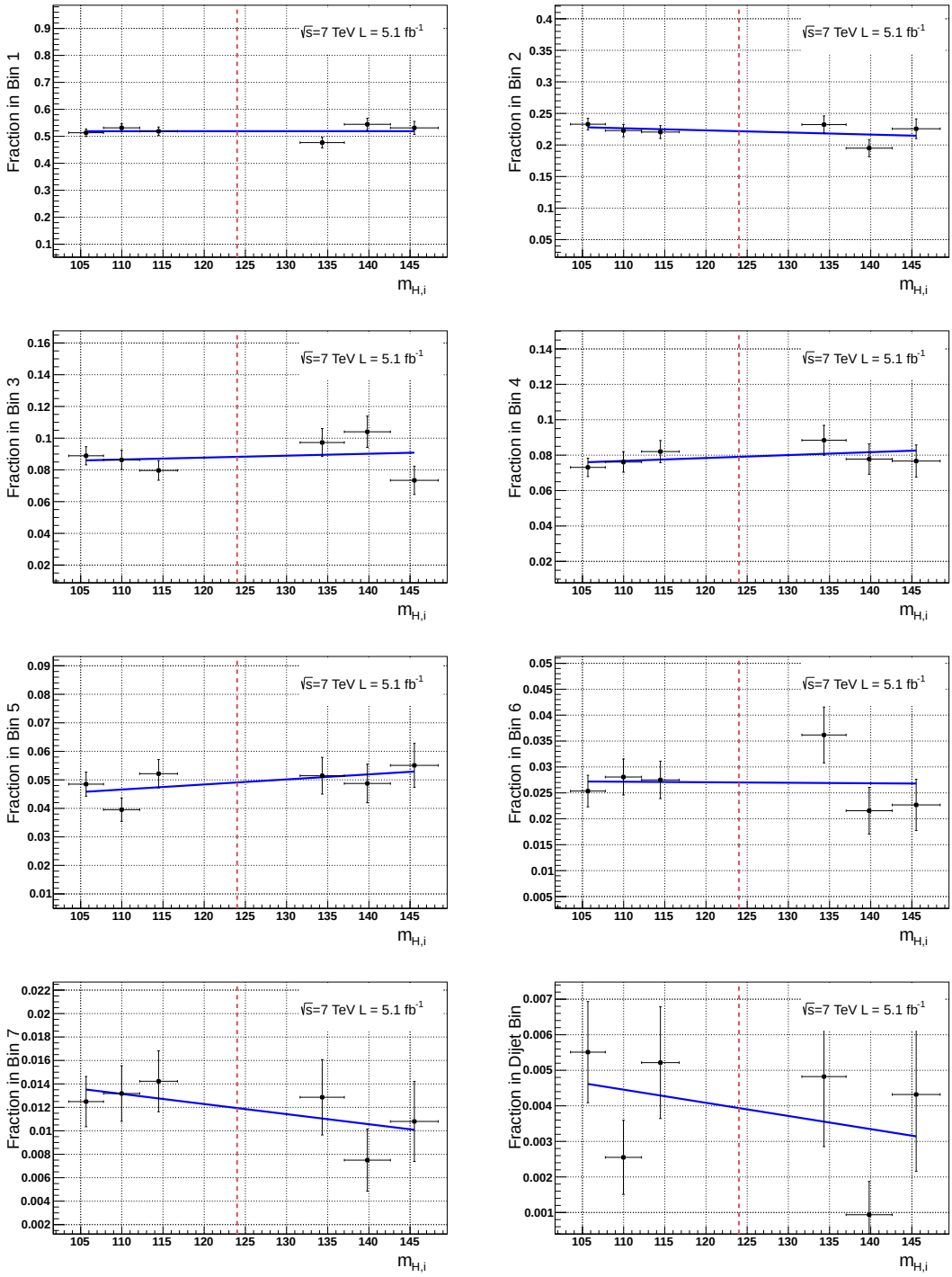


**Figure 4.24.:** Simultaneous fits to the six sidebands in data to determine the background shape for $m_H = 124$ GeV. There are eight panels showing the result in each of the seven BDT bins plus one for the dijet tagged bin. The six black points in each panel are the fractional populations of the data in each sideband. The blue line represents the linear fit used to determine the fraction of background in each bin.

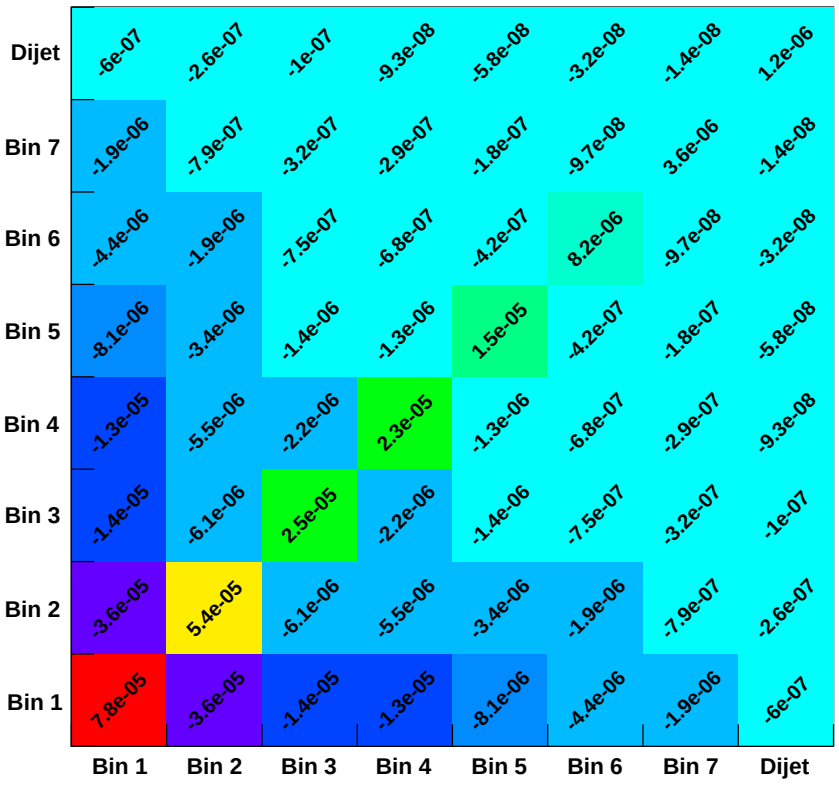


**Figure 4.25.:** Covariance matrix from the sideband fit to determine the background shape at $m_H = 124$ GeV. The covariance matrix includes the additional 20% systematic attributed to possible second order variations in the BDT output background distribution with mass.

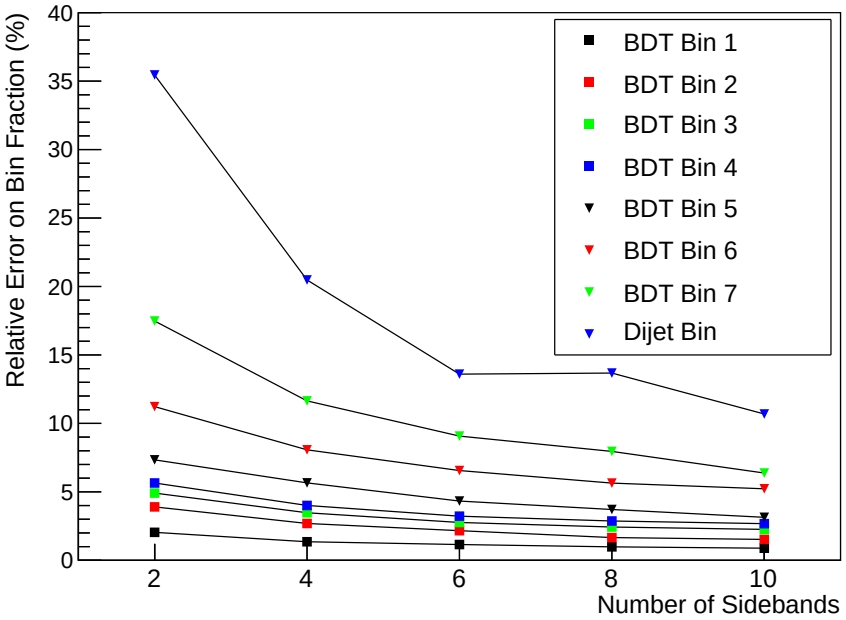


**Figure 4.26.:** Relative total fit uncertainty on the background model in each bin at $m_H = 130$ GeV as a function of the number of sidebands used in the fit to determine the shape of the background.

At most linear variations with mass are considered for the background BDT output distribution. This corresponds to evaluating the first term in a Taylor series for the true shape of the distribution about $m_H$. Higher terms can be introduced but the statistical precision of the fit will be reduced in doing so. To check for potential significant deviations in the data from linearity, pseudo-experiments were generated in which the expected fractions, $f_i$ are assumed to follow,

$$f_j = p_{0,j} + p_{1,j}(m_{H,i} - m_H) + \frac{1}{2}p_{2,j}(m_{H,i} - m_H)^2. \tag{4.13}$$

The parameter values, $p_{0,j}, p_{1,j}$ and $p_{2,j}$ and their uncertainties were determined by fitting over a larger number of sidebands for a particular mass hypothesis. This is done by extending the range of $j$ to allow any sideband which is contained inside the range $100 < m_{\gamma\gamma} < 180$ GeV. For most mass hypotheses, this corresponds to fifteen sidebands in total. For each pseudo-experiment, the parameters were varied within their uncertainties (accounting for correlations) thereby systematically altering the expectation value for the number of events in each bin before generating a Poisson toy for the observed number of events per bin in each sideband. The usual linear fit is then performed and the fraction of events in each bin for the signal region is extracted and compared to the true generating

fraction. The difference between these two values can be used to determine the total error under the assumption that a second term in the Taylor expansion is present in the data. This error is taken as the root mean square (RMS) around zero of the difference between the true and fitted values for $f_i$ in 10,000 pseudo-experiments. When compared to the error from the linear fits, it was found that the total uncertainty was covered by inflating the errors systematically by 20%. The value of 20% is a conservative choice being the largest value found when repeating the study over a range of mass hypotheses.

### 4.4.5. Signal Model

The signal model for the Higgs boson decay to two photons at a given mass is constructed by binning the BDT response from MC simulation of the four production processes, $ggH$, $qqH$, $VH$ and $ttH$. The simulation is corrected using auxiliary measurements from $Z \to e^+e^-$ events in data to account for imperfect modeling of the detector. These corrections are applied to the Monte Carlo event by event and can be categorized into photon and diphoton level corrections.

#### Photon Level Corrections

The energy resolution of the calorimeter is measured in data using $Z \to e^+e^-$ events in categories defined by the position and $r_9$ of the supercluster. Photons in the central region of the detector with $r_9 > 0.94$ are further divided into those whose supercluster seed lies close to a module boundary and those whose does not. The additional energy smearing required for the Monte Carlo in each category is determined by smearing $Z \to e^+e^-$ MC until the $e^+e$ invariant mass distribution matches that of the data. This additional resolution is included in the Higgs MC by scaling the energy of each photon by $G(1, \sigma_{cat})$ where $G$ is a Gaussian distributed random variable centered at 1, and $\sigma_{cat}$ is the additional resolution required to match the data in a particular category. The exact definitions of the photon-level categories and the additional resolution measured in each category are given in Table B.1.

The efficiency for a photon to pass the pre-selection is measured in $Z \to e^+e^-$ data in four categories. These are defined by whether or not the supercluster is in the ECAL barrel or either endcap and the value of $r_9$ being greater or less than 0.94. The ratio of the efficiency measured in data to that measured in MC provides a scale factor which is applied to the signal MC. Each signal event is reweighted by the product of the scale

factors for each photon in the selected diphoton pair. In addition to these corrections, the value of $\sigma_E$ and the photon ID BDT for each photon is shifted in each signal event to account for imperfections in detector simulations as described in Section 4.3.1.

### Diphoton Level Corrections

The efficiency to select the correct vertex in the event is measured using $Z \to \mu^+\mu^-$ events as a function of the boson $p_T$ as described in Section 4.2.3. Signal MC events are categorized by whether or not the selected vertex is within 10mm of the generated vertex. Each event is then re-weighted by the ratio of the probability that the event lies in a particular category as measured in $Z \to \mu^+\mu^-$ data ($\epsilon_{data}$) to that measured in $Z \to \mu^+\mu^-$ MC ($\epsilon_{MC}$). Figure 4.27 shows the weight, $\epsilon_{data}/\epsilon_{MC}$, applied to events in the signal MC in which the correct vertex is selected as a function of the Higgs boson candidate $p_T$ ($p_T^H$). Similarly, events in which this is not the case are reweighted by the ratio $(1-\epsilon_{data})/(1-\epsilon_{MC})$. The L1/HLT efficiency is measured in four diphoton categories depending on the maximum supercluster $\eta$ and minimum $r_9$ value of the two photons using $Z \to e^+e^-$ data. As the simulation does not include the trigger, the efficiency is applied directly as a weight to each MC event.

### Systematic Uncertainties

For each correction applied to the MC, the accuracy to which that correction is measured provides an estimate of the uncertainty present in the signal model. In the case of the energy scale measurement, no correction is applied to the MC although the uncertainty in that measurement is treated as a systematic on the per-photon energy in signal MC events. The systematic uncertainties that affect the shape of the signal are treated as correlated migrations across the BDT output bins. The effect of each systematic in each bin is derived by shifting the relevant quantity in the signal MC and recalculating the BDT output for each event. The difference between the signal yield in each bin after applying the shift quantifies the variation due to that uncertainty. In practise, these quantities are derived by applying shifts to the MC corresponding to $3\sigma$ variations of each uncertainty and interpolating the difference from the nominal values back to the $1\sigma$ level. This is done so that the evaluation of the variation in each bin is more robust for systematics which have a small effect on the BDT output and in signal processes with

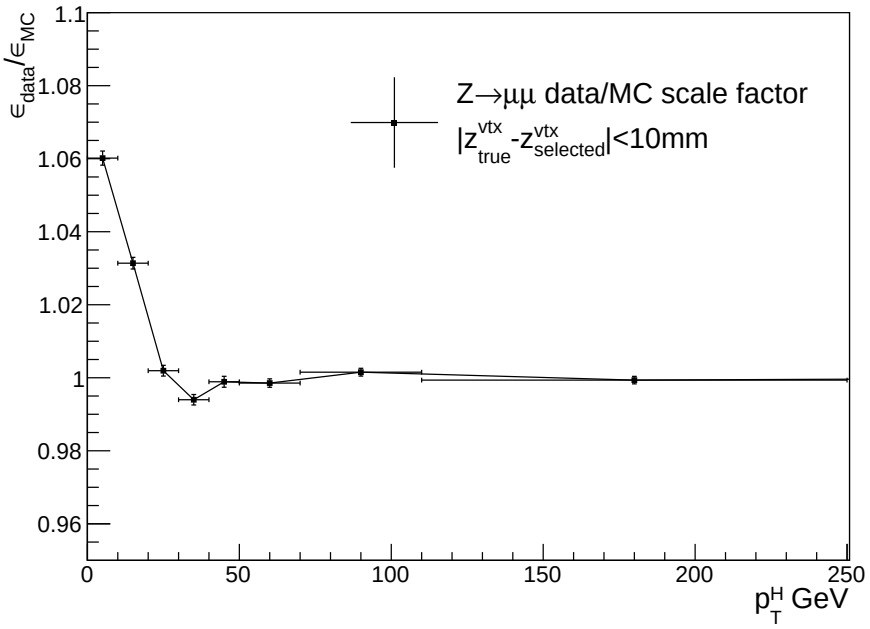


**Figure 4.27.:** Re-weighting applied to signal MC in which the $z$ position of the selected vertex is within 10mm of the true vertex as a function of $p_T^H$. The weights are derived from $Z \to \mu^+\mu^-$ events in data and MC.

fewer available MC events. Figure 4.28 shows the effect of the energy scale and resolution uncertainties on the BDT output of signal from gluon-gluon fusion production.

Imperfections in the simulation of the shower shape variables can cause discrepancies in the photon ID and $\sigma_E$ distributions obtained from the respective BDTs between data and MC. To account for this, systematic uncertainties are included corresponding to shifting or scaling the output of the photon ID BDT and regression BDT respectively and recalculating the BDT output for each event in signal MC. The size of the uncertainty is chosen to be that which covers the maximal difference in the ratio of each distribution for high $p_T$ photons. This is then validated using $Z \to e^+e^-$ events in which the electrons are reconstructed as photons. The overall efficiency×acceptance after applying these scale factors is shown as a function of $m_H$ in Figure 4.29. Due to the large variations observed when using different underlying event parton showering (UEPS) models for the two dominant production processes, systematics of 70% and 10% are included as the uncertainty in the fraction of gluon-gluon fusion and vector boson fusion events respectively which are expected to pass the dijet tag [46]. In addition to the shape systematics, theoretical errors on the SM Higgs boson cross-section are included due to uncertainties on the QCD scale and pdf variations of the various production modes [50].

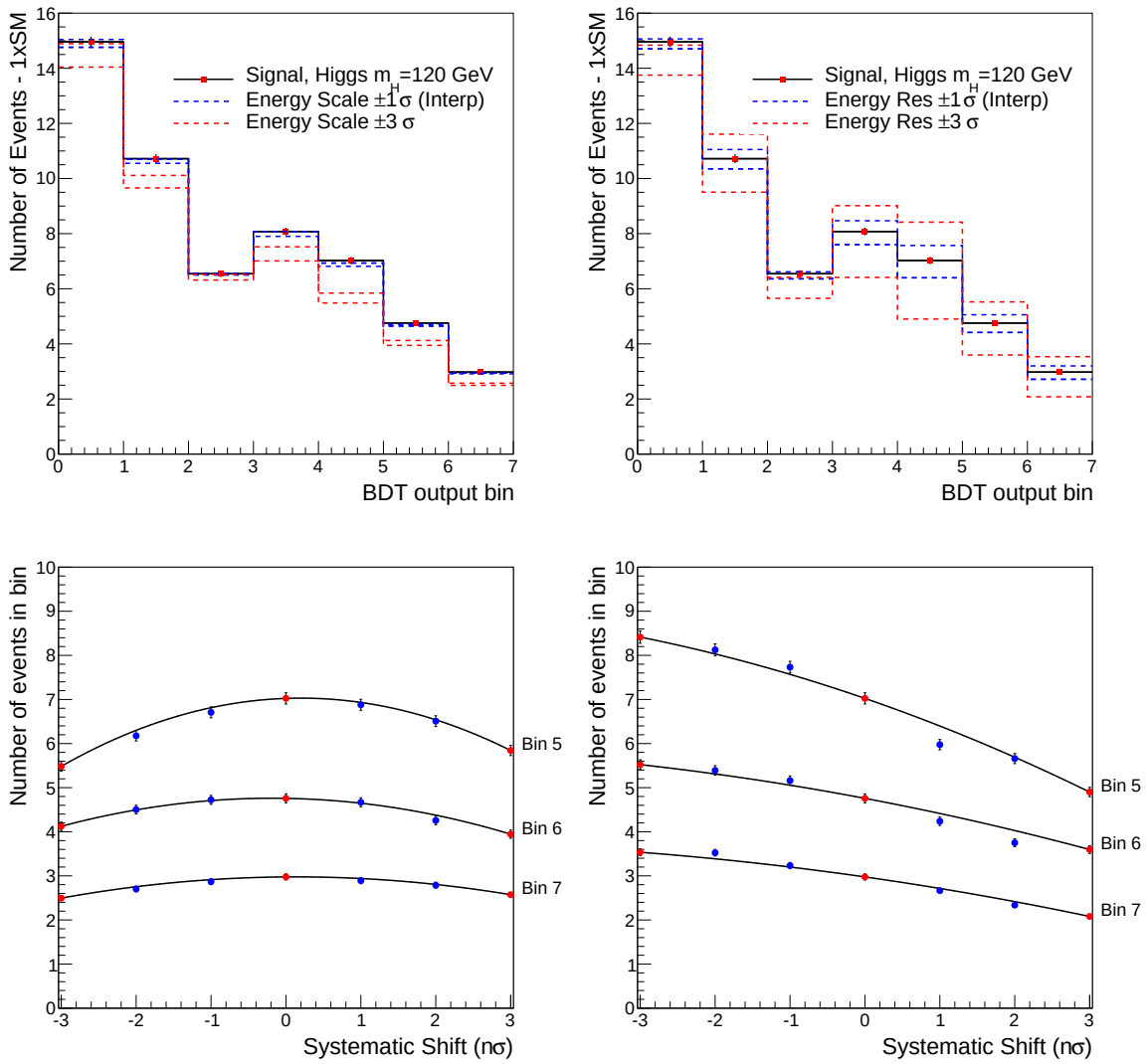


**Figure 4.28.:** Top: Energy scale (left) and resolution (right) uncertainties in the $ggH$ signal model. The effect of $\pm 3\sigma$ variations derived in MC are shown with red dashed lines while the interpolated $\pm 3\sigma$ are shown with blue. Bottom: Variation in bin content at different quantiles (number of standard deviations from the nominal) for the three highest $S/B$ BDT bins. The blue and red markers indicate the yields extracted directly from MC while the black line indicates the quadratic interpolation function used to derive the $\pm 1\sigma$ variations for the signal model.

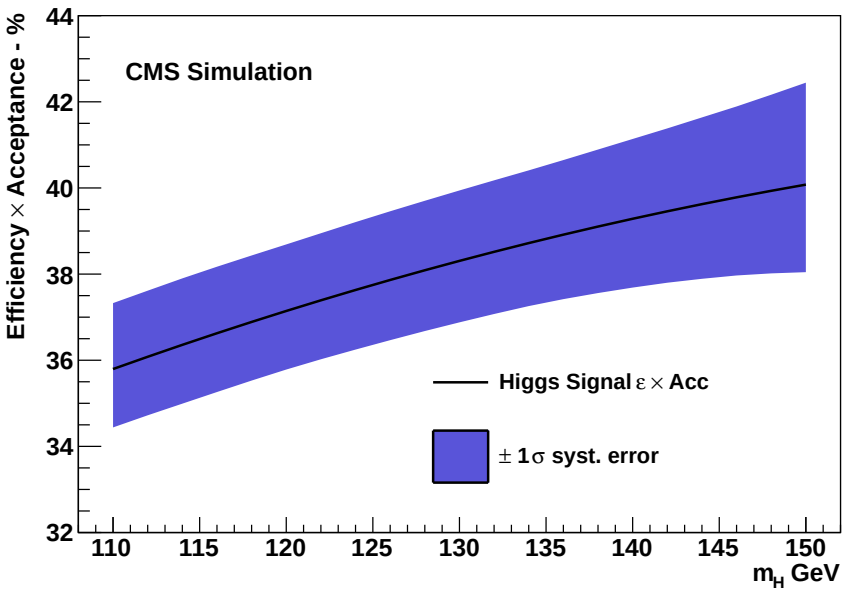


**Figure 4.29.:** Efficiency×acceptance for a SM Higgs boson as a function of its mass ($m_H$) after applying all of the corrections to the MC. The blue bands indicate the error from each source of systematic uncertainty on the signal model summed in quadrature.

A 2.2% luminosity error is also included as an uncertainty on the overall signal yield. A complete list of the systematics included in the signal model is given in Table 4.4.

### Interpolation to Intermediate Mass Points

Signal Monte Carlo is available in $m_H$ steps of 5 GeV in the range of 110 to 150 GeV. Due to the excellent resolution in the $H \to \gamma\gamma$ channel, it is necessary to interpolate between these generated mass points to construct the signal model at intermediate masses. As a result of selecting BDT input variables that are largely independent of the mass, the BDT output distribution in signal varies slowly and smoothly with $m_H$. This allows construction of the BDT output signal distribution at an intermediate mass point by performing a bin by bin vertical interpolation between the distributions from MC at neighboring mass hypotheses. The interpolation is performed separately for each signal production mode. The normalization at intermediate points is defined as the cross section times branching ratio, which is known for any $m_H$, for the intermediate mass multiplied by a linear interpolation of the acceptance times efficiency. A closure test on the interpolation procedure was performed by comparing the efficiency times acceptance

| **Source of systematic uncertainty** | | **Uncertainty** | |
|---|---|---|---|
| **Per photon** | | Barrel | Endcap |
| Photon identification efficiency | | 1.0% | 2.6% |
| Energy resolution | $r_9 > 0.94$ (low $\eta$, high $\eta$) | 0.22%, 0.61% | 0.91%, 0.34% |
| ($\Delta\sigma/E_{MC}$) | $r_9 < 0.94$ (low $\eta$, high $\eta$) | 0.24%, 0.59% | 0.30%, 0.53% |
| Energy scale | $r_9 > 0.94$ (low $\eta$, high $\eta$) | 0.19%, 0.71% | 0.88%, 0.19% |
| $(E_{data} - E_{MC})/E_{MC}$) | $r_9 < 0.94$ (low $\eta$, high $\eta$) | 0.13%, 0.51% | 0.18%, 0.28% |
| Photon identification MVA | | $\pm$0.025 (output shift) | |
| Photon energy resolution MVA | | 10% (output scaling) | |
| **Per Event** | | | |
| Integrated luminosity | | 4.5% | |
| Vertex finding efficiency | | $p_T^{\gamma\gamma}$-differential | |
| Trigger efficiency | either photon, $r_9 < 0.94$ in endcap | 0.4% | |
| | Other events | 0.1% | |
| Dijet-tagging efficiency | Vector boson fusion process | 10% | |
| Dijet-tagging efficiency | Gluon-gluon fusion process | 70% | |
| **Production cross-sections** | | Scale | PDF |
| Gluon-gluon fusion | | +12.5% -8.2% | +7.9% -7.7% |
| Vector boson fusion | | +0.5% -0.3% | +2.7% -2.1% |
| Associated production with W/Z | | 1.8% | 4.2% |
| Associated production with $t\bar{t}$ | | +3.6% -9.5% | 8.5% |
| **Scale and PDF uncertainties** | | $p_T^H$-differential | |

**Table 4.4.:** Sources of systematic uncertainties included in the signal model. Where a magnitude of the uncertainty from each source is given, the value represents a $\pm 1\sigma$ variation which is applied to the signal model.

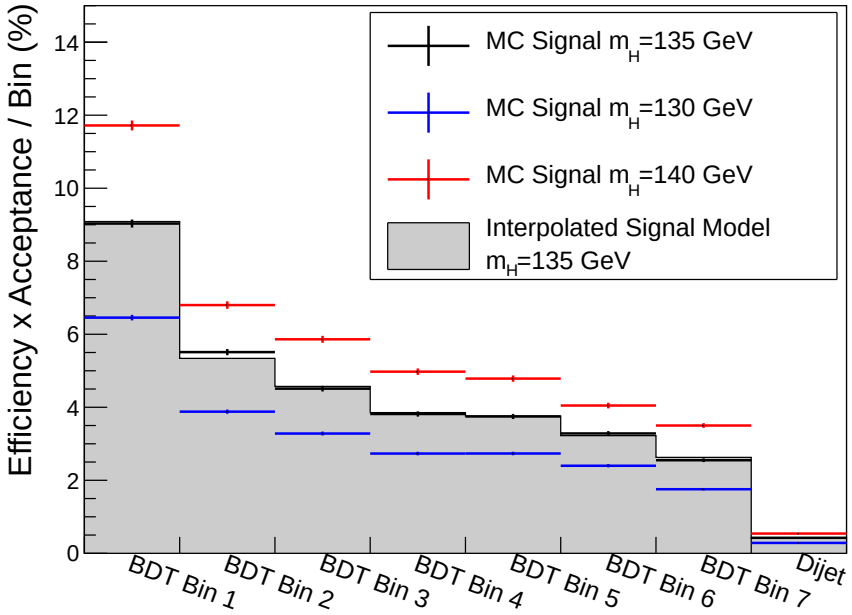


**Figure 4.30.:** Closure test for signal interpolation to intermediate mass points. The solid grey histogram is the result of a linear interpolation between the efficiency×acceptance in each bin of the blue ($m_H = 130$ GeV) and red ($m_H = 140$ GeV) histograms. The efficiency×acceptance from $ggH$ MC generated with mass 135 GeV is shown in black for comparison.

per bin at $m_H = 135$ GeV with one derived from gluon-gluon fusion MC generated with $m_H = 130$ GeV and $m_H = 140$ GeV (Figure 4.30). The closure test shows good agreement between the distributions; residual differences are negligible compared with the other systematics included in the signal model.

### Validation with $Z \to e^+e^-$ data

As with the other MVA discriminators in the $H \to \gamma\gamma$ analysis, the signal model is validated by running the BDT on both $Z \to e^+e^-$ MC and data with the electron veto inverted. A comparison of the data and MC is shown in Figure 4.31. Although the BDT output shape is not expected to be the same for $Z \to e^+e^-$ events as for $H \to \gamma\gamma$ events, the agreement seen between data and MC for $Z \to e^+e^-$ events indicates that the reconstruction and kinematics of a potential signal in data will be well modelled in the signal MC.

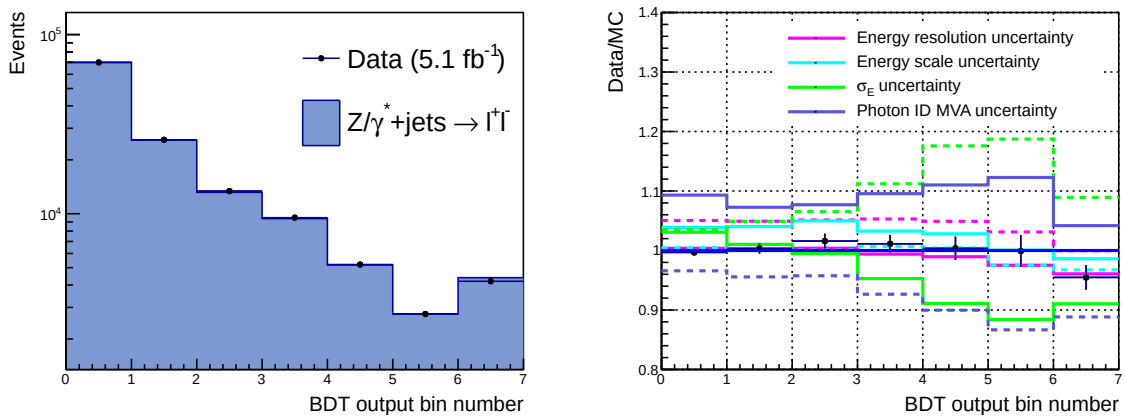


**Figure 4.31.:** BDT output distribution for $Z \to e^+e^-$ events in data and MC (left). Data/MC ratio for the BDT output distribution (right). The variation in MC due to the largest systematic uncertainties included in the signal model are shown for comparison.

### 4.4.6. Likelihood Model for Signal Extraction

The $H \to \gamma\gamma$ analysis was performed on the full 2011 dataset collected at CMS corresponding to $5.1 fb^{-1}$of proton-proton collision data at a centre of mass energy of 7 TeV. Figure 4.32 shows the observed number of events in data in each BDT output bin and from the dijet tagged events in the $\pm 2\%$ signal region centered on $m_H$ =125 GeV. The background model described in section 4.4.4 is shown in blue with the $\pm 1/2\sigma$ uncertainties represented by the coloured bands. The expected contribution from a SM Higgs boson with a mass of 125 GeV is shown in red. The full set of distributions for all mass hypotheses tested can be found online [64]. The signal to background ratio increases with higher BDT bin number, as shown in Figure 4.33, making the higher BDT bins more sensitive. The ratio is highest in the dijet tagged bin due to the additional suppression of the background by requiring two $qqH$ jets.

For the purposes of signal extraction, the analysis can be expressed in the form of a simple combination of counting experiments. The likelihood function (Equation 4.14) is proportional to a product of Poisson terms and parameterises the relative compatibility of the data with the signal and background models as a function of the signal strength $\mu$. The systematic uncertainties are included via the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}^s, \boldsymbol{\theta}^b)$ (nuisance

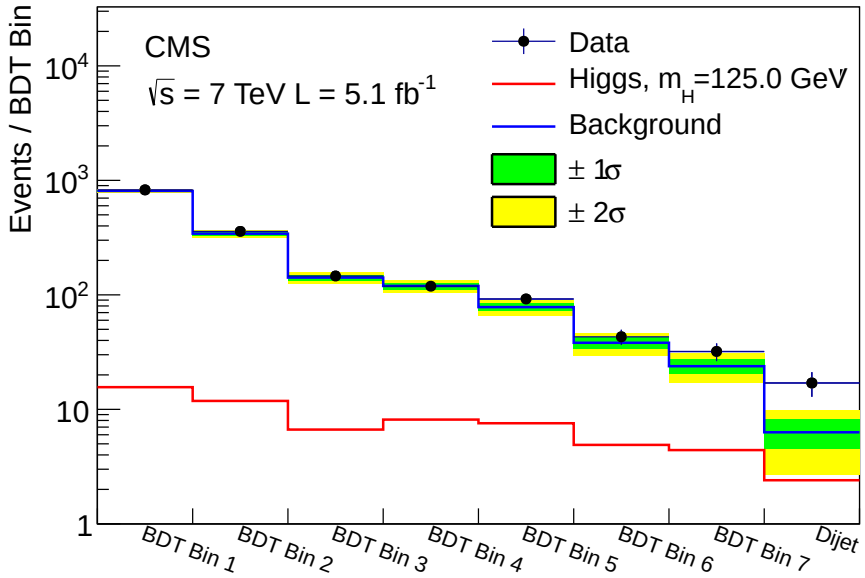


**Figure 4.32.:** Observed number of events in data for each of the seven BDT bins and dijet bin at $m_H = 125$ GeV. The background model is shown in blue along with the maximal $\pm 1/2\sigma$ variations. The expected contribution from a SM Higgs boson is shown in red [65].

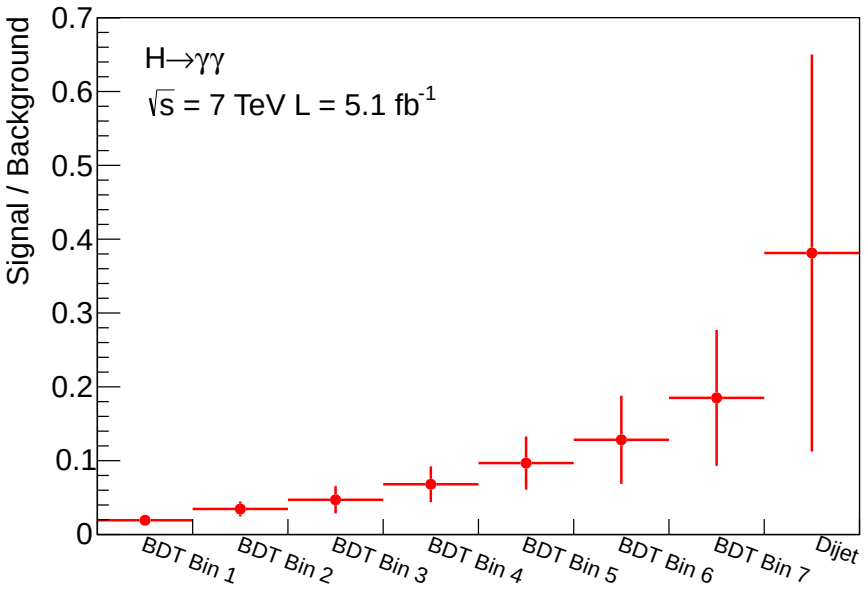


**Figure 4.33.:** Signal to background ratio in each of the seven BDT bins and dijet bin at $m_H = 125$ GeV. The expected background is taken from the data-driven model described in Section 4.4.4. The error bars represent the uncertainty in the ratio due to the uncertainties in the background model.

parameters) and $p$ is a product of unit width Gaussian distributions centered at $\boldsymbol{\theta}$.

$$\mathcal{L}(\text{data}|\mu,\boldsymbol{\theta}) = p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) \cdot \prod_{j=1}^{8} \text{Poisson}\left(d_j|\mu \sum_p s_j^p(\boldsymbol{\theta}) + b_j(\boldsymbol{\theta})\right) \tag{4.14}$$

The observed number of events in each bin, $d_j$, and expected contributions from each signal production process and background, $s_j^p$ ( $p\varepsilon\{ggH, qqH, VH, ttH\}$ ) and $b_j$, correspond to one mass hypothesis although the general form is applicable to all values of $m_H$.

To avoid cases in which the expectations for the contents of each bin become negative, the effect of each systematic on the signal or background is modelled using log-normal distributions. In this analysis, each systematic affects either the signal model or the background model. The functions $s_i(\boldsymbol{\theta}^s)$ and $b_i(\boldsymbol{\theta}^b)$ are given by Equations 4.15 and 4.16 respectively where $\boldsymbol{\theta}^s$ represents the nuisance parameters of the signal model and $\boldsymbol{\theta}^b = \left(\theta_N, \theta_1^b \ldots \theta_7^b\right)$ represent the eight independent nuisances of the background model.

$$s_j(\theta^s) = s_j^{p,mc} \cdot \prod_k \left(1 + \frac{\sigma_k^{s,p}}{s_j^{p,mc}}\right)^{\theta_k^s} \tag{4.15}$$

$$b_j(\theta^b) = N\left(1 + \frac{\sigma_N}{N}\right)^{\theta_N^b} \cdot f_j \prod_{k=1}^{7} \left(1 + \frac{\sqrt{\lambda_k} V_{kj}}{f_j}\right)^{\theta_k^b} \tag{4.16}$$

The values $s_j^{p,MC}$ in Equation 4.15 are the expected values for the signal from each of the four Higgs boson production processes ($ggH, qqH, wzH, ttH$) derived from the signal MC taking all MC to data corrections into account. The values of $\sigma_k^{s,p}$ are the correlated bin uncertainties of the signal model due to each independent source of uncertainty calculated using the quadratic interpolation described in Section 4.4.5. In practice, $\sigma_k^{s,p}$ has two values, one corresponding to positive values of $\theta_k^s$ and one for negative values. This is to account for asymmetric variations caused by uncertainties in the signal model such as that due to the energy scale. The values $V_{kj}$ and $\lambda_k$ in Equation 4.16 are the eigenvectors and corresponding eigenvalues of the covariance matrix determined in Section 4.4.4. Finally, $\sigma_N$ is the uncertainty on the background normalisation. This likelihood can be used as a statistical model of the data for the purpose of hypothesis testing, setting limits and quantifying excesses observed in the data.

# Chapter 5.

# Statistical Interpretations of the Data

When searching for new physics it is often desirable to do so in the context of some specific theoretical model or well motivated benchmark scenario. Where the theory provides well defined predictions, experimental data can be used to verify or reject the theory by means of hypothesis testing. This chapter describes the frequentist statistical procedures employed at CMS to perform these tests and provide quantitative, statistical interpretations of the data. In Section 5.1, an overview is given of the procedure by which $p$-values are calculated to test the compatibility of the data with a given hypothesis and the $CL_s$ technique for setting exclusion limits is introduced. The results of applying these procedures to the $H \to \gamma\gamma$ analysis described in Chapter 4 are given in Section 5.2 including updates for the 8 TeV using data taken in 2012 up to the ICHEP conference that year.

## 5.1. Hypothesis Testing

The goal is to use the data to reject one of two hypotheses, $H_0$ and $H_1$, known as the null and alternate hypotheses respectively. A function is defined, $t(data)$, which characterises the observed data as a single real value. When rejecting the hypothesis, $H_0$, the critical region, $w$, is defined as the set of possible values of $t$ which indicate that $H_0$ is not true. The probability then to observe $t \in w$ when $H_0$ is true ($\alpha$),

$$\alpha = P(t \in w | H_0) \tag{5.1}$$

is the probability that $H_0$ would be rejected even if it was true. The strength of a test (referred to as its power) is quantified by the probability, $1-\beta$ that $t \in w$ when $H_1$ is true.

$$1-\beta = P(t \in w|H_1) \tag{5.2}$$

In the case of the search of the SM Higgs boson, the two hypotheses can be parameterised in terms of a production cross-section relative to that predicted by the SM, $\sigma/\sigma_{SM}$. The null hypothesis ($H_0$) is then that under which no SM Higgs boson exists, $\sigma/\sigma_{SM} = 0$, while the alternate ($H_1$) is characterized by $\sigma/\sigma_{SM} = 1$. In this case, $H_0$ is referred to as the background-only hypothesis and $H_1$ the signal-plus-background hypothesis. The possible outcomes of $t$ are assumed to be random with a probability density function (pdf) $f_{\sigma/\sigma_{SM}}(t)$. The values of $\alpha$ and $1-\beta$ are the integral of the pdfs over the critical region;

$$1-\beta = \int_w f_1(t)dt \tag{5.3}$$

$$\alpha = \int_w f_0(t)dt. \tag{5.4}$$

It can be shown that the choice of $w$ which maximises the power of the test for a given value of $\alpha$ are the set of points for which,

$$q = \frac{f_1(t)}{f_0(t)} \geq c_\alpha, \tag{5.5}$$

where $c_\alpha$ is chosen such that Equation 5.4 holds [66]. In the search for the SM Higgs boson, the compatibility of the data with the presence of a Higgs boson is interpreted in terms of the continuous parameter, $\mu$, which scales the signal strength relative to that expected from the Standard Model. Again, the null hypothesis, $H_0$, is characterized by setting $\mu = 0$; however, an infinite number of alternate hypotheses exist for any value $\mu \geq 0$. The likelihood, $\mathcal{L}(t|\mu)$, is defined as a function of $\mu$ for a fixed realisation of the data and related to each pdf by a constant of proportionality. The quantity $q$ in Equation 5.5, known as the "test-statistic", is then the ratio of the likelihood at the two values $\mu = 1$ and $\mu = 0$.

The test-statistic for a given $\mu$ is defined as the ratio of likelihoods, $q_\mu$, in which the values for the nuisance parameters are taken from fits to the data (profiled), as given in

Equation 5.6.

$$q_\mu = \begin{cases} -2\ln\dfrac{\mathcal{L}(\text{data}|\mu,\hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} < 0 \end{cases}. \tag{5.6}$$

The values $\hat{\boldsymbol{\theta}}_\mu$ and $\hat{\boldsymbol{\theta}}$ are the values of the nuisance parameters for which the likelihood attains its maximum given a particular value of $\mu$ and letting $\mu$ float freely in the fit ($\mu = \hat{\mu}$). An immediate consequence of this definition is that the value attained by the test statistic is always positive. Small values of the test statistic indicate outcomes which are in favour of the signal-plus-background hypothesis, where large values indicate outcomes which disfavour it. Due to this, the critical region $w$ can always be defined as the right hand tail of the normalized distribution of the test-statistic $f(q_\mu|\mu)$,

$$w = \{q_\mu : q_{mu} \in (c_\alpha, +\infty)\}, \tag{5.7}$$

Commonly the integral of $f(q_\mu|\mu)$ above the observed value of the test-statistic in data ($q_\mu^{obs}$), known as a $p$-value, is calculated to provide a measure of how much the data disfavour a particular value of $\mu$.

### 5.1.1. Exclusion Limits

An upper limit ($\mu_{up}$) can be determined for $\mu$ such that the hypotheses represented by the set $\{\mu : \mu > \mu_{up}\}$ are nested (contained) within the hypothesis represented by $\sigma/\sigma_{SM} = \mu => \mu_{up}$. For the special case when $\mu_{up} = 1$, the presence of a SM Higgs boson is excluded (at some confidence level $c \in (0, 1)$) in favour of the background-only hypothesis. The constraint of $\hat{\mu} \leq \mu$ is imposed in the chosen test-statistic, $q_\mu$, when calculating upper limits, which forces the limit on $\mu$ to be one-sided. At CMS, upper limits are determined using the $CL_s$ which is designed to provide less stringent exclusion limits in analyses which are less sensitive to signal [67]. This procedure involves computing two $p$-values (tail probabilities) under two hypothesis, $\mu = 0$ and $\mu \neq 0$ given by,

$$CL_{s+b} = \int_{q_\mu^{obs}}^{\infty} f(q_\mu|\mu, \boldsymbol{\theta} = \boldsymbol{\theta}_\mu^{obs}) dq_\mu \tag{5.8}$$

$$CL_b = \int_{q_\mu^{obs}}^{\infty} f(q_\mu|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs}) dq_\mu, \tag{5.9}$$

where $q_\mu^{obs}$ is the value of the test-statistic in data. The smallest value of $\mu$ for which the ratio $CL_s = \frac{CL_{s+b}}{CL_b} < \alpha$ is quoted as the upper limit on $\mu$ with confidence level $1 - \alpha$. Typically, for exclusion, $\alpha$ is chosen as 0.05 so that when the upper limit on $\mu$ is less than one, all hypotheses of $\sigma/\sigma_{SM} \geq 1$ are excluded at the 95% confidence level or more. In the search for a SM Higgs boson, this corresponds to excluding a SM Higgs boson with a particular mass $m_H$ hypothesis under which the likelihood is constructed.

The distributions of the test-statistic, $q_\mu$, under the two hypotheses are generated by throwing pseudo-experiments using the signal and background models such as those derived in Section 4.4. First, the values of $\boldsymbol{\theta}_\mu^{obs}$ and $\boldsymbol{\theta}_0^{obs}$ are set by fitting the likelihood to the observed data at a particular value of $\mu$ and for $\mu = 0$ respectively. Pseudo-data, $d_j$, for each bin are generated according to a Poisson distribution with expectation value $\mu s_j(\boldsymbol{\theta}_\mu^{obs}) + b_j(\boldsymbol{\theta}_\mu^{obs})$. Pseudo-measurements for each nuisance parameter, $\tilde{\boldsymbol{\theta}}$, are then regenerated before evaluating the test-statistic $q_\mu$ in order to model the effect of systematic uncertainties. Examples of the normalised distributions of $q_\mu$ for $\mu = 0.6$ and $\mu = 0$ are shown in Figure 5.1.

### 5.1.2. Quantifying Excesses in the Observed Data

In the presence of a sizeable excess in data, the background-only hypothesis can be rejected in favour of a signal-plus-background-like one. Specifically, in the search for the SM Higgs boson, the excess will be compatible with the presence of a SM Higgs boson up to the rate at which it is produced. This is due to the inclusion of the signal model in the definition of the likelihood which typically includes the shape of the expected signal in some discriminating variable and the relative populations expected in different channels. In order to quantify the excess, the test-statistic is replaced with $q_0$ by setting $\mu = 0$ in Equation 5.6. The constraint $\hat{\mu} > 0$ ensures that only excesses in the data are considered significant. The background-only hypothesis is rejected in favour of a signal-plus-background one when the $p$-value $p_0$, given in Equation 5.10, is less than some pre-determined critical level $\alpha$.

$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs}) dq_0. \tag{5.10}$$

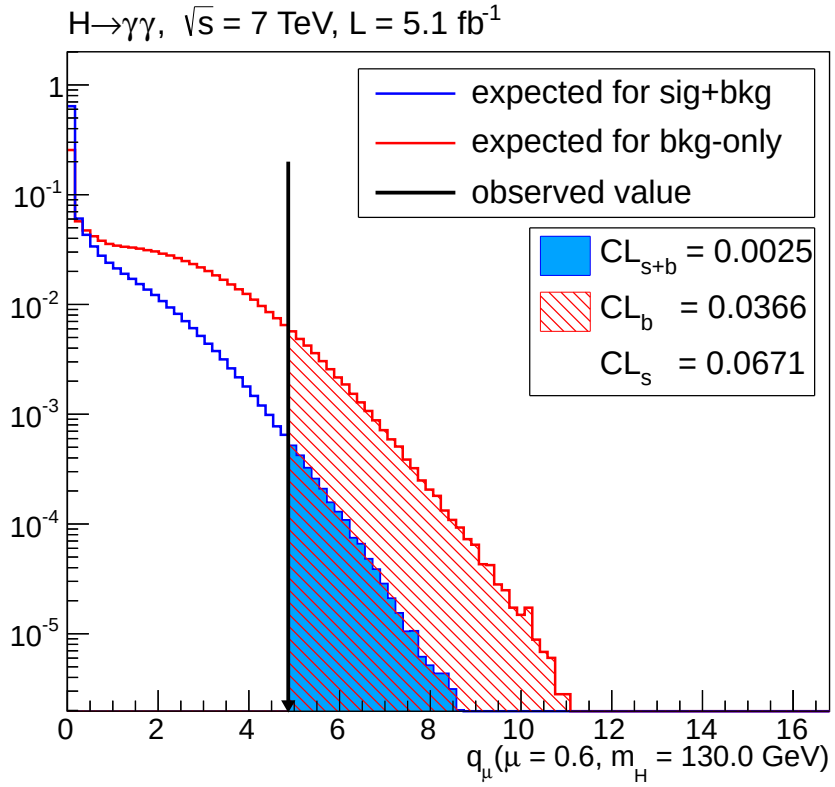


**Figure 5.1.:** Distributions of the test statistic $q_\mu$ under a background-only hypothesis ($\mu = 0$) and signal plus background hypothesis ($\mu = 0.6$) for a SM Higgs boson of mass 130 GeV. The distributions are normalised to unit area. The observed value of the test statistic from data is indicated by the black arrow.

Since $p_0$ is uniformly distributed between 0 and 1 under the hypothesis $\mu = 0$, $p_0$ is exactly the probability $\alpha$ of falsely rejecting the background-only hypothesis. The critical value for $\alpha$ is typically $2.87 \times 10^{-7}$ (corresponding to a significance of $5\sigma$) when searching for new physics.

In the case of the search for the SM Higgs boson, there is an implicit assumption that the test statistic is defined for a given value of $m_H$. The test statistic designed this way means that only excesses which are compatible in shape with that of a Higgs signal at some $m_H$ are considered significant. The $H \to \gamma\gamma$ signal is a narrow peak in the diphoton invariant mass distribution meaning only localised excesses in $m_{\gamma\gamma}$ are considered significant. The value of $p_0$ is therefore interpreted as the probability that the background can fluctuate to produce a localised excess; $p_0$ is termed the local p-value.

Analogous to calculating limits, the distribution $f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs})$ can be obtained either through generating toys or using an analytic form. Figure 5.2 shows the normalised distribution of $q_0$ under the background-only hypothesis generated from pseudo-experiments compared with the analytic form, in this case a $\chi^2$ distribution with a single degree of freedom, at $m_H = 124$ GeV.

## 5.2. $H \to \gamma\gamma$ Statistical Results

The likelihood in Equation 4.14 was coded using the `C++` based statistical package `RooFit/RooStats` version `5.3.0` [68]. A framework for automating the procedure of combining datasets, generating toys and evaluating likelihoods in the context of the combined search for the SM Higgs boson was developed within `CMSSW` under the package `HiggsAnalysis/CombinedLimit` [69]. All of the results shown in the following sections were obtained using this package.

The 95% confidence upper limits on $\sigma(H \to \gamma\gamma)/\sigma(H \to \gamma\gamma)_{SM}$ were determined using the full 2011 dataset for different values of $m_H$ in the range to which the channel $H \to \gamma\gamma$ is most sensitive. Since the resolution of the signal peak in the $H \to \gamma\gamma$ channel is of the order 1 GeV, the limit is calculated in 100 MeV steps in the range $110 < m_H < 150$ GeV. Figure 5.3 shows the expected and observed upper limit on the ratio $\sigma(H \to \gamma\gamma)/\sigma(H \to \gamma\gamma)_{SM}$ in that range. Where the observed line falls below the red line at one, a SM Higgs boson decaying to two photons, with mass $m_H$, is excluded at the 95% confidence level or more. The limits were calculated using an asymptotic

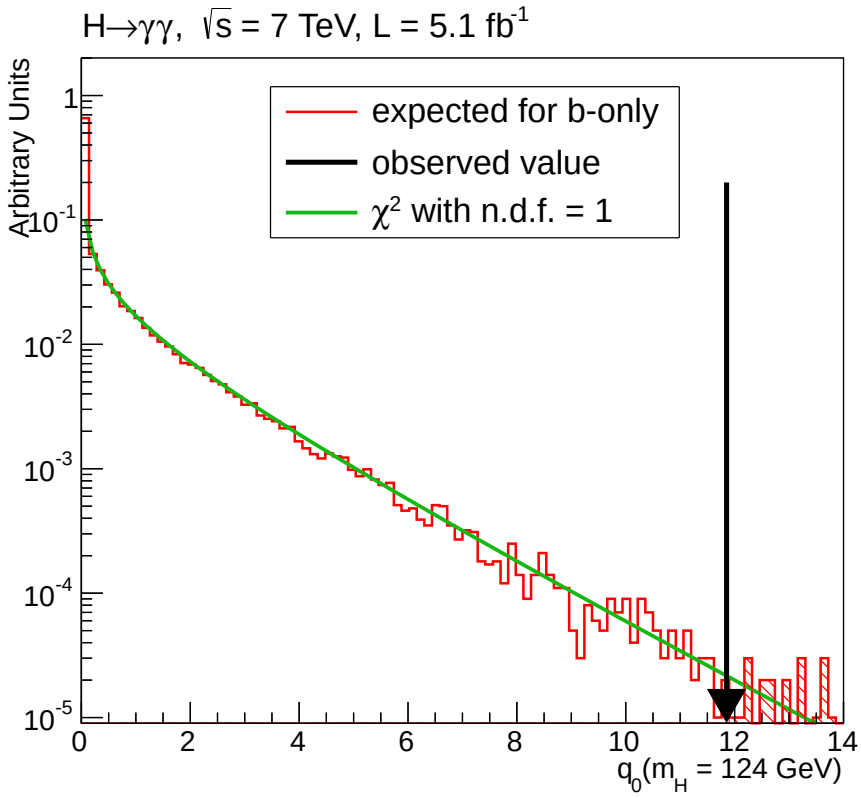


**Figure 5.2.:** Normalised distribution of $q_0$ at $m_H = 124$ GeV under the background-only hypothesis generated from toys (red histogram) and from the analytic form (green line). The observed value, $q_0^{obs}$, obtained from the data is indicated by the black arrow.

| | Toys | Asymptotic |
|---|---|---|
| $m_H = 120$ GeV | | |
| 2.5% | $0.534 \pm 0.044$ | 0.533 |
| 16% | $0.777 \pm 0.012$ | 0.778 |
| median | $1.175 \pm 0.020$ | 1.174 |
| 84% | $1.785 \pm 0.021$ | 1.795 |
| 97.5% | $2.592 \pm 0.213$ | 2.635 |
| $m_H = 130$ GeV | | |
| 2.5% | $0.629 \pm 0.051$ | 0.605 |
| 16% | $0.822 \pm 0.012$ | 0.798 |
| median | $1.149 \pm 0.019$ | 1.145 |
| 84% | $1.665 \pm 0.019$ | 1.663 |
| 97.5% | $2.349 \pm 0.192$ | 2.372 |
| $m_H = 140$ GeV | | |
| 2.5% | $0.855 \pm 0.070$ | 0.817 |
| 16% | $1.040 \pm 0.015$ | 1.001 |
| median | $1.361 \pm 0.022$ | 1.346 |
| 84% | $1.869 \pm 0.021$ | 1.849 |
| 97.5% | $2.540 \pm 0.208$ | 2.546 |

**Table 5.1.:** Comparison of expected median upper limit and quantiles obtained using the asymptotic calculation of $CL_s$ and toys. The error quoted in the toys column is the statistical uncertainty from only generating 1000 toys at each value of $\mu$. The comparison is made at three mass hypotheses in the range 120 to 140 GeV.

approximation for the distribution of $q_\mu$ thereby removing the need for generation of pseudo-experiments [70]. The procedure involving the generation of toys was however conducted for several mass hypotheses and found to agree with the asymptotic calculation. Table 5.1 shows this comparison for the median expected, 68% and 95% quantile ranges at different values of $m_H$.

The local p-value from the data is determined in steps of 100 MeV in the range $110 < m_H < 150$ GeV using the analytic expression $p_0 = \sqrt{q_0^{obs}}$ as shown in Figure 5.4. The expectation in the presence of a SM Higgs boson at each $m_H$ tested is shown in blue while the expectation from a SM Higgs boson with mass 125 GeV is shown in red. The largest excess in the range occurs near $m_H = 124$ GeV corresponding to a local significance of $3.4\sigma$. The excess is larger than expected in the presence of a SM Higgs

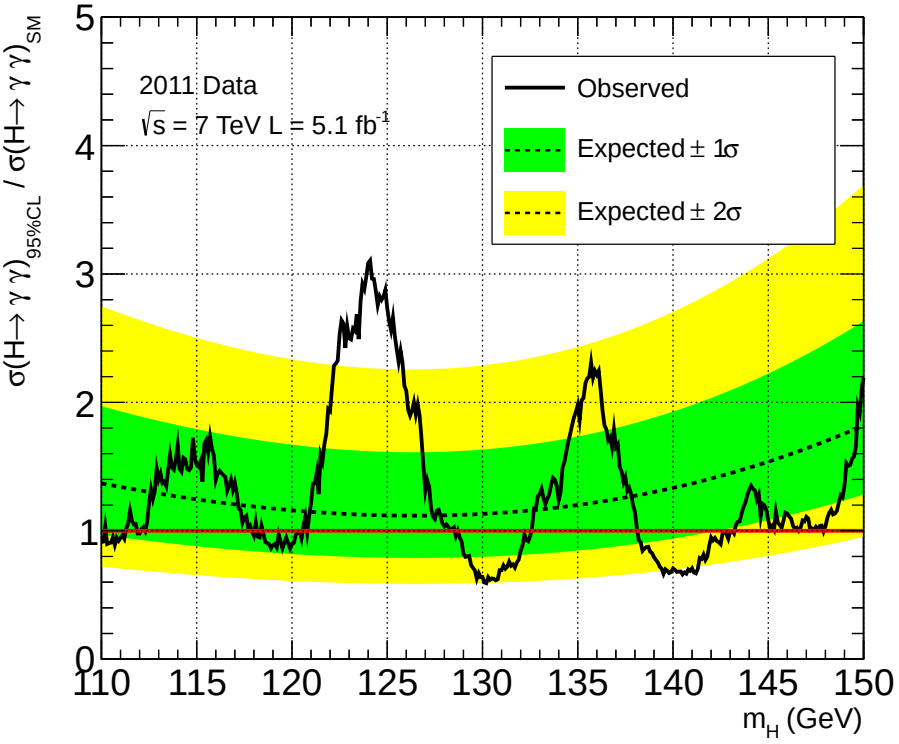


**Figure 5.3.:** Exclusion limits on SM Higgs boson production and subsequent decay to two photons in the range $110 < m_H < 150$ GeV. The black dashed line indicates the median expected value for the upper limit on $\mu$ given the size of the dataset while the green and yellow bands indicate the 68% and 95% quantile ranges respectively. The black solid line shows the observed upper limit extracted from the data at steps in $m_H$ of 100 MeV. Where this line falls below the red line at 1, a SM Higgs boson at that mass is excluded at the 95% confidence level or more.

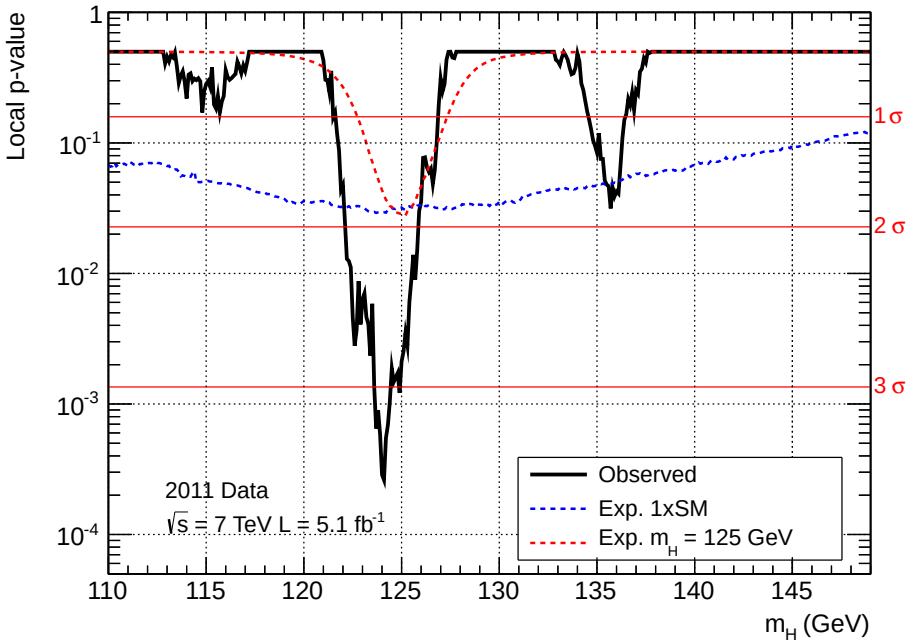


**Figure 5.4.:** Local p-value ($p_0$) calculated in steps of 100 MeV in the range $110 < m_H < 150$. The observed $p_0$ obtained from the data is shown in black while the expected value in the presence of a SM Higgs boson is given by the dashed blue line. The expectation from a Higgs boson with mass 125 GeV is shown as a red dashed line. The right hand scale shows the significance in standard deviations at each $m_H$.

boson near that mass. This is reflected in Figure 5.5 which shows the value of $\mu$ at which the likelihood attains its maximum, $\hat{\mu}$, as a function of $m_H$. The excess observed at 124 GeV corresponds to $\hat{\mu} = 1.93^{+0.67}_{-0.60}$, that is nearly twice the expectation from a SM Higgs boson.

### The Look-Elsewhere Effect

As the signal for the decay $H \to \gamma\gamma$ is a narrow mass peak, the probability to observe a local excess anywhere in the search range is much larger than the probability to find one at any particular $m_H$. This is an example of the look-elsewhere effect [71]. Due to this, the local p-value must be modified so as to express the probability to find an excess at least as significant as the one seen in data for all values of $m_H$. This is done by throwing background-only pseudo-experiments and finding the minimum $p_0$ across all values of $m_H$ searched over. The fraction of pseudo-experiments with a minimum $p_0$ less than the one observed in data is then the global p-value. Figure 5.6 shows the relationship

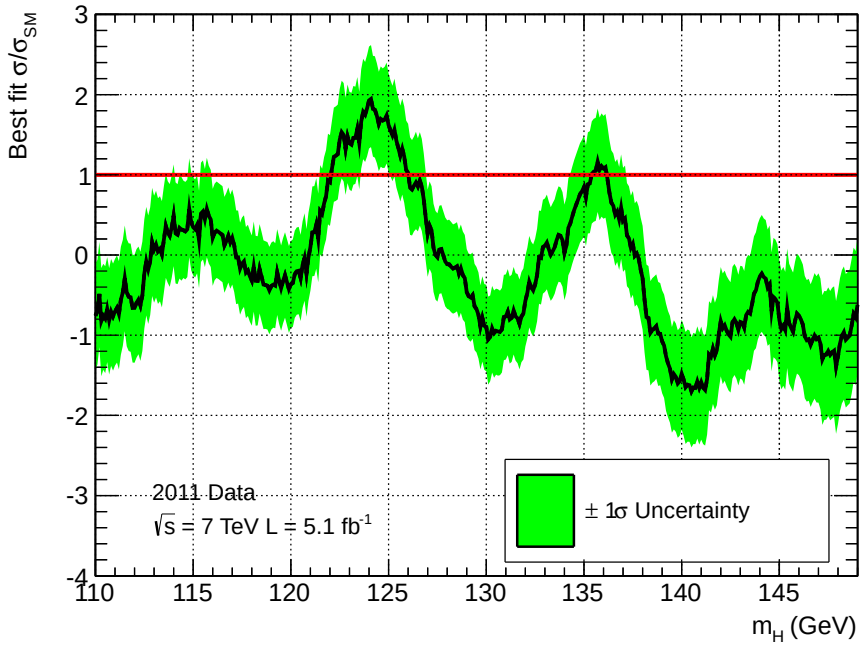


**Figure 5.5.:** Best fit for the signal strength, $\hat{\mu}$, in steps of 100 MeV in the range $110 < m_H < 150$. The green bands indicate the 68% uncertainty on $\hat{\mu}$ for a fixed $m_H$. The red line at 1 represents the expectation for a SM Higgs boson.

between local and global p-values. The red line shows a fit of the function,

$$p_{global} = p_{local} + Ce^{\frac{-Z^2}{2}}, \tag{5.11}$$

where $Z$ is the local significance and $C$ is a free parameter [72]. This function is then used to determine the look-elsewhere effect for larger significances. The excess observed at 124 GeV corresponds to a global significance of $2.4\sigma$.

In order to generate suitable background-only toys, pseudo-data are generated in two variables, $m_{\gamma\gamma}$ and the diphoton BDT output. The value of $m_{\gamma\gamma}$ for each event in the pseudo-data is generated from a sum of two power laws which is fit to the full $m_{\gamma\gamma}$ spectrum in data in the range $100 < m_{\gamma\gamma} < 180$ GeV. The value of the diphoton BDT is generated by fitting a kernel density estimator to the distribution in data. The value of $\Delta m/m_H$ is then calculated for each pseudo-event at every $m_H$ and the pseudo-dataset is analysed using the usual likelihood of Equation 4.14. This approach is necessary to maintain the correlations in the likelihood between neighbouring mass-hypotheses.

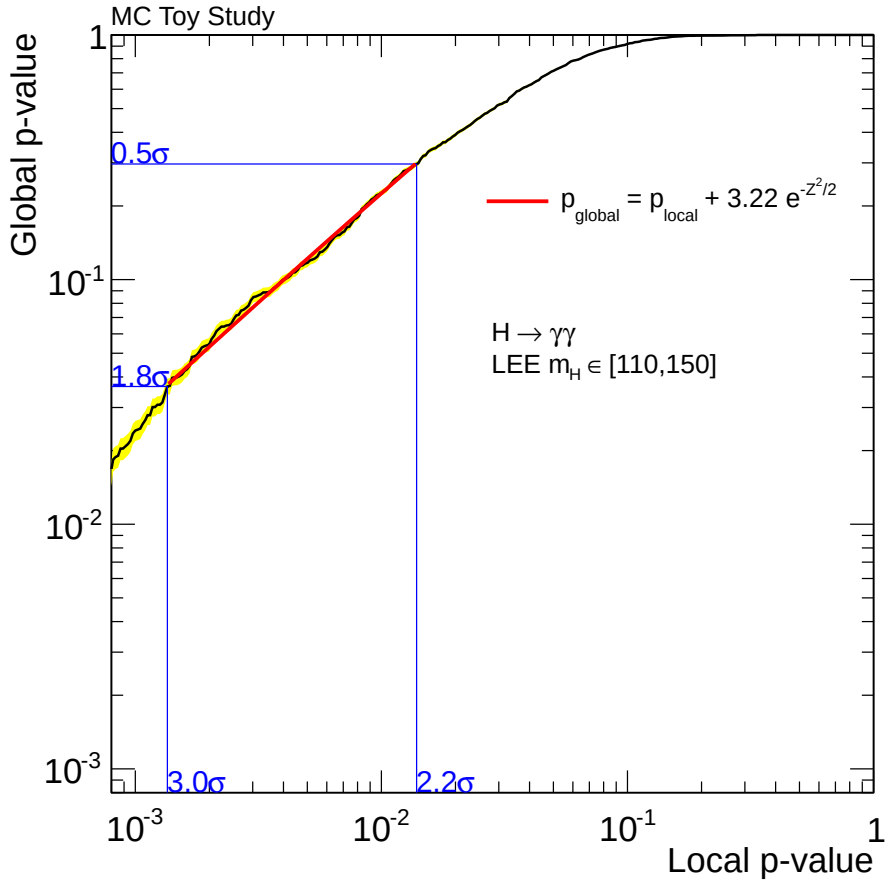


**Figure 5.6.:** Relationship between local and global p-values to determine the look-elsewhere effect in the $H \rightarrow \gamma\gamma$ search for the range 110 to 150 GeV. The yellow band indicates the statistical precision of the relationship due to the limited number of toys produced. The red line indicates a fit of an analytic relation between the two and is used to calculate the global p-value for larger local significances.

### 5.2.1. Inclusion of 2012 Data

The search described in Chapter 4 was repeated on data collected at CMS during the 2012 proton-proton run of the LHC, up to the time of the ICHEP conference in July 2012, at a center-of-mass energy of 8 TeV. The additional data were combined with the 7 TeV dataset as separate categories. The following section contains the results from the combined datasets corresponding to a total integrated luminosity of $10.4 fb^{-1}$ [73].

### 5.2.2. Updates for the 8 TeV Analysis

The majority of the analysis remains unmodified between the two data taking periods. Due to increased pileup conditions in the 2012 data, the regression BDTs and vertex BDTs were re-trained using MC weighted to a higher average number of pileup vertices. As a result of this, both the diphoton and event categorisation BDTs were re-trained to incorporate the changes. In addition, the slight variations in kinematics between centre-of-mass energies of 7 and 8 TeV are accounted for in the retraining. The isolation input variables to the photon ID BDT were modified, removing the correction for the average energy density in the event, $\rho$. Instead $\rho$ was introduced as an additional input variable so that the correlation between the number of pileup vertices and the isolation could be taken into account in the BDT training.

Both the energy scale and resolution were re-measured for the 2012 dataset and the corrections applied to data and MC as with the 2011 analysis. The invariant mass cut on the dijet system for the dijet tagged events category was reduced to 250 GeV, increasing the acceptance of $qqH$ production. The dijet events were further subdivided by separating events with a large reconstructed dijet mass, improving the sensitivity of the search. For the analysis described in Section 4.4, this results in two dijet bins with one being the tight class, containing dijet events with $m_{jj} > 500$ GeV and the other loose class containing all other dijet events. Figure 5.7 shows the observed number of events from the 2012 dataset in each of the BDT output bins and the two dijet categories for $m_H = 125$ GeV. The background model is derived using the procedure described in Section 4.4.4 using the 2012 dataset. The contribution expected from a SM Higgs boson is shown in red.

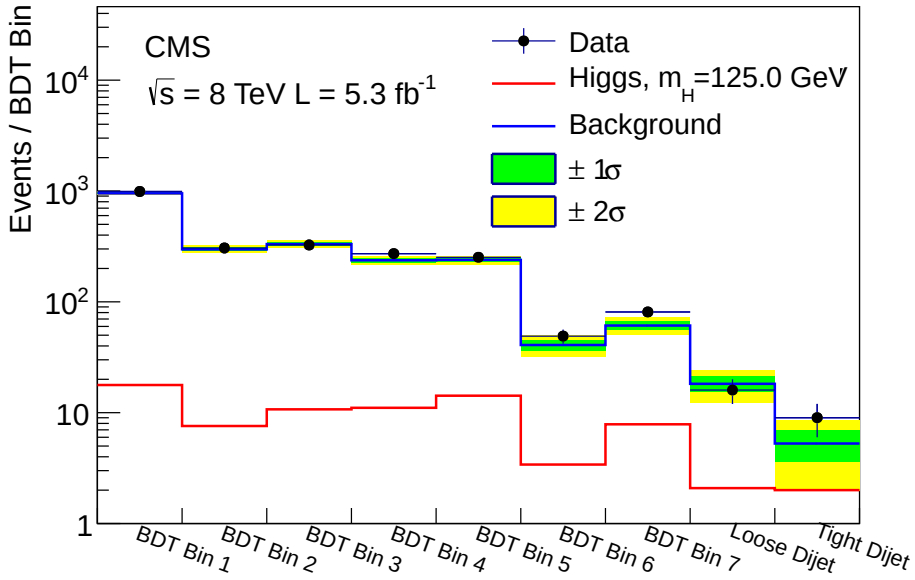


**Figure 5.7.:** Observed number of events in the 2012 dataset for each of the seven BDT bins and tight/loose dijet bins for $m_H = 125$ GeV. The background model is shown in blue along with the maximal $\pm 1/2\sigma$ variations. The expected contribution from a SM Higgs boson is shown in red [65].

### 5.2.3. Results from the Combined Datasets

The 2011 and 2012 datasets were combined statistically by extending the likelihood in Equation 4.14 to include a new set of categories which correspond to the updated analysis for the 2012 dataset. By including the additional data as separate categories, exclusion limits and p-values are calculated as described in Section 4.4.6. Figure 5.8 (left) shows the expected and observed 95% upper limits on $\sigma(H \to \gamma\gamma)/\sigma(H \to \gamma\gamma)_{SM}$ calculated in half GeV steps in $m_H$ from the combined datasets. The observed local p-value, $p_0$, determined for the 7 TeV, 8 TeV and combined datasets, as a function of $m_H$ is shown in Figure 5.8 (right). The largest excess is observed at $m_H = 124$ GeV, corresponding to a local significance of $4.8\sigma$. This is reduced to a global significance of $3.9\sigma$ when considering the look-elsewhere effect in the range 110 to 150 GeV.

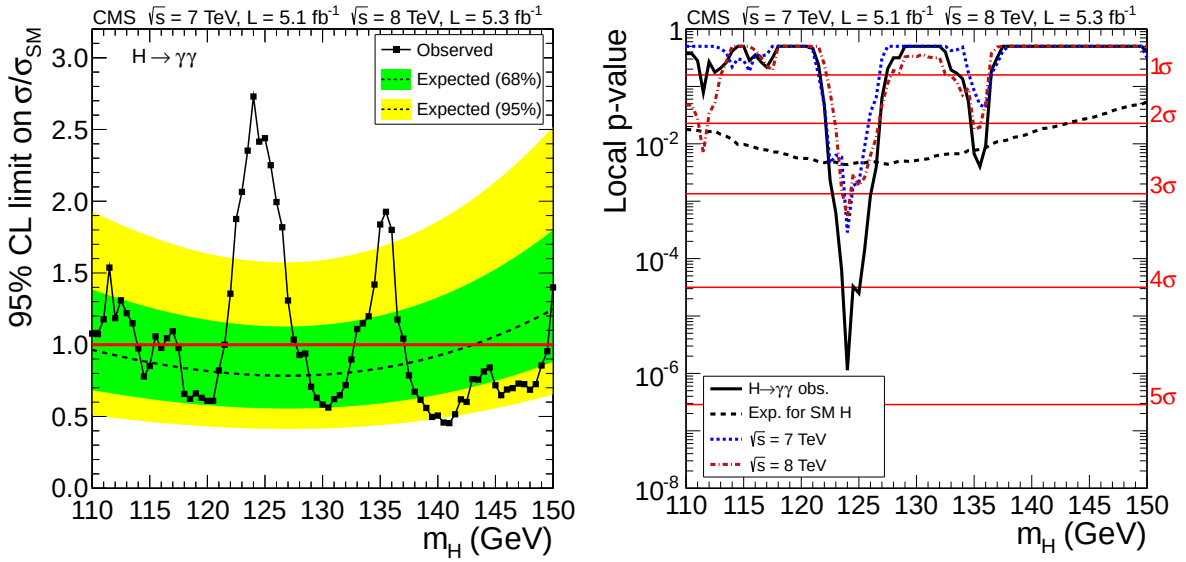


**Figure 5.8.:** Exclusion limits on SM Higgs boson production and subsequent decay to two photons (left) and local p-value, $p_0$ (right) in the range $110 < m_H < 150$ GeV from the combined 2011 (7 TeV) and 2012 (8 TeV) datasets. In the left figure, the black dashed lines indicates the median expected value for the upper limit on $\mu$ given the size of the dataset while the green and yellow bands indicate the 68% and 95% quantile ranges respectively. The black solid line shows the observed upper limit. In the right figure, the observed $p_0$ obtained from the combined datasets is shown in black while the expected value in the presence of a SM Higgs boson is given by the black dashed line. The observed $p_0$ from the 2011 (7 TeV) and 2012 (8 TeV) datasets individually are shown by the blue and red dashed lines respectively. The right hand scale shows the significance in standard deviations at each $m_H$ [65].

# Chapter 6.

# Higgs Combination and Properties

The sensitivity of the search for the SM Higgs boson depends not only on the production cross-section and branching ratio to a particular decay channel, but also on the efficiency of the selection, the experimental resolution and the relative proportions of signal to background processes. These quantities typically vary greatly as a function of $m_H$. By combining results from searches in many decay channels, over a large range in mass, the overall sensitivity is greatly improved. This chapter describes the combined searches for the SM Higgs boson and measurements of its properties. In Section 6.1, a short review on the techniques used for statistical combination of data from different analyses at CMS is provided and a set of diagnostic tools developed by the author are discussed. The results of the search using the ICHEP 2012 dataset, which led to the announcement of the discovery of a new particle by ATLAS and CMS in July 2012 [73], are included. Section 6.2 deals with early studies of the properties of the newly discovered particle presented at the Hadron Collider Physics (HCP) symposium in November 2012. This includes a discussion of the Feldman-Cousins technique which was implemented and performed by the author to extract information on the compatibility of the new state with the SM Higgs boson.

## 6.1. Combined Higgs Searches

In order to combine data from all decay channels relevant in the search for the SM Higgs boson, the likelihood for a particular outcome of the data given a particular value of $\mu$ is

the product of the individual likelihoods in each channel $i$,

$$\mathcal{L}(\text{data}|\mu,\boldsymbol{\theta}) = p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) \cdot \prod_i \mathcal{L}_i(\text{data}_i|\mu,\boldsymbol{\theta}) = \prod_i P(\text{data}|\boldsymbol{\theta}). \tag{6.1}$$

The relative signal strength $\mu$ is a single parameter which scales the signal yield in all sub-channels simultaneously. Systematic uncertainties in the signal and background models in each channel are modelled through the nuisance parameters, $\boldsymbol{\theta}$. Typically these nuisances will be constrained by some external measurements $\tilde{\boldsymbol{\theta}}$, such as the energy scale measured in $Z \to e^+e^-$ events in the two-photon decay channel described in Section 4.2.2. The term $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ is a product of the pdfs of each nuisance parameter, which are usually Gaussian distributions, for each independent source of systematic uncertainty. Although each event observed in data is exclusive to a particular channel, many sources of systematic uncertainty are common to several analyses. For this reason, some of the nuisance parameters are correlated between sub-channels.

The likelihood was coded using the `C++` based statistical package `RooFit/RooStats` version `5.3.0` [68]. As with those shown in Chapter 5, all of the results in the following sections were obtained using the `CMSSW` package `HiggsAnalysis/CombinedLimit` [69].

### 6.1.1. Diagnostics with Toy Datasets

Frequentist statistical techniques often involve generating many pseudo-datasets (toys) to build the distribution of a test-statistic. These distributions are used to set confidence intervals or determine the significance of some observed excess in experimental data. The combined Higgs searches at CMS employ the profiled likelihood test-statistic (Equation 5.6) in which the nuisance parameters, $\boldsymbol{\theta}$, are profiled. For calculating the significance of an excess in data, the distribution of the test-statistic $q_0$ under the background-only hypothesis is required. The procedure for determining this distribution proceeds as follows;

- Fit the observed data fixing $\mu = 0$. The values of the nuisance parameters at which the likelihood attains its maximum are denoted $\boldsymbol{\theta}_{obs}$.
- Generate a toy dataset under the background-only hypothesis. For the purposes of generating data, the nuisance parameters are fixed to $\boldsymbol{\theta} = \boldsymbol{\theta}_{obs}$.

- Fit the toy dataset twice, once fixing $\mu = 0$, $\mathcal{L}(\text{data}|0, \hat{\boldsymbol{\theta}}_0)$ and once more letting $\mu$ float freely, $\mathcal{L}(\text{data}|\hat{\mu}, \hat{\boldsymbol{\theta}})$. When evaluating the likelihood, the values of $\tilde{\boldsymbol{\theta}}$ are randomly generated from their pdfs (producing "pseudo-measurements") in order to model the systematic uncertainties.

As the values of the nuisances are profiled from the data, it is important to check for additional correlations between sub-channels which may not have been properly accounted for when building the background and signal models in each channel.

A realistic example of a search for an hypothesised particle, $H$, decaying to two $\tau$ leptons was produced in the form of a simple counting experiment. The search is performed as a combination of three channels arising from the possible subsequent decays of the two tau-leptons; $\tau_h e$, $\tau_h \mu$ and $e\mu$, where $\tau_h$ denotes a hadronically decaying tau-lepton. Events are categorised according to the channel in which they are reconstructed. In each category, the expected background is estimated either from simulation or some control region in data. The data are represented by the number of events observed in each category. Several sources of systematic uncertainty are included which affect the expected signal and background in one or more of the channels. Systematics are incorporated into the likelihood in the form of nuisance parameters as described previously. The analysis is summarized in Table 6.1 which details the number of expected events from each background and signal process in each channel as well as the observed count in data. The analysis is available from the `CMSSW` package `HiggsAnalysis/CombinedLimit` under the directory `data/tutorials/realistic-counting-experiment.txt` [69].

Around 90,000 toy datasets were generated under the background-only hypothesis as is appropriate for determining the distribution of $q_0$. Each toy dataset was fit twice, once fixing $\mu$ to zero and a second allowing $\mu$ to float freely. The results of the fits are used to diagnose the fits and highlight potentially problematic channels or nuisance parameters. Figure 6.1 shows a summary of the fit results in the nuisance parameter `lumi`, which models the systematic uncertainty associated with the total luminosity measurement. The upper left panel shows two pull distributions of the values from the fit. The entries are calculated as the difference between the value of the fitted parameter and the value from the best fit to data, $\theta_{obs}$, divided by the $1\sigma$ uncertainty on the parameter before fitting to the data. The blue histogram includes all toys while the red shows the results for toys in which the best fit signal strength is positive. Since the test-statistic $q_0$ is designed to report only excesses in the data, it is important to check that nuisance parameters correlated to the signal strength are well behaved.

| **channel** | $\tau_h - e$ | | | $\tau_h - \mu$ | | | $e - \mu$ | | |
|---|---|---|---|---|---|---|---|---|---|
| observed | 517 | | | 540 | | | 101 | | |
| expected | Sig | $Z \to \tau\tau$ | QCD | Sig | $Z \to \tau\tau$ | QCD | Sig | $Z \to \tau\tau$ | other |
| | 0.34 | 190 | 327 | 0.57 | 329 | 259 | 0.15 | 88 | 14 |
| **systematics** | | | | | | | | | |
| luminosity | 11% | - | - | 11% | - | - | 11% | - | 11% |
| tau ID | 23% | 23% | - | 23% | 23% | - | - | - | - |
| $Z \to ll$ norm | - | 4% | - | - | 4% | - | - | 4% | - |
| signal eff | 4% | 4% | - | 4% | 4% | - | 4% | 4% | 4% |
| $e$ fake rate | - | - | 20% | - | - | - | - | - | - |
| $\mu$ fake rate | - | - | - | - | - | 10% | - | - | - |
| other | - | - | - | - | - | - | - | - | 10% |

**Table 6.1.:** A realistic counting experiment across several channels. The number of observed events and that expected from signal and background processes are given per channel. Several sources of systematic are included which effect the expected rate of each signal or background process. Where a dash is entered, the systematic uncertainty has no effect on that process or channel.

The pull distributions are fitted with a Gaussian and the width and mean are reported in the upper right panel. Since the pseudo-measurements are generated around the best fit to data, the pull distributions are expected to be centered around 0. In general, nuisance parameters are constrained from external measurements so it is expected that the width of the pull distribution is 1. Nuisance parameters which are further constrained by the observed data will typically have a pull distribution with width less than unity. The parameter `lumi` does not show signs of being constrained by the data. This is reflected in the lower left panel which shows the correlation between the nuisance parameter and the value generated for the pseudo-measurement of this parameter (`lumi_In`) in each toy. This behaviour is expected since this nuisance parameter mostly affects the signal process and is correlated across all channels so that only the overall normalization is altered. Since these fits allow $\mu$ to float freely, any parameter which alters only the overall normalization of the signal should achieve the value generated for its pseudo-measurement at the maximum of the likelihood. The lower right panel shows the shape of the negative log-likelihood as a function of the nuisance parameter ($\theta$),

$$-\log \frac{\mathcal{L}(\text{data}|\mu = \hat{\mu}, \theta = \theta_{S+B})}{\mathcal{L}(\text{data}|\mu = \hat{\mu}, \theta)}, \tag{6.2}$$

near its minimum value ($\theta_{S+B}$). At each point, all other parameters are fixed to those of the best fit to the data (in this case, from the fit allowing $\mu$ to float). The likelihood is expected to be parabolic around its minima with no secondary (local) minima present. Degenerate minima, which can cause instabilities in the fitting procedure, will be visible in the shape of the negative log-likelihood. The diagnostic tools described here were applied to the ICHEP 2012 combination for $m_H = 125$ GeV as documented in a CMS Analysis note by the author [74]. The full set of diagnostic summary plots can be found online [75].

### 6.1.2. Higgs Search Combination

A search for the SM Higgs boson was performed by combining data recorded at CMS at a centre of mass energy of 7 and 8 TeV. The search was performed in five decay modes, $H \to \gamma\gamma$, $H \to ZZ$, $H \to WW$, $H \to \tau\tau$ and $H \to bb$ with datasets of integrated luminosities of $4.9-5.1\ fb^{-1}$ and $5.1-5.3\ fb^{-1}$ from the 2011 and 2012 data taking periods of the LHC respectively. The search is performed across a wide range in Higgs boson mass hypothesis ($m_H$) from 110 to 600 GeV. For $m_H > 150$ GeV, the $H \to \gamma\gamma$, $H \to \tau\tau$ and $H \to bb$ decay modes are not used as they are significantly less sensitive than the $H \to WW$ and $H \to ZZ$ channels. Figure 6.2 shows the relative sensitivities of each decay channel in terms of the expected exclusion limit for the size of the dataset used. The exclusive final state topologies in each of the five modes used in the combination, including the size of the dataset used and the mass range to which they are sensitive, are given in Table 6.2.

#### Combined Search Channels

The $H \to \gamma\gamma$ analysis is one of the most sensitive channels at low $m_H$. The analysis is that described in Chapter 4 with the exception of the signal extraction method; Method A is used in the combination. Events are categorized using the diphoton BDT into four classes chosen so as to optimise the search in terms of the expected limit at $m_H = 125$ GeV. The diphoton invariant mass spectrum in each category is fit with polynomial functions whose order is determined following a procedure designed to reduce any potential bias to less than 20% of the statistical uncertainty on the background [77]. The dijet selected events are categorised separately and treated in the same way as the inclusive events.

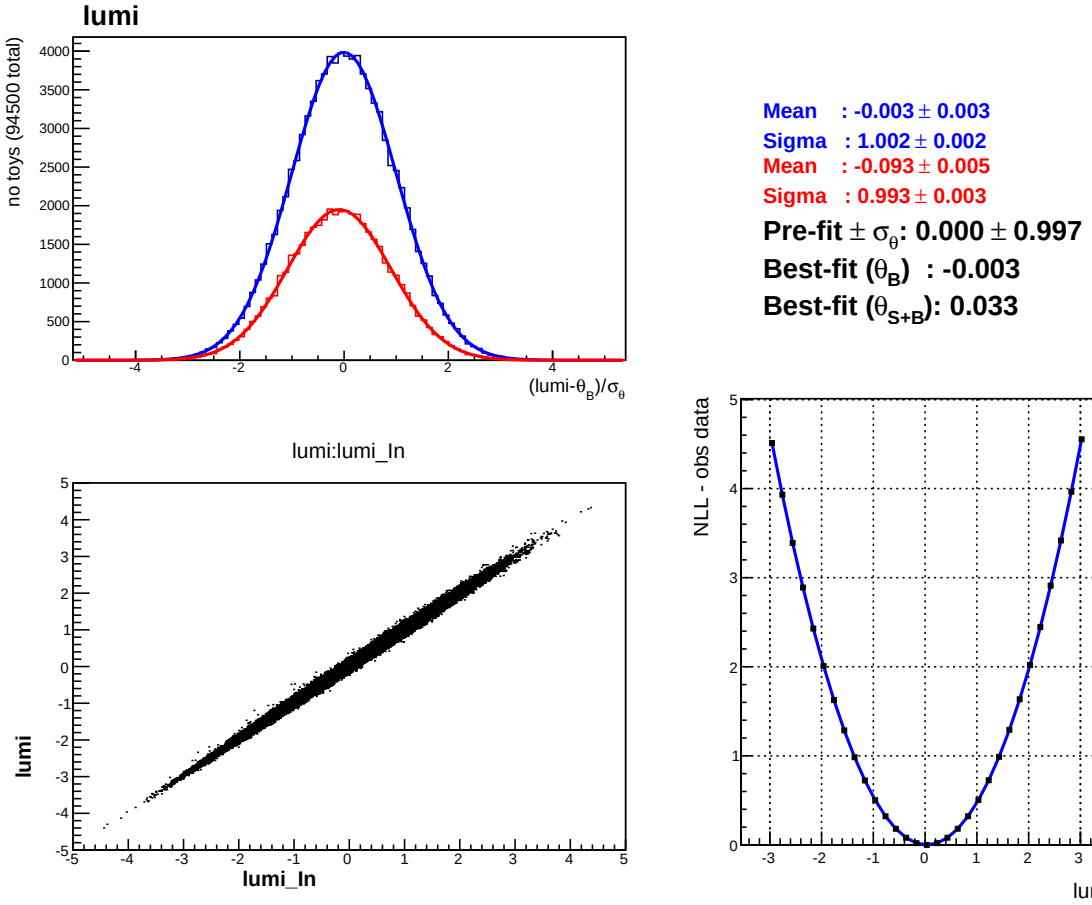


**Figure 6.1.:** Summary plots for the parameter `lumi` of the realistic counting experiment. The entries in the histograms are for fits to toys generated under the background-only hypothesis letting $\mu$ float freely. The red histogram includes only toys in which a positive signal strength is fitted. The bottom left panel shows the correlation between the value generated for the pseudo-measurement of the nuisance `lumi_In` and the fitted value of the parameter. The bottom right panel shows the shape of the negative log-likelihood (NLL) as a function of the nuisance parameter. The parameters of the fitted Gaussian for each histogram are given as the Mean and Sigma. The value and error of the nuisance parameter are given before fitting to the data (Pre-fit), followed by the best fit value of the parameter under the background-only and signal-plus-background hypotheses.

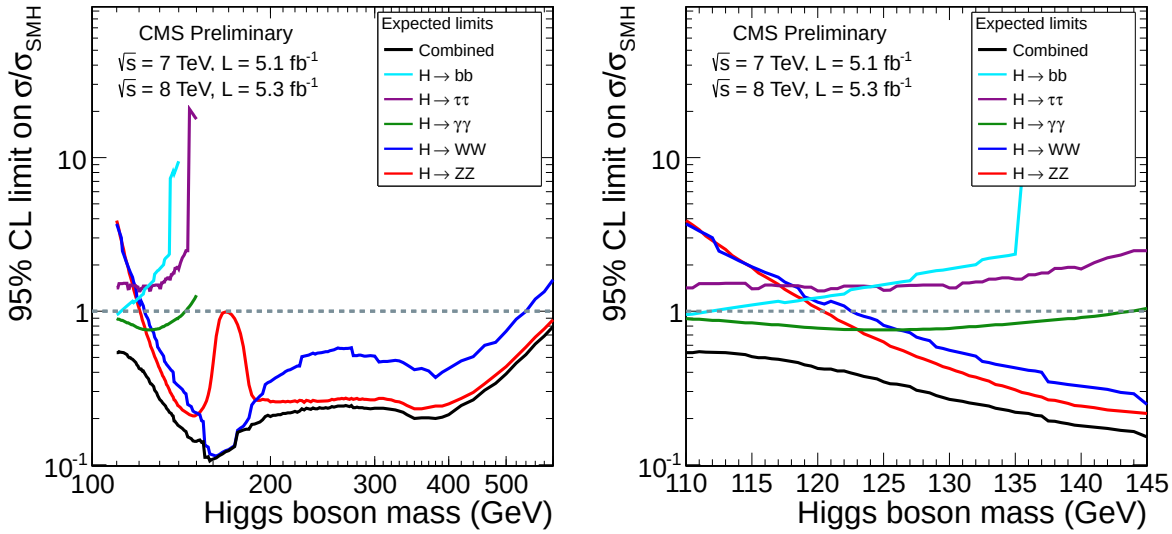


**Figure 6.2.:** Median expected 95% CL upper limits on $\mu = \sigma/\sigma_{SM}$ for the five Higgs boson decay channels and their combination in the absence of a Higgs boson as a function of $m_H$. The limits are given in the range 110-600 GeV (left) and 110-145 GeV (right). A channel which falls below 1, indicated by the dashed line, for some range is expected to exclude a Higgs boson in that range at the 95% CL or more using this dataset [76].

Due to the extremely high cross-section of $b\bar{b}$ production in p-p collisions, the $H \to bb$ search focuses on Higgs boson production in association with a $W$ or $Z$ boson, which are identified by the presence of leptons. In the case of neutrino final states such as $Z \to \nu\nu$, the missing transverse energy (${E_T}^{miss}$) of the event is required to be large, indicating momentum propagated out of the detector by the neutrinos [78]. This quantity is defined as the magnitude of the negative vector sum over all energy deposits, $E_k$, in the calorimeters projected into the transverse plane $(\hat{\mathbf{i}}, \hat{\mathbf{j}})$,

$$ {E_T}^{miss} = |\mathbf{E}_T^{miss}| = | \sum_{deposits} - \sin\theta_k \left( E_k \cos(\phi_k)\hat{\mathbf{i}} + E_k \sin(\phi_k)\hat{\mathbf{j}} \right) | \tag{6.3} $$

The Higgs boson candidate itself is reconstructed by looking for two $b$-tagged jets indicated by their production at secondary vertices. Events are categorized into those where the $W$ or $Z$ boson is recoiling away from the $b\bar{b}$ system with high momentum. The main backgrounds are from $W/Z$+jets and $t\bar{t}$ as well as from $WZ$ and $ZZ$ in which the $Z$ decays to a pair of $b$-quarks. The backgrounds are suppressed by use of a multivariate analysis technique trained on MC simulation. The search is also performed in events in which the Higgs boson is produced in association with a pair of top-quarks ($ttH$)

categorized into either lepton plus jet or dilepton final states [78]. This mode was not included in the 8 TeV dataset for the ICHEP 2012 combination.

In the $H \to \tau\tau$ decay channel, the search is performed using events with leptonic final states and events in which one of the tau-leptons decays hadronically ($\tau_h$) [79]. Events are divided into categories based on the number and type of jets in the event and by the transverse momentum of the visible part of the tau decay. A signal in this channel will be visible as a broad excess in the invariant mass of the $\tau\bar{\tau}$ system ($m_{\tau\tau}$). The main backgrounds are from $Z \to \tau\tau$ events and $W$+jet production. Production of $qqH$ is tagged by the association of two jets consistent with those resulting from vector boson fusion. Finally, the $VH$ modes are exploited by selecting events which have one or more additional leptons consistent with a $W$ or $Z$ boson decay.

The $H \to WW$ analysis is one of the most sensitive analyses at CMS for values of $m_H$ between 150 and 200 GeV [80]. The $WW \to 2l2\nu$ sub-channel consists of events with two oppositely charged leptons, a large ${E_T}^{miss}$ and up to two jets (to target $qqH$ production). The events are further divided into categories in which the two leptons are of the same or opposite flavour to exploit the different background contributions from $Z$ decays. For the 7 TeV analysis, an MVA classifier was trained on signal and background MC to separate signal from background. The search is conducted by looking for an excess of events in the output distribution of the MVA. In the $WW \to l\nu2q$ sub-channel, a broad excess is searched for in the four-body invariant mass spectrum [81]. The invariant mass is reconstructed from the lepton four-vector and $\mathbf{E}_T^{miss}$ assuming the mass of the $l\nu$ is that of a $W$ boson and choosing the neutrino's longitudinal component to be that which minimises the transverse momentum of the $l\nu$ system. Associated production of the Higgs boson with a $W$ boson is searched for by looking for an excess of events with three leptons and large ${E_T}^{miss}$ [82].

The $H \to ZZ$ analysis focuses on four final state topologies. The $ZZ \to 4l$ is a search for a narrow four-lepton invariant mass peak over a small background [83]. The kinematics of the $4l$ system are used to assign a probability that the event is from either a signal or background process to improve the sensitivity. For the lower mass region ($m_H < 180$ GeV), only one of the lepton pairs is required to have a mass consistent with an on-shell $Z$ boson. The $4e$, $4\mu$ and $2e2\mu$ sub-channels are categorised separately as the mass resolutions and the background rates differ between the three final states. In the $ZZ \to 2l2\tau$ and $ZZ \to 2l2q$ channels, a broader peak is searched for in the dilepton-ditau and dilepton-dijet mass spectrum respectively [83, 84]. The limited jet energy resolution and the effect of the neutrino escaping detection in leptonic tau decays degrades the

mass resolution in these channels compared to the $4l$ decay. The $ZZ \to 2l2\nu$ search looks for a leptonic $Z$ decay and a large ${E_T}^{miss}$ [85]. In this channel, the decay is not fully reconstructible and so a measure of the mass is given by the transverse mass, $m_T$, defined as

$$m_T = \left[2p_T^Z {E_T}^{miss}(1 - \cos(\Delta\phi))\right]^{\frac{1}{2}}, \tag{6.4}$$

where $p_T^Z$ is the transverse momentum of the dilepton system and $\Delta\phi$ is the angle in the transverse plain between that momentum vector and $\mathbf{E}_T^{miss}$. A broad excess of events in the $m_T$ distribution is used to signal the presence of a SM Higgs boson.

### Combined Results

The 95% upper limits on the signal strength $\mu = \sigma/\sigma_{SM}$ as a function of the hypothesised Higgs boson mass, $m_H$, are shown in Figure 6.3. The right hand figure is an enlargement of the region $110 < m_H < 145$ GeV. The median expected limit in the absence of a SM Higgs boson is less than 1 for the range $110 < m_H < 600$ GeV. The observed limits are consistent with statistical fluctuations given the size of the dataset in most of the range as indicated by the fact that the observed line lies within the 68% or 95% quantiles. However an excess of events is observed at low mass in the range $122.5 < m_H < 127$ GeV so that a SM Higgs boson with a mass in that range cannot be excluded at the 95% confidence level. The significance of the excess is quantified as a function of $m_H$ by calculating the local $p$-value, $p_0$ as shown in Figure 6.4. For the overall combination, the local $p_0$ is around $5.5 \times 10^{-7}$, equivalent to a significance of $4.9\sigma$. The test indicates that the observed excess is incompatible with the background-only hypothesis indicating the presence of a new state with a mass near 125 GeV. The largest contributions to the excess are from the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ channels, both of which have good mass resolutions and hence a good localisation of the excess. The combination of the two high mass resolution channels results in a local significance of $5.0\sigma$. Of the lower resolution channels, only $H \to WW$ shows an excess at 125 GeV. The inclusion of the $H \to bb$ and $H \to \tau\tau$ channels reduced the overall significance. The overall global $p$-value in the range 115-130 GeV is calculated by generating 10,000 pseudo-datasets and fitting for the constant $C$ in the relationship to the local $p$-value given in Equation 5.11 (see Figure 6.5). The look-elsewhere effect, calculated as the ratio between the local and global $p_0$, is around 11 such that the global significance of the full combination remains high at $4.4\sigma$.

| $H$ decay | $H$ prod | Final state | No. sub-chans | $m_H$ (GeV) | Lumi ($fb^{-1}$) 7/8TeV |
|---|---|---|---|---|---|
| $\gamma\gamma$ | untagged | $\gamma\gamma$ (kinematic classes) | 4 | 110-150 | 5.1/5.3 |
| | $qqH$-tag | $\gamma\gamma + jj$ ($m_{jj}$ classes in 8TeV) | 1 or 2 | 110-150 | 5.1/5.3 |
| $bb$ | $VH$-tag | $(\nu\nu, ee, \mu\mu, e\nu, \mu\nu + 2j_b)\times$(low/high $p_T^V$) | 10 | 110-135 | 5.0/5.1 |
| | $ttH$-tag | $l + (4, 5, \geq 6j) \times\ (3, \geq 4j_b),\ l + 4j + 2j_b,$ $ll + (2, \geq 3j_b)$ | 9 | 110-140 | 5.0/– |
| $\tau\tau$ | $0/1-$jets | $(e\tau_h, \mu\tau_h, e\mu, \mu\mu) \times$ (low/high $p_T^{\tau\tau}$) $\times\ (0/1j)$ | 16 | 110-145 | 4.9/5.1 |
| | $qqH$-tag | $(e\tau_h, \mu\tau_h, e\mu, \mu\mu) + jj$ | 4 | 110-145 | 4.9/5.1 |
| | $ZH$-tag | $(ee, \mu\mu) \times (\tau_h\tau_h, \mu\tau_h, \mu\tau_h, e\mu)$ | 8 | 110-160 | 5.0/– |
| | $WH$-tag | $ee\tau_h, \mu\mu\tau_h, e\mu\tau_h$ | 3 | 110-140 | 4.9/– |
| $WW \to l\nu l\nu$ | $0/1-$jets | $(ee/\mu\mu, e\mu) \times (0/1j)$ | 4 | 110-600 | 4.9/5.1 |
| | $qqH$-tag | $(l\nu l\nu + jj)$ (SF or OF $ll$ in 8TeV) | 1 or 2 | 110-600 | 4.9/5.1 |
| | $WH$-tag | $3l3\nu$ | 1 | 110-200 | 4.9/– |
| | $VH$-tag | $(l\nu l\nu + jj)$ (SF or OF $ll$) | 2 | 118-190 | 4.9/– |
| $WW \to l\nu qq$ | untagged | $(e\mu) \times (jj + 0/1j)$ | 4 | 170-600 | 5.0/5.1 |
| $ZZ \to 4l$ | untagged | $4e, 4\mu, 2e2\mu$ | 3 | 110-600 | 5.0/5.3 |
| $ZZ \to 2l2\tau$ | untagged | $(ee, \mu\mu) \times (\tau_h\tau_h, e\tau_h, \mu, \tau_h, e\mu)$ | 8 | 200-600 | 5.0/5.3 |
| $ZZ \to 2l2q$ | untagged | $(ee, \mu\mu) + jj(0, 1, 2j_b)$ | 6 | 200-600 | 4.9/– |
| $ZZ \to 2l2\nu$ | untagged | $(ee, \mu\mu) + \not{E}_T(0, 1, 2j)$ (not $qqH$ jets) | 6 | 200-600 | 4.9/5.1 |
| | $qqH$-tag | $(ee, \mu\mu) + \not{E}_T + jj$ | 2 | 200-600 | 4.9/5.1 |

**Table 6.2.:** Summary of analyses included in the ICHEP 2012 combination [76]. The column for $H$ prod indicates the production process targeted by the sub-channel. A label "untagged" indicates that the main contribution is from the $ggH$ production process. The final states for each channel are exclusive (no events lie in more than one sub-channel). The notations used here are: $jj$ indicating a dijet pair whether from a $W$, $Z$ boson decay or being consistent the vector-boson fusion process; $j_b$ denotes a jet which is identified as a $b$-jet; $l$ is either a muon ($\mu$) or electron ($e$); OF and SF are dilepton pairs with opposite flavour ($e\mu$) and same flavour ($ee$ or $\mu\mu$) respectively.

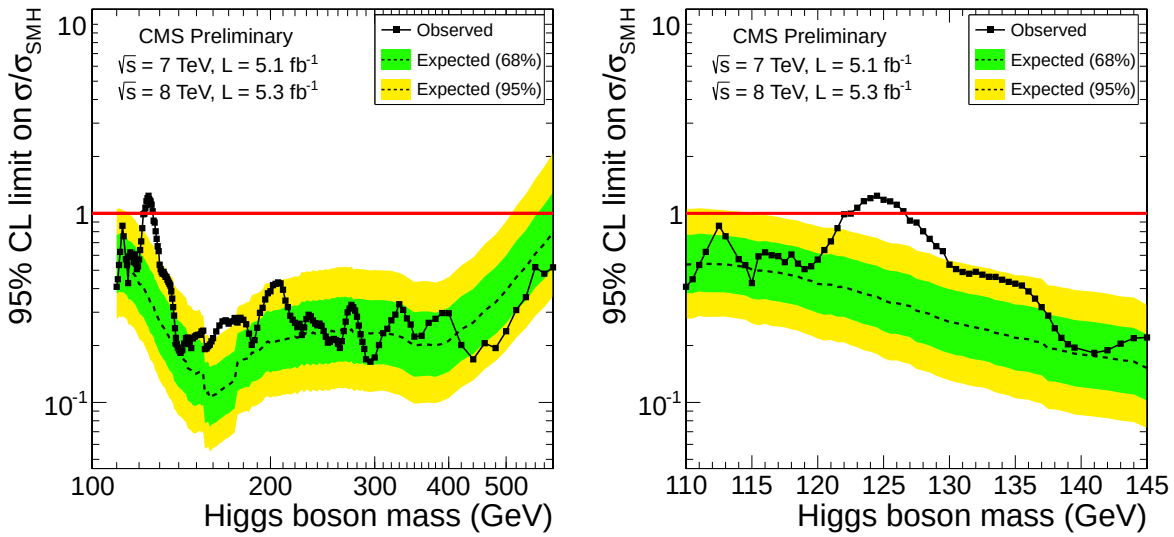


**Figure 6.3.:** Combined 95% upper limits on the production cross-section of Higgs boson production relative to that of the Standard Model in the $m_H$ ranges 110-600 GeV (left) and 110-145 GeV (right) [76]. The median upper limits expected in the absence of a SM Higgs boson are indicated by the dashed black line and the 68% and 95% quantiles by the green and yellow bands respectively. The observed upper limits from the combined ICHEP 2012 dataset is shown by the black solid line. Where the observed limit is lower than 1 (red line), a SM Higgs boson with that $m_H$ is excluded at the 95% confidence level.

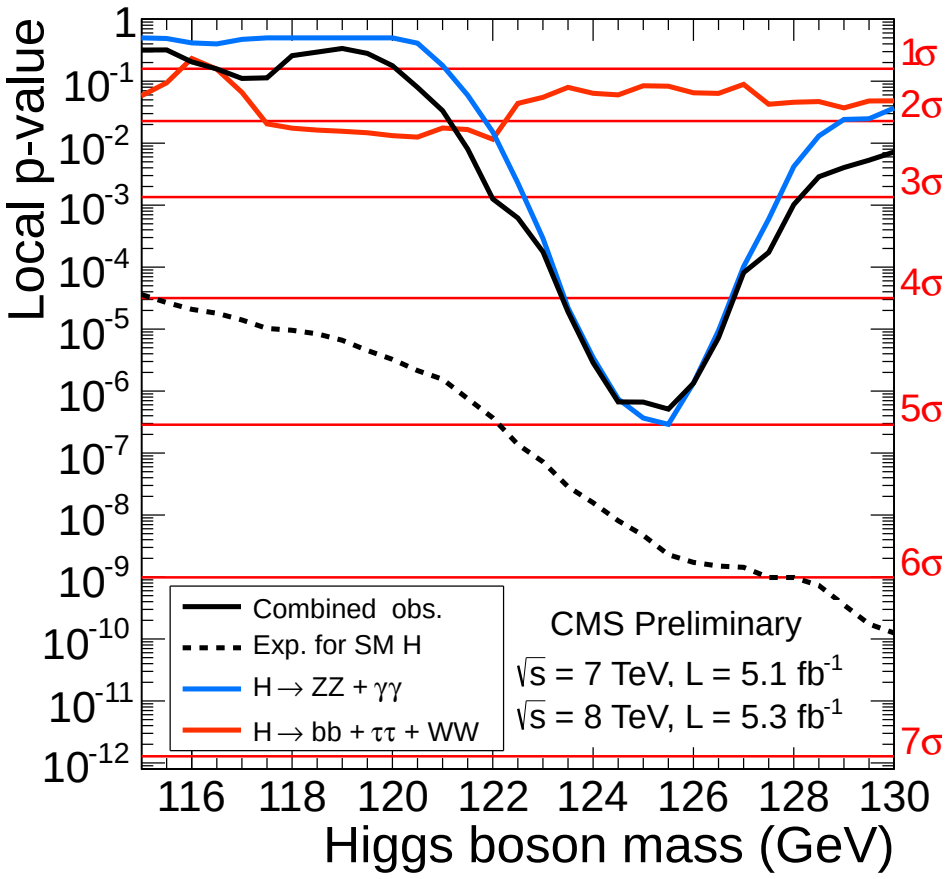


**Figure 6.4.:** The observed local $p$-value, $p_0$ for sub-combinations of the low and high resolution channels and the overall combination as a function of $m_H$. The dashed line shows the expected $p_0$ at each $m_H$ should a SM Higgs boson exist with mass $m_H$ [76].

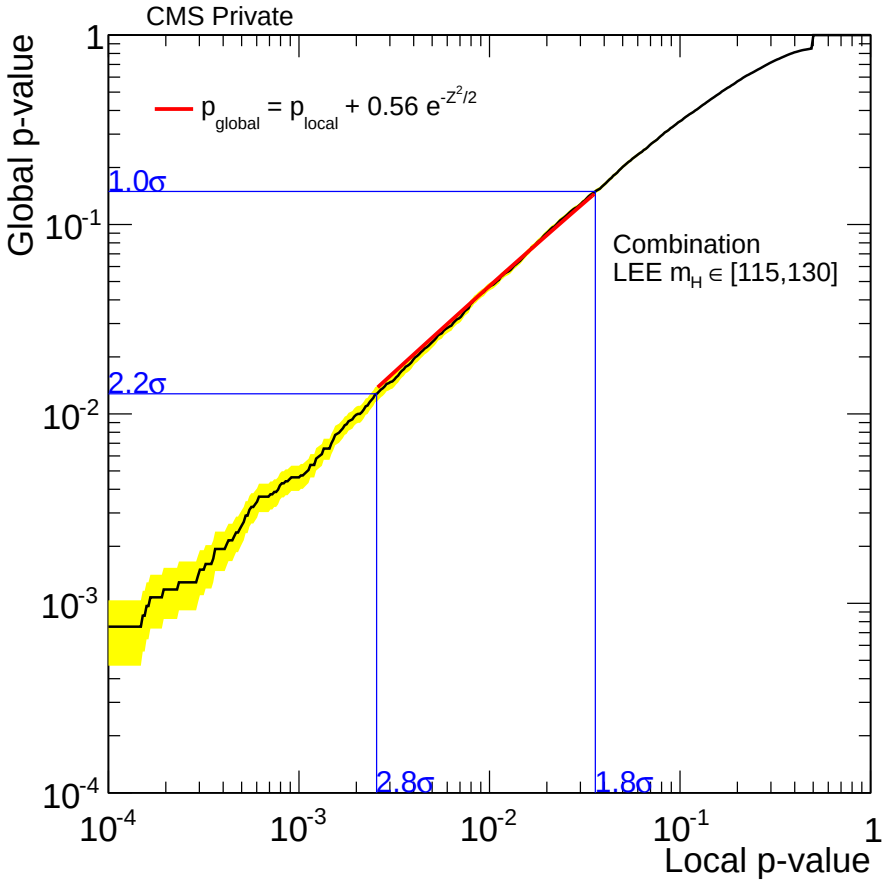


**Figure 6.5.:** Relationship between the local and global $p_0$ in the range 115-130 GeV. The red line indicates the analytic expression (shown) which is fit to the relationship derived from 10,000 pseudo-datasets.

## 6.2. Higgs Properties

With the announcement of the discovery of a new state near 125 GeV, attention at ATLAS and CMS turned to the characterization of the particle through measurements of its properties. In particular, emphasis is placed on ascertaining the compatibility of the new state with the SM Higgs boson. This section includes discussions of some of the techniques used at CMS to determine the properties of the newly discovered state and results presented at the HCP symposium in November 2012. All of the analyses described in Section 6.1.2, with exception of $H \to \gamma\gamma$, were updated to improve their sensitivity and include the additional data collected at CMS [86]. The total integrated luminosity of the 8 TeV data sample used is up to $12.2 fb^{-1}$ depending on the specific channel.

### 6.2.1. Extracting Signal Parameters

The best fit value for the signal strength is evaluated by scanning for the value of $\mu$ at which the likelihood (Equation 6.1) attains its maximum in data. This can be extended where more than one signal parameter is of interest by generalising to the profiled likelihood ratio,

$$q_{\mathbf{x}} = -2 \ln \frac{\mathcal{L}(\text{data}|\mathbf{x}, \hat{\boldsymbol{\theta}}_{\mathbf{x}})}{\mathcal{L}(\text{data}|\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})}, \tag{6.5}$$

where $\mathbf{x} = x_1, x_2, \cdots, x_N$ represents the $N$ parameters of interest in the signal model. The values of the nuisance parameters which maximise the value of the likelihood, first fixing the values of $\mu$ and then letting them float freely, are denoted $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$ and $\hat{\boldsymbol{\theta}}$ respectively. The values for which $q_{\mathbf{x}} = 0$ in the observed data are the best fit values. The contour defined by the set of points for which $C_N(q_{\mathbf{x}}) = 0.68$, where $C_N$ is the cumulative distribution function (cdf) of a chi-squared distribution with $N$ degrees of freedom, is interpreted as the 68% confidence contour. In one dimension, the values at which $q_{\mathbf{x}} = 1$ represent the usual 68% confidence interval. This method for extracting confidence intervals is known to fail when the best fit values lie within or near non-physical regions. It is necessary therefore to cross-check this method to avoid quoting non-physical results. This is achieved through the use of the Feldman-Cousins procedure.

### The Feldman-Cousins Procedure for Evaluating Confidence Intervals

For parameters such as the relative production cross-section, $\mu$, negative values are not considered physical. Constraints on the fit can be imposed to avoid quoting unphysical values. However, where the best-fit values for the signal model parameters lie outside physically allowed regions, the relationship between the values of $q_\mu$ and the 68% confidence interval no longer holds. In order to assign the correct confidence intervals, the Feldman-Cousins procedure is used [87]. The procedure involves throwing pseudo-datasets and evaluating a test-statistic to determine the compatibility of the data with each point in the $N$-dimensional parameter space. The test-statistic used in the one-dimensional case of the signal strength is defined using the ratio of profiled likelihoods, $q_\mu$ (Equation 5.6). The physical constraint on the parameter is imposed in this case by requiring that $\hat{\mu} \geq 0$. The probability to obtain a value of the test-statistic larger than the one observed in data ($CL_{s+b}$) is calculated as in Equation 5.8, where the distribution $f(q_\mu|\mu, \boldsymbol{\theta} = \boldsymbol{\theta}_\mu^{obs})$ is generated from evaluating the test-statistic in pseudo-datasets. As with calculating upper limits, for generating the pseudo-data, the nuisances ($\boldsymbol{\theta}$) are set to the values obtained from a fit to the data. Figure 6.6 shows an example of this distribution for two values of $\mu$ from the (0/1)-jet bin of the $H \to \tau\tau$ analysis and the values of $q_\mu^{obs}$ obtained from the observed data. The 68% confidence interval for $\mu$ is determined as the union of all values of $\mu$ for which $1 - CL_{s+b} < 0.68$. Figure 6.7 shows the values of $1 - CL_{s+b}$ for different values of $\mu$ in the 0/1 jet bin of the $H \to \tau\tau$ analysis. The vertical red line indicates $CL_{s+b} = 0.68$ and the values at which the curve crosses this line (indicated by the horizontal red lines) form the 68% confidence interval for $\mu$. The procedure is easily extended to a higher number of dimensions by exchanging the test-statistic for $q_{\mathbf{x}}$, given in Equation 6.5. Pseudo-datasets are generated and fit as before and the boundary of the union of points for which $1 - CL_{s+b} < 0.68$ defines a confidence-contour in an $n$-dimensional parameter space.

## 6.2.2. Combined Mass Measurement

The mass of the Higgs boson is a free parameter in the context of the Standard Model. The high resolution channels, $H \to \gamma\gamma$ and $H \to ZZ \to 4l$, provide the strongest constraint on the mass of the new particle as the signal is visible as a narrow peak in the invariant mass of its decay products. To measure the mass, $m_X$, of the particle in a model-independent way, the signal strengths for the $gg \to H \to \gamma\gamma$, $qq \to H \to \gamma\gamma$ and $H \to ZZ \to 4l$ processes are assumed to be independent and thus are treated as nuisance

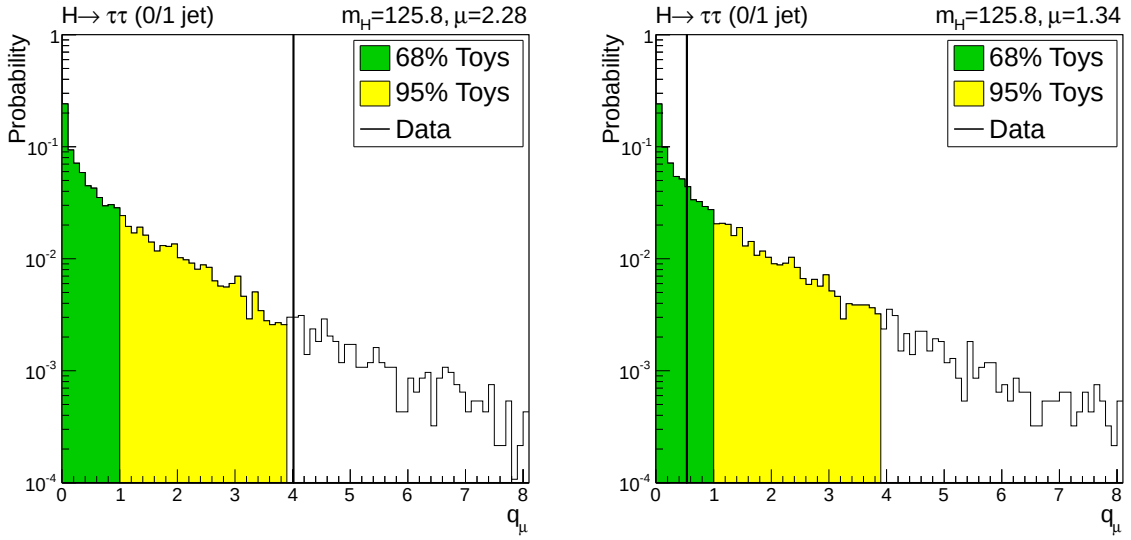


**Figure 6.6.:** Distributions of the test statistic $q_\mu$ for the 0/1 jet bin of the $H \to \tau\tau$ analysis at the combined best fit mass, $m_H = 125.8$ GeV. The green and yellow filled regions indicate the 68% and 95% quantiles of the distribution respectively. The left distribution is generated at $\mu = 2.28$ which lies outside of the 68% confidence interval while the right distribution is generated at $\mu = 1.34$ which lies inside the 68% confidence interval. The values of the test statistic obtained from the observed data, $q_\mu^{obs}$, are indicated by the solid vertical lines.

parameters in the likelihood. Each of the signals in these channels is assumed to be due to the presence of a single state with mass $m_X$. Figure 6.8 (left) shows the value of the test-statistic $q_{m_X}$ for the $H \to \gamma\gamma$, $H \to ZZ$ channels and their combination near the best fit points. From the combination, the mass is determined to be $m_X = 125.8 \pm 0.5$ GeV. The 68% confidence interval is determined from the values of $m_X$ at which the curve crosses the horizontal red line at 1. Large background fluctuations in the $H \to \gamma\gamma$ channel can result in large variations of the measured mass when the signal is small. Conversely, the kinematic constraints on the $4l$ system cause a large variation in the branching ratio of $H \to ZZ \to 4l$, and hence the expected signal yield, as a function of $m_H$. Figure 6.8 (right) shows the two-dimensional 68% confidence intervals in $m_X$ and $\sigma/\sigma_{SM}$ for the $H \to ZZ \to 4l$, $H \to \gamma\gamma$ and their combination. For this combination, the ratio of signal strengths between the two channels is kept fixed to the SM expectation; only the overall signal strength is left as a free parameter. The best fit value of $m_X$ is consistent with the value determined in the one-dimensional case. The best fit value for the combined signal strength relative to the Standard Model is $0.88 \pm 0.21$ for a mass of 125.8 GeV.

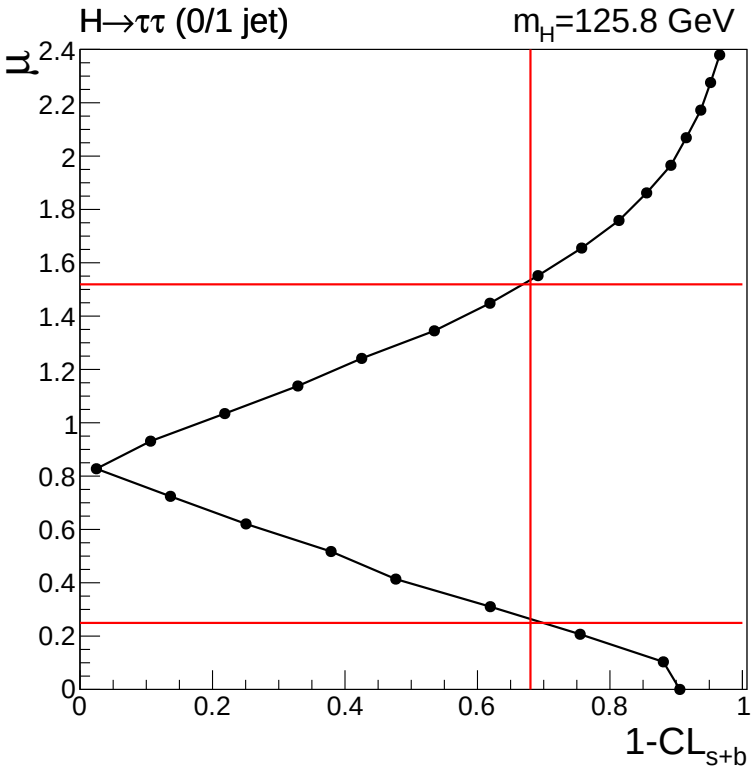


**Figure 6.7.:** Confidence level evaluation curve for the $H \to \tau\tau$ analysis in the (0/1) jet bin. At each point, pseudo-data are generated with signal injected at the given value of $\mu$ and its confidence level (CL) calculated. Linear interpolation between the generated points is used to determine the 68% confidence interval; the two values of $\mu$ (horizontal lines) which cross the curve at $1 - CL_{s+b} = 0.68$ (vertical red line).

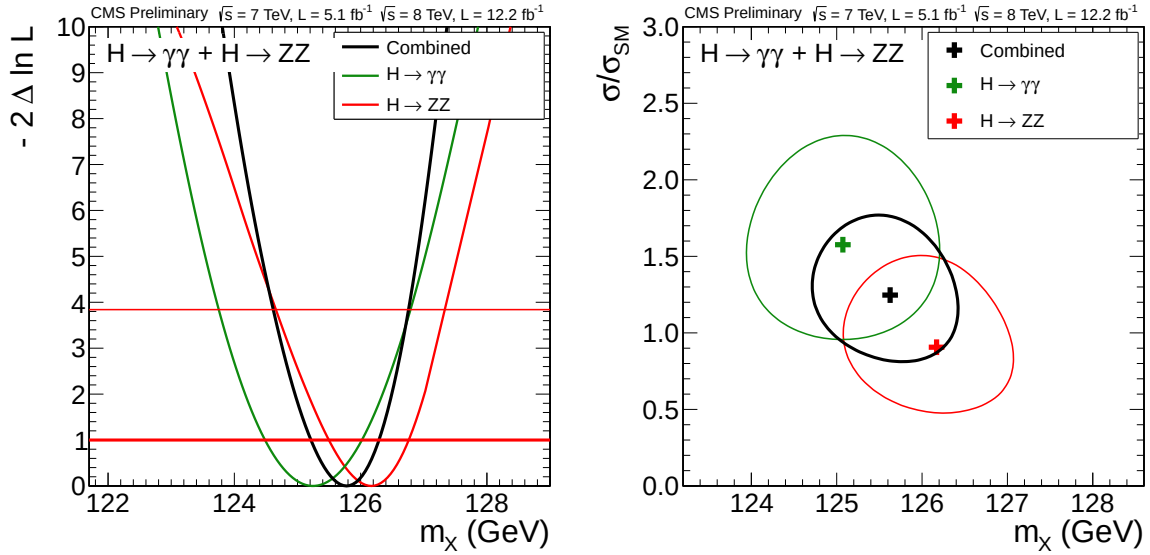


**Figure 6.8.:** Left: One-dimensional scan of $q_{m_x}$ for the $H \to \gamma\gamma$, $H \to ZZ$ channels and their combination. For the combination, the relative signal strengths between the channels are allowed to float. The 68% and 95% confidence intervals for $m_X$ are determined as the values at which the curves cross the horizontal red lines. Right: 68% confidence contours in $m_X$ and $\sigma/\sigma_{SM}$ for the $H \to \gamma\gamma$ and $H \to ZZ$ channels and their combination. For this combination, the relative signal strengths of the channels are kept fixed to the SM expectation [86].

### 6.2.3. Compatibility with the Standard Model

The Standard Model makes very precise predictions for the coupling of the Higgs boson to all of the known fundamental particles. These couplings directly influence the various rates of production and decay of the Higgs boson. Precise measurements of these rates in the combined search channels provide information on the couplings. Significant deviations from the values predicted by the SM would indicate the presence of new physics.

#### Channel Compatibility

When determining the preferred value of $\mu$ in the combined data, the ratios of decay rates to each contributing channel relative to that predicted by the SM are kept constant. By relaxing this constraint, the compatibility of the new state with the SM Higgs boson can be studied on a per-decay/per-production level. Due to the limited amount of data collected at CMS, some of the channels and sub-channels entering the combination have a negative value for the best fit signal strength ($\mu = \sigma/\sigma_{SM}$). In order to avoid quoting unphysical values in each channel, the Feldman-Cousins procedure is used to determine

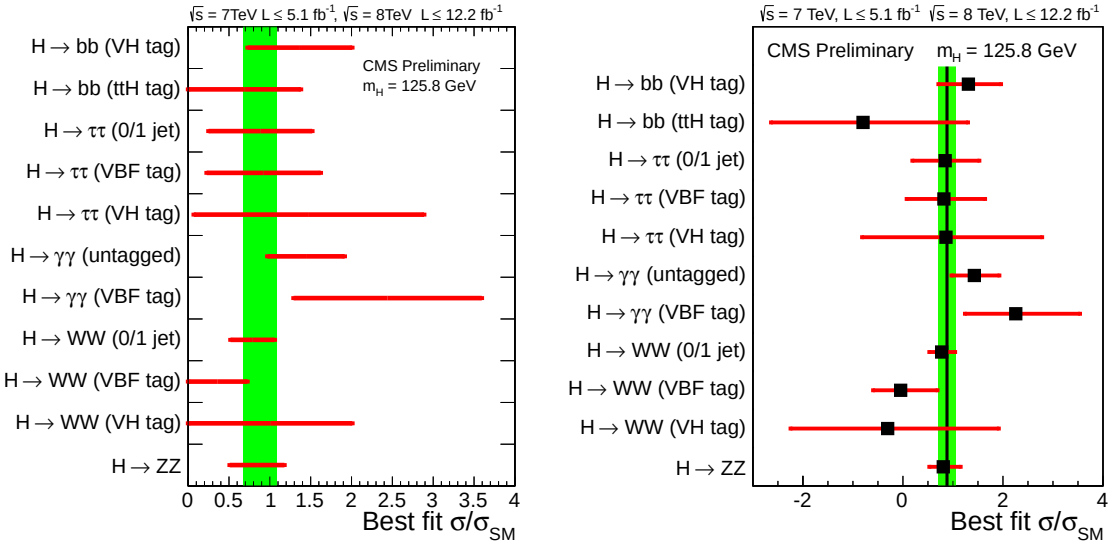


**Figure 6.9.:** 68% confidence intervals for $\mu = \sigma/\sigma_{SM}$ for individual channels or combination of sub-channels determined using the Feldman-Cousins procedure (left) and by scanning the likelihood (right). The value of $\sigma/\sigma_{SM}$ denotes the production cross-section times the relevant branching fraction for a given channel, relative to the SM. The green band indicates the 68% confidence interval on $\sigma/\sigma_{SM}$ for all channels combined. The intervals are determined at the best fit mass, $m_H = 125.8$ GeV [86].

68% confidence intervals for $\sigma/\sigma_{SM}$ separately in the different channels/sub-channels entering the combination. Figure 6.9 (left) shows the 68% confidence intervals on $\sigma/\sigma_{SM}$ for the sub-channels included in the combination obtained from the HCP dataset. The results are compared with the intervals determined directly from a scan of $q_\mu$, as shown in the same figure (right). The two methods are found to be in good agreement. The intervals are extracted for a Higgs boson mass $m_H = 125.8$ GeV (the overall best fit mass of the new state obtained from the same dataset). The 68% confidence interval on $\sigma/\sigma_{SM}$ for the full combination is indicated by the green band. With the exception of the dijet (VBF) tagged channel in the $H \to \gamma\gamma$ analysis, all of the intervals contain the value $\mu = 1$ which is the expected value for a SM Higgs boson.

Several of the analyses which are combined in the search for the Higgs boson use selections (tags) which are specifically designed to enhance the sensitivity to particular Higgs boson production topologies. The $H \to WW$, $H \to \tau\tau$ and $H \to \gamma\gamma$ analyses all include dijet (or VBF tagged) categories which are designed predominantly to select events produced via vector-boson fusion ($qqH$). Additional sensitivity is gained in the $H \to WW$, $H \to \tau\tau$ and $H \to bb$ channels by looking for additional leptons or $E_T$ in association with production of a vector boson ($VH$). The production rates associated

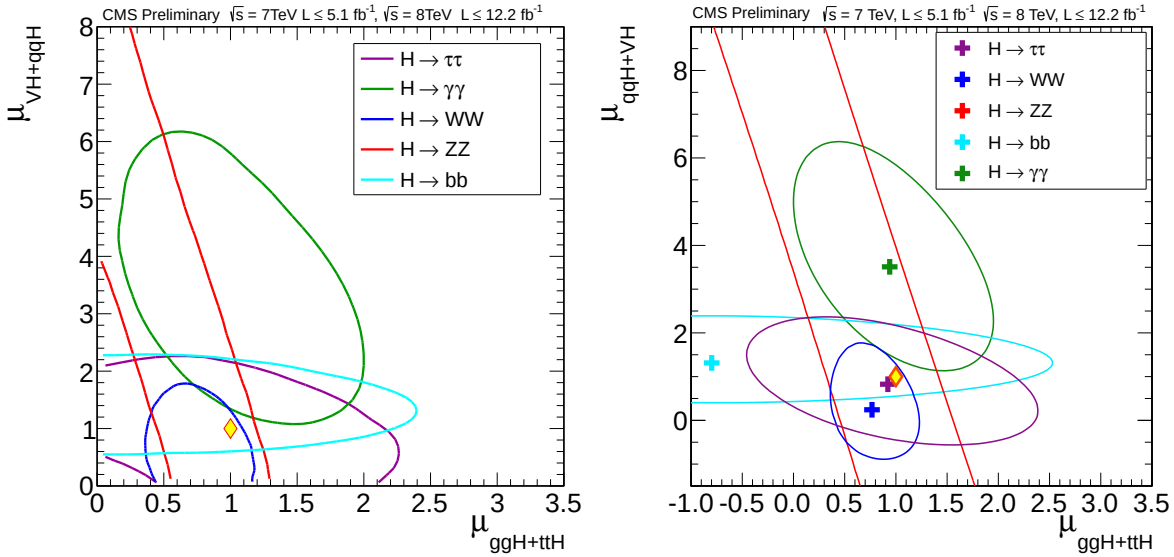


**Figure 6.10.:** 68% confidence contours for the production cross-section in $ggH$ and $ttH$ modes ($\mu_{ggH+ttH}$), and $VH$ and $qqH$ modes ($\mu_{VH+qqH}$), relative to the SM determined using the Feldman-Cousins procedure (left) and by scanning the likelihood (right). Each colour indicates the result by combining all sub-channels in a particular decay mode. The crosses indicate the best fit values of the two parameters. The yellow diamond at $(1, 1)$ indicates the SM values. The contours are determined at the best fit mass, $m_H = 125.8$ GeV [86].

to couplings with top-quarks ($ggH$ and $ttH$) and vector bosons ($qqH$ and $VH$) are determined by removing the requirement that the relative production cross-sections $\mu_{ggH+ttH}$ and $\mu_{VH+qqH}$ are equal. The compatibility of the rates observed in data with respect to those predicted by the Standard Model can be tested using the Feldman-Cousins procedure or scanning the test-statistic $q_{\mathbf{x}}$. The relative branching ratios to each of the five observable final states are left unconstrained. Figure 6.10 shows the 68% confidence contours for each of the five decay processes using the two methods. Good agreement is found when comparing the two methods. With the exception of the $H \to ZZ$ analysis, the explicit exploitation of the different production modes leads to elliptical contours. The SM point $(1, 1)$, indicated by the yellow diamond, is contained within the 68% confidence contours from each decay channel with the exception of $H \to \gamma\gamma$.

### Coupling Measurements

The compatibility of the couplings of the new particle with the SM cannot be directly ascertained in the experimental data. In order to extract the relevant information, the

rates of production and decay in the various channels must be interpreted in terms of the underlying couplings to the SM particles. For the purposes of evaluating the compatibility in the couplings, the following simplifications are made:

- Signals observed in each of the different search channels originate from a single resonance near 125 GeV.
- The natural width of the resonance is small enough to be neglected such that the cross-section of the signal in each channel can be expressed as

$$(\sigma \cdot BR)(ii \to H \to ff) = \frac{\sigma_{ii}\Gamma_{ff}}{\Gamma}, \tag{6.6}$$

  where $\sigma_{ii}$ is the production cross-section through the initial state $ii$, $\Gamma_{ff}$ is the partial decay width to the final state $ff$ and $\Gamma$ is the total width.
- Only modifications of the absolute values of the coupling strengths are allowed. The structure of the couplings is fixed to the SM, in particular this means the new state is assumed to be a CP-even scalar.

In general, no specific assumptions are made on any additional states of new physics which could influence the phenomenology of the 125 GeV state. A number of frameworks to investigate the coupling structure of the new particle are used at CMS [88]. The simplest of these is an unfolding of the production cross-section modifiers $\mu_{ttH+ggH}$, $\mu_{VH+qqH}$ by expressing them as functions of the couplings to fermions $\kappa_f$ and vector bosons $\kappa_V$ and is described here as an example. The decay rates to each channel are also expressed as functions of these parameters such that the overall yield in each channel relative to the SM expectation is parameterized. The ratio of the total width to that predicted by the SM is denoted $\kappa_H = \Gamma/\Gamma_{SM}$. Table 6.3 shows the parameterization of $(\sigma \cdot BR)(ii \to H \to ff)$ for each production/decay included in the combination. The parameters $\kappa_f$ and $\kappa_V$ are the couplings relative to the SM predictions for the Higgs boson such that the SM is recovered setting $\kappa_V = \kappa_f = 1$. The only non-trivial scaling is from the $H \to \gamma\gamma$ vertex (indicated by $\kappa_\gamma$) which is needed to account for the contribution from the $WW$ and $t\bar{t}$ loops. No invisible final states are assumed so that the total width, $\Gamma$, is a function of $\kappa_V$ and $\kappa_f$.

Figure 6.11 shows the best fit values in the observed data for $\kappa_V$ and $\kappa_f$ and the 68% confidence contours determined from a scan of $q_{\mathbf{x}}$. The values are extracted independently in each decay channel and from the full combination. In addition to the SM point, the

| | $H \to \gamma\gamma$ | $H \to ZZ/H \to WW$ | $H \to bb/H \to \tau\tau$ |
|---|---|---|---|
| $ggH/ttH$ | $\frac{\kappa_f^2\kappa_\gamma^2(\kappa_f,\kappa_V)}{\kappa_H(\kappa_f,\kappa_V)^2}$ | $\frac{\kappa_f^2\kappa_V^2}{\kappa_H(\kappa_f,\kappa_V)^2}$ | $\frac{\kappa_f^2\kappa_f^2}{\kappa_H(\kappa_f,\kappa_V)^2}$ |
| $qqH/VH$ | $\frac{\kappa_V^2\kappa_\gamma^2(\kappa_f,\kappa_V)}{\kappa_H(\kappa_f,\kappa_V)^2}$ | $\frac{\kappa_V^2\kappa_V^2}{\kappa_H(\kappa_f,\kappa_V)^2}$ | $\frac{\kappa_V^2\kappa_f^2}{\kappa_H(\kappa_f,\kappa_V)^2}$ |

**Table 6.3.:** Boson and fermion vertex scaling as a function of $\kappa_V$ and $\kappa_f$ for each production/decay included in the combination. Each cell represents the scaling factor applied to the production (row) decay (column) combination.

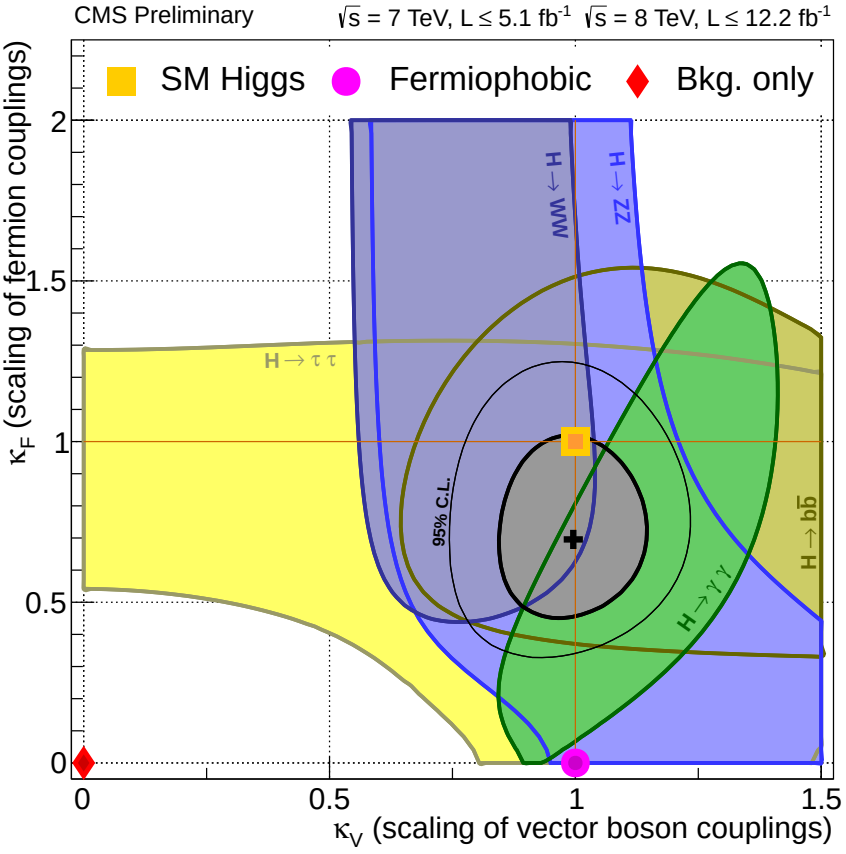


**Figure 6.11.:** The 68% confidence contours extracted from data in the individual decay channels (coloured regions) and the full combination (solid line). The yellow square shows the SM value, while the fermiophobic and background-only scenarios are indicated by the pink dot and red diamond respectively [86].

fermiophobic Higgs scenario, in which the Higgs boson does not couple to fermions, is indicated. The data are compatible with the expectation of a SM Higgs boson; the SM point ($\kappa_V = \kappa_f = 1$) lies within the 95% confidence contour defined by the data.

# Chapter 7.

# Conclusions and Outlook

The Standard Model of particle physics provides the most precise description of fundamental physics and remains the most experimentally verified model available. The mechanism by which electroweak symmetry breaking occurs in the Standard Model, giving rise to the masses of the fundamental fermions and bosons, predicts the existence of a new massive scalar boson, the Higgs boson. Such a particle should be experimentally observable, although prior to the LHC being turned on, no such particle had been discovered.

In this thesis, a search for this particle, through its decay to two photons, in proton-proton collisions recorded with the CMS detector, during 2011 at a centre of mass $\sqrt{s} = 7$ TeV has been described. The decay channel, despite having a relatively low branching ratio, is one of the most sensitive at CMS due to the high resolution of the electromagnetic calorimeter and the narrow invariant mass peak it provides. The analysis detailed employed the use of several multivariate analysis techniques in order to provide the greatest sensitivity to a potential signal. As the signal yield in the two photon decay channel is small, the search for $H \to \gamma\gamma$ is highly sensitive to the background modelling. The signal extraction technique described in this thesis was one which was developed by the author and served as a cross-check of the published result from the 2011 dataset. This allowed for additional scrutiny on the background modelling to which the search in this decay channel is so sensitive. When combined with additional data from 2012 at a centre of mass energy $\sqrt{s} = 8$ TeV, an excess near 125 GeV was observed with a significance of around $4\sigma$.

In order to maximise the sensitivity of the search for the Standard Model Higgs boson, data from several decay channels are combined at CMS using the methods described in this thesis. An excess compatible with a Standard Model Higgs boson with a mass

around 126 GeV was observed in the combined 2011 and 2012 datasets. The excess was significant enough so as to claim discovery at the $5\sigma$ level. The excess is driven by the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ channels, although additional evidence is found in the $H \to WW$ channels. With the data available by the Hadron Collider Physics symposium of November 2012, study of its couplings to Standard Model particles indicates that the new particle is consistent with the Standard Model Higgs boson, though additional data are required to make a definitive statement.

The discovery of the new particle is one of great significance to particle physics. Should the particle turn out to conform to the predictions of the Standard Model, its discovery will have provided a great step towards understanding the nature of electroweak symmetry breaking. However, if this turns out not to be the case, deviations from the predictions will indicate hints of potential new physics and serve as guidance in the search for physics beyond the Standard Model . Additional data will be taken once the LHC resumes collisions in 2015 with an increased centre-of-mass energy of $\sqrt{s} = 14$ TeV. With the additional data, stronger statements can be made as to the exact nature of the new particle and its interactions with Standard Model particles or potential other new particles. With this discovery in hand and the search potentially at an end, it is clear that a new window into fundamental physics has been opened, and the intriguing studies into what that physics is has now begun.

# Appendix A.

## A.1. L1 Jet Energy Correction Fits

The energy corrections applied online at L1 to jets are derived in 11 $|\eta|$ bins corresponding to the 11 regions of the GCT. The values of the fitted parameters in each bin for the parameterisation in Equation 3.5 are given in Table A.1. These parameter values are extracted from fitting the L1 response as a function of $E_T^{L1}$ as described in Section 3.3.1. Figures A.1 and A.2 show the results of those fits in each $|\eta|$ bin.

| | $p_0$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|---|
| $0 < \|\eta\| < 0.348$ | 1.114 | 2.297 | 5.959 | 1.181 | 0.7286 | 0.3673 |
| $0.348 < \|\eta\| < 0.695$ | 0.7842 | 4.331 | 2.672 | 0.5743 | 0.8811 | 0.4085 |
| $0.695 < \|\eta\| < 1.044$ | 0.961 | 2.941 | 2.4 | 1.248 | 0.666 | 0.1041 |
| $1.044 < \|\eta\| < 1.392$ | 0.6318 | 6.6 | 3.21 | 0.8551 | 0.9786 | 0.291 |
| $1.392 < \|\eta\| < 1.740$ | 0.3456 | 8.992 | 3.165 | 0.5798 | 2.146 | 0.4912 |
| $1.740 < \|\eta\| < 2.172$ | 0.8501 | 3.892 | 2.466 | 1.236 | 0.8323 | 0.1809 |
| $2.172 < \|\eta\| < 3.0$ | 0.9027 | 2.581 | 1.453 | 1.029 | 0.6767 | -0.1476 |
| $3.0 < \|\eta\| < 3.5$ | 1.117 | 2.382 | 1.769 | 0.0 | -1.306 | -0.4741 |
| $3.5 < \|\eta\| < 4.0$ | 1.634 | -1.01 | 0.7184 | 1.639 | 0.6727 | -0.2129 |
| $4.0 < \|\eta\| < 4.5$ | 0.9862 | 3.138 | 4.672 | 2.362 | 1.55 | -0.7154 |
| $4.5 < \|\eta\| < 5.0$ | 1.245 | 1.103 | 1.919 | 0.3054 | 5.745 | 0.8622 |

**Table A.1.:** Calibration coefficients used to parameterise the L1 jet correction function (Equation 3.5) for each of the 11 GCT regions.

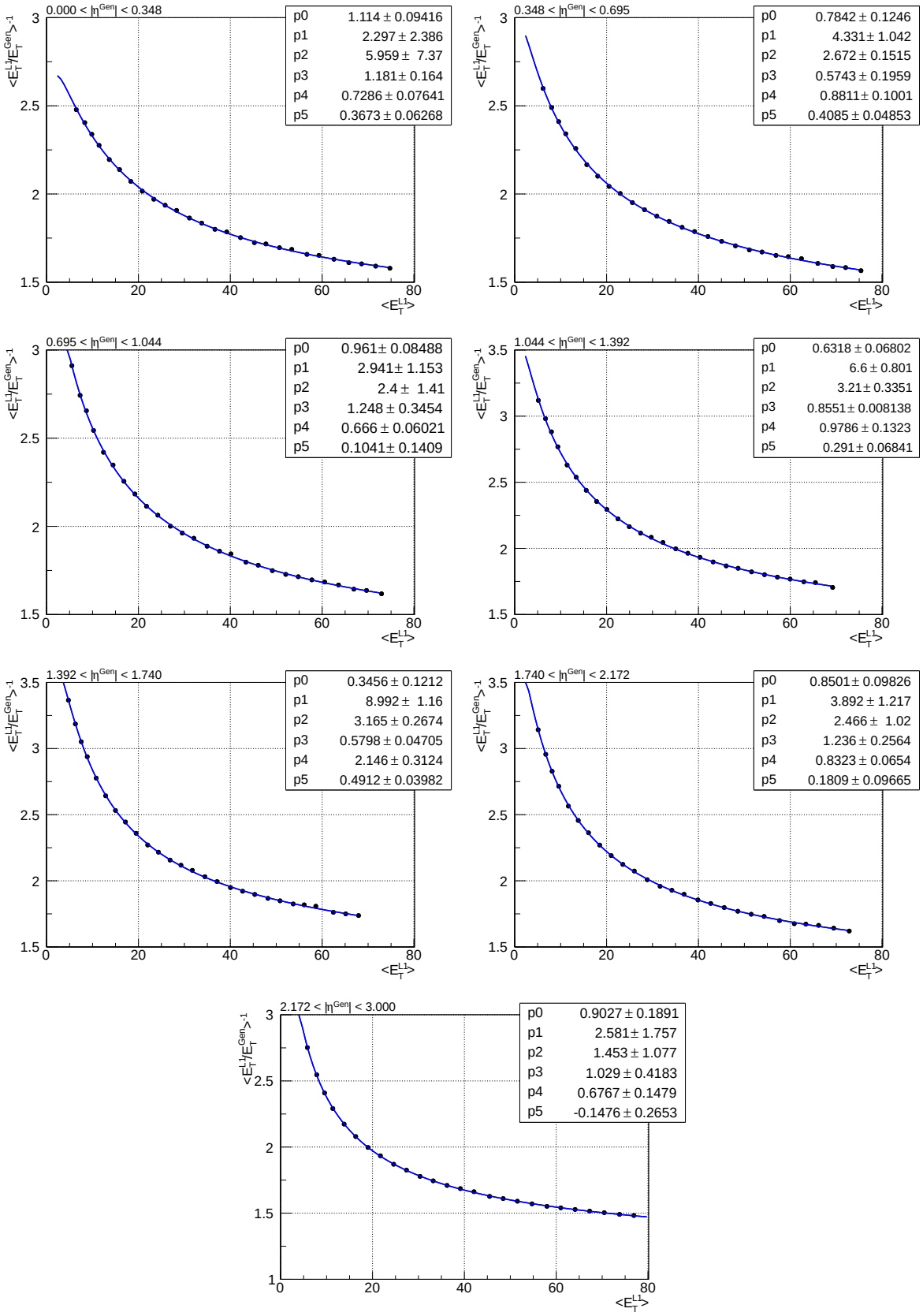


**Figure A.1.:** Fitted correction functions for each of the 7 GCT regions covered by the ECAL and HCAL. The points are fit with the function of Equation 3.5 to provide a parameterisation of the corrections to be applied to L1 jets.

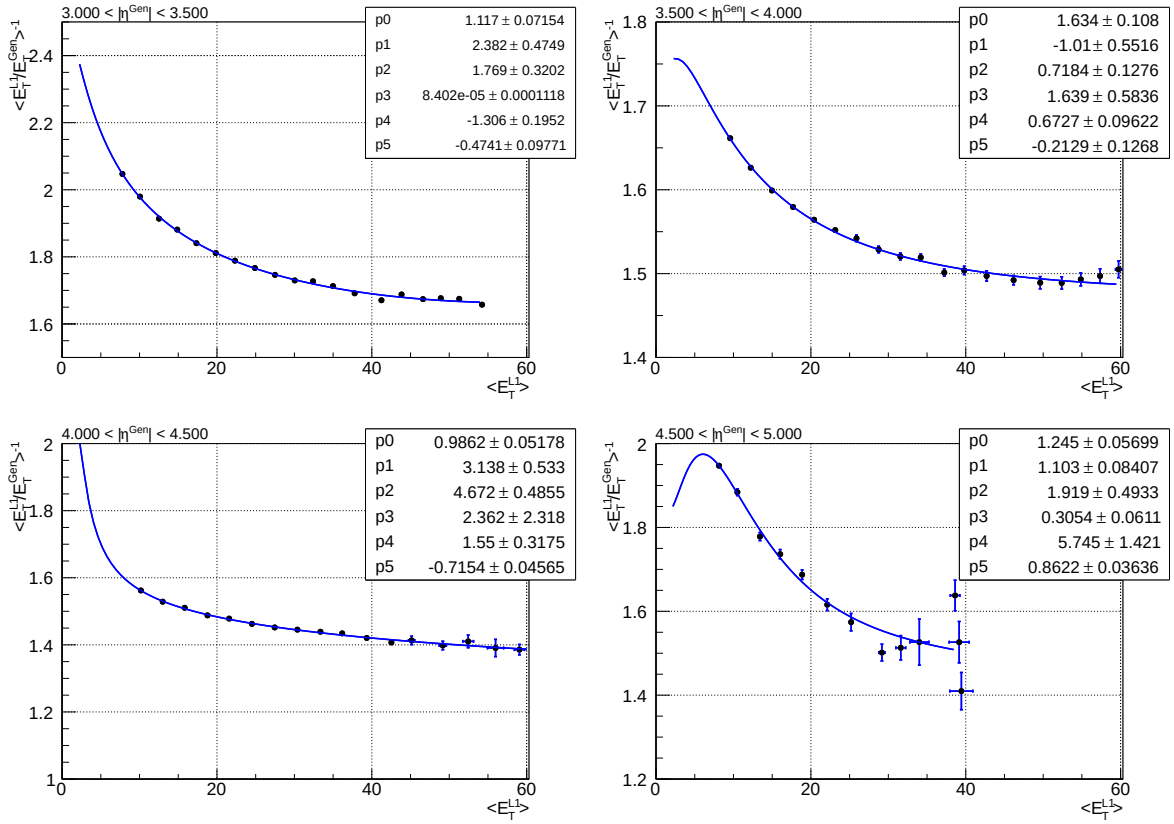


**Figure A.2.:** Fitted correction functions for each of the 4 GCT regions covered by the HF. The points are fit with the function of Equation 3.5 to provide a parameterisation of the corrections to be applied to jets online in the GCT.

## A.2. L1 Jet Resolution

The L1 jet resolution measured in MC was shown as a function of $E_T^{L1}$ before and after applying the derived jet energy calibrations in Figure 3.12. The value of the response in at each point is taken from a Gaussian fit to the distribution of $E_T^{L1}/E_T^{Gen}$ in bins of $E_T^{L1}$. Figures A.3 and A.4 show the fits before applying the corrections while Figures A.5 and A.6 show the fits after. The central value for the points in Figure 3.12 are taken from the width ($\sigma$) of the fitted Gaussian distributions. The error bars on this plot are taken from the error on the fitted value of $\sigma$ which is too small to be visible. The improvement in resolution after applying the corrections is clearly visible.

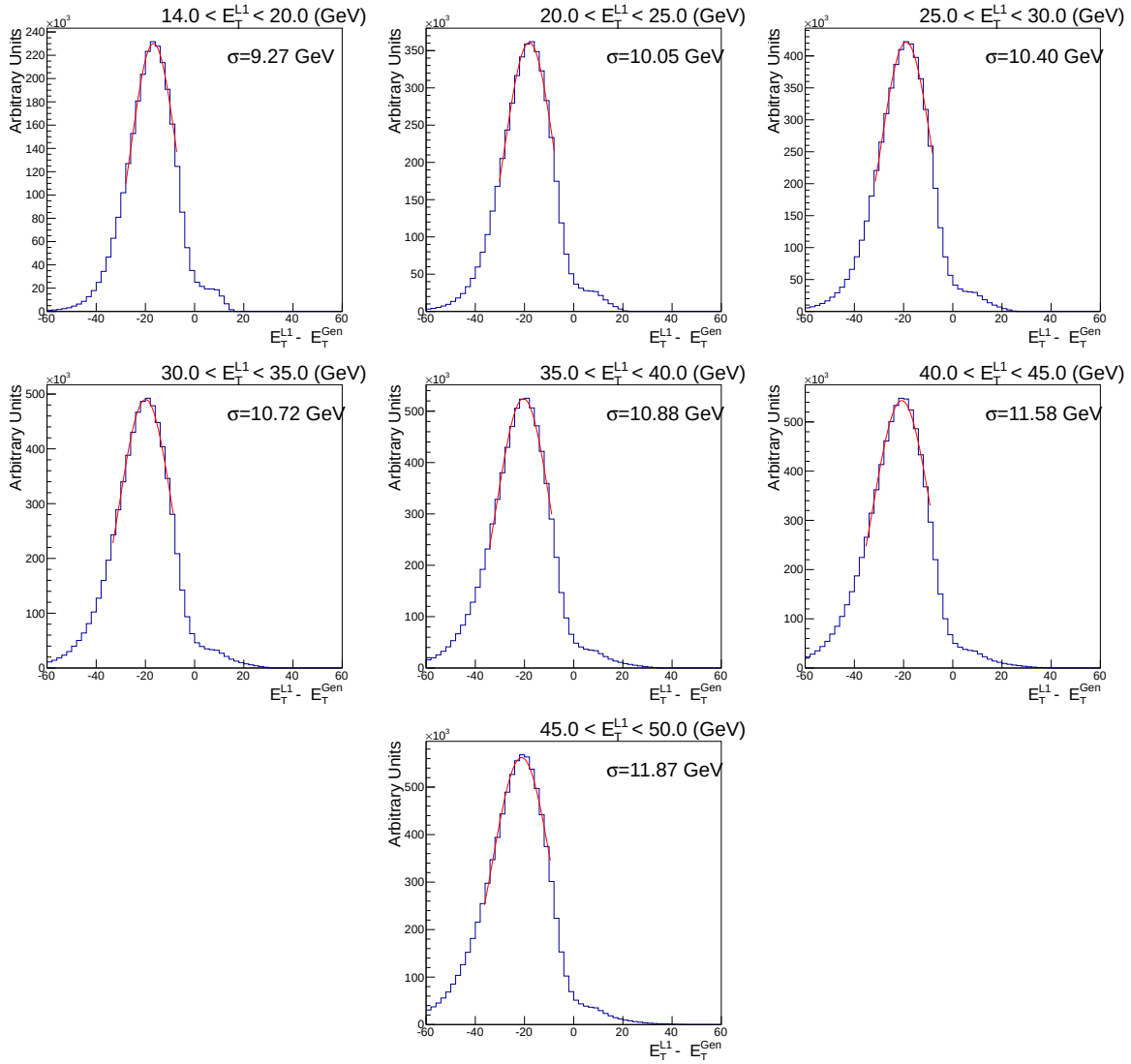


**Figure A.3.:** Part one of the distributions of $E_T^{L1} - E_T^{Gen}$ in bins of $E_T^{L1}$ of the uncorrected MC jets. The fitted Gaussian is used to extract the resolution as a function of $E_T^{L1}$.

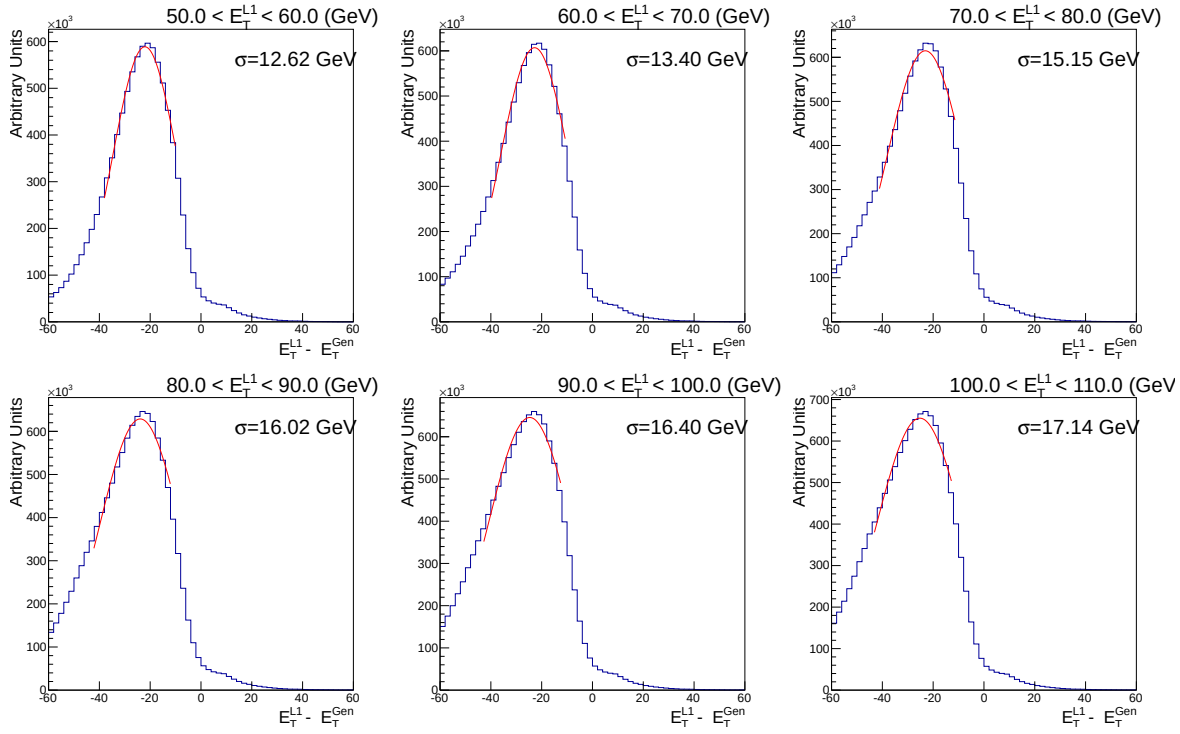


**Figure A.4.:** Part two of the distributions of $E_T^{L1} - E_T^{Gen}$ in bins of $E_T^{L1}$ of the uncorrected MC jets. The fitted Gaussian is used to extract the resolution as a function of $E_T^{L1}$.

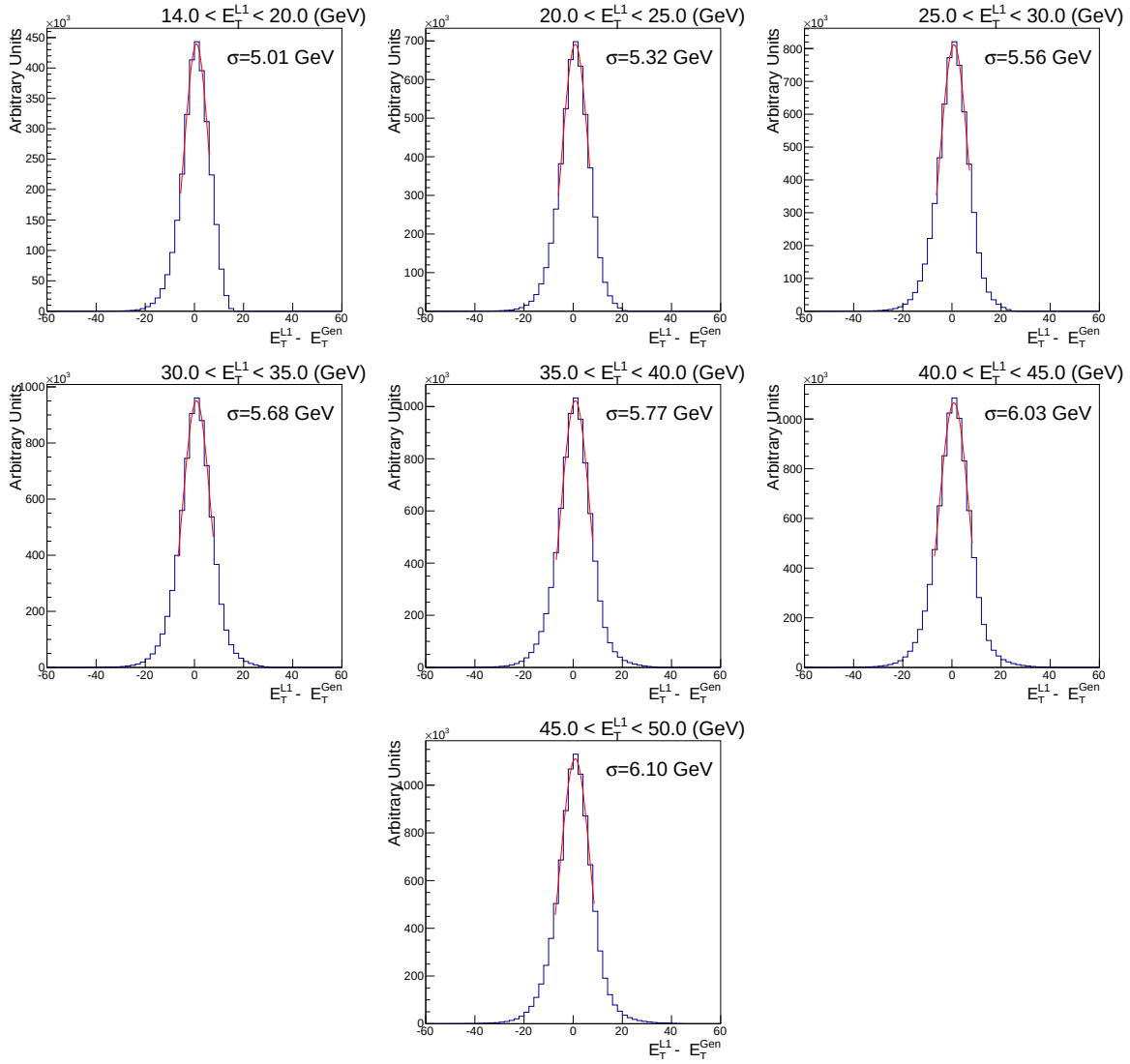


**Figure A.5.:** Part one of the distributions of $E_T^{L1} - E_T^{Gen}$ in bins of $E_T^{L1}$ of the corrected MC jets. The fitted Gaussian is used to extract the resolution as a function of $E_T^{L1}$.

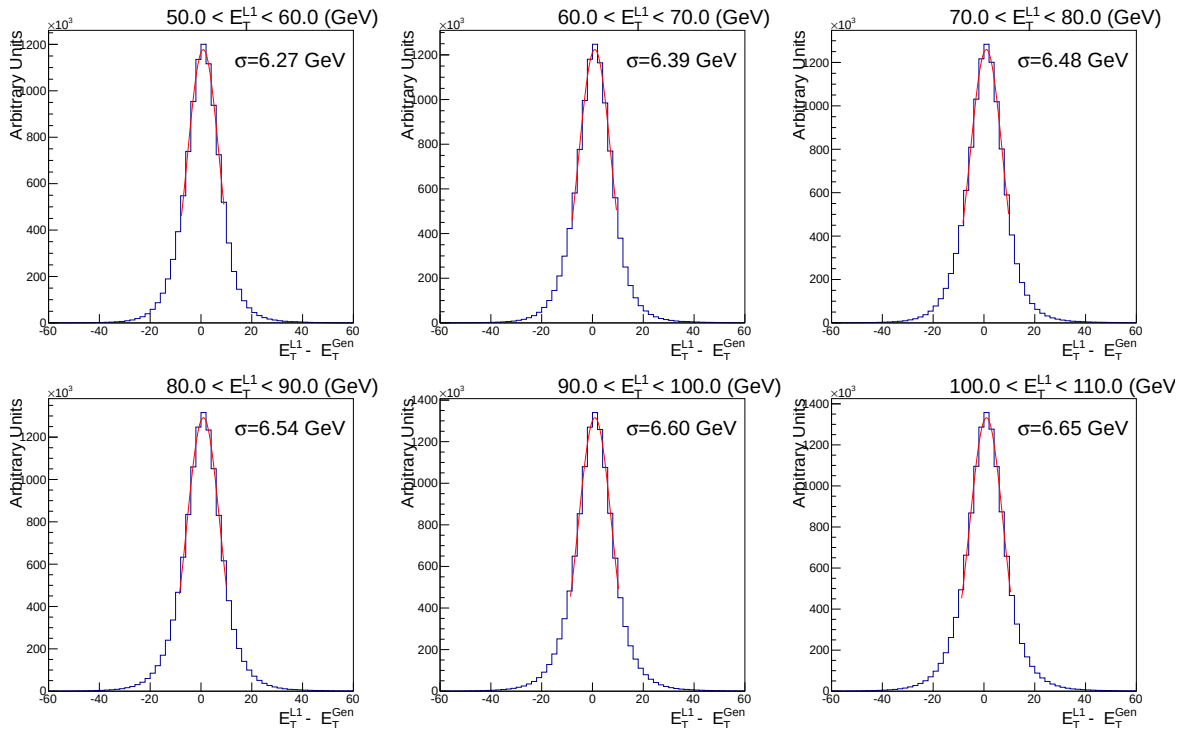


**Figure A.6.:** Part two of the distributions of $E_T^{L1} - E_T^{Gen}$ in bins of $E_T^{L1}$ of the corrected MC jets. The fitted Gaussian is used to extract the resolution as a function of $E_T^{L1}$.

# Appendix B.

## B.1. Energy Scale and Resolution Measurements

The energy scale and resolution is measured in the 2011 dataset using $Z \to e^+e^-$ events as described in Section 4.2.2. The additional resolution required to match the $Z \to e^+e^-$ peak in MC to that of the data (Table B.1) is used to correct the Higgs MC for modelling the signal in the $H \to \gamma\gamma$ analysis. The scale measurements (Tables B.2 and B.3 ) are used to correct the energy of the photons in data.

| Category | $\sigma_E/E$ (%) |
|---|---|
| EB, $\|\eta\| < 1$, $R9 > 0.94$, NOT GAP | $0.67^{+0.10}_{-0.33} \pm 0.22$ |
| EB, $\|\eta\| < 1$, $R9 > 0.94$, GAP | $0.77^{+0.06}_{-0.12} \pm 0.22$ |
| EB, $\|\eta\| < 1$, $R9 < 0.94$ | $0.96^{+0.05}_{-0.05} \pm 0.24$ |
| EB, $\|\eta\| > 1$, $R9 > 0.94$ | $1.41^{+0.15}_{-0.33} \pm 0.60$ |
| EB, $\|\eta\| > 1$, $R9 < 0.94$ | $1.96^{+0.06}_{-0.07} \pm 0.59$ |
| EE, $\|\eta\| < 2$, $R9 > 0.94$ | $2.68^{+0.15}_{-0.20} \pm 0.90$ |
| EE, $\|\eta\| < 2$, $R9 < 0.94$ | $2.79^{+0.09}_{-0.10} \pm 0.30$ |
| EE, $\|\eta\| > 2$, $R9 > 0.94$ | $2.93^{+0.08}_{-0.08} \pm 0.34$ |
| EE, $\|\eta\| > 2$, $R9 < 0.94$ | $3.01^{+0.11}_{-0.12} \pm 0.52$ |

**Table B.1.:** Additional energy resolution included in the $H \to \gamma\gamma$ signal model measured from comparison of $Z \to e^+e^-$ data and MC. The label "NOT GAP" indicates superclusters whose seed crystal is located more than 5 crystals away from an ECAL module boundary whereas the label "GAP" indicates superclusters whose seed crystal is within 5 crystals of an ECAL module boundary [56].

| Category | Run Range | $\Delta P$ |
|---|---|---|
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 160431 - 167913 | $-0.0004 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 170000 - 172619 | $-0.0016 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 172620 - 173692 | $-0.0017 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 175830 - 177139 | $-0.0021 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 177140 - 178421 | $-0.0025 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 < 0.94$ | 178424 - 180252 | $-0.0024 \pm 0.0002 \pm 0.0019$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 160431 - 167913 | $0.0059 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 170000 - 172619 | $0.0046 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 172620 - 173692 | $0.0045 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 175830 - 177139 | $0.0042 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 177140 - 178421 | $0.0038 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert < 1$, $r_9 > 0.94$ | 178424 - 180252 | $0.0039 \pm 0.0002 \pm 0.0013$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 160431 - 167913 | $-0.0045 \pm 0.0006 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 170000 - 172619 | $-0.0066 \pm 0.0008 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 172620 - 173692 | $-0.0058 \pm 0.0007 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 175830 - 177139 | $-0.0073 \pm 0.0006 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 177140 - 178421 | $-0.0075 \pm 0.0006 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 < 0.94$ | 178424 - 180252 | $-0.0071 \pm 0.0007 \pm 0.0071$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 160431 - 167913 | $0.0084 \pm 0.0013 \pm 0.0051$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 170000 - 172619 | $0.0063 \pm 0.0014 \pm 0.0051$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 172620 - 173692 | $0.0071 \pm 0.0013 \pm 0.0051$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 175830 - 177139 | $0.0056 \pm 0.0013 \pm 0.0051$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 177140 - 178421 | $0.0054 \pm 0.0013 \pm 0.0051$ |
| EB, $\lvert\eta\rvert > 1$, $r_9 > 0.94$ | 178424 - 180252 | $0.0058 \pm 0.0013 \pm 0.0051$ |

**Table B.2.:** Relative energy scale difference in data and MC ($\Delta P$) in the ECAL barrel, measured in $Z \to e^+e^-$ data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the $r_9$ distributions between electrons and photons [56].

| Category | Run Range | $\Delta P$ |
|---|---|---|
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 160431 - 167913 | $-0.0082 \pm 0.0008 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 170000 - 172619 | $-0.0025 \pm 0.0011 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 172620 - 173692 | $-0.0035 \pm 0.0010 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 175830 - 177139 | $-0.0017 \pm 0.0009 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 177140 - 178421 | $-0.0010 \pm 0.0009 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 < 0.94$ | 178424 - 180252 | $0.0030 \pm 0.0009 \pm 0.0088$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 160431 - 167913 | $-0.0033 \pm 0.0010 \pm 0.0018$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 170000 - 172619 | $0.0024 \pm 0.0012 \pm 0.0018$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 172620 - 173692 | $0.0014 \pm 0.0011 \pm 0.0018$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 175830 - 177139 | $0.0032 \pm 0.0010 \pm 0.0018$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 177140 - 178421 | $0.0040 \pm 0.0010 \pm 0.0018$ |
| EE, $\lvert\eta\rvert < 2$, $r_9 > 0.94$ | 178424 - 180252 | $0.0079 \pm 0.0010 \pm 0.0018$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 160431 - 167913 | $-0.0064 \pm 0.0008 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 170000 - 172619 | $-0.0046 \pm 0.0009 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 172620 - 173692 | $-0.0029 \pm 0.0009 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 175830 - 177139 | $-0.0040 \pm 0.0009 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 177140 - 178421 | $-0.0050 \pm 0.0008 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 < 0.94$ | 178424 - 180252 | $-0.0059 \pm 0.0009 \pm 0.0019$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 160431 - 167913 | $0.0042 \pm 0.0006 \pm 0.0028$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 170000 - 172619 | $0.0060 \pm 0.0008 \pm 0.0028$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 172620 - 173692 | $0.0077 \pm 0.0007 \pm 0.0028$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 175830 - 177139 | $0.0067 \pm 0.0007 \pm 0.0028$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 177140 - 178421 | $0.0056 \pm 0.0007 \pm 0.0028$ |
| EE, $\lvert\eta\rvert > 2$, $r_9 > 0.94$ | 178424 - 180252 | $0.0047 \pm 0.0007 \pm 0.0028$ |

**Table B.3.:** Relative energy scale difference in data and MC ($\Delta P$) in the ECAL endcaps, measured in $Z \to e^+e^-$ data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the $r_9$ distributions between electrons and photons [56].

## B.2. Binning Algorithm Optimisation

The optimisation procedure used to select the bin boundaries of the $H \to \gamma\gamma$ categorisation BDT involves a full scan over all combinations of bin boundaries. As this scan can be very slow, the procedure is separated into two parts, first a broad scan in large steps to find the region containing the optimum point then using small steps to refine the scan. The first step in the binning procedure is designed to ensure that at least 20 background events are expected in every bin. This gives a total of $B$ bins at a given luminosity. To maintain this feature, only boundaries which match any of the $B-1$ bin edges (remembering -1 and 1 are fixed boundaries) are scanned. The step size of the scan is therefore expressed as a step in number of bins so that for a given BDT output range, $(b_i, b_j)$ includes an integer number of the $B$ bins. The fine scan is defined to have a step size of 1, being the minimum step size defined this way. The step size for the broad scan, $P$, can be chosen to reduce the total time taken for the scan. For $N$ BDT boundaries, the scan is $N$-dimensional and the total number of points to scan (combinations of bin boundary values) assuming the two step procedure is given by,

$$\frac{1}{2^{N-1}}\left[\left(\frac{B}{P}\right)^N + (2P)^N\right] \tag{B.1}$$

imposing the condition $b_1 < b_2 < \cdots < b_N$. Figure B.1 shows the total number of iterations required to perform the full scan for different numbers of boundaries as a function of the broad step size $P$. The value, $P_{min}$, which minimises the total number of iterations is the same for any value of $N$ and is given by,

$$P_{min} = e^{\frac{1}{2}\ln(B/2)} \tag{B.2}$$

The scan is repeated, increasing the number of boundaries until the improvement in terms of the maximum expected significance in the presence of a SM Higgs boson is less than 0.1%. Figure B.2 shows the additional sensitivity gained as the number of final BDT output bins is increased for different starting values of $B$. The red curve is representative of the actual scan performed for the 2011 analysis.

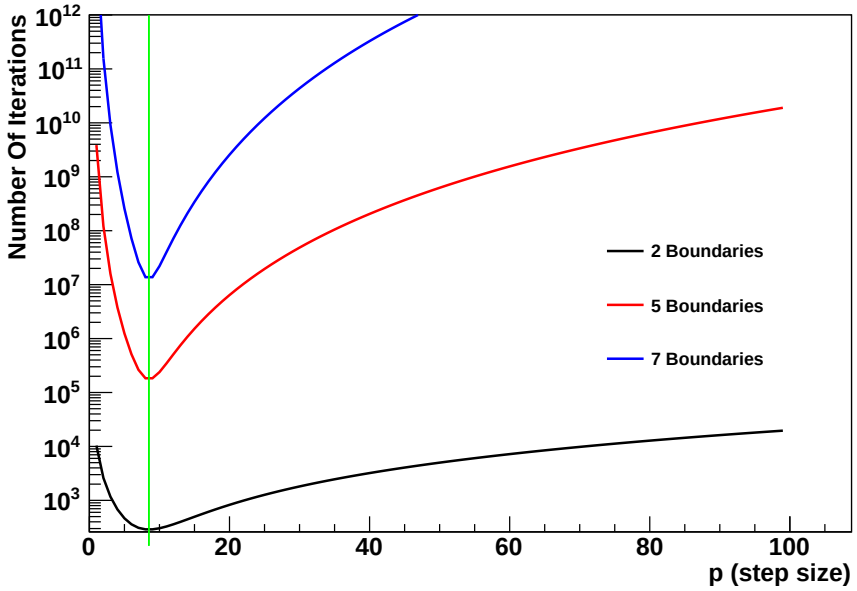


**Figure B.1.:** Total number of iterations in the binning optimization scan as a function of the broad step size $P$. The curve is shown for different numbers of final BDT boundaries. The minimum always occurs at the same value of $P$ as indicated by the green vertical line.

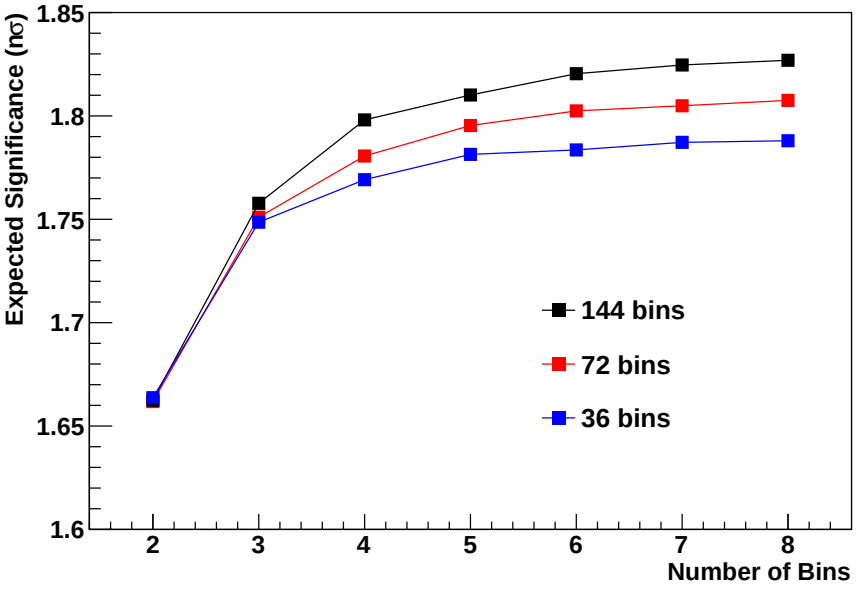


**Figure B.2.:** Increase in expected significance in the presence of a SM Higgs boson as the number of final BDT output bins is increased. The three curves show the improvement for different numbers of initial bins, $B$. The red curve is representative of the result obtained from performing the optimization procedure in the 2011 analysis.

## B.3. Signal Systematics

The treatment of systematic variations in the signal modelling for the $H \to \gamma\gamma$ analysis described in Chapter 4 is the same for all uncertainties except those due to the theoretical uncertainty on the Higgs boson production cross-sections and the integrated luminosity measurement. For each uncertainty, the relevant quantity in the MC is varied by $3\sigma$ and the resulting BDT distributions are compared to the nominal one. The three "templates" (corresponding to nominal and $\pm 3\sigma$ variations) are used to determine the $1\sigma$ variations of the $j$-th BDT bin of the signal model due to the $k$-th signal systematic ($\sigma_k^{s,p}$ used in Equation 4.15). The procedure is performed for each signal process, $p$, separately. The value for the $1\sigma$ variation in each bin is given by,

$$\sigma^{\pm} = a \pm b + c, \tag{B.3}$$

where $\sigma^+$ is the value of $\sigma_k^{s,p}$ used for positive values of the associated nuisance parameter and $\sigma^-$ is for negative values. The parameters $a$ and $b$ are determined for a particular bin by solving the set of simultaneous equations;

$$\begin{pmatrix} s^{-3\sigma} \\ s^{mc} \\ s^{3\sigma} \end{pmatrix} = \begin{bmatrix} 9 & -3 & 1 \\ 0 & 0 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \tag{B.4}$$

where $s^{mc}$ is the nominal value for the signal in that bin and $s^{\pm 3\sigma}$ are the values determined from the $\pm 3\sigma$ templates.

Figure B.3 shows the $\pm 3\sigma$ and $1\sigma$ variations of the BDT distribution expected from the $ggH$ production process calculated from MC and using the interpolation procedure respectively. The distributions are normalised to the expectation in $5.1 fb^{-1}$. The energy scale and resolution uncertainties can be found in Section 4.4.5 (Figure 4.28).

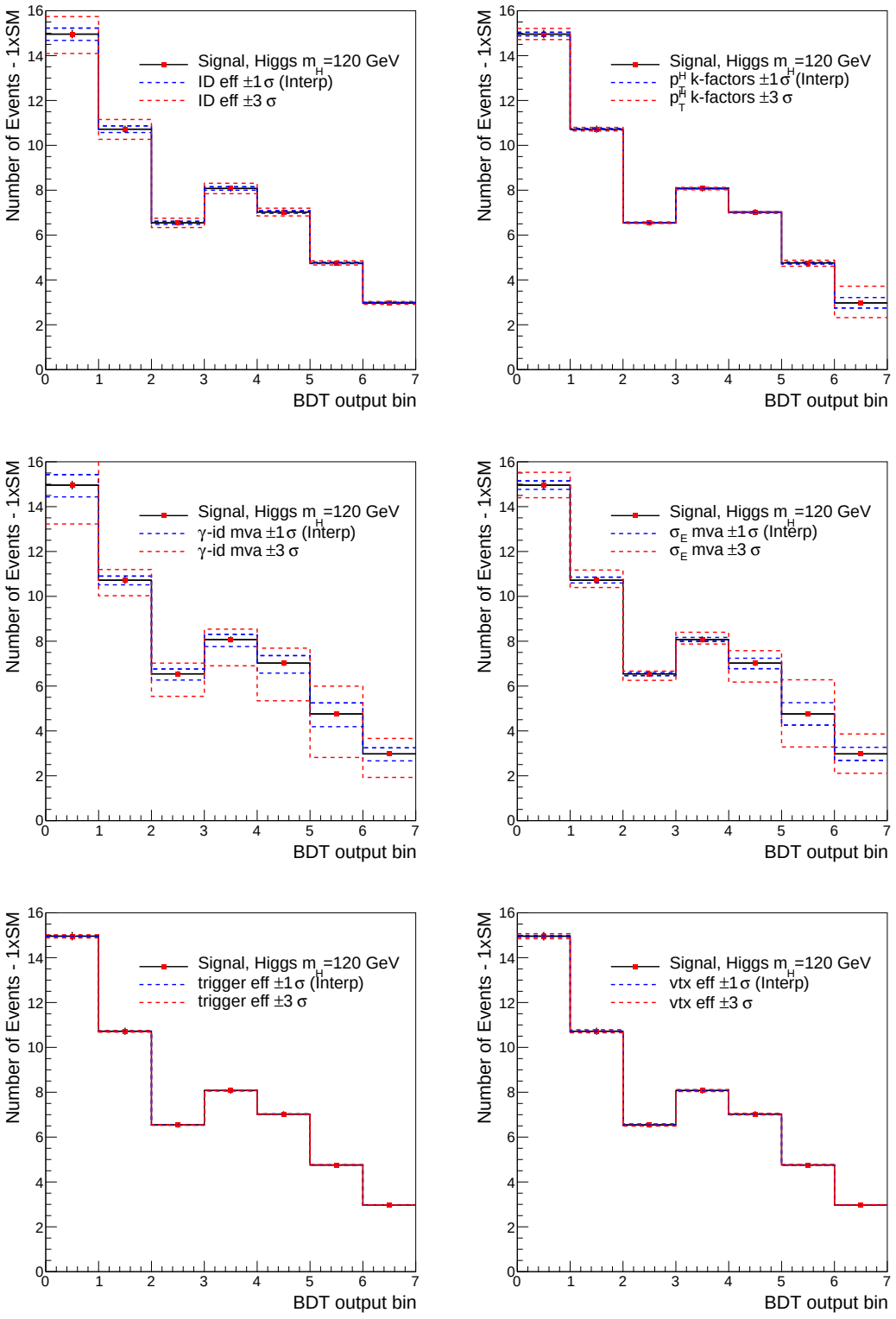


**Figure B.3.:** Systematic uncertainties on the $ggH$ signal model. The effects of $\pm 3\sigma$ variations derived in MC is shown with red dashed lines while the interpolated $\pm 3\sigma$ are shown with blue.

# Appendix C.

## C.1. Per-event Log-likelihood Ratio

The combined observed significance from the ICHEP 2012 dataset, corresponding to $4.9\sigma$, is driven largely by the two high resolution channels $H \to \gamma\gamma$ and $H \to ZZ \to 4l$. Those two channels alone combine provide a significance of $5.0\sigma$ at $m_H = 125$ GeV. Before entering the likelihood of Equation 6.1, the contribution from the events in each channel are first combined. For each channel, the sum over events $i$,

$$\log \mathcal{L}_{\text{channel}}(\text{data}|\mu) = \sum_i \log \mathcal{L}(i|\mu), \tag{C.1}$$

represents summing the per-event log-likelihood at a given value of $\mu$. As usual, the value of $m_H$ is implicitly assumed in the definition of the likelihood, $\mathcal{L}$. Here the nuisance parameters are not explicitly indicated, although they are profiled in the usual way at a given value of $\mu$. The test-statistic appropriate for determining significances, $q_0$, can be expressed as the difference in the negative log-likelihoods ($\Delta(nll)$) for $\mu = 0$ and $\mu = \hat{\mu}$ (the best fit signal strength),

$$q_0 = -2\left[\log \mathcal{L}(i|\mu = \hat{\mu}) - \log \mathcal{L}(i|\mu = 0)\right] = 2\Delta(nll). \tag{C.2}$$

The individual contribution from each event in data in each channel can therefore be determined by considering the per-event delta log-likelihood. Figure C.1 shows the distribution of the per-event log-likelihood in data for the $H \to \gamma\gamma$ and $H \to ZZ$ channels. The distributions expected under the background-only and signal-plus-background hypotheses, where $\mu$ is set to the best fit value, are also shown. In these two channels, there is an additional term in the likelihood which represents the normalization of the signal plus background model as the likelihood in these cases is unbinned. It

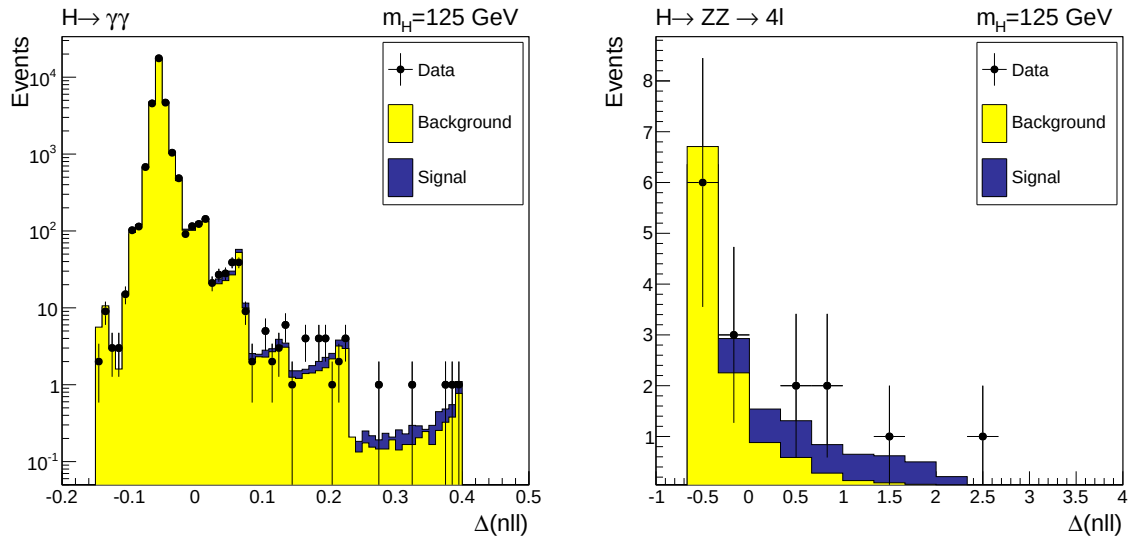


**Figure C.1.:** Per-event delta negative log-likelihood ($\Delta nll$) distributions for the background-only and signal-plus-background hypotheses in the ICHEP 2012 $H \to \gamma\gamma$ (left) and $H \to ZZ \to 4l$ (right) analyses. The distributions for the observed events from each channel are indicated by the black points. The likelihoods are evaluated for $m_H = 125$ GeV at the best fit values of $\mu$ from the combination of these two channels only.

should be noted that the best fit value for $\mu$ is evaluated from the combined data in the $H \to \gamma\gamma$ and $H \to ZZ$ channels so that the individual contributions from each datum can be positive or negative.

## C.2. Feldman-Cousins Boundary Effects

The Feldman-Cousins procedure used to check the compatibility of the new observed particle with the Standard Model Higgs boson typically produces the same 68% confidence contours as obtained from scanning $q_{\mathbf{x}}$. Disagreement between the two methods is usually observed where the best fit value is outside the physically allowed region. However, for contours which are close to the boundaries of the physical region, the two methods will yield different results even if the best fit point is inside the allowed region. A simple demonstration of this effect can be seen in Figure C.2 which shows two contours in the $\mu_{VH+qqH}$, $\mu_{ggH+ttH}$ plane obtained from data in the $H \to \gamma\gamma$ analysis on the ICHEP dataset. The signal extraction technique used here is the binned technique described in Section 4.4. The two contours shown are those at the 50% and 75% confidence levels from each method. These contours are chosen specifically in this case to demonstrate the

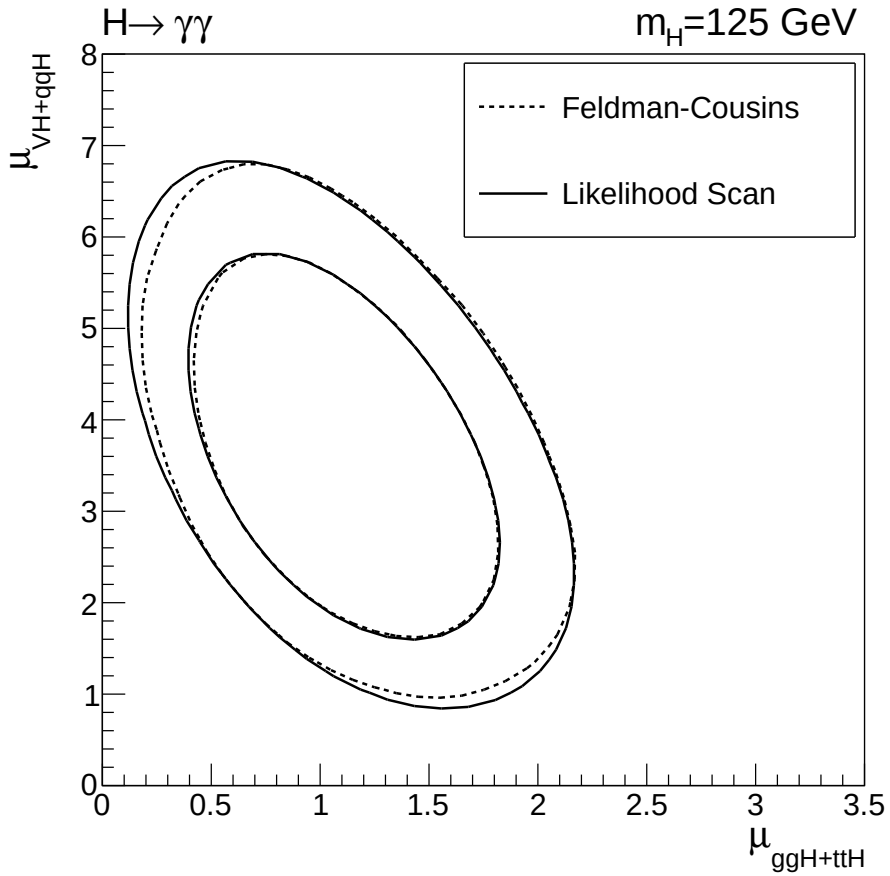


**Figure C.2.:** Comparison between 50% (inner) and 75% (outer) contours in data from the $H \to \gamma\gamma$ channel as determined using the Feldman-Cousins and a scan of $q_{\mathbf{x}}$ (labelled "Likelihood Scan"). In the Feldman-Cousins technique, the constraints, $\mu_{ggH+ttH} \geq 0$ and $\mu_{VH+qqH} \geq 0$ are imposed.

effect of the boundaries at $\mu_{VH+qqH} = 0$ and $\mu_{ggH+ttH} = 0$. Although the 50% contours agree well between the two methods, disagreement can be seen between the 75% contours where the contour is close to one of the boundaries.

# Bibliography


[1] The CMS Collaboration. Measurement of the Inclusive $W$ and $Z$ Production Cross Sections in $pp$ Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 1110:132, 2011. doi: 10.1007/JHEP10(2011)132.

[2] N. Wardle, D. Futyan, J. Hays, N. Rompotis, C. Seez, T. Virdee, and D. Wardrope. Extraction of the $w \to e\nu$ signal yield in pp collisions at CMS using the ABCDE method. *CMS-AN-11/009*, 2011.

[3] R. C. Lopes de Sa. Precise measurements of the $W$ mass at the Tevatron and indirect constraints on the Higgs mass. *ArXiv e-prints: 1204.3260*, 2012.

[4] J. Beringer and others. (Paricle Data Group). Review of Particle Physics. *Phys. Rev. D*, 86:010001, Jul 2012. doi: 10.1103/PhysRevD.86.010001. URL `http://link.aps.org/doi/10.1103/PhysRevD.86.010001`.

[5] E. Noether. Invariant variation problems. *Transport Theory and Statistical Physics*, 1:186–207, 1971. doi: 10.1080/00411457108231446.

[6] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl.Phys.*, 22:579–588, 1961. doi: 10.1016/0029-5582(61)90469-2.

[7] S. Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19: 1264–1266, Nov 1967. doi: 10.1103/PhysRevLett.19.1264. URL `http://link.aps.org/doi/10.1103/PhysRevLett.19.1264`.

[8] A. Salam. Weak and Electromagnetic Interactions. *Conf.Proc.*, C680519:367–377, 1968.

[9] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental Test of Parity Conservation in Beta Decay. *Phys. Rev.*, 105:1413–1415, Feb 1957. doi: 10.1103/PhysRev.105.1413. URL `http://link.aps.org/doi/10.1103/PhysRev.105.1413`.

[10] I. J. R. Aitchison and A. J. G. Hey. *Gauge Theories in Particle Physics, 2 Volume Set*. Taylor & Francis, 3 edition, January 2004. ISBN 0750309822. URL http://www.worldcat.org/isbn/0750309822.

[11] F. Halzen and A. D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. Wiley, February 1984. ISBN 0471887412. URL http://www.worldcat.org/isbn/0471887412.

[12] P. W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys.Lett.*, 12: 132–133, 1964. doi: 10.1016/0031-9163(64)91136-9.

[13] T. W. B. Kibble. Symmetry breaking in non-abelian gauge theories. *Phys. Rev.*, 155:1554–1561, Mar 1967. doi: 10.1103/PhysRev.155.1554. URL http://link.aps.org/doi/10.1103/PhysRev.155.1554.

[14] P. W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys.Rev.Lett.*, 13:508–509, 1964. doi: 10.1103/PhysRevLett.13.508.

[15] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global Conservation Laws and Massless Particles. *Phys.Rev.Lett.*, 13:585–587, 1964. doi: 10.1103/PhysRevLett.13.585.

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

[17] D. M. Webber et al. Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision. *Phys. Rev. Lett.*, 106:041803, Jan 2011. doi: 10.1103/PhysRevLett.106.041803. URL http://link.aps.org/doi/10.1103/PhysRevLett.106.041803.

[18] J. Ellis, M. K. Gaillard, and D. V. Nanopoulos. A Historical Profile of the Higgs Boson. *ArXiv e-prints: 1201.6045*, 2012.

[19] U. M. Heller, M. Klomfass, H. Neuberger, and P. Vranas. Numerical analysis of the Higgs mass triviality bound. *Nuclear Physics B*, 405:555–573, September 1993. doi: 10.1016/0550-3213(93)90559-8.

[20] J. Beringer and others (Particle Data Group). Higgs Bosons: Theory and Searches. *Phys. Rev. D*, 86:010001, 2012.

[21] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, and LEP Working Group For Higgs Boson Searches. Search for the Standard Model Higgs boson at LEP. *Physics Letters B*, 565:61–75, July 2003. doi: 10.1016/S0370-2693(03)00614-2.

[22] The TEVNPH Working Group and CDF and D0 Collaborations. Combined CDF and D0 Search for Standard Model Higgs Boson Production with up to 10.0 $fb^{-1}$ of Data. *ArXiv e-prints:1203.3774*, March 2012.

[23] LEPEWWG. The LEP Electroweak Working Group. 2012. http://lepewwg.web.cern.ch/LEPEWWG/.

[24] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.). Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. *CERN-2011-002*, CERN, Geneva, 2011.

[25] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.). Handbook of LHC Higgs Cross Sections: 2. Differential Distributions. *CERN-2012-002*, CERN, Geneva, 2012.

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

[27] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08003, 2008. URL `http://stacks.iop.org/1748-0221/3/i=08/a=S08003`.

[28] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3:S08004, 2008. doi: 10.1088/1748-0221/3/08/S08004.

[29] The LHCb Collaboration. The lhcb detector at the lhc. *Journal of Instrumentation*, 3(08):S08005, 2008. URL `http://stacks.iop.org/1748-0221/3/i=08/a=S08005`.

[30] CERN. CMS Compact Muon Solenoid. Feb 2010. http://public.web.cern.ch/public/Objects/LHC/CMSnc.jpg.

[31] A. Tricomi. Performances of the ATLAS and CMS silicon tracker. *The European Physical Journal C - Particles and Fields*, 33:s1023–s1025, 2004. ISSN 1434-6044. doi: 10.1140/epjcd/s2004-03-1801-1. URL `http://dx.doi.org/10.1140/epjcd/s2004-03-1801-1`.

[32] M. Weber. Calibration, alignment and tracking performance of the cms silicon strip tracker. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 628(1): 59 – 63, 2011. ISSN 0168-9002. doi: 10.1016/j.nima.2010.06.284. URL http://www.sciencedirect.com/science/article/pii/S0168900210014749. Proceedings of the 12th International Vienna Conference on Instrumentation.

[33] CMS Collaboration. Tracking and Primary Vertex Results in First 7 TeV Collisions. *CMS-PAS-TRK-10-005*, 2010.

[34] The CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006.

[35] P. Adzic and others. Energy resolution performance of the CMS electromagnetic calorimeter. *CMS-AN-06/140*, 2006.

[36] E. Meschi, T. Monteiro, C. Seez, and P. Vikas. Electron Reconstruction in the CMS Electromagnetic Calorimeter. *CMS-AN-01/034*, 2001.

[37] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois. Electron reconstruction in CMS. *The European Physical Journal C*, 49:1099–1116, 2007. ISSN 1434-6044. doi: 10.1140/epjc/s10052-006-0175-5. URL http://dx.doi.org/10.1140/epjc/s10052-006-0175-5.

[38] W. Adam et al. Electron reconstruction in CMS. *CMS-AN-09/164*, 2009.

[39] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov. Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9, 2005.

[40] The CMS Collaboration. ECAL Detector Performance, 2011 Data. *CMS-DP-2012-007*, May 2012.

[41] The CMS Collaboration. The trigger and data acquisition project technical design report, volume 1, the level-1 trigger. *CERN/LHCC 2000-038, CMS TDR 6.1*, 2000.

[42] The CMS Collaboration. The trigger and data acquisition project technical design report, volume 2: Data acquisition and high-level trigger. *CERN/LHCC 2002-026, CMS TDR 6.2*, 2002.

[43] M. Cacciari, G. P. Salam, and G. Soyez. The anti-$k_t$ jet clustering algorithm. *Journal of High Energy Physics*, (04):12, 2008. URL http://arxiv.org/abs/0802.1189.

[44] The CMS Collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6(11):P11002, 2011. URL `http://stacks.iop.org/1748-0221/6/i=11/a=P11002`.

[45] J. Brooke, B. Mathias, A. Tapper, and N. Wardle. Calibration and Performance of the Jets and Energy Sums in the Level-1 Trigger. *CMS-INTERNAL-NOTE*, 2012.

[46] The CMS Collaboration. Search for the standard model Higgs boson decaying into two photons in pp collisions at CMS. *Physics Letters B*, 710(3): 403 – 425, 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.03.003. URL `http://www.sciencedirect.com/science/article/pii/S0370269312002547`.

[47] A. Bengalia et al. Search for a Standard Model Higgs boson decaying into two photons employing multivariate methods. *CMS-AN-12/048*, 2012.

[48] C. Oleari. The POWHEG BOX. *Nuclear Physics B Proceedings Supplements*, 205: 36–41, August 2010. doi: 10.1016/j.nuclphysbps.2010.08.016.

[49] G. Bozzi, S. Catani, D. de Florian, and M. Grazzini. The qT spectrum of the Higgs boson at the LHC in QCD perturbation theory. *Physics Letters B*, 564 (12):65 – 72, 2003. ISSN 0370-2693. doi: 10.1016/S0370-2693(03)00656-7. URL `http://www.sciencedirect.com/science/article/pii/S0370269303006567`.

[50] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. *ArXiv e-prints*, January 2011.

[51] T. Sjöstrand, S. Mrenna, and P. Skands. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics*, 5:026, May 2006. doi: 10.1088/1126-6708/2006/05/026.

[52] S. Agostinelli, J. Allison, K. Amako, et al. Geant4 - a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003. ISSN 0168-9002. doi: 10.1016/S0168-9002(03)01368-8. URL `http://www.sciencedirect.com/science/article/pii/S0168900203013688`.

[53] The CMS Collaboration. CMS Software Page. 2013. https://cms-cpt-software.web.cern.ch/cms-cpt-software/General.

[54] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis. *ArXiv e-prints:0703039*, March 2007.

[55] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29:1189–1232, 2000. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.9093`.

[56] A. Benaglia et al. Search for a Standard Model Higgs boson decaying into two photons. *CMS-AN-12/160*, 2012.

[57] P. D. Dauncey, M. Kenzie, and C. Seez. Residual photon energy corrections and resolution from simulation. *CMS-AN-11/343*, 2011.

[58] T. Skwarnicki. *A study of the radiative cascade transitions between the Upsilon-prime and Upsilon resonances*. PhD thesis, Institute of Nuclear Physics, Krakow, 1986. http://inspirehep.net/record/230779/files/230779.pdfDESY-F31-86-02.

[59] W. Erdmann. Offline Primary Vertex Reconstruction with Deterministic Annealing Clustering. *CMS-IN-11/014*, 2011.

[60] The Egamma ID Group. Tag and probe methodology for analyses using electrons and photons. *CMS-AN-12/116*, 2012.

[61] The CMS Collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6:11002, November 2011. doi: 10.1088/1748-0221/6/11/P11002.

[62] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput.Phys.Commun.*, 10:343–367, 1975. doi: 10.1016/0010-4655(75)90039-9.

[63] H. Abdi and L. J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. ISSN 1939-0068. doi: 10.1002/wics.101. URL `http://dx.doi.org/10.1002/wics.101`.

[64] N. Wardle. Higgs to two photon binned bdt distribution plots for the 2011 dataset. 2013. http://nckw.web.cern.ch/nckw/hgg-fmva-2011/data-model.

[65] The CMS Collaboration. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *CMS-PREPRINT (submitted to JHEP)*, 2013.

[66] F. E. James. *Statistical Methods in Experimental Physics: 2nd Edition*. World Scientific Publishing, February 2006.

[67] A. L. Read. Presentation of search results: the CLs technique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693, 2002. URL `http://stacks.iop.org/0954-3899/28/i=10/a=313`.

[68] W. Verkerke and D. P. Kirkby. The RooFit toolkit for data modeling. *eConf*, C0303241:MOLT007, 2003.

[69] G. Petrucciani. Documentation of the RooStats-based statistics tools for Higgs PAG. 2013. https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideHiggsAnalysisCombinedLimit.

[70] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71: 1–19, 2011. ISSN 1434-6044. doi: 10.1140/epjc/s10052-011-1554-0. URL `http://dx.doi.org/10.1140/epjc/s10052-011-1554-0`.

[71] L. Lyons. Open statistical issues in particle physics. *ArXiv e-prints: 0811.1663*, November 2008.

[72] E. Gross and O. Vitells. Trial factors for the look elsewhere effect in high energy physics. *European Physical Journal C*, 70:525–530, 2010. doi: 10.1140/epjc/s10052-010-1470-8.

[73] The CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1): 30 – 61, 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.021. URL `http://www.sciencedirect.com/science/article/pii/S0370269312008581`.

[74] G. Davies, J. Hays, and N. Wardle. Diagnostic tools for Higgs searches and properties measurements at CMS. *CMS-AN-12/317*, 2012.

[75] N. Wardle. Higgs Combination Diagostic summary plots for the ICHEP 2012 dataset. 2013. http://nckw.web.cern.ch/nckw/combination-diagnostics-ichep2012.

[76] The CMS Collaboration. Observation of a new boson with a mass near 125 GeV. *CMS-PAS-HIG-12-020*, 2012.

[77] The CMS Collaboration. Evidence for a new state decaying into two photons in the search for the Standard Model Higgs boson in pp collisions. *CMS-PAS-HIG-12-015*, 2012.

[78] The CMS Collaboration. Search for the Standard Model Higgs boson produced in association with W or Z bosons, and decaying to bottom quarks for ICHEP 2012. *CMS-PAS-HIG-12-019*, 2012.

[79] The CMS Collaboration. Search for a Standard Model Higgs boson decaying to tau pairs in pp collisions. *CMS-PAS-HIG-12-018*, 2012.

[80] The CMS Collaboration. Search for the Standard Model Higgs boson decaying to a W pair in the fully leptonic final state in pp collisions at $\sqrt{s} = 8$ TeV. *CMS-PAS-HIG-12-017*, 2012.

[81] The CMS Collaboration. Search for the Standard Model Higgs boson in the H to WW to lvjj decay channel. *CMS-PAS-HIG-12-021*, 2012.

[82] The CMS Collaboration. Study of associated Higgs boson (WH) Production in the three leptons nal state at 7 TeV. *CMS-PAS-HIG-11-034*, 2012.

[83] The CMS Collaboration. Evidence for a new state in the search for the Standard Model higgs boson in the H to ZZ to 4 leptons channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *CMS-PAS-HIG-12-016*, 2012.

[84] The CMS Collaboration. Search for the Standard Model Higgs Boson in the decay channel H to ZZ(*) to q-qbar $l^-l^+$ at CMS. *CMS-PAS-HIG-11-027*, 2011.

[85] The CMS Collaboration. Search for the standard model higgs boson in the $h \to zz \to 2l2\nu$ channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *CMS-PAS-HIG-12-023*, 2012.

[86] The CMS Collaboration. Combination of Standard Model Higgs boson searches and measurements of the properties of the new boson with a mass near 125 GeV. *CMS-PAS-HIG-12-045*, 2012.

[87] G. J. Feldman and R. D. Cousins. Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, 57:3873–3889, Apr 1998. doi: 10.1103/PhysRevD.57.3873. URL `http://link.aps.org/doi/10.1103/PhysRevD.57.3873`.

[88] LHC Higgs Cross Section Working Group, A. David, A. Denner, M. Duehrssen, M. Grazzini, et al. LHC HXSWG interim recommendations to explore the coupling structure of a Higgs-like particle. 2012.

# List of Figures

# List of Tables