MEASUREMENT OF THE HIGGS BOSON CROSS-SECTION AND COUPLINGS IN THE TWO PHOTON DECAY CHANNEL

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Michael David Planer

Colin Jessop, Co-Director

Nancy Marinelli, Co-Director

Graduate Program in Physics
Notre Dame, Indiana
July 2018
This document is in the public domain.
The results of the analysis of the Higgs boson decaying into two photons with the 2016 dataset are described. The analysis is performed using the dataset recorded by the CMS experiment at the LHC from pp collisions at center-of-mass energies of 13 TeV corresponding to an integrated luminosity of 35.9 fb$^{-1}$. Events were classified to maximize signal efficiency and to study gluon fusion, vector boson fusion, vector boson associated production, and top fusion Higgs boson production modes. The observed significance at $m_H = 125.09$ GeV is $1.16^{+0.15}_{-0.14}$ times the standard model expectation. Signal strengths for the different Higgs boson production modes, coupling modifiers to bosons and fermions, and effective couplings to photons and gluons are also presented.
To my parents, Norman and Sharon.
CONTENTS

FIGURES ................................................................. iv

TABLES ................................................................. v

ACKNOWLEDGMENTS ................................................... vi

CHAPTER 1: INTRODUCTION ........................................... 1

CHAPTER 2: THEORY ................................................... 3
  2.1 The Standard Model Higgs boson ................................ 3
    2.1.1 Overview .................................................. 3
    2.1.2 Example of local gauge invariance: QED .................. 5
    2.1.3 Spontaneous electroweak symmetry breaking and the Higgs boson 7
  2.2 Higgs production and decay modes ............................ 9
  2.3 Higgs boson properties study in the two photon final state .... 13

CHAPTER 3: THE DETECTOR AND EXPERIMENTAL APPARATUS .... 15
  3.1 The Large Hadron Collider .................................... 15
  3.2 The Compact Muon Solenoid .................................. 18
    3.2.1 Coordinate system ...................................... 19
    3.2.2 The magnet .............................................. 20
    3.2.3 The tracking system .................................... 20
    3.2.4 The electromagnetic calorimeter ......................... 23
    3.2.5 The hadronic calorimeter ............................... 27
    3.2.6 The muon system ........................................ 28
      3.2.6.1 Drift tubes ....................................... 29
      3.2.6.2 Cathode strip chambers ......................... 30
      3.2.6.3 Resistive plate chambers ....................... 31
    3.2.7 The trigger system .................................... 32
      3.2.7.1 The level 1 trigger ............................ 32
      3.2.7.2 The high level trigger ......................... 32
    3.2.8 CMS simulation ......................................... 33
<table>
<thead>
<tr>
<th>FIGURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Particle content of the SM ..................</td>
</tr>
<tr>
<td>2.2 Higgs field potential ........................</td>
</tr>
<tr>
<td>2.3 Higgs boson production modes ................</td>
</tr>
<tr>
<td>2.4 Higgs boson decay to two photons ............</td>
</tr>
<tr>
<td>2.5 Diagrams of backgrounds to the Higgs boson decay into two photons</td>
</tr>
<tr>
<td>3.1 LHC Complex ...................................</td>
</tr>
<tr>
<td>3.2 CMS diagram ...................................</td>
</tr>
<tr>
<td>3.3 CMS Tracker Diagram ...........................</td>
</tr>
<tr>
<td>3.4 Tracker Material Budget .......................</td>
</tr>
<tr>
<td>3.5 CMS ECAL diagram ................................</td>
</tr>
<tr>
<td>3.6 CMS HCAL diagram ................................</td>
</tr>
<tr>
<td>3.7 CMS Drift Tube Diagram ........................</td>
</tr>
<tr>
<td>3.8 CMS CSC Diagram ................................</td>
</tr>
<tr>
<td>4.1 ECAL Pulse Shape ................................</td>
</tr>
<tr>
<td>4.2 Shower Shape Correction (R9) ..................</td>
</tr>
<tr>
<td>4.3 Energy Scale vs Run Before Correction .......</td>
</tr>
<tr>
<td>4.4 Energy Scale vs Run After Correction ........</td>
</tr>
<tr>
<td>4.5 Energy Scale and Smearing Corrections ........</td>
</tr>
<tr>
<td>4.6 Energy Scale and Smearing Corrections ........</td>
</tr>
<tr>
<td>5.1 Distribution of number of reconstructed vertices</td>
</tr>
<tr>
<td>5.2 Level 1 trigger efficiency vs $p_T$ ..........</td>
</tr>
<tr>
<td>5.3 Level 1 trigger efficiency vs $\eta$ ..........</td>
</tr>
<tr>
<td>5.4 HLT trigger efficiency vs $p_T$ ..............</td>
</tr>
<tr>
<td>5.5 HLT trigger efficiency vs $\eta$ .............</td>
</tr>
<tr>
<td>5.6 Photon BDT distribution ........................</td>
</tr>
<tr>
<td>5.7 Validation of Photon ID with $Z \rightarrow ee$</td>
</tr>
</tbody>
</table>
TABLES

2.1 STANDARD MODEL 125 GeV HIGGS BOSON PRODUCTION CROSS SECTION ........................................... 11
2.2 STANDARD MODEL 125 GeV HIGGS BOSON DECAY MODES ................................................................. 13
5.1 LIST OF DATA SAMPLES USED IN THE ANALYSIS ................................................................. 48
5.2 LIST OF SIMULATED SIGNAL SAMPLES .................................................................................. 52
5.3 LIST OF SIMULATED BACKGROUND SAMPLES ........................................................................ 53
5.4 SELECTION FOR THE DIPHOTON TRIGGER ......................................................................... 57
5.5 PRESELECTION REQUIREMENTS ............................................................................................... 60
5.6 PRESELECTION EFFICIENCY MEASUREMENT ........................................................................ 61
5.7 CONVERSION-SAFE ELECTRON VETO EFFICIENCY MEASUREMENT .............................................. 62
7.1 EXPECTED YIELDS .................................................................................................................. 135
7.2 ENERGY SCALE UNCERTAINTIES .......................................................................................... 139
7.3 SYSTEMATIC CONTRIBUTIONS ............................................................................................... 141
ACKNOWLEDGMENTS

I would like to acknowledge my two advisors, Prof. Colin Jessop and Research Associate Prof. Nancy Marinelli, through their support, I was able to contribute to one of the most impressive accomplishments of human science and engineering to date. I also wish to thank Federico Ferri, Seth Zenz, Martina Malberti, Shervin Nourbakhsh (Analysis conveners) throughout my time on the analysis, as well as the rest of the Higgs to diphoton analysis team. A big thanks to the members of my defense committee (Prof. Antonio Delgado, Prof. Kevin Lannon, And Prof. Anna Simon) for reading my thesis, which was no small task. My many office mates in 512: Dr Matthias Wolf, Dr Geoff Smith, Fanbo Meng, and Dr Yutaro Iiyama provided both technical support and an atmosphere of fun. My lunchtime group, which changed considerably over my time at CERN: Prof. Ted Kolberg, Prof. Rachel Yohay, Dr David Morse, Dr Bingxuan Liu, Dr Paul Lujan, Juliana Froggatt, Dr Kelly Beernaert, Dr Jess Brinson, and Dr Justin Pilot contributed to my happiness and offered sound advice. My friends at Notre Dame were important to me as well, but of particular import is Christopher Wotta. A more trustworthy friend is hard to come by and knowing that I always had a place to sleep in South Bend was of no small comfort. During my time at CERN, I was blessed with profound friendships which helped me progress both in physics and life. Dr Sarah Boutle was, for a time, my only officemate and helped make the first year at CERN a delightful experience. Dr Vera Chetvertkova was a wonderful companion on the bus and through many adventures in the Franco-Swiss region. Dr Juliette Alimena’s friendship and patience meant a lot to me during the final year that I spent at CERN. Dr Charlie Mueller was a brilliant
friend throughout graduate school both at Notre Dame and at CERN. And finally, I thank my parents, Norm and Sharon, for all their support over the years. They provided me with every opportunity to succeed, while giving me the space to create and solve many of my own challenges. Through their love and guidance I’ve grown from a small child who kept falling out of trees while wondering how the world works to an adult who quantifies the workings of the world (and who only falls out of trees when he intends to).
CHAPTER 1

INTRODUCTION

The Standard Model of particle physics has accurately predicted many measurements, the last confirmed piece being the discovery of the Higgs boson in 2012 [8], [16]. There are still open questions in physics which cannot be answered by the Standard Model, however, by measuring precisely the parameters of the Higgs boson, it may be possible to constrain new physics and help focus the development of new theoretical models.

While the branching ratio of the Higgs boson through the $H \rightarrow \gamma\gamma$ decay channel is small ($\approx 0.2\%$) the two photon final state presents a clean invariant mass peak that can be precisely reconstructed. This channel was one of the most important in the initial discovery of the Higgs boson and continues to be one of the best tools to measure the properties of the Higgs boson. As more data are recorded, it becomes possible to analyze the different production modes of the Higgs boson at the LHC as it decays into two photons. This allows measurement of the couplings of the Higgs boson to fermions and bosons, and the effective couplings to gluons and photons.

This thesis corresponds to the a preliminary $H \rightarrow \gamma\gamma$ analysis covering the 2016 dataset documented in [18] by the CMS group and presents measurements of the Higgs boson. These include the signal strength relative to Standard Model prediction, the coupling modifiers between the Higgs boson and fermions and bosons, and the effective coupling between the Higgs boson and photons and gluons. Earlier versions of this analysis on the 2012 dataset corresponding to the discovery of the Higgs boson were also published by CMS [13].
First, the theory of the Standard Model of particle physics will be discussed in general, focusing on the Higgs boson and relevant interactions for its measurement at the LHC in the diphoton decay channel. A discussion of the LHC and the CMS detector, with specific focus on those elements which are central to this analysis, will follow. Presented next are details of the selection requirements to identify signal events and sort the analysis into categories to improve the overall performance of the analysis. Finally, a discussion of the statistical and systematic treatment of the data are described and the final results are presented.

My main contributions to the analysis were the design and testing of the analysis trigger (described in section 5.3), as well as characterizing the $p_T^{\text{miss}}$, and defining the VH categories (focusing on the $E_T^{\text{miss}}$ category described in section 6.2.2).
2.1 The Standard Model Higgs boson

Nearly all of the predictions of the Standard Model of particle physics (SM) have been confirmed by experimental evidence, the latest being the discovery of the Higgs boson [8] [16] in 2012, the last building block of the theory. However, there are still several open questions about the nature of the Higgs boson and its coupling to SM particles.

2.1.1 Overview

The SM describes the interaction of all known particles through the electromagnetic, weak, and strong forces [25] [37] [33]. The SM is a mathematical representation of quantum field theory (QFT) governing those three forces. At present, gravitation is not included in the SM and several observations (dark matter, dark energy) are not represented by the SM.

Fermions and bosons make up the particle content of the SM. Fermions are half-integer spin particles that comprise six leptons and six quarks. Three leptons are neutral, called neutrinos and the other three are charged (±1). Three quarks have two-thirds charge (±2/3) while the remaining three have one-third charge (±1/3). Bosons are integer spin particles that comprise 12 gauge bosons: eight gluons, three weak gauge bosons, one photon and one Higgs boson. The gauge bosons are excitations of their fields that allow the interactions of the respective forces, while the
Higgs field gives mass to the fermions and bosons. Figure 2.1 shows a diagram of the particles of the SM.

**Figure 2.1.** A summary of all elementary particles and their interactions in the Standard Model [31].

The Standard Model is a fully renormalizable field theory given by the symmetry group:

\[
\text{SU}(3)_C \otimes \text{SU}(2)_W \otimes \text{U}(1)_Y
\]

(2.1)

where the SU(3)$_C$ group is the color symmetry group that defines quantum chromodynamics (QCD), SU(2)$_W$ group is the gauge group that defines weak isospin, and
U(1)\gamma group defines the hypercharge. The electromagnetic and weak interactions are unified by the electroweak gauge group \(SU(2)_W \otimes U(1)_Y\). The Standard Model is generally presented using the Lagrangian formalism based on the kinematic principle of least action [26].

QCD provides eight spin one gluons as excitations of the field. Only gluons and quarks carry color charge and can interact via QCD. The electroweak gauge group provides three spin one bosons from the SU(2) group and one from the U(1) group. Through the Weinberg-Salam model, these produce the \(W^\pm\) and \(Z\) bosons as well as the photon. Any charged particle can interact with a photon through the electromagnetic force, while any fermions can interact with \(W^\pm\) or \(Z\) bosons through the weak force.

The addition of the Higgs scalar field breaks the electroweak symmetry and grants the bosons mass while maintaining renormalizability and unitarity. This scalar field also interacts with leptons and gives them mass through Yukawa coupling. This chapter will go into somewhat further discussion of electroweak symmetry breaking as it is of particular relevance to the analysis.

2.1.2 Example of local gauge invariance: QED

Any quantum field theory must be locally gauge invariant: the Lagrangian must be invariant under group transformations that depend on position in spacetime. This naturally gives rise to massless gauge bosons defined by the group symmetry (\(N^2-1\) bosons for SU(\(N\)) symmetry groups).

A brief explanation from QED showing how the photon arises from local gauge invariance on the U(1) symmetry group follows.

The basic Lagrangian for QED is represented with an arbitrary spin 1/2 field:

\[
\mathcal{L} = \bar{\Psi} \left( i \gamma^\mu \partial_\mu - m \right) \Psi
\]  

(2.2)
where $\Psi$ is an arbitrary fermionic free particle field, and $\gamma^\mu$ are the four Dirac matrices. The Lagrangian is invariant under the global transformation:

$$\Psi \rightarrow e^{i\theta} \Psi \quad (2.3)$$

But if $\theta$ is dependent on position in spacetime (local transformation), it is clear that $\Psi$ is not invariant.

In order to make the free particle satisfy local invariance an additional term must be added to the Lagrangian:

$$\mathcal{L} = \bar{\Psi} (i\gamma^\mu \partial_\mu - m) \Psi - e\bar{\Psi}\gamma^\mu \Psi A_\mu \quad (2.4)$$

with $A_\mu$ transforming as:

$$A_\mu \rightarrow A_\mu - \frac{\partial \theta(x)}{e} \quad (2.5)$$

By convention, the new term is absorbed into the kinetic term of the Lagrangian. The new $\partial_\mu$ operator is the gauge covariant derivative:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu - ieA_\mu \quad (2.6)$$

$A_\mu$ corresponds to the four-vector potential of the QED field and $e$ is the charge of the free particle field. The Lagrangian for this field theory than becomes:

$$\mathcal{L} = \bar{\Psi} (i\gamma^\mu D_\mu - m) \Psi - \frac{1}{4} F^\mu_\nu F^{\mu\nu} \quad (2.7)$$

where $F^\mu_\nu = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the EM field tensor (representing the kinematic properties of the EM gauge field).

By enforcing local invariance on the U(1) symmetry group, the properties of QED can be represented. For instance the final term in equation 2.4 is the interaction term
between the photon and fermions, it is only non-zero for charged particles. This can
be extended to the SU(3) group with 8 gluons to cover the color interaction of QCD.
This formalism can also be applied to SU(2)⊗U(1) to generate electroweak QFT, but
requires the addition of a scalar (Higgs) field with a non-zero expectation value to
grant mass to the bosons of the theory while maintaining local invariance.

2.1.3 Spontaneous electroweak symmetry breaking and the Higgs boson

In order to preserve gauge invariance of the SM while allowing gauge bosons to
have mass, it is necessary to break the symmetry of the electroweak sector of the
theory (SU(2)⊗U(1)) by adding a complex scalar Higgs doublet. A full explanation
of spontaneous symmetry breaking in the electroweak sector can be found in [27]. This
formulation produces the 3 massive gauge bosons (W and Z bosons), one massless
gauge boson (photon), and a massive scalar boson (Higgs boson).

The scalar Higgs doublet is added to the electroweak sector of the Standard Model
(SU(2)W ⊗ U(1)Y):

\[
\phi \equiv \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}
\] (2.8)

The gauge is chosen such that the Higgs field obtains a vacuum expectation value:

\[
< \phi > = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}
\] (2.9)

Starting with the Lagrangian for the electroweak sector below:
\[ \mathcal{L} = -\frac{1}{4} W_{\mu \nu}^i W^{a \mu \nu} - \frac{1}{4} F_{\mu \nu} F^{\mu \nu} \]

where:

\[ W_{\mu \nu}^i = \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g e^{ijk} W_\mu^j W_\nu^k \]

\[ F_{\mu \nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \]

\( W^i \) correspond to the three component SU(2) group

\( B \) corresponds to the single component U(1) group

Then by adding the kinetic and subtracting the potential terms from the scalar field addition:

\[ \mathcal{L}(\phi) = \frac{1}{2} D_\mu \phi^\dagger D^\mu \phi + \mu^2 \phi^\dagger \phi - \frac{1}{2} (\phi^\dagger \phi)^2 \]

where:

\[ D_\mu \phi = [\partial_\mu - (i/2)g \sigma_i W_\mu^i - (i/2)g' B_\mu] \phi \]

One can produce a set of degenerate minima for \( \mu^2 > 0 \) and \( \lambda > 0 \). A schematic of this potential is shown in figure 2.2. This so-called Mexican hat potential grants a non-zero vacuum expectation value (\( v \)) for the Higgs field:

Figure 2.2. Shape of the Higgs potential of equation 2.11. [38]
(0|\Phi|0) = \frac{v}{\sqrt{2}}; \text{ where: } v = \sqrt{\frac{\mu^2}{\lambda}} \tag{2.12}

and breaks the electroweak gauge symmetry. This grants the bosons (at lowest order perturbation theory) masses:

\begin{align*}
M_H &= v\sqrt{2\lambda} \\
M_W &= \frac{v}{\sqrt{2}} g = \frac{e v}{2\sin\theta_W} \\
M_Z &= \frac{v}{\sqrt{2}} \sqrt{g^2 + g'^2} = \frac{M_W}{\cos\theta_W} \\
M_\gamma &= 0
\end{align*} \tag{2.13}

where g and g’ are the couplings with the SU(2) and U(1) sectors.

In order to obtain a mass for the fermions, one can couple the left-handed fermionic doublet ($\psi^L$) with the right handed singlet ($\psi^R$) and the Higgs doublet by introducing a Yukawa interaction term to the Lagrangian [33].

The excitation of the Higgs field is the massive scalar Higgs boson. This particle is a natural product of the Standard Model and will couple directly to all particles that gained mass through the addition of the Higgs field. The amplitude of the Higgs-vector boson vertex is ($\frac{2M_H^2}{v}$), while the fermion amplitude is ($\frac{m_f}{v}$). This means that the Higgs couples most strongly to the highest mass particles below half the Higgs boson mass. The photons and gluons are suppressed through a fermion or boson loop because they cannot couple directly to the Higgs boson. By measuring the strength of the coupling between the Higgs boson and fermions and bosons, one can further test the validity of the Standard Model.

2.2 Higgs production and decay modes

Proton-proton collisions, in the Standard Model, can produce the Higgs boson via four main mechanisms: gluon fusion ($gg \rightarrow H$), vector boson fusion ($qq \rightarrow H+2$
jets), associated production of a Higgs boson with a W or Z boson (Higgs-strahlung), and associated production with a $t\bar{t}$ pair ($t\bar{t}$ fusion). Due to the proton’s parton distribution function at LHC energies, gluon fusion production is by far the most common Higgs boson production mode. Two gluons from the colliding protons interact via a quark loop to produce a Higgs boson. Even with the suppression by the quark loop, this mode is roughly 9 times more common than the other modes combined. The other production modes produce the Higgs boson in association with other particles as shown in figure 2.3. The cross sections of each Higgs production mode as calculated by the LHC Higgs Cross Section Working Group [22] are shown in table 2.1.

Figure 2.3. Higgs boson production modes at the LHC: gluon fusion $gg\rightarrow H$ (upper left), associated production $gg\rightarrow t\bar{t}H$ (upper right), Higgs-strahlung $q\bar{q} \rightarrow W(Z)H$ (lower left), and vector boson fusion $qq \rightarrow qqH$ (lower right).
## TABLE 2.1

**STANDARD MODEL 125 GeV HIGGS BOSON PRODUCTION CROSS SECTION**

<table>
<thead>
<tr>
<th>Cross section (in pb) for $m_H=125$ GeV with 13 TeV center of mass energy [22]</th>
<th>ggH</th>
<th>VBF</th>
<th>WH</th>
<th>ZH</th>
<th>$t\bar{t}H$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$48.6^{+5.6%}_{-7.2%}$</td>
<td>$3.78^{+2%}_{-2%}$</td>
<td>$1.37^{+2%}_{-2%}$</td>
<td>$0.88^{+5%}_{-5%}$</td>
<td>$0.51^{+9%}_{-13%}$</td>
<td>$50.6$</td>
</tr>
</tbody>
</table>
The SM Higgs boson can decay into pairs of fermions, W bosons, or Z bosons at leading order. The coupling of the Higgs boson to fermions is proportionate to the fermion’s mass, so the branching fraction is proportional to the square of its mass. This means that when the Higgs boson decays fermionically, it primarily decays into $b\bar{b}$ pairs. For bosonic decays, the Higgs boson can decay into a pair of off-shell W or Z bosons, which themselves decay into two pairs of fermions.

Finally, the Higgs boson can decay into a pair of photons or gluons via a single loop. These decays are more relevant due to the relatively low mass of the Higgs boson (125 GeV). The two photon decay mode is addressed in this analysis because despite its relatively low branching fraction, it has a very clear experimental signature with a narrow peak in the diphoton invariant mass spectrum. The two photon decay mainly arises from a top quark loop or W boson loop as shown in figure 2.4. The branching fractions of the 125 GeV SM Higgs boson can be seen in table 2.2.

![Higgs decay diagrams](image)

Figure 2.4. A diagram of the Higgs boson decay into a pair of photons through a quark loop or W boson loop.
### Table 2.2

**Standard Model 125 GeV Higgs Boson Decay Modes**

<table>
<thead>
<tr>
<th>Decay Channel</th>
<th>Branching Ratio</th>
<th>Rel. Uncertainty [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \to b\bar{b}$</td>
<td>$5.84 \times 10^{-1}$</td>
<td>$+3.2%$ $-3.3%$</td>
</tr>
<tr>
<td>$H \to W^+W^-$</td>
<td>$2.14 \times 10^{-1}$</td>
<td>$+4.3%$ $-4.2%$</td>
</tr>
<tr>
<td>$H \to \tau^+\tau^-$</td>
<td>$6.27 \times 10^{-2}$</td>
<td>$+5.7%$ $-5.7%$</td>
</tr>
<tr>
<td>$H \to ZZ$</td>
<td>$2.62 \times 10^{-2}$</td>
<td>$+4.3%$ $-4.1%$</td>
</tr>
<tr>
<td>$H \to \gamma\gamma$</td>
<td>$2.27 \times 10^{-3}$</td>
<td>$+5.0%$ $-4.9%$</td>
</tr>
<tr>
<td>$H \to Z\gamma$</td>
<td>$1.53 \times 10^{-3}$</td>
<td>$+9.0%$ $-8.9%$</td>
</tr>
<tr>
<td>$H \to \mu^+\mu^-$</td>
<td>$2.18 \times 10^{-4}$</td>
<td>$+6.0%$ $-5.9%$</td>
</tr>
</tbody>
</table>

### 2.3 Higgs boson properties study in the two photon final state

The Higgs boson appears to match expectations of the Standard Model of particle physics based on current measurements, but more precision is required. By measuring the ratio of the measured signal yield divided by the expected signal yield from simulation ($\sigma/\sigma_{SM}$) of the Higgs boson for different decay channels, it is possible to verify the Standard Model (and other potential theories) which investigate the coupling of the Higgs boson to fermions and bosons. Several analyses are ongoing utilizing the CMS and ATLAS detectors at the LHC focusing on different decay modes of the Higgs boson. Analyses are further categorized to identify the production mode (based on the particles produced in association with the Higgs boson as shown in figure 2.3).

This analysis searches for Higgs bosons which decay into two photons; this signal has a relatively large background. But this background falls smoothly as a function of the invariant mass of the diphoton system. This means that the sharp mass peak
of the signal from Higgs boson decay can be easily identified over the background due to the excellent photon energy resolution of the CMS detector.

The main contribution to the diphoton background comes from the QCD Born and Box diagrams as shown in figure 2.5. Additionally, QCD can produce events which have one real photon and one or two jets that are hard to distinguish from real photons. Finally, Z bosons decaying into two electrons can appear as two photons if the detector fails to properly associate the electron track with an energy deposit. These backgrounds are all produced at much higher rates at the LHC than the Higgs decaying to two photons, but they produce a smooth, continuous invariant mass distribution. Due to the excellent photon energy resolution, the signal photons produce a sharp peak in the two photon invariant mass distribution allowing this analysis to be one of the most sensitive at the LHC.

![Diagram](image_url)

Figure 2.5. A diagram of the background channels to the analysis.
CHAPTER 3

THE DETECTOR AND EXPERIMENTAL APPARATUS

The European Organization for Nuclear Research (CERN) has housed colliders and high energy physics experiments since its founding in 1954 near Geneva, Switzerland. The organization is currently home to 22 member states, 7 associate members (or working towards the status), and 5 observer states (including the United States of America). Many more countries have scientific contracts or co-operation agreements with CERN. The international collaboration at CERN has contributed many discoveries in the field of particle physics since its creation. Currently, the largest experiments at CERN use the Large Hadron Collider (LHC). There are 2 general purpose detectors, A Toroidal LHC ApparatuS (ATLAS) and the Compact Muon Solenoid (CMS). Both are located on opposing sides of the LHC ring. Additional experiments: Large Hadron Collider beauty, A Large Ion Collider Experiment, Total Elastic and diffractive cross section Measurement, the Large Hadron Collider forward, and Monopole and Exotics Detector At the LHC perform various specific measurement during the LHC’s operation. The following analysis focuses only on the CMS detector near Cessy, France.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a 26.7 km circumference proton-proton collider operated by CERN on the border of France and Switzerland. The tunnel for the previous CERN accelerator, LEP, was upgraded to house the LHC. The LHC was designed to operate at a luminosity of $10^{-34}\text{cm}^{-2}\text{s}^{-1}$ with an energy of 7 TeV.
per proton, giving a center of mass energy of 14 TeV. The LHC can hold a total
of 2808 proton bunches with a spacing of 25ns. During the course of run1, a total
of 6.1 fb\(^{-1}\) data were taken with a center of mass energy of 7 TeV and 23.3 fb\(^{-1}\) at
8 TeV. The milestone measurement from run1 was the discovery of the Higgs boson,
made concurrently by CMS and ATLAS on July 4th 2012. After a hiatus to upgrade
the LHC and detectors, the LHC resumed operation with center of mass collisions of
14 TeV in 2015 and collected a total of 2.7 fb\(^{-1}\) in 2015 and 35.9 fb\(^{-1}\) in 2016.

In order to produce and collide the high energy protons required at the LHC, a
complex of accelerators is required. The protons originate in a tank of hydrogen gas
that is ionized by the Duo-plasmatron source. A radio-frequency cavity accelerates
the protons to 750 keV. The proton beam is then sent to the Linac 2, which accelerates
it to 50 MeV. From there, the beam is injected into the Proton Synchrotron Booster,
bringing the beam energy up to 1.4 GeV. The Proton Synchrotron then accelerates
the protons up to 25 GeV. The beam is then sent to the Super Proton Synchrotron
where they are brought up to 450 GeV. From here, the protons are transferred into
the two beampipes of the LHC. It takes 4 minutes and 20 seconds to inject the protons
into the LHC ring. This must be done once into the clockwise beampipe and once into
the counterclockwise beampipe. Then a total of 20 minutes is required to accelerate
the protons up to the current maximum energy of 6.5 TeV. A full schematic of the
LHC accelerator complex can be seen in figure 3.1.

Typically, the beams circulate and collide inside the detectors for several hours
before the beams degrade sufficiently to require a beam dump and restart of the
injection process.

A slightly different chain of accelerators is required to accelerate the lead ions and
is not discussed here.

A magnetic field of 8.33 T is required to maintain a 7 TeV proton beam in the
LHC. The magnets are constructed from NbTi superconducting wire and cooled by
Figure 3.1. The LHC is the last ring (dark gray line) in a complex chain of particle accelerators. The smaller machines are used in a chain to boost the protons to their final energies. [5]
liquid helium at 1.9 K. Dipole magnets are used to bend the path of the protons to follow the ring. There are 1232 superconducting dipole magnets in the LHC. Quadrupole magnets are used to focus the proton bunches, keeping a tight beam. They alternate between squeezing horizontally and vertically. Finally, there are magnets which focus the beams much tighter at the collision points, inside the LHC experiments.

3.2 The Compact Muon Solenoid

The Compact Muon Solenoid is one of the all-purpose detectors at the LHC. It is designed to measure proton-proton collisions at the LHC. The CMS experiment comprises a full silicon tracker, a crystal Electromagnetic Calorimeter (ECAL), a Hadron Calorimeter (HCAL), encased in a superconducting solenoid. Outside of the magnet is an iron return yoke which is embedded with the muon chambers. The detector measures 14.6 m in height, 21.6 in length, and weighs 12500 t.
A more detailed description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, can be found in [5].

3.2.1 Coordinate system

CMS uses a right-handed coordinate system, with the origin at the nominal interaction point, the x-axis pointing to the center of the LHC, the y-axis pointing up (perpendicular to the LHC plane), and the z-axis along the anticlockwise-beam direction. The polar angle, $\theta$, is measured from the positive z-axis and the azimuthal angle, $\phi$, is measured in the x-y plane.
3.2.2 The magnet

Part of CMS namesake comes from the superconducting solenoid magnet which makes up much of its mass. The magnet generates a 3.8 T constant axial magnetic field inside the solenoid and a 1.9 T field in the outer part of the detector. This constant magnetic field causes charged particles to curve as they pass through the detector and is an integral to the momentum measurement of particles.

The magnet is 12.5 m in length with a stored energy of 2.6 GJ at full current. It has a bore radius of 6.3 m and weighs 220 tonnes. An iron return yoke is composed of 5 wheels and 2 endcaps (with 3 layers each) and weighs a total of 10000 tonnes. The wheels and layers are interspersed between the layers of the muon chambers.

To generate the 3.8 T magnetic field 41.7 MA-turns are produced by 4 layers of NbTi Rutherford-type cables with aluminum insert to support the unprecedented shear stress in a magnet of this type.

The cold mass of the solenoid is maintained at 1.8 K, cooled by liquid helium refrigeration. Full details of the CMS magnet can be found in [5].

3.2.3 The tracking system

The tracking system [7] is the innermost part of CMS and consists of 2 different silicon detectors: an inner pixel tracker and an outer strip tracker. These work in tandem to reconstruct charged particle tracks with high precision and the primary and secondary vertices present in each event. The LHC during the 2015-2016 run period operated at around $10^{-34}$cm$^{-2}$s$^{-1}$ which leads to roughly 1 MHz-mm$^{-2}$ at a radius of 4 cm. The design requirements for CMS require a pixel occupancy of less than 1%. This requirement allows the use of the more course strip tracker at a radius of 10 cm. The pixel detector has a size of 100x150 $\mu$m$^2$ in r-\(\phi\) and z and gives an occupancy on the order of $10^{-4}$ per pixel in 2015-2016 operating conditions. Three cylindrical layers of pixels at 4.4, 7.3, and 10.2 cm surround the interaction point.
Additionally, there are 2 pixel endcap layers at each end of CMS at \(|z|\) 34.5 and 46.5 cm. The pixel tracking is designed to allow for 3 point tracking for nearly the entire range \(-2.5 < \eta < 2.5\).

The strip tracker functions from an inner radius of 20 cm to an outer radius of 110 cm. The strip size is typically 10 cm x 80 \(\mu\)m up to 55 cm from the beamline. The strip size is increased up to the outer radius of the tracker to reduce cost and readout channels. An occupancy of 2-3% per strip is maintained. The strip tracker has a more complex layout and the entire tracker schematic can be seen in figure 3.3. The strip tracker is split into the Tracker Inner Barrel (TIB), Tracker Outer Barrel (TOB), Tracker Inner Disk (TID), and Tracker Endcap (TEC).

Figure 3.3. Schematic cross section through the CMS tracker. Each line represents a detector module. Double lines indicate back-to-back modules which deliver stereo hits.[5]
In order to maintain low noise from dark current in the silicon, the tracker must be kept between -10 C and -27 C, depending on the radiation damage. The infrastructure to cool, support, and read out the signals from these 70 million channels is considerable and presented a challenge in design. This material contributes up to 2 radiation lengths of material in front of the ECAL, which reduces the possible resolution of the ECAL. The material budget can be seen in figure 3.4.

![Tracker Material Budget](image)

Figure 3.4. Material budget in units of radiation length as a function of pseudo-rapidity $\eta$ for the different sub-detectors (left) and broken down into the functional contributions (right).[5]

The CMS Tracker is an important part of the $H \rightarrow \gamma \gamma$ analysis. It is used to properly reconstruct the primary vertex, which is needed to properly calculate the opening angle of the diphoton pair and thus the Higgs mass. Additionally, photons can interact with the tracker material and convert into an electron-positron pair before reaching the ECAL. This gives more information about the primary vertex, but also changes the cluster shape in ECAL. More on vertex selection and converted
photons in sections 5.6 and 4.5.1. Finally, the tracker is used to measure charged particle isolation. Hadronic showers produced by QCD backgrounds leave a signal in the tracking system. This signal is different from that produced by a photon which converts in the tracker. As such, energy deposits in ECAL overlaid by tracker hits are generally associated with QCD jets, not real photons. Section 5.5 details photon isolation and fake photon rejection.

3.2.4 The electromagnetic calorimeter

The most important detector in the $H \rightarrow \gamma\gamma$ analysis is the Electromagnetic Calorimeter (ECAL). The SM Higgs boson at 125 GeV has an intrinsic width of roughly 4 MeV. This means that the $H \rightarrow \gamma\gamma$ mass peak is expected to appear as a sharp signal peak over a smoothly falling background. Since it is not possible to measure perfectly the energy of the incoming photons, the measured width of this peak will come almost entirely from the detector energy resolution of the photons from the Higgs decay. The photon energy is measured by the ECAL and thus must have excellent energy reconstruction and calibration.

The electromagnetic calorimeter consists of 75,848 lead tungstate crystals, which provide coverage in pseudo-rapidity $|\eta| < 1.48$ in a barrel region (EB) and $1.48 < |\eta| < 3.0$ in two endcap regions (EE). Preshower detectors consisting of two planes of silicon sensors interleaved with a total of 3$X_0$ of lead are located in front of each EE detector.

The lead tungstate crystals have a radiation length of 0.89 cm and a Moliere radius of 2.2 cm. This allows for a compact, fine granularity detector. The scintillation decay time of the crystals is of the same order as the bunch crossing time in the LHC; 80% of the scintillation light is emitted within 25 ns. The crystal length corresponds to 25.8 radiation lengths.

The light output varies by 2.1% C$^{-1}$ at the operational temperature of 18C.
Figure 3.5. Layout of the CMS electromagnetic calorimeter showing the arrangement of crystal modules, supermodules and endcaps, with the preshower in front.[5]
Roughly 4.5 photoelectrons per MeV are collected in both the APD and VPT photosensors. The crystals are transparent to visible light and the scintillation light has a broad optical spectrum centered about 420 nm.

The performance of the ECAL decreases with radiative fluence in the crystals; ionizing radiation creates color centers due to oxygen vacancies and impurities in the crystal lattice which causes absorption bands. This produces a dose and time dependent loss of light in ECAL that is somewhat recovered by a spontaneous recovery due to thermal fluctuations. The crystal transparency is monitored by a laser injection system and calibrated as a function of time.

The barrel portion of ECAL covers the pseudo-rapidity range $|\eta| < 1.479$ and has an angular granularity in $\phi$ and $\eta$ of 1 degree. The crystal faces are roughly 1 Moliere radius to maximize position sensitivity while minimizing the energy losses from sharing the energy shower among many crystals. The crystals are mounted in a quasi-projective geometry to point 3 degrees (in both $\eta$ and $\phi$) from the interaction point. This is done to prevent the trajectory of a particle from aligning with the gaps between the crystals. The gap between crystals is .35 mm thick and made of aluminum facing the crystal and two layers of glass fiber-epoxy resin. The interior radius of the ECAL is 129 cm from the beamline. The light is collected at the back end of each crystal by two Hamamatsu S8148 reverse structure avalanche photon diode (APD). They APDs are operated at a bias voltage between 340 and 430 V to maintain an optical gain of roughly 50. The two APDs have a light collection area of 5x5 mm$^2$ and the signal is read out in parallel and then summed to measure the scintillation light produced in the crystal.

The ECAL endcap increases coverage of the ECAL to $|\eta| < 3.0$ which are placed 315.4 cm from the interaction point along the z-axis. The endcaps are arranged in an xy-grid and focused 130 cm beyond the interaction point to prevent a particle’s trajectory from aligning with the gaps between the crystals. The endcap crystals have
a slightly larger cross-section than the barrel crystals with a front facing cross-section corresponding to roughly 1.3 Moliere radii to reduce cost. The endcap crystals are 220 mm or 24.7 $X_0$ in length, the reduced length is due to a lead preshower in front of the ECAL.

The scintillation photons produced in the ECAL endcap are collected by PMT188 vacuum photo triodes (VPT) instead of the APDs used in the barrel.

The preshower comprises 2 alternating layers of lead and silicon and was originally designed to differentiate $\pi_0$ particles from prompt photons with fine position resolution. The sensors consist of 32 strips and both layers are oriented orthogonally to get an x-y position. The preshower detector covers the $|\eta|$ range of 1.653 to 2.6. The preshower along with its support material corresponds to 2.3 $X_0$ of material in front of the ECAL endcap.

The ECAL was designed to provide good energy resolution for photons for $H \rightarrow \gamma\gamma$. This energy resolution can be parameterized in equation 3.1. Where $S$ is the stochastic term, $N$ the noise term, and $C$ the constant term.

\[
\frac{\sigma}{E} = \frac{S}{\sqrt{E}} \oplus \frac{N}{E} \oplus C \quad (3.1)
\]

The stochastic term arises from lateral shower containment, photo-statistics, and interaction with preshower (where present). The constant term arises from non-uniformity of the longitudinal light collection, intercalibration errors, and leakage of energy through the back of the crystal. The noise term comprises the electronics noise, digitization noise, and pileup noise (other event contamination).

In the barrel section of the ECAL, an energy resolution of about 1% is achieved for unconverted photons in the tens of GeV energy range. The remaining barrel photons have a resolution of about 1.3% up to a pseudo-rapidity of $|\eta| = 1$, rising to about 2.5% at $|\eta| = 1.4$. In the endcaps, the energy resolution of unconverted photons is about 2.5%, while the remaining endcap photons have a resolution between 3 and
4% [30].

3.2.5 The hadronic calorimeter

The hadronic calorimeter is important in measuring hadron jets and neutrinos resulting in apparent missing transverse energy. Four separate subdetectors compose the hadronic calorimeter to provide coverage out to $|\eta| < 5.2$ with two inside the magnetic coil and two outside.

![Figure 3.6. Longitudinal view of the CMS detector showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters.](image)

The Barrel Hadronic Calorimeter (HB) is placed inside the magnetic coil and covers $|\eta| < 1.3$, roughly corresponding to the ECAL barrel subdetector in position. The Endcap Hadronic Calorimeter (HE) is also inside the magnetic coil and comprises
two endcaps that extend the coverage up to $|\eta| < 3$. Both of these subdetectors are constructed with layers of flat brass absorber plates parallel to the beam line. The layers are staggered to minimize the amount of dead material a single particle could pass through. Plastic scintillator is interleaved between each absorber layer and are segmented so that each channel corresponds to a resolution of $(\Delta \eta, \Delta \phi) = (0.087, 0.087)$. The resolution is $(\Delta \eta, \Delta \phi) = (0.17, 0.17)$ for $|\eta| > 1.6$. The absorber consists of 14 brass layers. The first and last absorber plates are stainless steel for structural support. The total absorber thickness is 5.82 interaction lengths in the central barrel and increases to roughly 10 interaction lengths at high $|\eta|$.

Outside the magnet covering the HB region is the Outer Hadronic Calorimeter (HO). This is used to extend the depth of the calorimeter and catch the tails of any hadronic showers which may have penetrated the magnetic coil. The granularity is the same as the HB, but only 2 layers of scintillators are present on either side of a layer of magnetic return iron. This extends the HCAL containment to 11.8 interaction lengths. Finally, the Forward Hadronic Calorimeter (HF) is placed outside the magnetic coil and extends the calorimeter’s coverage to $|\eta| < 5.2$. These detectors are much more radiation hard and quartz fibers are used to measure the Cherenkov light of the incident particles.

For all but the HF, plastic wavelength shifting fibers send signals to be read out by hybrid photodiode (HPD). The HF uses radiation hard quartz fiber and photomultiplier tubes to measure the Cherenkov light.

3.2.6 The muon system

Muon detection is useful for tagging the various production modes of the Higgs boson. Additionally, the decay of the Z boson into two muons is used for calibration of the vertexing efficiency as described in section 5.6. The muon system covers up to $|\eta| < 2.4$ and is composed of drift tubes (DTs), cathode strip chambers (CSCs),
and resistive plate chambers (RPCs) [6]. Combining the muon system with the tracker system provides high efficiency to reconstruct muons while getting precise measurements of their momenta.

3.2.6.1 Drift tubes

The drift tubes (DTs) cover the barrel of the CMS detector up to $|\eta| < 1.2$. Drift tubes can be used because of the distance from the interaction point, the particle density is low and the magnetic field is low enough. The 21 mm thick DTs are filled with a gas mixture of Ar and CO$_2$ (85% and 15% respectively). A layer of the iron magnetic return yoke sits between each of the 4 DTs layers. In each layer, there are drift tubes aligned with the beam axis measure the $\phi$ position, then tubes orthogonal to the beam line to measure the $z$ position, then another set of DTs to measure the $\phi$ position. Each set of DTs in a layer are offset to minimize dead space and each layer of DTs are offset slightly so the dead space from infrastructure is not aligned.
3.2.6.2 Cathode strip chambers

The cathode strip chambers (CSCs) are multiwire proportional chambers comprising 6 anode wire planes interleaved among 7 cathode panels. They are used to detect muons in the CMS endcap ($0.9 < |\eta| < 2.4$). The anode wires are used to measure the $\eta$ position of the muons, while the induced charge on the orthogonal
cathode strips is used to measure the $\phi$ position. The gas mixture consists of 50% CO$_2$, 40% Ar, and 10% CF$_4$. There are a total of 4 layers of CSCs with the iron magnetic return yoke between each layer. The CSCs are arranged so that a muon must cross 3 to 4 stations as they pass through the endcap.

![Figure 3.8. Quarter-view of the CMS detector. Cathode strip chambers of the Endcap Muon system are highlighted.][5]

3.2.6.3 Resistive plate chambers

The resistive plate chambers (RPCs) are gaseous parallel-plate detectors and are used to match a particular muon to a given bunch crossing. The RPCs consist of 2 gaps operating in avalanche mode to minimize the their response time. In the barrel RPCs are mounted on either side of the interior two DTs. The outer two DTs have an RPC mounted on their interior side. In the endcap there are a total of three
layers of RPCs in front of each CSC station (up to $|\eta| < 1.6$). The RPCs provide very low spatial resolution and are used mainly for triggering and matching muons to the correct bunch crossing.

3.2.7 The trigger system

During the 2016 run period, protons were collided at a rate of 40 MHz. Events recorded by CMS have a typical size of 1MB for proton-proton collisions; it is not possible to record every collision. To reduce the bandwidth that CMS must record and analyze, a two-tiered trigger system is used [28]. The first level (L1), custom hardware processors, uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a time interval of less than 4 $\mu$s. The second level, known as the high-level trigger (HLT), consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to around 1 kHz before data storage.

3.2.7.1 The level 1 trigger

The level 1 trigger is designed to reduce the event rate from 40MHz to less than 100kHz. It consists of programmable electronics which analyzes coarsely segmented data from the calorimeters and muons systems, while holding the high-resolution data in a pipeline on the front-end electronics.

3.2.7.2 The high level trigger

Events that pass the level 1 trigger are then analyzed by the high level trigger (HLT). A farm of processors compose the HLT and are capable of fully reconstructing objects in an event. These objects use nearly the same reconstruction techniques as the full offline analysis and select roughly 1kHz of interesting physics events based on the particle type and kinematic properties of the event.
3.2.8 CMS simulation

Monte Carlo (MC) simulations are necessary to understand the dynamics and detector behavior of high energy physics collisions.

The simulation begins with the parton-parton interaction (generation step) and continues through the interaction of the final state particles with a simulation of the CMS detector. This produces data in the same format as recorded by CMS and can then be processed using the same framework used to process and analyze the real collision data. Depending on the processes being produced, different generators may be used for simulation of the hard scattering process than for the showering of the final state particles and the decay of unstable particles.

Various MC generators were used to produce the samples in this analysis and further details can be found in [1].
4.1 Energy measurement in ECAL

ECAL uses a 12-bit ADC to sample analog signals from the APDs and VPTs at 40 MHz. Due to the scintillation mechanism in the crystals and shaping in the electronics, 10 consecutive samples of the signals are stored for each crystal (figure 4.1). The signal pulse starts from the 4th sample and the pedestal value is estimated from the first 3. The pulse shape is consistent across ECAL for electromagnetic showers and the pulse amplitude in ADC counts is multiplied by an ADC-to-GeV conversion factor after subtracting the pedestal value. Due to variation from channel to channel and with time, an intercalibration constant is also applied to properly estimate the energy deposited in a crystal. [10]

The ECAL crystal size was designed such that a typical electromagnetic shower would spread over only a few crystals. Roughly 95% of the shower is contained in a 3x3 matrix of crystals. In some regions of CMS, interaction with the tracker material and services can contribute up to 2 radiation lengths upstream of ECAL. This causes the electrons to undergo bremsstrahlung and photons to convert into an electron-positron pair. The magnetic field further bends the electrons and positrons, creating a larger spread in the final electromagnetic shower detected in ECAL. A clustering algorithm is designed to account for the shower shapes which are recorded by CMS [30]. Additional corrections are applied for imperfect clustering and geometry effects.
Figure 4.1. Typical pulse shape measured in the ECAL, as a function of the difference between the time \( T \) of the ADC sample and the time \( T_{\text{max}} \) of the maximum of the pulse. The dots indicate ten discrete samples of the pulse, from a single event, with pedestal subtracted and normalized to the maximum amplitude. The solid line is the average pulse shape, as measured with a beam of electrons triggered asynchronously with respect to the digitizer clock phase. [10]
4.2 ECAL noise and simulation

Using the laser monitoring system described in subsection 4.4 the noise of each channel is measured using the variation of the first three samples from the ECAL pulse shape (which have no signal present). Dark current in the APDs is the main component of single channel noise in the ECAL barrel. Dark current increases with neutron radiation causing defects in the silicon lattice of the APDs. This dark current noise corresponds to approximately 60 MeV. In the ECAL endcaps, the VPTs maintain a nearly constant level of electronic noise, but radiation damage to the scintillation crystals reduces light yield. This means that the gain must be increased, generating an effective increase in single channel noise. The noise in the ECAL endcap depends strongly on fluence and thus $\eta$, but corresponds to roughly 150 MeV for the majority of channels. Noise measurements from data are applied in the simulation of ECAL used in this analysis.

4.3 Clustering algorithms

In CMS a dynamic algorithm is used to fully reconstruct the energy of electromagnetic deposits in ECAL. Due to bremsstrahlung and photon conversions from tracker material, the showers in ECAL can be spread out azimuthally and consist of deposits from more than one particle. This algorithm forms a cluster by starting with the amplitude described in section 4.1 from a high energy seed crystal. Then all contiguous crystals with energies at least two standard deviations above the noise are added to the cluster. These clusters are designed to collect the energy deposits from a single particle impinging on ECAL. These clusters can overlap and energy is shared between overlapping crystals based on the relative energy of the seed crystal. Finally, Super Clusters (SCs) are created by combining clusters from multiple impacts of the same originating photon or electron from photon conversion and bremsstrahlung deposits.
Using these clustering algorithms improves the shower containment and minimizes additional detector noise (from adding extra channels). [30]

4.4 Energy calibration

In order to measure an absolute energy, the ECAL must be calibrated using well understood physics. The crystals must be corrected on a single channel level for time variations, then the crystals are intercalibrated so they each crystal provides a similar energy response, and finally the absolute energy scale is tuned.

The main cause of time variations on the single crystal level is due to discoloration of the ECAL crystals from radiation. The crystals recover some transparency via spontaneous recovery due to thermal annealing when no radiation from collisions is present. In situ, this variation is measured by a laser calibration system. Laser light of 447nm is injected through optical fibers in each crystal. The laser calibration system cycles through each crystal every 40 minutes which provides sufficient time resolution to correct the change in response.

Variation in response between different channels arises predominantly from crystal light yield differences in the ECAL barrel. In the endcap, differences in the gain of VPT are also relevant. Inter-calibrations have been derived using collision data. Using the $\phi$ symmetry method, each ring of crystals can be inter-calibrated (the energy deposited in each crystal averaged over many minimum bias collisions depend on eta and are independent of $\phi$). The invariant mass of photon pairs from $\pi^0$ and $\eta$ decays, the $E/p$ ratio of isolated electrons from W and Z boson decays, and the invariant mass of the $Z \rightarrow ee$ are used to intercalibrate the crystals in the eta direction as well as to calculate the absolute energy scale. [30]
4.4.1 Multivariate electron and photon energy corrections

With the full energy corrections and clustering algorithms applied, the energy can still be mismeasured due to the position of the impinging particle on the crystal face, the upstream material, and the variation of energy absorbed before reaching ECAL. Corrections to these effects are derived using a multivariate regression technique known as a boosted decision tree (BDT) implemented with the TMVA package [35]. The regression takes several kinematic and shower shape variables as inputs and estimates the fraction of true energy to uncorrected energy of the supercluster. The regression is trained separately for photons and electrons using simulated samples. [30]

4.5 Photon reconstruction

The photon follows the standard ECAL supercluster reconstruction and corrections detailed previously in section 4.3. An additional requirement that the transverse energy of the supercluster is greater than 10 GeV is required.

4.5.1 Reconstruction of conversions

Photons which convert into electron-positron pairs in the tracker produce showers with a larger cross-section in ECAL, resulting in worsened energy resolution with respect to unconverted photons. Using the tracker information, they can be reconstructed to help with identification of the correct Higgs interaction vertex. Roughly 1/4 of events entering this analysis have at least one photon reconstructed as a conversion. Electron-positron pairs are reconstructed using three different tracking algorithms: the default iterative tracking algorithm, the GSF electron tracks (passing a Gaussian sum filter), and ECAL seeded tracks (tracks seeded from the ECAL supercluster). Further details on track reconstruction can be found in section 4.6. Tracks are required to have nHits > 4 and $\chi^2 < 10$. 
Tracks from the prompt collision are removed by requiring the conversion tracks to have a distance of minimum approach greater than 0. The angular separation of the oppositely charged tracks must be $\Delta \cot(\theta) < 0.1$. The $\Delta z$ of the innermost hits of the conversion tracks must be less than 5 cm and the vertex must be inside the tracker volume ($r < 120$ cm and $z < 300$ cm). The track pair is then constrained to a single vertex and the tracks are recalculated using the new vertex. The fit to the single vertex must converge with a $\chi^2 > 10^{-6}$ and the transverse momentum of the pair must be greater than 10 GeV. If the conversion satisfies these requirements, it is then matched to an ECAL supercluster requiring the $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2} < 0.1$; the minimal $\Delta R$ combination is selected if a pair of tracks satisfies the requirement.

4.6 Electron reconstruction

Electrons are used in the analysis both for energy calibration (using $Z \rightarrow ee$) and to tag production modes for the Higgs boson when produced alongside the Higgs boson. Electrons objects are similar to photon objects as they use the same ECAL clusters, but since electrons are charged, an associated track must also be found. Electrons can either be seeded by ECAL superclusters and then matched to tracks, or seeded by tracks and then matched to superclusters.[29]

4.6.1 ECAL seeded electron reconstruction

This algorithm starts with a supercluster of $E_T > 4$ GeV. The cluster position and energy is used along with the assumption that the deposit was produced by either a positron or an electron. The algorithm can then extrapolate the most likely trajectory back to the primary vertex (done for both electron and positron assumptions). The intersections of the helical trajectory of the electron and the innermost layers of the tracker are then compared to tracker seeds. The tracker seeds consist of 2 or 3 hits with vertices from pixel tracks. If the first two seeds fall inside the predicted
spatial range from the supercluster trajectory estimate, the track is matched to the
supercluster. The efficiency for electrons from Z boson decay is 92% from simulation
estimates. [29]

4.6.2 Tracker seeded electron reconstruction

For events where the electron \( p_T \) is low or the electrons are contained inside
jets, the efficiency to reconstruct the helical trajectory from ECAL superclusters is
low. In this case, the electron reconstruction starts from tracker seeds rather than
ECAL superclusters. Tracks are reconstructed using the standard Kalman filter (KF)
algorithm [7]. In the case with negligible bremsstrahlung, the track will match with
the corresponding ECAL supercluster and provide a good momentum measurement.

For tracks with considerable bremsstrahlung the KF algorithm is unable to follow
the path of the electron. In this case, the tracks are refit using a dedicated Gaussian
sum filter.

Finally, a multivariate analysis is used to select the tracker seed as an electron
seed. This takes as input variables estimating the quality of the KF and GSF tracks
as well as the geometrical and energy matching of ECAL and tracker information.
[29]

4.6.3 List of photon variables

Defined here are several variables used in the reconstruction and identification
of photons. Shower shape variables describe how the energy recorded by CMS is
distributed through the supercluster by leveraging the granularity of ECAL. Isolation
variables look at energy recorded by CMS which overlaps or is immediately adjacent
to the supercluster object.

Shower shape variables:

- \( S_4: \frac{E_{2x2}}{E_{5x5}} \). This is the ratio of the energy in the 2x2 grid crystal array which
gives the maximum energy and the energy in the 5x5 crystal array centered on the seed crystal.

- $R_6$: $E_{3x3}/E_{SC}$. This is the ratio of the energy sum in the 3x3 crystal array centered on the seed crystal and the supercluster energy.

- $\sigma_{\eta\eta}$. This is the log-weighted standard deviation of the single crystal $\eta$ in the 5x5 crystal grid centered on the seed crystal. The weight is 4.7 plus the log of the ratio of the single crystal energy and the $E_{5x5}$.

- $\text{cov}_{\eta\phi}$. This is the log-weighted covariance of single crystal $\eta$ $\phi$ within the 5x5 crystal grid centered on the seed crystal. The weighting is the same as for $\sigma_{\eta\eta}$ above.

- $\sigma_{\eta}$. The standard deviation of single crystal $\eta$ within the supercluster weighted by the single crystal energy over supercluster energy.

- $\sigma_{\phi}$. The standard deviation of single crystal $\phi$ within the supercluster weighted by the single crystal energy over supercluster energy.

- $\sigma_{RR}$. The quadratic sum of the energy weighted standard deviation in strip index in the x and y planes for the preshower detector. This is only used for photons reconstructed to overlap with the preshower detector.

Isolation variables:

- $PF\text{PhotonIso}$: The transverse energy sum of all particles identified as photons within a $R = 0.3$ cone centered on the candidate photon’s direction. An inner veto of $R = 0.015(0.07)$ for photons in the barrel (endcap) is used to remove any energy from the candidate photon/electron which may bias the isolation estimate.

- $PF\text{ChargedIso}$: The transverse energy sum of all particles identified as charged hadrons within a $R = 0.3$ cone centered on the candidate photon’s direction. An inner veto of $R = 0.02$ is used to remove any energy from the candidate photon/electron which may bias the isolation estimate.

Kinematic variables:

- $SC \; \eta$: The $\eta$ position corresponding to the reconstructed photon as calculated from the CMS detector origin to the reconstructed supercluster location in the electromagnetic calorimeter.

- $SC \; E_{Raw}$: The sum of the energy recorded by CMS in each crystal of the supercluster corresponding to the reconstructed photon.
• $\rho$: The estimate of transverse energy per unit area due to pileup interactions and underlying event content recorded by CMS. The energy density is the median transverse energy density of all anti-$kT$ reconstructed jets in the event.
4.7 Energy scale and resolution with $Z \rightarrow ee$ events

In order to verify the corrections described in 4.4, to apply scale corrections to the photon energy, and to properly compare data and simulation $Z \rightarrow ee$ events are used. It has been shown that electrons and photons of the same $R_0$ value shower similarly in ECAL and thus have the same energy performance [29]. A large number of $Z \rightarrow ee$ events are produced at the LHC; due to the similarities of the ECAL response to electrons and photons, these events can be used to estimate the performance of photons from a Higgs boson decay and to validate the simulation used in this analysis. Careful weighting of events is applied to correct for the difference in boson mass between the Higgs and $Z$ bosons for the study.

4.7.1 $Z \rightarrow ee$ event selection

Electrons for energy scale and resolution studies are selected by looking for objects reconstructed as photons which are matched to a pixel seed. The mass of the electron-positron pair must have $70 < m_{ee} < 110$. Additionally, a trigger requirement is applied in both data and simulation (at least one HLT electron with $p_T > 27$) and the two fully reconstructed electrons are required to have $p_T > 40$ and 30 GeV respectively.

4.7.2 Comparison between data and simulation samples

Several shower shape variables used in the analysis show discrepancies between data and simulation. A shifting of these variables on a per event basis such that the simulation matches the data distribution is performed to account for these discrepancies.

The shift is calculated based on the percentile of the events in the simulation distribution with a new value assigned based on the same percentile for the data distribution. Much better agreement is seen after the corrections are applied.
Figure 4.2. The discrepancy between data and simulation was most prevalent in the $R_9$ shower shape variable. It is corrected by a per event shift of the based on the difference of the distribution in simulation and data.

4.7.2.1 Energy scale

The supercluster energy scale in data is finally corrected by varying the scale to match that of simulated $Z \rightarrow ee$ electron-positron pair. A convolution of Breit-Wigner and a Crystal Ball are fit to the $Z \rightarrow ee$ peak from ECAL superclusters in data and simulation. The difference between these two is then used to calculate the scale difference. The energy scale varies as a function of time, $\eta$, and $R_9$. The first step is to fit this mean parameter for various run ranges. After this, a set of smearing corrections are derived for 8 categories ($4\eta \times 2R_9$) minimizing the likelihood between the smeared simulation and the data. Finally, the 8 categories are corrected with a residual energy shift correction to align the data to the simulation.
Figure 4.3. $Z \to ee$ line-shape peak values in several run ranges for 4 $\eta$ categories: 2 categories in EB, 2 categories in EE. Simulation value is displayed as a continuous line.

Figure 4.4. $Z \to ee$ line-shape peak values in several run ranges after applying the time-dependent scale corrections to data. Simulation value is displayed as a continuous line.
Figure 4.5. Agreement between data and simulation before and after the energy scale and smearing corrections for $Z \rightarrow ee$ for the case where both electrons satisfy $|\eta| < 1R_9 > 0.94$ (left) and $1.566 < |\eta| < 2R_9 < 0.94$ (right).

Figure 4.6. Agreement between data and simulation after the energy scale and smearing corrections for $Z \rightarrow ee$ for the case where both electrons are in the EB (left), 1 electron in EB and 1 in EE (middle), and both in the EE (right).
CHAPTER 5

EVENT SELECTION

To maximize the analysis significance, events must be selected to have two high quality photons which could come from Higgs boson decay while rejecting signal like background. These criteria are developed using mainly simulated samples. A loose selection is applied at the trigger level, with a tighter selection applied offline on the fully reconstructed objects.

5.1 Data samples

The samples recorded by CMS and used in this analysis correspond to an integrated luminosity of 35.9 fb$^{-1}$ recorded through 2016 with proton-proton collisions at a center-of-mass energy of 13 TeV. Table 5.1 displays the samples used in the analysis along with their corresponding integrated luminosity.
<table>
<thead>
<tr>
<th>Samples</th>
<th>$\mathcal{L}$ [fb$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>/DoubleMuon/Run2016B-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016B-23Sep2016-v3/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016C-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016D-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016E-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016F-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016G-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016H-PromptReco-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleMuon/Run2016H-PromptReco-v3/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016B-03Feb2017-v2-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016C-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016D-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016E-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016F-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016G-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016H-03Feb2017_ver2-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/DoubleEG/Run2016H-03Feb2017_ver3-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016B-03Feb2017_v2-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016C-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016D-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016E-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016F-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016G-03Feb2017-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016H-03Feb2017_ver2-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleElectron/Run2016H-03Feb2017_ver3-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016B-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016B-23Sep2016-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016B-23Sep2016-v3/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016C-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016D-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016E-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016F-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016G-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016H-PromptReco-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016H-PromptReco-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SingleMuon/Run2016H-PromptReco-v3/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016B-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016B-23Sep2016-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016B-23Sep2016-v3/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016C-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016D-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016E-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016F-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016G-23Sep2016-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016H-PromptReco-v1/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016H-PromptReco-v2/MINIAOD</td>
<td>35.9</td>
</tr>
<tr>
<td>/SinglePhoton/Run2016H-PromptReco-v3/MINIAOD</td>
<td>35.9</td>
</tr>
</tbody>
</table>
5.2 Simulated samples

A description of how the simulated samples were generated was provided in section 3.2.8. These all used CUETP8M1 PYTHIA8 tuning and the events are propagated through the CMS detector simulation using the GEANT 4 package.

For a given recorded event (bunch crossing) several proton collisions will occur inside the CMS detector. Typically, only one of those collisions will be a hard scattering event producing interesting physics that caused the CMS trigger system to record the event. In this case, the final state particles from the other collisions constitute a background for the analysis. The number of collisions for a given event is known as pileup (PU).

There is a difference in the average number of collisions per event (average pileup) between the simulation and data samples. Having less pileup in the simulation means that the occupancy in the simulated detector is lower than in data and can affect the signals and backgrounds measured in simulation. To properly account for this, simulated events are weighted such that the pileup distributions are the same in simulation as data. This correction was validated on a $Z \rightarrow ee$ sample and is shown in figure 5.1.

5.2.1 Signal samples

Signal samples for the four Higgs boson production modes described in section 3.2.8 are generated with POWHEG and MADGRAPH aMC@NLO, although MADGRAPH aMC@NLO is used in the analysis. These parton level samples are then passed to PYTHIA8 to perform the parton showering and hadronization. The LHC Cross-Section Working Group calculations for cross-sections and branching ratios are used [21]. Signal samples range over the mass values 120, 123, 124, 125, 126, 127, 130 GeV are used for signal modeling. The full suite of signal samples used by the analysis are listed in Table 5.2 with their corresponding equivalent integrated luminosity.
5.2.2 Background samples

PYTHIA8, the MADGRAPH matrix-element generator interfaced with PYTHIA8, and the matrix element Sherpa generator were used to generate different background samples for this analysis (Sherpa was also used to generate particle showers).

The Sherpa generator was used to model the diphoton prompt-prompt background, it includes the Born processes with up to 3 additional jets at lowest-order (LO) as well as the box processes at LO.

PYTHIA8 was used to model both the prompt-fake and fake-fake backgrounds. By applying a “double EM-enriched” filter during the production of the QCD and $\gamma +$ jet sample, it is possible to reduce computing resources needed to produce and store the background samples required for this analysis. The filter requires an electromagnetic shower that could be reconstructed as a photon coming from photons, electrons, or neutral hadrons with $p_T > 15$ GeV. The signal must also have no more than two charged particles within $\Delta R < 0.2$, approximating our offline tracker iso-
lation requirement. Only electrons, muons, taus, pions, and kaons of $p_T > 1.6$ GeV and $\eta < 2.2$ are considered.

A combination of MADGRAPH and MADGRAPH aMC@NLO are used to produce the $Z \to ee$ samples, while only aMC@NLO is used to produce the diboson samples. The details of these backgrounds are in table 5.3.
TABLE 5.2

LIST OF SIMULATED SIGNAL SAMPLES

<table>
<thead>
<tr>
<th>Samples</th>
<th>( \mathcal{L} ) [ fb (^{-1}) ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>/GluGluHToGG_M125_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/ //MINIAODSIM</td>
<td>67.8</td>
</tr>
<tr>
<td>/GluGluHToGG_M-125_13TeV_powheg_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>32.2</td>
</tr>
<tr>
<td>/VBFHToGG_M125_13TeV_amcatnlo_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>516</td>
</tr>
<tr>
<td>/VBFHToGG_M-125_13TeV_powheg_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>266.6</td>
</tr>
<tr>
<td>/VHToGG_M125_13TeV_amcatnloFXFX_madspin_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>128</td>
</tr>
<tr>
<td>/VHToGG_M120_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>786.4</td>
</tr>
<tr>
<td>/VHToGG_M130_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>43.3</td>
</tr>
<tr>
<td>/ttHJetToGG_M130_13TeV_amcatnloFXFX_madspin_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>210.8</td>
</tr>
<tr>
<td>/ttHJetToGG_M130_13TeV_amcatnloFXFX_madspin_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>72.2</td>
</tr>
<tr>
<td>/ttHJetToGG_M130_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>351.7</td>
</tr>
<tr>
<td>/GluGluHToGG_M130_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>39.2</td>
</tr>
<tr>
<td>/VBFHToGG_M130_13TeV_amcatnloFXFX_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>229.7</td>
</tr>
<tr>
<td>/VHToGG_M130_13TeV_amcatnloFXFX_madspin_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>97.8</td>
</tr>
<tr>
<td>/ttHJetToGG_M130_13TeV_amcatnloFXFX_madspin_pythia8/ RunII Spring16DR80X-2_0-0-25ns-0-0-RunII Spring16MiniAODv1-PUSpring16RAWAODSIM_80X_nucRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>448.8</td>
</tr>
<tr>
<td>Samples</td>
<td>$\mathcal{L}$ [fb$^{-1}$]</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
</tr>
<tr>
<td>GJet_Pt-20to40_DoubleEMEnriched_MGG-80to80_TuneCUETP8M1_13TeV_Pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>27.4</td>
</tr>
<tr>
<td>GJet_Pt-40to40_DoubleEMEnriched_MGG-80to80_TuneCUETP8M1_13TeV_Pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>20.8</td>
</tr>
<tr>
<td>QCD_Pt-30to40_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>0.2</td>
</tr>
<tr>
<td>QCD_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>0.045</td>
</tr>
<tr>
<td>DiPhotons_Box_MGG-80toInf_Sherpa/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>178.5</td>
</tr>
<tr>
<td>DYToEE_NNPDF30_13TeV-powheg-pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>1294.3</td>
</tr>
<tr>
<td>TTGJets_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>37.5</td>
</tr>
<tr>
<td>WToLL_nuG_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISpring16DR80X-2.0-0.25ns-2.0-0.0-v0-RunIISpring16MiniAODv1-PU80X_mcRun2_asymptotic_2016_v3-v1/MINIAODSIM</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**Used in vertex study Section 5.6**

/DYJetsToLL\_M-50\_TuneCUETP8M1\_13TeV-amcatnloFXFX-pythia8/AllVertices\_CollFiltered\_de2bf50dde802a810f416955c53a2cde/USER

**Used in vertex study Section 5.6**

/GJets\_HT-40To100\_TuneCUETP8M1\_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PU80X\_mcRun2\_asymptotic\_2016\_miniAODv2\_v0-v1/MINIAODSIM | 0.2 |
| GJets\_HT-100To200\_TuneCUETP8M1\_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PU80X\_mcRun2\_asymptotic\_2016\_miniAODv2\_v0-v4/MINIAODSIM | 0.5 |
| GJets\_HT-200To400\_TuneCUETP8M1\_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PU80X\_mcRun2\_asymptotic\_2016\_miniAODv2\_v0-v4/MINIAODSIM | 4.5 |
| GJets\_HT-400To600\_TuneCUETP8M1\_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PU80X\_mcRun2\_asymptotic\_2016\_miniAODv2\_v0-v1/MINIAODSIM | 8.7 |
| GJets\_HT-600ToInf\_TuneCUETP8M1\_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PU80X\_mcRun2\_asymptotic\_2016\_miniAODv2\_v0-v1/MINIAODSIM | 26.3 |
5.3 Trigger selection

The trigger system is described briefly in section 3.2.7. The trigger used in this analysis is part of a suite of triggers used across all CMS analyses that must reduce the event rate from 40 MHz to 1 kHz by selecting events to be written to disk for the analysis. To meet these specifications, a fairly tight selection was required at the trigger level. The analysis trigger was designed to have less stringent requirements than the offline selection with these limitations in mind. The analysis trigger selected two photons with tight shower shape and isolation requirements and a cut on the invariant mass of the two photons. These reject many events which do not have two good-quality photons allowing a large rejection of background (and thus bandwidth reduction) while having minimal effect on the main signal region of the analysis.

Events which would otherwise pass the analysis selection can potentially fail the diphoton trigger decision due to the differences in object reconstruction between fast trigger reconstruction and full analysis reconstruction. The trigger efficiency is defined as the fraction of events passing the analysis selection that pass the trigger decision over the number of events passing only the analysis requirements. The trigger efficiency is measured with $Z \rightarrow ee$ events collected at the LHC and comparing to simulation and is described in section 5.3.3. A correction for this small loss of signal events is applied as a scale factor on the final signal yields, the uncertainty from this estimate is applied as a systematic by shifting the trigger efficiency scale factor.

5.3.1 Level 1 trigger

Each high level trigger (HLT) diphoton path is seeded by at least one hardware level 1 (L1) electromagnetic candidate. In principle any photon candidate in ECAL could produce an L1 seed, so the overall efficiency of a diphoton HLT path seeded by a single L1 seed is higher than requiring two L1 seeds. Due to bandwidth limitations at the L1, $p_T > 40$ GeV was the lowest momentum threshold for a single electromagnetic
candidate. Due to low-level nature of the L1 objects, the $p_T$ can be mismeasured causing an inefficiency around the $p_T$ threshold of the L1 trigger. To maintain good efficiency for events where the lead photon is not well measured by the L1 trigger, the HLT paths are also seeded by a suite of L1 pairs with varying transverse momenta. The lowest thresholds were set at 22 GeV and 15 GeV for the respective lead and sublead $p_T$ requirements to trigger the L1. Efficiency of the L1 triggers measured on data using $Z \rightarrow ee$ events by the tag and probe (described in detail in section 5.3.3) method are shown in figures 5.2 and 5.3.

![Figure 5.2](image)

Figure 5.2. Trigger efficiency measured on data for $Z \rightarrow ee$ events using the tag and probe method. Efficiency with respect to offline probe $p_T$. Efficiency of photons in the 4 cut-based analysis categories are shown. The plots correspond to L1 objects from EG 22 (left) and EG 10 (right) which are used to seed the diphoton HLT trigger.

5.3.2 High level trigger

The objects flagged by the L1 trigger are used as input seeds to the HLT. The HLT does a much more robust software reconstruction, very similar to the offline selection...
Figure 5.3. Trigger efficiency measured on data for $Z \rightarrow ee$ events using the tag and probe method. Efficiency with respect to offline probe $\eta$. Curves correspond to L1 objects matched to offline photons in the cut-based analysis categories listed. L1 objects EG 22 (left) and EG 10 (right) are the seeds for diphoton HLT. An additional selection of $p_T^{\text{lead}} > 41.67$ GeV and $p_T^{\text{sublead}} > 31.25$ GeV based on scaling $p_T$ cuts for diphotons with invariant mass equal to the Higgs boson was applied.

to further reduce the bandwidth of CMS. The HLT selection criterion can be grouped into general, isolation plus calorimeter identification (Iso+CaloId), and $R_9$. The HLT selection applies the global criterion to all objects then either the Iso+CaloId or the $R_9$ selection. The variables consist of HLT versions of the following (based on the variables defined in section 4.6.3:

- **General variables**: $p_T$ of both photons, $m_{\gamma\gamma}$, HE (with cone size $R = .14$)
- **Variables used in CaloId+Iso filters**: $\sigma_{\text{min}}$, $\rho$-corrected ECAL PF Cluster ISO, Tracker isolation in a hollow cone (with $R = .29$ outer and $R = 0.06$ inner radii)
- **Variables used in $R_9$ filters**: $R_9$

The so-called L1-seeded leg of the HLT is required to have $E_T > 30$ GeV. The cluster must be at $|\eta| < 2.5$ and have an $R_9 > 0.5(0.8)$ in EB(EE). A baseline selection of $H/E < 0.12(0.1)$ EB(EE) is additionally applied. Each object is then required to satisfy either $R_9 > .85(0.9)$EB(EE). If they pass these tight $R_9$ requirements, they are allowed to fail the CaloId+Iso requirements. If they fail the tight $R_9$ requirements,
they must pass the CaloID+Iso filters to continue to the unseeded step. These consist of $\sigma_{\eta\eta} < 0.015(0.035)EB(EE)$ and ECAL isolation $< 6.0 + 0.002E_T$.

If at least one cluster passes the above criteria, the entire ECAL is clustered using the full HLT clustering algorithm (not just those seeded by L1 objects as in the L1-seeded leg). Two clusters are now required to pass the unseeded requirements (one of these clusters will be the object which passed the L1-seeded HLT requirements above). An $E_T > 18$ GeV requirement is used in place of the 30 GeV threshold for the L1-seeded HLT. The remainder of the selections are exactly the same as for the L1-seeded leg, with the addition that both objects must pass satisfy track isolation $> 6.0 + 0.002E_T$.

Finally, the mass of the diphoton object is required to be above 80 GeV. The full set of trigger requirements can be found in table 5.4.

\begin{table}[h]
\centering
\small
\begin{tabular}{|c|c|c|c|c|}
\hline
& H/E & $\sigma_{\eta\eta}$ & R$_{9}$ & ECAL iso. & Track iso. \\
\hline
EB; $R_{9} > 0.85$ & $< 0.12$ & -- & $> 0.5$ & -- & -- \\
\hline
EB; $R_{9} \leq 0.85$ & $< 0.12$ & $< 0.015$ & $> 0.5$ & $< 6.0 + 0.012p_T$ & $< 6.0 + 0.002p_T$ \\
\hline
EE; $R_{9} > 0.90$ & $< 0.1$ & -- & $> 0.8$ & -- & -- \\
\hline
EE; $R_{9} \leq 0.90$ & $< 0.1$ & $< 0.035$ & $> 0.8$ & $< 6.0 + 0.012p_T$ & $< 6.0 + 0.002p_T$ \\
\hline
Other trigger requirements & & & & & \\
\hline
HLT seeded $p_T > 30$ GeV & & & & & \\
\hline
HLT unseeded $p_T > 18$ GeV & & & & & \\
\hline
$\text{m}_{\gamma\gamma} > 90$ GeV & & & & \\
\hline
\end{tabular}
\caption{SELECTION FOR THE DIPHOTON TRIGGER}
\end{table}
5.3.3 Trigger performance

The thresholds for the HLT object selection were chosen to maintain sustainable bandwidth of the HLT while still accepting as many signal events as possible, especially for the more performant categories of the analysis.

The tag and probe (described below) technique was used to measure the efficiency of the trigger selection with a selection of $Z \rightarrow ee$ events from data. Using the event selection described in section 4.7.1 it is possible to select high purity $Z \rightarrow ee$ events from data. Since the trigger used to select $Z \rightarrow ee$ events only require one electron, it is possible to measure the triggering efficiency of the non-triggered electron in an unbiased way.

One of the objects reconstructed as tight photons which passed the $Z \rightarrow ee$ requirements is matched to an HLT object and caused the tag and probe trigger to fire. This way, any additional reconstructed photons in the event are completely unbiased by the trigger selection. All non-tagged photons are then required to pass the loose analysis photon ID cut and defined as probes (Photon ID is described in detail in section 5.5). The collection of probes passing this selection become the denominator of the HLT efficiency for the seeded leg of the diphoton HLT path. If the probe can be matched geometrically (within $R < 0.3$) to the seeded leg of the diphoton HLT path, an entry is added to the numerator. The same analysis is done separately for the unseeded leg of the diphoton HLT path, with lower preselection $p_T$ requirement being applied to the probe as well as the requirement that the reconstructed tag match the seeded leg of the diphoton HLT path. This full procedure is duplicated for all combinations of photons which can satisfy the tag and probe requirements for a given event to avoid any selection bias.

HLT efficiency of both the seeded and unseeded legs of the diphoton trigger with respect to photon $p_T$ and $\eta$ can be found in figures 5.4 and 5.5 respectively.
Figure 5.4. Trigger efficiency measured on data for $Z \rightarrow ee$ events using the tag and probe method. The four untagged cut-based categories are shown for the seeded (left) and unseeded (right) HLT leg of the diphoton trigger. The analysis requires $p_{T_{lead}} > m_{\gamma\gamma}/3$ and $p_{T_{sublead}} > m_{\gamma\gamma}/4$ with $m_{\gamma\gamma} > 100$.

Figure 5.5. Trigger efficiency measured on data for $Z \rightarrow ee$ events using the tag and probe method. The four untagged cut-based categories are shown for the seeded (left) and unseeded (right) HLT leg of the diphoton trigger.

An additional cut corresponding to $p_{T_{lead}} > 125/3$ GeV and $p_{T_{sublead}} > 125/4$ GeV based on scaling mass cuts at the Higgs boson mass was applied. The inefficiency in the $|\eta| \sim 1.5$ comes from an $R_9$ turn on.
5.4 Event preselection

All photons entering the analysis are subject to a loose preselection before the analysis level selection is applied. The first requirement of the preselection is to remove events which were filtered out in the simulation (due to computing requirements). The second is to apply selection slightly tighter than the trigger (to minimize bias from any mismodeling of the trigger in simulation). The final requirement is to remove photons which will never enter the analysis. The preselection was optimized and validated on both data and simulation.

All photons must have either $R_9 > 0.8$, PF Charged Iso $< 20$ GeV, or PF Charged Iso/p_T $< 0.3$. This mimics the generator level cuts applied during the final steps of event generation from simulation.

The photons then must satisfy the preselection based on the HLT selection. Applying these cuts reduces the systematic dependence of the analysis on the precise understanding of the trigger efficiency turn-ons. The variables are defined in section 4.6.3 and the requirements for those variables are shown in figure 5.5. The diphoton pair must also satisfy $m_{\gamma\gamma} > 100$ GeV. With the exception of the $R_9$ requirement, these are all strictly tighter than the trigger selection.

<table>
<thead>
<tr>
<th>TABLE 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRESELECTION REQUIREMENTS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H/E</th>
<th>$\sigma_{\eta \eta}$ (5x5)</th>
<th>$R_9$ (5x5)</th>
<th>pfPhoIso</th>
<th>TrackerIso</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB; $R_9 &gt; 0.85$</td>
<td>&lt;0.08</td>
<td>–</td>
<td>&gt;0.5</td>
<td>–</td>
</tr>
<tr>
<td>EB; $R_9 \leq 0.85$</td>
<td>&lt;0.08</td>
<td>&lt;0.015</td>
<td>&gt;0.5</td>
<td>&lt;4.0</td>
</tr>
<tr>
<td>EE; $R_9 &gt; 0.90$</td>
<td>&lt;0.08</td>
<td>–</td>
<td>&gt;0.8</td>
<td>–</td>
</tr>
<tr>
<td>EE; $R_9 \leq 0.90$</td>
<td>&lt;0.08</td>
<td>&lt;0.035</td>
<td>&gt;0.8</td>
<td>&lt;4.0</td>
</tr>
</tbody>
</table>

60
Each photon must also be within the geometrical acceptance of the ECAL, so a requirement of $|\eta| < 2.5$ and $1.442 < |\eta| < 1.566$ is applied.

In order to measure the efficiency of the preselection, the tag and probe method on the $Z \rightarrow ee$ sample is used and the data and simulation results are compared. A scale factor is derived to correct the simulation to match the data and is shown in table 5.6. This scale factor is then shifted by its uncertainty and included as a systematic uncertainty on the signal yields.

**TABLE 5.6**

PRESELECTION EFFICIENCY MEASUREMENT

<table>
<thead>
<tr>
<th>Data</th>
<th>Simulation</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrel; R9 &gt; 0.85</td>
<td>0.9423</td>
<td>0.0004</td>
</tr>
<tr>
<td>Barrel; R9 &lt; 0.85</td>
<td>0.8225</td>
<td>0.0012</td>
</tr>
<tr>
<td>Endcap; R9 &gt; 0.90</td>
<td>0.9153</td>
<td>0.0007</td>
</tr>
<tr>
<td>Endcap; R9 &lt; 0.90</td>
<td>0.5011</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

5.4.1 Electron veto efficiency and scale factors

In order to reject electrons from the photons in the analysis, an electron veto is applied. It is similar to inverting the pixel seed described in section 4.6.1. It is, by definition, not possible to measure the efficiency of the pixel seed veto using electrons. We use the real photon from a selection of events from a $Z \rightarrow \mu\mu\gamma$ sample to measure the efficiency of the electron veto on photons.
This high purity (98%) sample from data came from the dimuon primary dataset with the full integrated luminosity of 35.9 fb$^{-1}$. The simulated sample was Drell-Yan decaying to two leptons. Events in both simulation and data are required to pass at least one of a suite of double muon triggers as recommended by the CMS Muon group. If at least 2 muons pass the “tight” muon ID [6] with a $p_T > 10$ GeV and $|\eta| < 2.4$ with the $(PFChargedHadronIso)/p_T < 0.2$ the event is stored. The muons must be oppositely signed with a dimuon mass $m_{\mu \mu} > 35$ GeV. If a photon is found satisfying the preselection criteria and is within $\Delta R < 0.8$ of the nearest muon and the 3-body invariant mass $60 < m_{\mu \mu \gamma} < 120$ GeV and $m_{\mu \mu} + m_{\mu \mu \gamma} < 180$ GeV events will enter the analysis.

The photon is then used as a probe to measure the electron veto efficiency. The ratio of the number of photons passing the electron veto divided by the total number of preselected photons are given in table 5.7 in simulation and data along. Efficiencies and the ratios between simulation and data are given in four categories based on $R_9$ and $\eta$.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
 & Data & & & Simulation & & \\
\hline
Barrel; $R_9 > 0.85$ & 0.9932 & 0.0005 & & 0.9972 & 0.0008 & 0.9960 & 0.0010 \\
Barrel; $R_9 < 0.85$ & 0.9750 & 0.0018 & & 0.9824 & 0.0041 & 0.9924 & 0.0045 \\
Endcap; $R_9 > 0.90$ & 0.9851 & 0.0013 & & 0.9852 & 0.0025 & 0.9999 & 0.0029 \\
Endcap; $R_9 < 0.90$ & 0.9525 & 0.0071 & & 0.9658 & 0.0113 & 0.9862 & 0.0137 \\
\hline
\end{tabular}
\caption{Conversion-Safe Electron Veto Efficiency Measurement}
\end{table}
5.5 Photon identification

In order to select events when Higgs boson decay into two photons, one must be able to separate the peak in the diphoton mass distribution on top of a smoothly falling background. The background can be broken up into two components, “irreducible” and “reducible”. The irreducible component consists of two prompt photons that come from the hard scattering vertex. The reducible component has at least one object which is misidentified as a photon. These events with “fake” photons arise from dijet and $\gamma + j$ events. If at least one jet fakes a photon (typically when the jet fragments into a $\pi^0$ or $\eta$ meson which decay into two photons) it is possible that this event will appear to have 2 photons coming from the hard scattering vertex. Since we are looking at high energy photons, these photons from jet fragmentation tend to be tightly collimated and appear very similar to a single prompt photon.

Photon identification in CMS is designed to reduce these fake photons while selecting as many prompt photons as possible. A photon ID BDT is trained using several shower shape and isolation variables to select prompt photons and reject jets faking photons.

The shower shape allows discrimination of closely aligned photons produced in neutral meson decays and single prompt photons produced in the primary interaction vertex. Since a neutral meson decays into two photons, the profile in ECAL is wider than the typical prompt photon. Photons from Higgs boson decays are isolated (not accompanied by other particles and associated energy nearby). Measuring the isolation can reject events where photons arise from neutral hadrons in jets.

In order to effectively reject fake photons and select prompt photons a classification BDT was trained. The photon ID BDT training used a signal and background sample with the full set of variables listed in section 4.6.3. The prompt (signal) and fake (background) photons were taken from the same $\gamma + j$ simulated sample. Events with one prompt and at least one fake photon from the sample are considered. The
signal sample is reweighted such that the supercluster $p_T$ and supercluster $\eta$ are the same as for the reconstructed fake photon background. Both reconstructed photons are required to satisfy the analysis preselection criteria 5.4 before being considered in the BDT training. Half of the events in the sample are used to train the photon ID BDT, while the other half are used to test the output.

After the final training and testing on the $\gamma$+jet sample, the BDT output was checked on the Higgs boson signal samples and the remaining analysis backgrounds. Figure 5.6 shows the photon identification BDT score of the lower scoring photon of diphoton pairs in the analysis mass window ($100 < M_{\gamma\gamma} < 180$) for events satisfying the analysis preselection criteria. The agreement between the sum of all background and data is consistent.

In the analysis, photons are required to have a photon ID BDT score of greater than -0.9. This guarantees 99% efficiency on signal photons.

5.5.1 Validation with $Z \rightarrow ee$ events

Once the training for the photon ID has been completed (with all corrections applied to input variables) validation is performed on $Z \rightarrow ee$ events using the event selection described in this section. Figure 5.7 shows the score from the photon ID BDT for $Z \rightarrow ee$ data and simulation. Some systematic uncertainties are expected due to the differences in data and simulation of the input variables. The systematic uncertainty on the photon ID BDT is given as a shift in the photon ID BDT output of $\pm 0.03$ along with a linearly increasing term to account for the tail seen at low photon ID BDT scores.
Figure 5.6. Photon identification BDT score of lower scoring photon of diphoton pairs in the analysis mass window ($100 < m_{\gamma\gamma} < 180$ GeV). Histograms for the sum of the simulated background as well as broken up into components is shown along with events from actual data.
Figure 5.7. Photon ID MVA output distribution for $Z \rightarrow ee$ events in data and simulation, for photon candidates in the ECAL barrel (left) and endcaps (right). The event selection consists of the preselection (including the requirement ID MVA > 0.9) with the electron veto inverted. The systematic uncertainty applied to the shape from the simulation, corresponding to a shift of 0.03 in the value of the MVA output along with a linearly increasing term for the low values is shown by the shaded region.
5.6 Vertex selection

In order to calculate the invariant mass of a diphoton pair, the opening angle between the two photons must be properly measured. The ECAL has no pointing power to determine the position of the vertex (and thus the opening angle between the photons). Normally, photons leave no signal in the tracker, making determination of the vertex challenging (events had an average of $\approx 20$ reconstructed vertices from pileup from which to choose).

It is possible to use the recoiling tracks to determine the vertex position; an even more precise measurement can be obtained when a photon pair produces (or “converts”) in the tracker. The vertex selection can have a large impact on the mass resolution of the Higgs boson; selecting the wrong vertex has an average worsening of 1 GeV in situ. By calculating a per-event probability of selecting the correct vertex the analysis can estimate which events have worsened mass resolution due to vertex mismeasurement. Full details of the standard CMS vertex identification can be found in [7].

5.6.1 Vertex selection performance of non-converted photons

It has been shown that given the energy resolution of ECAL and the typical kinematics of the Higgs boson decay into two photons, the diphoton opening angle is a negligible contribution to the mass resolution if the chosen vertex is within 1 cm of the true Higgs boson vertex [17]. The efficiency of selecting a vertex within 1 cm of the true simulated vertex is shown in figure 5.8 as a function of $p_T$ of the diphoton pair and number of reconstructed vertices on the simulated Higgs boson signal samples.

In addition to the usual event reweighting so that the simulated pileup distribution matches that of data, the events are reweighted so that the simulated physical vertices distribution (beamspot) is the same as in data. The overall efficiency of identifying the Higgs boson vertex within 1 cm of the true vertex is 81% after the analysis
Figure 5.8. Efficiency of choosing the vertex within 1 cm of the true vertex from $H \rightarrow \gamma \gamma$ simulation is shown as a function of the pT of the diphoton pair and number of vertices in the event. The average vertex probability is superimposed with its uncertainty derived from Z validation. The plot are reweighted to match the final dataset PU and from beam spot size differences between data and simulation.

The performance is then further validated on data recorded by CMS. By using $Z \rightarrow \mu \mu$ events, one can very accurately identify the vertex of the Z boson (between 30 $\mu$m and 60 $\mu$m). Since the algorithm for identifying the correct vertex uses only the recoiling tracks of the Higgs boson vertex, the same procedure can be performed on $Z \rightarrow \mu \mu$ events. To properly control the sample, the tracks associated with the muons from Z decay are removed before running the vertex ID BDT described earlier in this section.

A dimuon trigger was used to collect events used for this study, corresponding to 35.9 fb$^{-1}$.

Events are required to have at least two muons passing the tight identification requirements with a mass between 70 and 110 GeV. Using this selection, good agreement was found between data and simulation as shown for the input variables in figure 5.9. The final efficiency of finding the correct vertex is compared in figure 5.10.
between data and simulation.

Figure 5.9. Distributions of inputs variables for data and simulation, together with their ratios for signal vertices (right vertex: matched to the muons) and background vertices (wrong vertex: not matching the muons).
Figure 5.10. Efficiency as a function of the second BDT score MVAProb to find the vertex within 1 cm of the true one using $Z \rightarrow \mu\mu$ events for data and simulation.

5.6.2 Vertex selection performance of converted photons

For events where at least one photon “converts” into an electron-positron pair there will be tracks in the detector which can be used to point to the Higgs boson vertex. The vertex ID algorithm can be verified with a $\gamma + \text{jet}$ sample. The photon and jet are identified and the tracks associated with the jet are removed from the event in order to mimic the diphoton system. Events are selected where the true photon is tagged as having been converted. This way, the tracks associated with the converted photon can be used to select the vertex. The vertex selected by the ID BDT can then be compared with the vertex tagged by the jet reconstruction.

A single photon trigger with $p_T > 50$ GeV was used to collect events for this
study. Due to limitations at data taking, this collection corresponds to a much smaller luminosity than the $Z \rightarrow \mu \mu$ study, but statistics are sufficient. Events were further required to contain a photon with a $p_T > 55$ GeV and at least one jet with $p_T > 30$ GeV and $\eta < 2.5$. A combination of photon and jet must also be non-overlapping (not contained inside the same cone with radius: $R(\gamma,\text{jet}) > 0.4$).

The tracks within $R < 0.4$ of the jet are ignored by the vertex ID BDT so that the jet appears like an unconverted photon. Events are selected if the true photon is tagged as a converted photon. The vertex determined by the vertex ID BDT is compared to the vertex associated with the fully reconstructed jet. The efficiency of this identification algorithm comparing data and simulation is shown in figure 5.11. With the additional information from the photon conversion tracks, the efficiency to reconstruct the vertex within 1 cm of the true vertex is very high.

5.6.3 Per-event vertex probability

It is possible to estimate the probability that the vertex selected by the identification BDT described in this section correctly identified the vertex on an event by event level. A BDT is trained to determine if the vertex selected by the vertex ID BDT is within 1 cm of the true vertex along the beam-axis. Figure 5.12 shows the output MVA distribution for correctly identified as well as incorrectly identified vertices.

For events where at least one photon converts inside the tracker, the vertex probability must be analyzed separately. Using the same $\gamma + \text{jet}$ sample, figure 5.13 shows the per event probability that the ID BDT selects a vertex within 1 cm of the jet identified (true) vertex along the beam-axis.
Figure 5.11. The efficiency of selection of the correct vertex in $\gamma + \text{jet}$ events with converted photons as a function of the pT of the photon-jet system (left) and as a function of the number of vertices in the event (right) in data and simulated events and their ratio.
Figure 5.12. Normalized distributions of the second BDT score “MVAProb” for selected and misassigned vertices in data and simulation in \(Z \rightarrow \mu \mu\) events.
Figure 5.13. The efficiency of selection of the correct vertex in $\gamma + \text{jet}$ events with converted photons as a function of the pT of the photon-jet system (left) and as a function of the number of vertices in the event (right) in data and simulated events and their ratio.
In order to improve the overall sensitivity of the analysis events are categorized by mass resolution, signal-to-background ratio, and production mechanism. While gluon-gluon fusion (ggH) events produce no additional particles, the remaining production mechanisms can be selected by associated final state particles. If the Higgs boson is produced in associated production with a Z or W boson (VH), additional requirements for missing transverse energy, leptons, or jets are applied. To select events where the Higgs boson is produced by vector boson fusion (VBF), additional requirements for a pair of jets separated by a large rapidity angle are applied. To select Higgs bosons produced via $t\bar{t}H$ fusion additional requirements for b-quarks along with leptons or additional jets are applied.

The categories with object requirements beyond two photons are called exclusive categories. The addition of the exclusive categories allow for measurement of the coupling of the Higgs boson to top quarks as well as vector bosons while reducing the background in these categories due to the constraints from additional objects in the final state.

The categories which require only two photons are called inclusive categories and correspond to Higgs bosons produced by ggH fusion. These categories are defined by the mass resolution of the diphoton pair.

In the case where an event satisfies the requirements of more than one category, the highest priority category is selected. If multiple diphotons in the event satisfy the requirements of the highest priority tag, the diphoton pair with the highest $p_T$ is
selected by the analysis. The classification priority is described below from highest priority to lowest:

- Events with leptons from leptonic or semi-leptonic top decays. Details in section 6.2.1.1.
- Remaining events with leptons from ZH and WH production. Details in section 6.2.2.2.
- Remaining events with jets from hadronic top decays. Details in section 6.2.1.2.
- Remaining events with two forward jets from VBF production. Details in section 6.2.3.
- Remaining events with large missing energy from ZH events where Z decays to neutrinos. Details in section 6.2.2.3.
- Remaining events with two jets from VH where the vector boson decays hadronically. Details in section 6.2.2.4.
- Remaining events from ggH production. Details in section 6.1.

Several of the categories listed have additional sub-categorization to improve overall signal-to-noise and will be described in their respective sections. During the optimization of these categories, care was taken to avoid looking at events in the signal region of the data. This was typically done by excluding events with diphoton invariant mass $115 < m_{\gamma\gamma} < 135$ during optimization and validation between data and simulation.
6.1 Inclusive categorization

If an event contains a diphoton pair which satisfies the preselection criteria described in section 5.4 and does not satisfy any of the exclusive categories selection, it falls into one of the inclusive categories. These are designed to select Higgs boson decays to two photons from ggH production. While this is the most common production mechanism, it also has the largest background of all the categories. In order to improve the sensitivity of this category, events are divided into four sub-categories such that the higher performing categories have a higher signal-to-background or better invariant mass resolution.

6.1.1 Multivariate diphoton categorization

A multivariate technique is used to break the inclusive categories up to improve the overall analysis performance. The diphoton classifier used in this analysis is a boosted decision tree using the TMVA package [35]. This BDT is set up to give a higher score to events with more signal-like characteristics and with good invariant mass resolution. The classifier must not preferentially select events based on the invariant mass of the diphoton system; the analysis must remain unbiased to the Higgs boson mass used in the simulation used for training.

Once an event passes the preselection, an additional loose cut on the photon ID MVA output is applied $> -0.9$. Finally, the photon transverse momentum must be above a mass-dependent threshold ($p_T > 1/3m_{\gamma\gamma}$ for the leading and $p_T > 1/4m_{\gamma\gamma}$ for the subleading photon).

The classifier takes as input the photon ID BDT output, an estimate of the diphoton mass resolution, and various kinematic properties of the diphoton system. The inputs are carefully selected to minimize the ability of the classifier to determine the invariant mass of the diphoton pair. Below is the list of inputs to the diphoton event classifier:
• transverse momentum for both photons divided by mass of the system: $p_T/m_{\gamma\gamma}$
• pseudo-rapidity of both photons: $\eta$
• cosine of angle between two photons in transverse plane: $\cos(\Delta\phi)$
• ID BDT output of both photons
• per-event relative mass resolution estimate, assuming the correct vertex was reconstructed: $\sigma_{rv}$
• per-event relative mass resolution estimate, assuming the incorrect vertex was reconstructed: $\sigma_{wv}$
• per-event probability the correct primary vertex was selected to reconstruct the mass: $p_{vtx}$

The per-event relative mass resolution estimate is calculated from the propagation of the photon energy resolution estimates assuming Gaussian resolutions:

$$\sigma_{rv} = \frac{\sigma_{m}^{right}}{m_{\gamma\gamma}} = \frac{1}{2} \sqrt{\left(\frac{\sigma_{E1}}{E_1}\right)^2 + \left(\frac{\sigma_{E2}}{E_2}\right)^2}$$  \hspace{1cm} (6.1)

In order to estimate the energy and associated resolution, we use a regression trained on simulated events. The $\sigma_E$ from simulation is then corrected to match the estimate of the resolution in data. This is accomplished by adding a smearing term in quadrature to the simulated photon energy after regression. In the case where the correct vertex has been chosen, this is simply propagated through to the diphoton mass. In the case where we assume the incorrect primary vertex has been selected, we must apply additional steps. The regression BDT also takes as input the displacement between the correct and selected vertex, $\sigma_{vtx}$.

The distance between the selected vertex and the true vertex is a Gaussian with a width of $\sqrt{2}\sigma_{Z}^{beamspot}$. Using the impact positions of the two photons in the calorimeter we calculate $\sigma_{vtx}$. The mass resolution in the case that the wrong vertex is selected is therefore given as:

$$\sigma_{wv} = \frac{\sigma_{m}^{wrong}}{m_{\gamma\gamma}} = \sqrt{\left(\frac{\sigma_{m}^{right}}{m_{\gamma\gamma}}\right)^2 + \left(\frac{\sigma_{vtx}}{m_{\gamma\gamma}}\right)^2}$$  \hspace{1cm} (6.2)
The BDT training must have the correct ratio of signal-to-background events with respect to the mass resolution. We weight the signal events by the below formula to accomplish this:

\[ w_{\text{sig}} = \frac{p_{\text{vtx}}}{\sigma_{rw}} + \frac{1 - p_{\text{vtx}}}{\sigma_{wv}} \]  

(6.3)

This weighting allows the BDT to classify events with better mass resolution higher than those with poor mass resolution.

Several simulated samples of background and signal processes are used to train the diphoton classification BDT at 13 TeV. Signal samples consist of the gluon fusion, VBF, VH, and t\(\bar{t}\)H production modes, weighted based on their respective cross-sections, with a mass of 125 GeV. These samples have been described in section 5.2.1. For the background samples prompt-prompt (diphoton+jets), prompt-fake (photon+jet and QCD dijet samples), and fake-fake (QCD dijet samples) are also weighted by respective cross-sections. For training, 3/4 of the events are used, while the remaining 1/4 of events are used for validation. Due to limitations during simulation, the QCD dijet samples have low statistics and high weights after our event selection. To avoid biases brought on by these few high weight background events, we reduce the weight of the QCD dijet events (fake-fake and prompt fake) by a factor of 25 during training. This reduced weighting generates a mildly sub-optimal training of the BDT, but removes the strong bias that could arise from single events with high weights.

The VH signal samples contain a small fraction of events with negative weights. It is complex to properly treat these using the standard TMVA implementation and can generate instabilities in the training. We chose to ignore those events in the BDT training. This provides a somewhat sub-optimal categorization, but since we use these events throughout the rest of the analysis, the final results are unbiased.
Figure 6.1. Comparison between data and simulation for leading photon kinematic variables, mass resolution estimates, and angle separating chosen photons. The cross-section weighted combination of simulated events are stacked and compared to the data, normalized to the area. Below each plot is the ratio of data over simulation.
Figure 6.2. Comparison between data and simulation for subleading photon kinematic variables and event vertex probability. The cross-section weighted combination of simulated events are stacked and compared to the data, normalized to the area. Below each plot is the ratio of data over simulation.
Figure 6.3. The mass resolution of the diphoton objects binned in signal efficiency bins of the diphoton BDT from simulation. Background events in the left plot and signal in the right.

Figure 6.4. Diphoton BDT output ROC curve for background vs signal simulation in the left plot. Diphoton BDT output shape for signal and background simulation in the right plot.
6.1.2 Validation of diphoton BDT with $Z \rightarrow ee$

Simulation accurately predicts the diphoton system arising from Higgs boson decay, but some inputs to the diphoton BDT are sensitive to the detector conditions. In order to properly validate the performance of the BDT with respect to these variables, $Z \rightarrow ee$ events can be used to test the BDT performance.

The $Z \rightarrow ee$ events are useful to validate the BDT performance and the bias of input variables arising from differences in simulation and the real detector. Since the $Z$ boson is a spin 1 particle, the kinematics of the decay particles are different from the spin 0 Higgs boson. Additionally, the natural width of the $Z$ boson is comparable to the mass resolution of the CMS detector, whereas the Higgs boson natural width is negligible when compared to the detector mass resolution. The $Z \rightarrow ee$ studies will therefore be somewhat less sensitive to imperfections in the modeling of the mass resolution.

The selection follows the procedure described in section 4.7.1; the signal model for $Z \rightarrow ee$ was constructed in the same way as the full Higgs boson analysis with
the full scale and smearing corrections applied.

Figures 6.6-6.8 show validation of the inputs to the diphoton BDT on $Z \rightarrow ee$ events. The final BDT output validation is shown in the right plot of figure 6.9 and the data-simulation comparison of the $Z \rightarrow ee$ mass peak is shown in the left plot of figure 6.9. Overall the differences between the data and simulation shown by $Z \rightarrow ee$ events is contained by the statistical and systematic uncertainties on the measurement.

6.1.3 Systematic uncertainties on diphoton BDT output

The largest sources of systematic uncertainty due to the diphoton BDT arise from the photon ID MVA and the per-photon energy resolution estimate. This only affects the signal samples, since the background modeling is completely data driven.

The photon ID MVA output ranges from -1 to 1 trained such that prompt photons provide an output closer to 1. To calculate the systematic uncertainty on the diphoton BDT from this, the signal photon ID MVA is shifted for every photon by $\pm 0.03$ with an additional linear increase of the uncertainty for events with a low single photon ID MVA value. This systematic effect shifts both input photons in the same direction, providing maximal shift in the diphoton BDT output (since both photons either appear more or less signal like). The updated diphoton MVA output is then used to determine the migration in the final signal categories and the signal yield is updated.

Systematic uncertainty in the per-photon energy resolution arises from imperfect modeling of the shower in the ECAL detector. This affects the mass resolutions input to the diphoton BDT. The resolution was smeared by $\pm 5\%$ for each photon in the same direction and propagated through to the diphoton BDT to give maximal category migration.

The systematic uncertainty on the diphoton BDT output can be seen in the error bands of figure 6.9.
Figure 6.6. Comparison between data and simulation for estimated per-object energy resolution, compared on electrons reconstructed as photons from Z boson decay, after all corrections are applied and propagated through the energy regression. For every variable, the comparison between the simulation (normalized to the product of process cross section and integrated luminosity) and data is shown, together with the ratio plot.
Figure 6.7. Comparison between data and simulation for BDT input variables (leading photon kinematic variables, mass resolution estimates under correct and incorrect vertex hypotheses, cosine of the angle in between the two photons) on electrons reconstructed as photons from Z boson decay. For every variable, the comparison between the simulation (normalized to the product of process cross section and integrated luminosity) and data is shown, together with the ratio plot.
Figure 6.8. Comparison between data and simulation for BDT input variables (kinematic variables for the subleading photon, per-event vertex probability) on electrons reconstructed as photons from Z boson decay. For every variable, the comparison between the simulation (normalized to the product of process cross section and integrated luminosity) and data is shown, together with the ratio plot.
(a) Diphoton BDT $Z \rightarrow ee$ validation mass distribution

(b) Diphoton BDT output distribution using $Z \rightarrow ee$

Figure 6.9. Comparison between data and simulation for electrons reconstructed as photons from $Z$ boson decay. The dielectron invariant mass distribution (on the left) and the diphoton BDT output (on the right) show good agreement between data and simulation.
6.1.4 Category optimization with diphoton BDT

In order to leverage the separation power of the diphoton BDT output between signal and background, the events are categorized based on the BDT output. Each category is then fit using the signal plus background models based on the diphoton mass spectrum. Four categories based on diphoton BDT output were used in this analysis. Splitting into more categories was found to improve expected significance of the analysis by only 1.2% as shown in figure 6.10.

![Figure 6.10. Expected significance vs the number of inclusive categories derived from diphoton BDT output. The expected significance is approximated using a simplified signal+background fit. The significance gained by adding an additional category is roughly 1% when adding a 5th category and continues to drop as the number of categories increases.](image-url)
6.2 Exclusive categories

The exclusive categories are designed to select events where the Higgs boson that decays into two photons comes from a production mode different than gluon-gluon fusion. Higgs boson production via top fusion, vector boson fusion, and in association with a W or Z boson have additional final state particles that can be used to drastically reduce the background component when properly identified.

The standard analysis selection is applied for all the exclusive categories with additional object requirements as described in the following sections.

6.2.1 t\bar{t}H categories

If a diphoton event is found to have two b-jets with additional leptons or jets in the final state the Higgs boson was likely to have been produced in associated production with a top quark-antiquark (t\bar{t}H) pair. The cross-section for this production mode is much lower than that of the ggH production, but with appropriate selection, the background can be reduced by a much larger factor. This channel is additionally interesting because of the large mass of the top quark. Measuring the coupling of the Higgs boson to top quark might provide further insights into electroweak symmetry breaking.

The top quarks quickly decay into a W boson and a b-quark. Additional requirements beyond the default diphoton selection are applied to select only events with this final state. A special b-tagging algorithm is used to identify jets from b quark hadronization. Then the W decays are identified by their final state consisting of a single charged lepton and missing transverse energy from the neutrino, or by the decay to two jets with an invariant mass consistent with the W boson. Events are categorized based on the decay of the W into either the t\bar{t}H hadronic or t\bar{t}H leptonic categories. The performance of these categories is shown in table 7.1.

Both leptonic and hadronic selections require the presence of jets and b-tagged
jets; ak4PFCHS jets, as described in section 6.2.3.1 are required with $p_T > 25$ GeV and $|\eta| < 2.4$. Jets must not overlap with the photons from Higgs boson decay and a requirement in $\phi - \eta$ space, $R(jet, \gamma) > 0.4$, is applied. The Combined Secondary Vertex (CSVv2) algorithm was used to identify b-jets [9].

The $t\bar{t}H$ leptonic category requires there to be 2 leptons (in this case either muon or electron) with $p_T > 20$ GeV and $|\eta| < 2.4$.

A special isolation variable, called mini-isolation ($I_{mini}$), is used to make sure leptons aren’t misidentified due to pileup. $I_{mini}$ for muons is defined as the scalar sum of the transverse momenta of all particle candidates, excluding the muon, within a $p_T$ dependent cone size of radius, $R$, centered on the muon position divided by the lepton $p_T$. The $p_T$ dependent cone size, $R$, is given as follows:

$$R = \frac{10}{\min(\max(p_T(\mu), 50), 200)}$$  \hspace{1cm} (6.4)

The isolation sum is corrected for pileup by estimating the energy deposited in the isolation cone using the tracks originating from pileup vertices. The variable cone size used by mini-isolation is more effective than using a fixed cone size for busy $t\bar{t}H$ events. Differences between data and simulation for muon selection using mini-isolation are calculated using tag and probe and $Z \rightarrow \mu\mu$ events. Scale factors were close to one in all cases and systematic uncertainties of 0.5% covered any differences found after the scale factors were applied. An additional 1% uncertainty was applied to cover the difference in scale factors between the $Z$ (where the scale factors were derived) and $t\bar{t}H$ (used in the analysis) environments.

Electrons within $|\eta| < 1.4442$ or $1.566 < |\eta| < 2.5$ that pass the loose cut based E-Gamma electron working point enter into the analysis. The details of the electron working point can be found in [11].

The background estimation for the $t\bar{t}H$ categories is defined by a control sample from data to avoid any biases from simulation. The control sample is selected with
a pair of photons where only one photon satisfies the preselection criteria, and the other has an inverted photon ID BDT requirement. There is additionally the signal sideband region, which uses the same event selection as the signal region, but excludes events with $115 < m_{\gamma\gamma} < 135$ GeV. The events in the control region are then weighted based on photon $p_T$ and $\eta$ to match the signal sideband region. This gives a statistically independent control sample from data with similar kinematic properties to the actual signal region. The diphoton mass distribution for various jet requirements are shown in figure 6.11 while the number of jets per event are shown in figure 6.12.

6.2.1.1 t\bar{t}H leptonic tag

The leptonic categories are designed to collect both semi-leptonic and leptonic decays of $t\bar{t}$ from t\bar{t}H events. $t\bar{t} \rightarrow b\bar{b}\nu_1\nu_2$ and $t\bar{t} \rightarrow b\bar{b}\nu_1\nu'_2$ where $l$ is either a muon or electron. In order to reduce the background, the lepton is required to have $p_T > 20$. The additional requirements beyond the standard diphoton selection for the t\bar{t}H leptonic categories are listed below:

- tighter photon $p_T$ requirements: $p_T^1/m_{\gamma\gamma} > 1/2$ and $p_T^2/m_{\gamma\gamma} > 1/4$
- at least one selected lepton with $p_T > 20$
- the lepton is separated from the photons: $R(l, \gamma) > 0.35$
- for electron channel, make sure we aren’t faking signal from Z-decay $|m_{e,\gamma} - m_Z| > 5$ GeV
- at least 2 good jets: $p_T > 25$ GeV, $|\eta| < 2.4$, $R(jet, \gamma) > .4$, and $R(jet, l) > 0.4$
- at least 1 good b-tagged jet: $p_T > 25$ GeV, b-tagged with medium working point CSVv2

6.2.1.2 t\bar{t}H hadronic tag

The hadronic categories are designed to collect hadronic $t\bar{t}$ decays from t\bar{t}H events. The strategy implemented in this analysis leverages a BDT discriminator was trained
Figure 6.11. Distributions of the invariant diphoton mass from the control sample (blue) and the data sidebands (black) for different criteria: at least 2 jets (up left), at least 3 jets (up right), at least 3 jets and 1 btag (bottom left), at least 4 jets and 1 btag (bottom right).
Figure 6.12. Jet multiplicity (left) and the b-jet multiplicity (right), defined with the medium working point of the CSVv2 tagger, for events with at least two jets and at least one lepton. Data from the signal region sidebands and data from the control sample are compared to simulated $t\bar{t}H$ events. All contributions are normalized to the integral of the signal region sidebands.
to select these events using various jet based quantities as inputs.

The inputs to the $t\bar{t}H$ hadronic BDT are described below:

- **nJets**: the number of jets with $p_T > 25$ GeV
- **maxBTag**: the CSVv2 discriminator value for the most likely $b$-jet
- **secondMaxBTag**: the CSVv2 discriminator value for the second most likely $b$-jet
- **leadJetPt**: the leading jet $p_T$

After standard diphoton selection is applied (with a tighter diphoton BDT cut), the dominant background is the diphoton background. The $t\bar{t}H$ hadronic BDT was trained with $t\bar{t}H$ simulation as signal and Standard Model diphoton simulation for background. Before training, an additional selection requiring at least 3 total jets, of which at least 1 must be a loosely $b$-tagged jet. The standard scale factors from the $b$-tagging POG for Moriond 2017 were applied to the simulation before the training as well.

The distributions of the input variables can be seen in figures 6.13. A validation of the $t\bar{t}H$ hadronic BDT output is shown in figure 6.14.

In order to determine where the category boundaries should be based on the output of the $t\bar{t}H$ hadronic BDT score, the expected signal and background yields were used to calculate significance. The signal yields for all signal modes are estimated using simulation (with all final analysis weights). The background estimates come from three different sources: sidebands (mass) from data, diphoton background from simulation, and the control sample region.

The background mass distributions are modeled with an exponential function. For the diphoton simulation and control sample, a scale factor is applied to normalize the total background count to that of the data (mass) sidebands when no cut on the $t\bar{t}H$ hadronic BDT score is applied. To estimate the significance an effective $\sigma$ is
Figure 6.13. Jet multiplicity, leading jet pT, maxBTag, and secondMaxBTag for events with at least two jets and no leptons. Data from the signal region sidebands and data from the control sample are compared to simulated ttH events. All contributions are normalized to the integral of the signal region sidebands.
calculated for the mass distribution of the signal and the $2\sigma$ window is used to collect counts for signal and background used in the significance calculation:

$$s_{\tilde{t}\tilde{H}} = \frac{0.945N_{\tilde{t}\tilde{H}}}{\sqrt{2\sigma_{eff}N_{bkg}/GeV + 0.945N_{sig}}}$$ (6.5)

Where $N$ corresponds to the number of events from the different sources inside the $2\sigma$ window for $\tilde{t}\tilde{H}$, all Higgs boson production modes, and whichever background is being analyzed. There is fairly good agreement between the background distributions between the data sidebands and the control sample used for training in the $\tilde{t}\tilde{H}$ hadronic BDT output as shown in figure 6.15. The $\tilde{t}\tilde{H}$ hadronic BDT output provides good separation power between $\tilde{t}\tilde{H}$ hadronic signal and background.

The final requirements (in addition to the standard diphoton selection) for selecting the $\tilde{t}\tilde{H}$ hadronic categories are:

- tighter photon $p_T$ requirements: $p_T^{\gamma_1}/m_{\gamma\gamma} > 1/3$ and $p_T^{\gamma_2}/m_{\gamma\gamma} > 1/4$
- at least 3 good jets: $p_T > 25$ GeV, $|\eta| < 2.4$, $R(jet, \gamma) > 0.4$, and $R(jet, l) > 0.4$
Figure 6.15. Distribution of the t\bar{t}H hadronic MVA classifier for events with at least two jets and no leptons (left) and for events with at least one b-loose-jet and at least 3 jets (right). Data from the signal region sidebands and data from the control sample are compared to simulated t\bar{t}H events.

All contributions are normalized to the integral of the signal region sidebands.
- at least 1 good b-tagged jet: \( p_T > 25 \text{ GeV} \), b-tagged with loose working point CSVv2
- \( \text{ttHHadMVA} > 0.75 \) and diphoton BDT > 0.4

6.2.1.3 \( \tilde{t}\tilde{H} \) systematic uncertainties

Sources of systematic uncertainty that are specific to the \( \tilde{t}\tilde{H} \) categories are detailed in this section.

**Jet energy scale and resolution:** The systematic error on the jet energy scale is estimated by shifting the scale up or down by the full jet energy scale uncertainty as given by the CMS JetMET group. The systematic error on the jet energy resolution is estimated by varying the resolution according to the disagreement between the resolution measured in data and simulation.

**Lepton identification efficiency:** The difference in signal efficiency when shifting the data/simulation scale factor by its uncertainty is taken as the uncertainty in lepton ID efficiency. This is roughly 1%.

**b-tagging efficiency and shape:** For the leptonic categories, the b-tagging efficiency uncertainty is evaluated by shifting the b-tagging scale factors by one standard deviation of the uncertainty given by the CMS JetMET group. For the hadronic category, the shape of the b-tagging discriminant is relevant. To properly account for this, the various inputs to the b-tagging discriminant are varied by one standard deviation individually. The signal yields change by less than 5% in the hadronic channel and less than 2% in the leptonic channel.

**gluon-fusion contamination:** While the simulated signal samples for Higgs boson production from gluon-gluon fusion are well modeled for most of the relevant phase space, there are systematic uncertainties that become relevant when a large number of real jets are produced. When \( N_{\text{jets}} > 4 \) radiate from the gluon-gluon initial state (most common additional signal entering the \( \tilde{t}\tilde{H} \) categories) the following
systematics are applied:

• Limited size of simulation (statistics): 10%

• Mismodeling of parton showering: 15% for 2 jet bin (leptonic categories), 35% for 5 jet bin (hadronic category). This is taken from data and simulation discrepancy in jet multiplicity in tt+jets events.

• Mismodeling of gluon splitting: 50% on gluon yield in t\(\bar{t}\)H categories. The difference in ratio between t\(\bar{t}\)bb and t\(\bar{t}\)jj cross-sections in data and simulation is used to estimate the mismodeling of real b-jets coming from gluon-gluon fusion.
6.2.2 VH categories

When the Higgs boson is produced in association with a vector boson (V), the event can be tagged based on the final state decay products of the V boson. In VH where V=W,Z, the possible cases are: $W \rightarrow l\nu$ and $Z \rightarrow l^+l^-$ which are tagged by requiring at least one lepton, $Z \rightarrow \nu\nu$ which is tagged with the presence of high $\vec{p}_T^{\text{miss}}$, and W/Z decaying to hadrons which are tagged by the presence of jets.

6.2.2.1 MET definition

The $\vec{p}_T^{\text{miss}}$ vector is defined as the projection onto the plane perpendicular to the beam axis of the negative energy vector sum of the momenta of all reconstructed particle-flow objects in an event. Its magnitude is referred to as $E_T^{\text{miss}}$. In the case of VH decays, the $E_T^{\text{miss}}$ originates from the neutrino(s) of the final state decay from the V boson. This signal is very small with respect to gluon fusion and Standard Model backgrounds; the dataset used in this analysis consists mainly of processes involving no real $E_T^{\text{miss}}$. The majority of events with high measured $E_T^{\text{miss}}$ that pass our selection have no real $E_T^{\text{miss}}$ and the fake signal comes from detector noise, inaccurate jet energy reconstruction, and relative misalignment of sub-detectors. Further information on $E_T^{\text{miss}}$ reconstruction and performance is available here [14].

The $E_T^{\text{miss}}$ performance was first validated on $Z \rightarrow ee$ events using the standard selection for $Z \rightarrow ee$ selection described in section 4.7.1. The most updated jet energy corrections were applied and the $E_T^{\text{miss}}$ was recomputed. A modulation in $\phi$ was observed, which was mismodeled in the simulation as shown in the left plot of figure 6.16. The true $E_T^{\text{miss}}$ distribution should be flat with respect to $\phi$, but since all of the $E_T^{\text{miss}}$ in the analysis backgrounds comes from detector effects, it is very difficult to correct. The analysis itself does not use $\phi$ from $E_T^{\text{miss}}$, but rather the separation between the diphoton object and the $\vec{p}_T^{\text{miss}}$ ($\Delta \phi(\gamma\gamma, \vec{p}_T^{\text{miss}})$). For events with true $E_T^{\text{miss}}$ the $\phi$ value and thus the $\Delta \phi$ value is well reconstructed. For background events, the
\(\phi\) modulation is unimportant and has the effect of smearing the \(\Delta\phi\) by \(< 0.1\) radians. The right plot of figure 6.16 shows a much smaller discrepancy between data and simulation for this derived quantity than the \(\phi\) distribution. Ultimately, we use the \(\Delta\phi(\gamma\gamma, p_T^{\text{miss}})\) from data for background to avoid this bias. The comparison between data and simulation is shown in figure 6.17: they agree for \(E_T^{\text{miss}} < 100\) GeV.

Figure 6.16. \(\phi(p_T^{\text{miss}})\) distribution (on the left) and \(\Delta\phi(\gamma\gamma, p_T^{\text{miss}})\) distribution (on the right) for \(Z \to e e\) events in data and simulation. The \(\Delta\phi(\gamma\gamma, p_T^{\text{miss}})\) distribution is used in the analysis, and gives better performance than shown here as described later in this section.

A sample of \(Z \to \mu\mu\) was additionally used to verify that the \(\phi\) modulation did not have any effect on events with true \(p_T^{\text{miss}}\). The muons can be removed to emulate true \(E_T^{\text{miss}}\) in \(Z \to \mu\mu\) events from data and simulation. These objects were required to pass the tight muon ID and have an invariant mass between 70 and 110 GeV. Since \(Z \to \mu\mu\) events will have no true \(E_T^{\text{miss}}\), the \(p_T^{\text{miss}}\) is recalculated excluding the two selected muons. If the \(E_T^{\text{miss}}\) is recomputed well for events with true \(E_T^{\text{miss}}\), the \(p_T^{\text{miss}}\)
Figure 6.17. $E_T^{\text{miss}}$ distribution for $Z \rightarrow \text{ee}$ events in data and simulation. Even though $Z \rightarrow \text{ee}$ does not produce real $E_T^{\text{miss}}$, the agreement for fake $E_T^{\text{miss}}$ is good below 100 GeV.

will correspond to the $\slashed{p}_T$ of the dimuon system. For low $E_T^{\text{miss}}$ there is a correlation between the $\slashed{p}_T^{\text{miss}}$ and dimuon $\slashed{p}_T$, but it is overshadowed by the background noise and $\phi(\slashed{p}_T^{\text{miss}})$ modulation. At $E_T^{\text{miss}}$ above 50 GeV there is a very strong correlation between $\slashed{p}_T^{\text{miss}}$ and the dimuon object. These results are summarized in figure 6.18 and figure 6.19. Additionally, the $\phi(\slashed{p}_T^{\text{miss}})$ modulation is nearly gone in both data and simulation.

Finally, the $\phi(\slashed{p}_T^{\text{miss}})$ modulation was measured in signal simulation. For gluon-gluon fusion events with high $E_T^{\text{miss}}$ (which is fake), there is significant modulation. For $ZH$ events with real $E_T^{\text{miss}}$ the $\phi(\slashed{p}_T^{\text{miss}})$ modulation is within the statistical uncertainty of the measurement. For events with real $E_T^{\text{miss}}$, $\phi(\slashed{p}_T^{\text{miss}})$ can be used safely, especially at higher $E_T^{\text{miss}}$. 
Figure 6.18. $Z \rightarrow \mu\mu$ events have no real $E_T^{\text{miss}}$, but we were able to emulate real $E_T^{\text{miss}}$ by removing the muon tracks from the event and recalculating the $p_T^{\text{miss}}$. When the dimuon $p_T$ is high, there is large “true” $E_T^{\text{miss}}$ and the agreement between data and simulation is good. When the dimuon $p_T$ is low, the “true” $E_T^{\text{miss}}$ is low and there is poor agreement between data and simulation.

6.2.2.2 VH leptonic categories

It is possible to exploit the exclusive selection of diphoton events produced in association with at least one high $p_T$ lepton originating from the leptonic decay of vector bosons in the VH mechanism.

In the presence of the tagged lepton, QCD background is strongly suppressed. The main remaining backgrounds are electroweak processes with photons where a lepton is produced in a $Z$ or $W$ decay.

Three VH leptonic categories were designed:

- ZHLeptonic: two leptons and no requirements on $E_T^{\text{miss}}$
- WHLeptonic: one lepton and $E_T^{\text{miss}} > 45$ GeV
- VHLeptonicLoose: one lepton and $E_T^{\text{miss}} < 45$ GeV

ZHLeptonic category: Designed to capture high purity ZH events by requiring
two well-reconstructed leptons.

- two leptons of the same flavor (muons or electrons only) with $p_T > 20 \text{ GeV}$; muons are required to pass the Tight Muon ID criteria, to have relative isolation $< 0.25$, and $|\eta| < 2.4$; electrons are required to pass the cut based E-Gamma ID working point and $|\eta| < 1.4442$ or $|1.566 < |\eta| < 2.5$

- the invariant mass between the two leptons must be between 70 and 110 GeV

- additional photon requirements: $p_T^{\gamma_1}/m_{\gamma\gamma} > 0.375$ and $p_T^{\gamma_2}/m_{\gamma\gamma} > 0.25$ $\Delta R(\gamma, l) > 0.5(1.0)$ for events with two muons (electrons) electron veto for the electron tag channel $\Delta R(\gamma, GSFtrack) > 0.4$

- a loose selection on the diphoton MVA variable is applied: diphoton MVA $> 0.405$. The chosen value corresponds to the lowest boundary of the untagged categories. The statistics are quite low in this region and the purity is quite high

**WHLLeptonic category:** Designed to capture high purity WH events by requiring one well-reconstructed lepton.
• at least one muon or electron with $p_T > 20$ GeV; muons are required to pass the Tight Muon ID criteria, to have relative isolation $< 0.25$, and $|\eta| < 2.4$; electrons are required to pass the cut based E-Gamma ID working point and $|\eta| < 1.4442$ or $|1.566 < |\eta| < 2.5$

• additional photon requirements: $p_T^1/m_{\gamma\gamma} > 0.375$ and $p_T^2/m_{\gamma\gamma} > 0.25$ $\Delta R(\gamma, l) > 1.0$ electron veto for the electron tag channel $\Delta R(\gamma, GSFtrack) > 0.4$

• $E_T^{\text{miss}} > 45$ GeV

• a selection on the diphoton BDT variable is applied: diphoton MVA $> 0.0$. The value was chosen to optimize significance.

**VHLeptonicLoose category:**

• at least one muon or electron with $p_T > 20$ GeV; muons are required to pass the Tight Muon ID criteria, to have relative isolation $< 0.25$, and $|\eta| < 2.4$; electrons are required to pass the cut based E-Gamma ID working point and $|\eta| < 1.4442$ or $|1.566 < |\eta| < 2.5$

• additional photon requirements: $p_T^1/m_{\gamma\gamma} > 0.375$ and $p_T^2/m_{\gamma\gamma} > 0.25$ $\Delta R(\gamma, l) > 1.0$ electron veto for the electron tag channel $\Delta R(\gamma, GSFtrack) > 0.4$

• $E_T^{\text{miss}} > 45$ GeV

• a selection on the diphoton BDT variable is applied: diphoton MVA $> 0.0$. The value was chosen to optimize significance.

To reduce contamination from $t\bar{t}H$ events, WHLeptonic and VHLeptonicLoose categories reject events with more than 2 jets. Jets considered for this veto must satisfy the below requirements:

• jet $p_T > 20$ GeV

• jet $|\eta| < 2.4$

• $\Delta R(\text{jet, lepton}) > 0.5$

• $\Delta R(\text{jet, } \gamma) > 0.5$

Systematic uncertainties specific to the VH leptonic tags arise from the uncertainty on the lepton identification and isolation efficiencies as well as the missing transverse energy. For electrons and muons, the uncertainty on the identification is
computed varying the scale factor between data and simulation by its uncertainty. The difference in signal efficiency from simulation is roughly 1%. The $E_T^{\text{miss}}$ systematic uncertainty is accounted for by shifting the $E_T^{\text{miss}}$ by shifting the $p_T$ of each type of particle flow candidate by their associated uncertainties. This is then propagated through to the $E_T^{\text{miss}}$. This effect only causes migration between the two leptonic categories.

6.2.2.3 VH missing transverse energy category

It is possible to exploit the exclusive selection of diphoton events produced in association with large $p_T^{\text{miss}}$ originating from the leptonic decay involving neutrinos from the vector boson in the VH mechanism. This category is designed to select events where the lepton from W decay is not reconstructed due to inefficiencies or when the lepton falls outside the detector acceptance. Events in which the Z decays to two neutrinos are also selected by this category.

In the presence of the tagged $E_T^{\text{miss}}$, QCD background is strongly suppressed. The main background comes from QCD events where detector inefficiencies produce fake $p_T^{\text{miss}}$.

The VH $E_T^{\text{miss}}$ category is designed as follows:

- additional photon requirements: $p_T^{\gamma_1}/m_{\gamma\gamma} > 0.375$ and $p_T^{\gamma_2}/m_{\gamma\gamma} > 0.25$
- $E_T^{\text{miss}} > 85$ GeV
- a selection on the diphoton BDT variable is applied: diphoton MVA > 0.6. The value was chosen to optimize significance.
- $\Delta\phi(\gamma\gamma, p_T^{\text{miss}}) > 2.4$

The agreement between data and simulation was shown to be reasonable for $E_T^{\text{miss}}$, but the statistics are too low in the high $E_T^{\text{miss}}$ region to optimize background using simulated events. An iterative tuning the selection one variable at a time was performed using a control sample in data.
The control region was defined and compared with the signal sidebands (100 < m_{\gamma\gamma} < 115 || 135 < m_{\gamma\gamma} < 180). An exponential function was used to estimate the background at 125 GeV. The same strategy was adopted for the control region. The control region was then normalized to the sideband and the used to estimate the expected background for the signal region during tuning of the selection criteria.

The signal yield was measured by applying the full VH E_{T}^{miss} category selection, but varying the requirement on the variable of interest. A Gaussian signal model was used to fit the mass distribution and the 1\sigma width was extracted. For this simplified significance optimization, the other signal modes were taken as background. The final background was estimated by counting the number of background events contained within 1\sigma of the signal mass fit.

The significance was given as:

\[ \sqrt{2((signal + background) \ln(1 + signal/background) - signal)} \]  \hspace{1cm} (6.6)

The first optimization performed was on the diphoton BDT output by optimizing the above criterion as described and the results are shown in figure 6.20. The control region for this selection used an inversion of a much looser \( \Delta\phi(\gamma\gamma, p_{T}^{miss}) < 2.0 \). This was validated to have no effect on the diphoton BDT distribution using data sidebands and thus safe to use for optimization.

To optimize the E_{T}^{miss} requirement, the diphoton BDT was inverted (0.0 < BDT < 0.6) as a control region. Unfortunately, this BDT requirement does bias the E_{T}^{miss} distribution and the statistics get too low for reliable background fits at E_{T}^{miss} > 100 GeV. Because of these limitations, E_{T}^{miss} > 85 GeV was chosen for the tag requirement even though the optimal cut found was somewhat higher as shown in figure 6.21. Overall, this gives slightly worse significance in the VH E_{T}^{miss} category, but was selected as the safest option.

To further reduce background contamination, one can reject background events
Figure 6.20. The left plot shows the signal diphoton BDT distribution, signal sideband, and control region in data (adding ggH, VBF, and t\(\bar{t}\)H signals) as background (\(E_{T}^{\text{miss}} > 70, \Delta\phi(\gamma\gamma, \mathbf{p}_{T}^{\text{miss}}) > 2.0\)). The right plot shows the significance while tuning the diphoton BDT selection.

using event topology. The direction of the diphoton and \(E_{T}^{\text{miss}}\) will be nearly back-to-back in VH events. In the absence of additional jets produced via radiation, the Higgs boson and vector bosons are balanced in the transverse plane. The selection for this category was optimized so that the difference in \(\Delta\phi(\gamma\gamma, \mathbf{p}_{T}^{\text{miss}})\) was greater than 2.4. The results of this optimization can be seen in figure 6.22. This criterion reduces contamination from background events and events produced by gluon-gluon fusion. The same inversion of the diphoton BDT was used as control region for this optimization.

The systematic uncertainty coming from the \(E_{T}^{\text{miss}}\) measurement is accounted for by shifting the \(E_{T}^{\text{miss}}\) by the systematic shift in reconstructed \(p_{T}\) of each type of particle flow candidate and propagating this through the \(E_{T}^{\text{miss}}\) calculation. This causes some event migration between the VH \(E_{T}^{\text{miss}}\) category and the untagged categories. Because the direction of the systematic shift in input collections is largely uncorrelated with the \(E_{T}^{\text{miss}}\) direction, the overall change in expected yields is very small.
6.2.2.4 VH hadronic category

It is possible to exploit the exclusive selection of diphoton events produced by a Higgs boson decay in association with a hadronically decaying vector boson. This category is designed to fully reconstruct the final state and requires the presence of two photon and two jets.

In the presence of two jets the background is greatly reduced, consisting mainly of diphoton QCD events where two hard jets arise from pileup, radiation or underlying event activity.

The VH hadronic category is selected as follows:

- additional photon requirements: \( p_T^{\gamma_1}/m_{\gamma\gamma} > 0.5 \) and \( p_T^{\gamma_2}/m_{\gamma\gamma} > 0.25 \)
- at least two jets with \( p_T > 40 \text{ GeV} \) && \( \eta < 2.4 \)
- dijet mass \( 60 < m_{jj} < 120 \text{ GeV} \)
- \( |\cos(\theta^*)| < 0.5 \)
- diphoton transverse momentum \( p_T(\gamma\gamma) < m_{\gamma\gamma} \text{ GeV} \)
Figure 6.22. The left plot shows the signal $\Delta \phi(\gamma\gamma, p_T^{\text{miss}})$ distribution, signal sideband, and control region in data (adding ggH, VBF, and ttH signals) as background ($E_T^{\text{miss}} > 70$, diphoton BDT $> 0.6$). The right plot shows the significance while tuning the $\Delta \phi(\gamma\gamma, p_T^{\text{miss}})$ selection.

The mass of the vector boson and Higgs boson set a higher energy scale (and thus diphoton $p_T$) than the SM diphoton background. Additionally, having all final state particles allows for fine kinematic requirements to reject more background events. The two jets from the decay of the vector boson have an invariant mass resonance at the mass of the vector boson, this isn’t true for the background events and can be used to reject background.

Finally, leveraging the angle between the decay products of the vector boson can allow further background rejection. The VH production mechanism arises from $V^* \rightarrow VH$. The direction of the diphoton and dijet system will be back-to-back in the rest frame. The same is not true of the backgrounds, where the photons and jets are typically from unrelated processes. The variable $\theta^*$ is defined as the angle between the direction of $V^*$ in the laboratory frame and its decay products in the $V^*$ rest frame. The distribution of this variable should be flat for two-body decays from signal, but due to the high boost of the $V^*$ frame, the background events tend
toward $\theta^* = 0$. The separation power of these variables can be seen in figure 6.23.

Figure 6.23. Dijet invariant mass (left) and $\cos \theta^*$ (right) distributions. The expected shape of VH events is compared to the shape of gluon fusion Higgs boson production and the major background: diphoton continuous production. All distributions are normalized to unity.

To optimize the selection given above, the precision on the measurement of the Higgs boson coupling to vector bosons was used. Signal events were estimated from simulation and background was estimated using a control sample orthogonal to the signal sample. The control sample was given by requiring one of the two photons in the diphoton system to have a single photon ID MVA value $< -0.9$. The sidebands of the diphoton invariant mass distribution (excluding $115 < m_{\gamma\gamma} < 135$) were used to normalize the control sample. The mass distribution of the normalized control sample is then used to estimate the background contained in the mass window $[125,130] \text{ GeV}$ by fitting an exponential to the smoothly falling background. A cut
and count experiment is then performed for each set of selections and the signal and background yields are used to calculate the precision on the Higgs boson coupling. 6.24. The optimization of the selection optimization can be seen in figure 6.25.

![Diphoton invariant mass distribution shape comparison between control sample and signal sample.](image)

Figure 6.24. Diphoton invariant mass distribution shape comparison between control sample and signal sample.

In order to optimize the diphoton BDT criterion, a different control sample is necessary because the diphoton BDT is strongly dependent on the single photon MVA output. For the control sample, the requirement on $|\cos(\theta^*)|$ was inverted. Otherwise, the same optimization procedure as above was performed using the final selection criteria for this channel to obtain the diphoton BDT criterion for this channel.
Figure 6.25. Results of the selection optimization based on the expected precision on the Higgs boson coupling ($\sigma_{RV}$) as a function of $p_T(\gamma\gamma)/m_{\gamma\gamma}$ (left) and $|\cos\theta^*|$ (right). The red lines indicate the cut used in Run 1 analysis.

The systematic uncertainties specific to the VH hadronic category are related to the uncertainty on the jet energy scale and jet energy resolution. The systematic uncertainty on the jet energy scale is estimated by shifting the scale by $1\sigma$. The systematic uncertainty on jet energy resolution is estimated by varying the resolution according to the level of disagreement seen between data and simulation.
6.2.3 VBF categories

If a diphoton event is found to have two forward jets, it is likely to be from vector boson fusion (VBF). Additional jets are often present, but typically the two highest \( p_T \) correspond to the scattered quarks from VBF and are used for these categories. While the cross-section for this production mode is roughly 10 times smaller than that of ggH, the presence of two forward jets allow selection of events with a very high signal-to-background ratio. In the most performant VBF category, 77\% of events are estimated to be true VBF events. The performance of these categories is shown in table 7.1.

The requirements for the VBF categories are designed to identify jets from VBF Higgs boson production and ignore real jets from other signal processes as well as Standard Model backgrounds. They also aim to reject fake jets arising when several low energy hits from pileup are clustered together. The additional preselection for VBF categories is listed below:

- tighter photon \( p_T \) requirements: \( p_T^{j1}/m_{\gamma\gamma} > 1/3 \) and \( p_T^{j2}/m_{\gamma\gamma} > 1/4 \)
- two jets passing the pileup jet ID described in section 6.2.3.1 and \( p_T^{j1} > 40 \) and \( p_T^{j2} > 30 \)
- \( m_{jj} > 250 \) GeV
- both photons have photon ID MVA > -0.2

Since the diphoton BDT is not trained exclusively using VBF as signal, there is additional contamination of prompt-fake photon events and can be seen in figure 6.26. In VBF events, diphoton pairs can receive a high diphoton MVA score, even if the single photons are assigned low MVA scores. Applying a tighter single photons ID requirement improves performance significantly; the exact value was chosen to maximize significance when combining all three VBF categories.

The tighter photon \( p_T \) cuts are applied to improve agreement between data and simulation. Lowering the \( p_T \) cuts by a total of 10 GeV would increase the combined
Figure 6.26. Distributions of leading and subleading photon ID BDT score after VBF preselection without ID cuts. We note a population of prompt-fake background events in the low ID score region. Data are taken from sideband regions 10 GeV either side of the signal region.
expected significance by 5%, but brings in much larger systematic uncertainties.

A special dijet BDT was trained to reject ggH events while selecting VBF events. Following that, the dijet BDT is combined with the standard diphoton BDT to reject background events. The dijet BDT leverages kinematic variables from the dijets to separate real VBF events from ggH and irreducible backgrounds. The combined MVA then takes the score from the dijet BDT, the diphoton BDT, and the diphoton \( p_T \) divided by diphoton mass. Three categories are defined by optimizing boundaries in the combined MVA output to maximize overall VBF significance.

6.2.3.1 Jet definition

Jets in this analysis are reconstructed by applying the anti-\( k_T \) algorithm [3] to all particle candidates (excluding charged particles not reconstructed from the selected vertex). In order to be included in the analysis, all jets must have \( |\eta| < 4.7 \).

Particles produced from pileup interactions can include in the jet clustering algorithm, appearing as very high \( p_T \) jets. A special algorithm was designed to remove these pileup jets from the jet collection used in this analysis. Pileup-Jet Identification (PUJID) uses the jet shape and compatibility of the jet’s tracks with the primary vertex to identify reconstructed jets which are likely from pileup interactions [15]. Most jets arising from pileup occur at \( |\eta| > 2.5 \), where the jet constituents cannot be associated to tracks.

There is also a Jet ID discriminant trained by the CMS JetMET group to help reject misreconstructed jets. The loose working point for this variable is used for all jets in this analysis, except for the jets in the VBF category, where the tight working point is used.
6.2.3.2 Dijet BDT

The dijet BDT is optimized to select events with two real jets as are typical of VBF events. The inputs to the BDT are:

- transverse momenta of each photon in the diphoton object divided by their invariant mass: $p_{\gamma 1}^T/m_{\gamma\gamma}$ and $p_{\gamma 2}^T/m_{\gamma\gamma}$
- transverse momenta of the 2 highest $p_T$ jets: $p_{j1}^T$ and $p_{j2}^T$
- invariant mass of the dijet: $m_{j1j2}$
- difference in pseudo-rapidity of the two jets: $\Delta \eta_{j1j2}$
- difference in azimuthal angle between dijet and diphoton: $\Delta \phi_{(j1j2,\gamma\gamma)}$
- centrality variable:
  
  $$C_{\gamma\gamma} = \exp\left(-\frac{4}{(\eta_1 - \eta_2)^2(\eta_{\gamma\gamma} - \frac{\eta_1 + \eta_2}{2})^2}\right)$$

  where $\eta_1$, $\eta_2$, and $\eta_{\gamma\gamma}$ are the correspond respectively to the lead jet, sublead jet, and diphoton.

- difference in azimuthal angle between two leading jets: $\Delta \phi_{jj}$
- minimum distance between leading or subleading jet and leading or subleading photon: $\min \Delta R(\gamma, jet)$

The kinematic variables described above are combined in a boosted decision tree using gradient boosting implemented using scikit-learn [32]. The BDT is trained on simulated events with VBF Higgs boson decaying into two photons with mass 125 GeV used as the signal. The full suite of background sample used in the analysis (excluding Drell-Yan) are used for background training. The Higgs boson to two photon from gluon-gluon fusion signal sample is used as an additional background to improve the purity of the dijet BDT.

6.2.3.3 Combined MVA and category optimization

The Combined Dijet MVA is also a gradient-boosted decision tree which took as input the Dijet BDT score, the Diphoton BDT score, and the $p_{T}^{\gamma}/m_{\gamma\gamma}$. This MVA
Figure 6.27. Kinematic dijet BDT input variable distributions. These plots show normalized distributions of the dijet BDT input variables and demonstrate the separation between signal and background with ggH also shown. These plots are produced with the loose VBF preselection used for the training described.
Figure 6.28. Combined Dijet MVA score distribution before and after photon ID MVA score cuts. Data are taken from sideband regions 10 GeV either side of the signal region. We note that the population of prompt-fake background is much reduced. ID cut plot is normalized only with the events of weight less than one to better show the agreement.
was optimized to separate the dijet signal from background processes. The Dijet BDT is trained using VBF $H \rightarrow \gamma\gamma$ with $m_H = 125$ GeV as signal and all standard backgrounds for the analysis. The gluon fusion is not used in training to avoid rejecting real signal diphotons (those with a high diphoton BDT score). Fortunately, the Combined Dijet MVA gives reasonable discrimination between VBF and gluon-fusion Higgs boson production. The distributions and separation power of the inputs to the VBF Combined Dijet MVA are shown in figure 6.29.

![Figure 6.29](image)

Figure 6.29. Input variables of the VBF Combined Dijet MVA. From the left to the right, diphoton MVA, kinematic dijet BDT output score, and the diphoton $p_T(\gamma\gamma)/m_{\gamma\gamma}$. These plots are produced with the loose VBF preselection.

Three VBF categories are chosen simultaneously in much the same way the untagged categories were defined. Three categories with boundaries based on the Combined Dijet MVA score. For $N$ categories, the boundaries are allowed to float and the significance by combining all $N$ categories is calculated. The highest overall significance is then chosen. A double Gaussian model is fit to the mass distribution
of the signal and the peak position and width are measured. The background mass distribution is fit with an exponential. The significance within $2 \sigma$ of the signal peak is then calculated.

Three categories were chosen as it was shown that there was little gain in significance with 4 or more categories. The purity and efficiency of these categories can be seen in table 7.1.

6.2.3.4 Validation of MVAs using $Z \rightarrow ee$ events

The Dijet BDT and Combined Dijet MVAs were validated on $Z \rightarrow ee$ events that contained two jets. The standard $Z \rightarrow ee$ selection for this analysis as described in section 4.7.1 was applied with the additional requirement of $p_T(j_1) > 40, p_T(j_2) > 30, m_{jj} > 250$ GeV, and $p_T(\gamma_1) > 40$.

The output of the MVAs are compared between data and simulation as well as the input variables. While the jet rate disagrees by more than 30% at high $|\eta|$, it is within the band given by the jet systematic uncertainties.

Finally, the tight Jet ID working point was determined optimal for including jets in this analysis. This variable was designed to reject incorrectly reconstructed jets. While the tighter working point has minimal effect on the overall significance of the VBF categories, it improves agreement between data and simulation considerably by rejecting more pileup jets. This is illustrated by figure 6.30.

6.2.3.5 Systematic uncertainties

The systematics uncertainties specific to the VBF topology arise solely from jets and are described below:

- Jet energy corrections are applied to data and simulation. The uncertainties on these corrections are applied to simulation and used to derive the magnitude of category migration between the three VBF tags and the untagged categories.
Figure 6.30. Data-simulation agreement in $\eta$ for lead (sublead) jet is shown on top (bottom). From left to right, the jets have no PUJID, loose, medium and tight PUJID requirement. The component of DY corresponding to both jets matching a true jet (not pileup) is shown, along with the components for the lead jet and sublead jet only being matched, and the component for both jets being matched. The tight PUJID requirement is necessary to improve agreement.
• Jet energy resolution corrections are applied to simulation using the recommended scale factors. The results of these shifts cause shifts between categories as in the jet energy corrections.

• The pileup jet ID distribution in data is compared to that in simulation for $Z \rightarrow ee$ events with 1 jet and calculating the shift needed to match simulation to data. These shifts are binned in $\eta$ and $p_T$ of the jet. The shifts are typically larger for low $p_T$ high $\eta$ jets. The results of these shifts appear as category migrations as well.

Theoretical uncertainty on the jet kinematics are calculated following the Stewart-Tackmann procedure [36].
CHAPTER 7
STATISTICAL ANALYSIS

Events are classified into 14 analysis categories which loosely correspond to the Higgs boson production modes as described in chapter 6. Simulated signal samples corresponding to the SM Higgs boson production processes (t\bar{t}H, WH, ZH, VBF, and ggH) are used to produce a signal model. The signal samples used in this analysis can be found in section 5.1. The analysis selection as described throughout this document is applied for simulations at 7 different mass values of the Higgs boson (120, 123, 124, 125, 126, 127, 130 GeV). Using these 7 different simulated signal samples, we produce parametric signal models for each category as described in section 7.1.

The background model is extracted using data recorded by CMS, not simulation. The data is run through the entire analysis as previously described in this paper. The background is then parameterized for each analysis category using a discrete profiling method described in section 7.2.

A description of the systematic effects on the analysis are described in section 7.3 along with an explanation of how they are propagated to the final result. The final measurements are presented in section 7.4 where a signal is extracted by performing a simultaneous likelihood fit to the diphoton invariant mass distributions for each of the analysis categories.

7.1 Signal modeling

An accurate signal model that can fit an arbitrary Higgs boson mass needs to be built. This signal model must account for the efficiency×acceptance as a function
of \( m_H \) as well as the shape of the observed mass distribution in each of the analysis categories. A parameterized model of the signal shape based on the Higgs boson mass is derived for each analysis category between 120 and 130 GeV. The systematic effects listed in section 7.3 that smear and shift the mass distribution are all applied before computation of the models. Each category is fit separately using a double Crystal Ball (DCB)\(^1\) + single Gaussian (Gaus) functional form. A simultaneous fit of the seven mass points is performed, assuming each parameter of the DCB+Gaus changes linearly with the mass of the Higgs boson. This limits the number of free parameters to fit each mass point and interpolate the signal shapes for an arbitrary Higgs boson mass.

This signal modeling is performed separately for each Higgs boson production process, analysis category, and RV or WV case. Additional signal models were constructed for bbH, tHq, and tHW signal production processes, but only mass 125 GeV was available from simulation, and the parametric model for these signals is flat with respect to \( m_H \). To determine signal model from the SM in a given analysis category the shapes are added together based on the expected efficiency in a given category. For instance, the final signal shape for the VH \( E_T^{\text{miss}} \) category will consist of mainly VH Higgs boson production samples, but have additional nearly 25\% from gluon fusion.

The correct identification of the vertex as described in section 5.6 has a large effect on the width of the Higgs boson mass peak. The final signal shape for a given analysis category also combines the signal shapes for the right and wrong vertex interpretations based on the measured vertex selection efficiency from simulation. The vertex efficiency is also interpolated between mass points when producing the parametric model for each category and process. Figures 7.1-7.3 show the output of

\(^{1}\) The Crystal Ball function was developed for the Crystal Ball experiment and consists of a Gaussian core with a power law to account for a low mass tail [34].
the parametric models after the final combination of signal shapes has been applied. Figure 7.4 (left) shows the output from the parametric fit for one of the untagged categories assuming a Higgs boson mass of 125 GeV as a demonstrative example. Figure 7.4 (right) shows the efficiency times acceptance of the full event selection as a function of the Higgs boson mass.

Table 7.1 summarizes the fits signal fits for each signal category and provides the final expected signal yields and purity. Figure 7.6 displays the same information graphically.

Systematic uncertainties corresponding to the smearing and scale of the individual photon energies, the fraction of events where the correct vertex was identified, and the material corrections to ECAL crystal light yields are incorporated as nuisance parameters into the signal models.
Figure 7.1. Parametric signal models constructed from simulated signal events in each of the inclusive categories. Different production mechanisms have been summed according to their relative SM cross sections and selection efficiency.
Figure 7.2. Parametric signal models constructed from simulated signal events in each of the VBF and t\bar{t}H categories. Different production mechanisms have been summed according to their relative SM cross sections and selection efficiency.
Figure 7.3. Parametric signal models constructed from simulated signal events in each of the VH categories. Different production mechanisms have been summed according to their relative SM cross sections and selection efficiency.
Figure 7.4. Full parameterized signal shape (left) for the untagged 1 analysis category and the $m_H=125$ GeV at $\sqrt{s} = 13$ TeV. The $\sigma_{\text{eff}}$ (half width of narrowest interval containing 68.3% of the invariant mass distribution) is shown along with the FWHM. Efficiency $\times$ acceptance (right) of the signal model as a function of $m_H$ for all categories combined.
7.2 Background modeling

The background model is produced by fitting several analytic functions to the $m_{\gamma\gamma}$ distribution from data between 100 and 180 GeV. The data are classified into the analysis categories to allow a unique background fit for each analysis category.

The shape of the background distribution is not known and may be further shaped by detector effects, this means that a theoretically driven background shape will not perform well. However, if the wrong background shape is used, a bias will be introduced to the analysis. In order to counteract that, we test a host of candidate functions.

The selection of a particular analytic function when fitting the background distribution is a source of systematic uncertainty and steps have been taken to minimize and properly calculate it. We treat the choice of function as a discrete parameter in the likelihood fit used to produce the final results. The resulting systematic uncertainty is then calculated by varying over this parameter as is done for the other systematics in the analysis. In order for this procedure to work properly, one must have a complete set of candidate functions.

- sums of exponentials:
  \[ f_N(x) = \sum_{i=1}^{N} p_{2i} e^{p_{2i+1}x} \] (7.1)

- sums of polynomials (in the Bernstein basis):
  \[ f_N(x) = \sum_{i=0}^{N} p_i b_{i,N}, \text{where } b_{i,N} := \binom{N}{i} x^i (1-x)^{N-i} \] (7.2)

- Laurent series:
  \[ f_N(x) = \sum_{i=1}^{N} p_i x^{-4+\sum_{j=1}^{i-1} (-1)^j (j-1)} \] (7.3)

- sums of power-law functions:
  \[ f_N(x) = \sum_{i=1}^{N} p_{2i} x^{-p_{2i+1}} \] (7.4)

Each of the above functions is fit to the background mass distribution and the
value of twice the negative logarithm of the likelihood (2NLL) is minimized. A penalty based on the number of floating parameters of the candidate function is applied to 2NLL. For each functional form listed above, the lowest order function is tested first. An F-test [24] is applied to determine if the next order polynomial improves the goodness of fit by a significant amount. If so, the next order version of the function is used instead.

This method determines the envelope of the lowest values of 2NLL profiled as a function of the parameter of interest. This envelope provides a broader curve than the 2NLL would have for a single functional form. Examples of the background fits from this method are shown in figure 7.5 and is discussed in more detail in [20].

The expected background around the signal peak and the signal to background ratio for each category can be seen in table 7.1. And a graphical version of the signal to background ratio can be seen in figure 7.6.
Figure 7.5. Set of functions chosen to fit the background of untagged category 1. An F-test selects the representative order for a given functional family. The choice of function is treated as a discrete nuisance parameter in the final 2NLL minimization; taking the minimum envelope naturally provides a systematic uncertainty on the choice of background function.
<table>
<thead>
<tr>
<th>Event Categories</th>
<th>SM 125GeV Higgs boson expected signal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Untagged Tag 0</td>
<td>45.33</td>
</tr>
<tr>
<td>Untagged Tag 1</td>
<td>474.27</td>
</tr>
<tr>
<td>Untagged Tag 2</td>
<td>661.93</td>
</tr>
<tr>
<td>Untagged Tag 3</td>
<td>603.12</td>
</tr>
<tr>
<td>VBF Tag 0</td>
<td>9.95</td>
</tr>
<tr>
<td>VBF Tag 1</td>
<td>8.56</td>
</tr>
<tr>
<td>VBF Tag 2</td>
<td>27.38</td>
</tr>
<tr>
<td>TTH Hadronic Tag</td>
<td>5.37</td>
</tr>
<tr>
<td>TTH Leptonic Tag</td>
<td>3.51</td>
</tr>
<tr>
<td>ZH Leptonic Tag</td>
<td>0.49</td>
</tr>
<tr>
<td>WH Leptonic Tag</td>
<td>3.47</td>
</tr>
<tr>
<td>VH LeptonicLoose Tag</td>
<td>2.68</td>
</tr>
<tr>
<td>VH Hadronic Tag</td>
<td>9.49</td>
</tr>
<tr>
<td>VH Met Tag</td>
<td>4.10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1859.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event Categories</th>
<th>signal width</th>
<th>Bkg</th>
<th>Bkg</th>
<th>naive expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_{II}$</td>
<td>$\sigma_{HM}$</td>
<td>(GeV$^{-1}$)</td>
<td>(GeV$^{-1}$ fb$^{-1}$)</td>
</tr>
<tr>
<td>Untagged Tag 0</td>
<td>1.32</td>
<td>1.24</td>
<td>21.92</td>
<td>0.61</td>
</tr>
<tr>
<td>Untagged Tag 1</td>
<td>1.47</td>
<td>1.32</td>
<td>924.21</td>
<td>25.74</td>
</tr>
<tr>
<td>Untagged Tag 2</td>
<td>1.94</td>
<td>1.68</td>
<td>2419.53</td>
<td>67.40</td>
</tr>
<tr>
<td>Untagged Tag 3</td>
<td>2.62</td>
<td>2.28</td>
<td>4855.01</td>
<td>135.24</td>
</tr>
<tr>
<td>VBF Tag 0</td>
<td>1.51</td>
<td>1.30</td>
<td>1.60</td>
<td>0.04</td>
</tr>
<tr>
<td>VBF Tag 1</td>
<td>1.66</td>
<td>1.38</td>
<td>3.25</td>
<td>0.09</td>
</tr>
<tr>
<td>VBF Tag 2</td>
<td>1.61</td>
<td>1.36</td>
<td>18.89</td>
<td>0.53</td>
</tr>
<tr>
<td>TTH Hadronic Tag</td>
<td>1.48</td>
<td>1.30</td>
<td>2.40</td>
<td>0.07</td>
</tr>
<tr>
<td>TTH Leptonic Tag</td>
<td>1.61</td>
<td>1.36</td>
<td>1.50</td>
<td>0.04</td>
</tr>
<tr>
<td>ZH Leptonic Tag</td>
<td>1.65</td>
<td>1.43</td>
<td>0.12</td>
<td>0.00</td>
</tr>
<tr>
<td>WH Leptonic Tag</td>
<td>1.64</td>
<td>1.43</td>
<td>2.09</td>
<td>0.06</td>
</tr>
<tr>
<td>VH LeptonicLoose Tag</td>
<td>1.67</td>
<td>1.56</td>
<td>3.50</td>
<td>0.10</td>
</tr>
<tr>
<td>VH Hadronic Tag</td>
<td>1.39</td>
<td>1.35</td>
<td>7.22</td>
<td>0.20</td>
</tr>
<tr>
<td>VH Met Tag</td>
<td>1.43</td>
<td>1.35</td>
<td>3.49</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1.96</td>
<td>1.62</td>
<td>8264.74</td>
<td>230.22</td>
</tr>
</tbody>
</table>
Figure 7.6. The expected number of signal events per category and the breakdown per production mode. $\sigma_{\text{eff}}$ is also provided as an estimate of the $m_{\gamma\gamma}$ resolution. The ratio of the number of signal events (S) to the number of signal plus background events (S+B) is shown (right).

7.3 Summary of systematic uncertainties

The majority of systematic uncertainties affect the signal model, since the background model is derived using a data driven technique. The signal systematic uncertainties affect the analysis by modifying the shape of the $m_{\gamma\gamma}$, changing the magnitude of the $m_{\gamma\gamma}$ distribution, or altering an input to the classification MVAs (and the yield in a given category). Those which affect the shape of $m_{\gamma\gamma}$ are applied directly to the signal model as a parametric nuisance parameter. Those which don’t affect the shape of $m_{\gamma\gamma}$ are treated as log-normal uncertainties. Those that systematically alter inputs to the classification MVAs provide a correlated log-normal uncertainty on category
yield (shift events from one analysis category to another).

- **Theory systematics**
  
  - *parton distribution functions (PDF) uncertainties*: the uncertainty from the choice of PDF is calculated by estimating the yield variation in each process and category when reweighting the input samples by the PDF4LHC15 combined PDF set and NNPDF30 [23] using the MC2hessian procedure [4]. The effect on the total yield is handled separately from the shifting of events to different categories. Category migrations are below 1%. The overall normalization variation is taken from [21].
  
  - *$\alpha_s$ uncertainty*: the uncertainty on the value of the strong force coupling constant $\alpha_s$ is evaluated following the PDF4LHC guidelines [2]. The effect on the total yield is handled separately from the shifting of events to different categories. The overall variation in the relative event yield due to the $\alpha_s$ uncertainty has been computed to be 2.6%.
  
  - *QCD scale uncertainty*: related to varying the renormalization and factorization scales. The effect on the total yield is handled separately from the shifting of events to different categories. For the total yields, separate scalings are implemented for each production mechanism (ggH, VBF, t\bar{t}H and VH). Three category migrations arise from three scenarios with up/down variations of the renormalization and factorization scales.
  
  - *Underlying event and parton shower uncertainty, corresponding to the choice and tuning of the generator*: The two variations are treated separately, with dedicated ggH and VBF samples for each. The shift in yields to each VBF and Untagged category is included with correlations.
  
  - *Uncertainty on the $H \to \gamma\gamma$ branching fraction*: estimated to be about 2% [21].
  
  - *Gluon fusion contamination in VBF and t\bar{t}H tagged categories*: the theoretical predictions for gluon fusion are not well estimated when the Higgs boson is produced in association with a large number of jets. The uncertainty on the yield of gluon fusion events in the VBF tagged classes has been estimated using the Stewart-Tackmann procedure as given by LHC Higgs Cross Section Working Group [36]. The overall normalization has been found to vary by 29% while migrations between the VBF categories are small. The systematic uncertainty on the gluon fusion contamination in the t\bar{t}H tagged classes have been estimated taking into account several contributions:
    
    * uncertainty due to the limited size of the simulated sample 10%
    * uncertainty from the parton shower modeling. This uncertainty is estimated as the observed difference in jet multiplicity between simulation predictions and data in $t\bar{t}$+jets events (which are dominated by
gluon fusion production \( gg \to t\bar{t} \), with fully leptonic \( t\bar{t} \) decays. This uncertainty is about 15% for the 2 jet bin corresponding to the leptonic category selections and about 35% for the 5 jet bin that is taken as a conservative estimate for the hadronic category.\[12\]

* uncertainty on the gluon splitting modeling. It is estimated by scaling the fraction of events from gluon fusion with real b-jets by the observed difference between data and simulation in the ratio \( \sigma(t\bar{b}b)/\sigma(t\bar{t}jj) \) at 13 TeV. This uncertainty constitutes a variation of about 50% in the yield of gluon fusion events.

• Integrated luminosity: estimated from data, provides 2.5% uncertainty on signal yield.

• Trigger efficiency: measured from \( Z \rightarrow ee \) events using the tag and probe technique. The size and uncertainty varies between the photon categories, which are binned in \( R_9, \eta, \) and \( p_T \). Variations are typically well below 1% and provide event yield variations of at most 0.1%.

• Photon preselection: taken as the uncertainty on the ratio between efficiency measured in data and in simulation. Ranges from 0.2% to 0.5% depending on the analysis category

• Vertex finding efficiency: the main contribution to this uncertainty arises from the modeling of the underlying event. The uncertainty on the measurement of ratio between data and simulation is second, followed by the uncertainty on the Higgs boson \( p_T \). This is treated as an additional nuisance parameter built into the signal model allowing the fraction of event with the right/wrong vertex scenario to change. The uncertainty on this parameter is 2%.

• Energy scale and resolution: the difference between the interactions of electrons and photon with material upstream of ECAL is the dominant factor. By changing the \( R_9 \) distribution, the electron selection, and the regression training the systematic effect from using electrons to perform the energy scale corrections for photons can be estimated. The uncertainties are added as dedicated nuisance parameters to the signal models and amount to less than 0.5%.

• Photon identification BDT score: the uncertainties on the signal yields in different analysis categories are estimated conservatively by propagating the discrepancies between data and simulation as systematic uncertainties.

• Per photon energy resolution estimator: parameterized as a rescaling of the resolution estimate by \( \pm 5\% \) about the nominal value.

• Jet energy scale and smearing corrections: implemented as migration within VBF categories, within \( tttH \) categories and from tagged to untagged categories. Jet energy scale corrections (JEC) account for a 8 to 18% migration within VBF categories and 11% from VBF to untagged categories. Migration due to energy
scale in H categories is about 5%. The jet energy resolution (JER) has an impact on the event migration of less than 3%.

- **b-tagging efficiency**: evaluated by varying the ratio between the measured b-tagging efficiency in data and simulation within their uncertainty. This causes a shift of 2% events in the lepton category and 5% in the hadronic category into the untagged categories.

- **Lepton identification**: computed by varying the ratio of the efficiency measured in data and simulation by its uncertainty. This causes a shift of less than 1% of events in the t\bar{t}H categories into the untagged categories.

- **Background modeling**: The method of selecting background functions is described in section 7.2 and the uncertainty is accounted for there.

- **Tracker material**: Electron and photon showers in ECAL have differing longitudinal maxima for the same energy, therefore any difference between the simulation of tracker material and the real detector will affect photon and electrons differently. To account for this, simulation of Drell-Yan and ggH samples with ±10% and ±5% the tracker material were tested. And the uncertainty on the energy scale was measured. A summary can be seen in table 7.2.

### TABLE 7.2

<table>
<thead>
<tr>
<th>Photon category</th>
<th>E-scale Uncertainty (×10^{-2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central EB (η &lt; 1.0) R_9 &lt; 0.85</td>
<td>0.035</td>
</tr>
<tr>
<td>Central EB (η &lt; 1.0) R_9 &gt; 0.85</td>
<td>0.033</td>
</tr>
<tr>
<td>Outer EB (η &gt; 1.0) R_9 &lt; 0.85</td>
<td>0.23</td>
</tr>
<tr>
<td>Outer EB (η &gt; 1.0) R_9 &gt; 0.85</td>
<td>0.11</td>
</tr>
<tr>
<td>EE R_9 &lt; 0.85</td>
<td>0.22</td>
</tr>
<tr>
<td>EE R_9 &gt; 0.85</td>
<td>0.34</td>
</tr>
</tbody>
</table>

- **Light collection non-uniformity**: must be quantified because electrons penetrate the ECAL crystals an average of one radiation length less than photons. We propagate the measurements from run1 forward with a slight increase due to
the loss in light yield during the recent run period (increase from 0.06% to 0.07%).

- **Shower shape modeling**: Running the photon energy regression with modified inputs the ratio of the relative response variation for electrons and photons was computed. This difference is taken as the uncertainty on the shower shape modeling.

A summary of all the systematic uncertainties and their effects on the analysis is given in table 7.3.
**TABLE 7.3**

SYSTEMATIC CONTRIBUTIONS

<table>
<thead>
<tr>
<th>Systematic</th>
<th>$\mu$</th>
<th>$\mu_{ggH}$</th>
<th>$\mu_{qqH}$</th>
<th>$\mu_{VH}$</th>
<th>$\mu_{ttH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photon identification</td>
<td>3.40%</td>
<td>4.59%</td>
<td>5.39%</td>
<td>2.05%</td>
<td>2.74%</td>
</tr>
<tr>
<td>Per photon energy resolution estimate</td>
<td>2.68%</td>
<td>4.20%</td>
<td>4.85%</td>
<td>1.47%</td>
<td>0.80%</td>
</tr>
<tr>
<td>Trigger efficiency</td>
<td>0.25%</td>
<td>0.04%</td>
<td>0.11%</td>
<td>0.12%</td>
<td>0.10%</td>
</tr>
<tr>
<td>Diphoton preselection</td>
<td>0.53%</td>
<td>0.23%</td>
<td>0.24%</td>
<td>0.40%</td>
<td>0.28%</td>
</tr>
<tr>
<td>Electron veto</td>
<td>0.46%</td>
<td>0.12%</td>
<td>0.21%</td>
<td>0.23%</td>
<td>0.22%</td>
</tr>
<tr>
<td>Vertex finding efficiency</td>
<td>0.56%</td>
<td>1.22%</td>
<td>0.28%</td>
<td>0.52%</td>
<td>0.42%</td>
</tr>
<tr>
<td>Photon energy scale and smearing</td>
<td>2.55%</td>
<td>3.02%</td>
<td>6.62%</td>
<td>3.21%</td>
<td>3.19%</td>
</tr>
<tr>
<td>Nonlinearity of detector response</td>
<td>0.40%</td>
<td>0.27%</td>
<td>0.26%</td>
<td>0.33%</td>
<td>0.15%</td>
</tr>
<tr>
<td>Nonuniformity of light collection</td>
<td>0.31%</td>
<td>0.20%</td>
<td>0.03%</td>
<td>0.16%</td>
<td>0.26%</td>
</tr>
<tr>
<td>Shower shape corrections</td>
<td>0.44%</td>
<td>0.36%</td>
<td>0.14%</td>
<td>0.55%</td>
<td>0.27%</td>
</tr>
<tr>
<td>Modelling of material budget</td>
<td>0.25%</td>
<td>0.08%</td>
<td>0.12%</td>
<td>0.13%</td>
<td>0.29%</td>
</tr>
<tr>
<td>Modelling of detector response in GEANT4</td>
<td>1.17%</td>
<td>0.32%</td>
<td>0.11%</td>
<td>0.17%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Jet energy scale and resolution</td>
<td>1.70%</td>
<td>2.53%</td>
<td>25.82%</td>
<td>2.09%</td>
<td>2.17%</td>
</tr>
<tr>
<td>ggF contamination in VBF categories</td>
<td>0.59%</td>
<td>1.12%</td>
<td>13.58%</td>
<td>1.27%</td>
<td>0.05%</td>
</tr>
<tr>
<td>UE and PS</td>
<td>0.53%</td>
<td>1.87%</td>
<td>8.88%</td>
<td>0.51%</td>
<td>0.32%</td>
</tr>
<tr>
<td>Lepton reconstruction and btag efficiencies</td>
<td>0.18%</td>
<td>0.07%</td>
<td>0.06%</td>
<td>1.08%</td>
<td>2.56%</td>
</tr>
<tr>
<td>MET</td>
<td>0.08%</td>
<td>0.08%</td>
<td>0.11%</td>
<td>0.66%</td>
<td>0.11%</td>
</tr>
<tr>
<td>ggF contamination in ttH categories</td>
<td>0.40%</td>
<td>0.09%</td>
<td>0.10%</td>
<td>0.23%</td>
<td>5.83%</td>
</tr>
<tr>
<td>Integrated luminosity</td>
<td>2.13%</td>
<td>2.83%</td>
<td>5.77%</td>
<td>3.24%</td>
<td>2.67%</td>
</tr>
<tr>
<td>Branching ratio</td>
<td>1.65%</td>
<td>1.96%</td>
<td>5.12%</td>
<td>2.69%</td>
<td>1.97%</td>
</tr>
<tr>
<td>QCD scale yield</td>
<td>2.63%</td>
<td>4.42%</td>
<td>0.29%</td>
<td>1.41%</td>
<td>10.38%</td>
</tr>
<tr>
<td>PDF and alphaS yield</td>
<td>2.13%</td>
<td>3.29%</td>
<td>5.13%</td>
<td>2.19%</td>
<td>4.08%</td>
</tr>
<tr>
<td>QCD scale migrations</td>
<td>1.75%</td>
<td>1.53%</td>
<td>8.21%</td>
<td>6.76%</td>
<td>1.45%</td>
</tr>
<tr>
<td>PDF migrations</td>
<td>0.46%</td>
<td>0.22%</td>
<td>0.73%</td>
<td>0.40%</td>
<td>0.79%</td>
</tr>
<tr>
<td>AlphaS migrations</td>
<td>0.62%</td>
<td>0.05%</td>
<td>1.03%</td>
<td>0.33%</td>
<td>0.17%</td>
</tr>
<tr>
<td>Total</td>
<td>7.35%</td>
<td>10.27%</td>
<td>34.38%</td>
<td>9.80%</td>
<td>14.24%</td>
</tr>
</tbody>
</table>
7.4 Results

Figure 7.7 (left) the expected uncertainty the measurement of the signal strength ($\mu = \sigma / \sigma_{SM}$) from simulation with a 125 GeV Higgs boson is presented. This was calculated by performing a likelihood scan of the signal strength modifier on an Asimov dataset [19] to give the expected uncertainty on the measurement. Figure 7.7 (right) shows the observed value of the signal strength which is $1.16^{+0.15}_{-0.14}$ times the Standard Model expectation.

The signal strength is then broken into the various production modes; the expected measurements are displayed in figure 7.8 (left) while the measured values are shown in figure 7.8 (right). Finally, the measured strength of the signal can be seen divided into all analysis categories in figure 7.9.

By measuring the coupling of the Higgs boson to various Standard Model particles it is possible to further validate the Standard Model. The expected signal strength from the vector boson production modes vs the fermion production modes is shown in figure 7.10 (left) while the measured signal strength is shown in figure 7.10 (right). Similarly, the coupling of the Higgs boson to vector bosons vs the coupling to fermions is shown in figure 7.11 (left) while the effective coupling of the Higgs boson to gluons vs photons is shown in figure 7.11 (right).
Figure 7.7. The expected uncertainty on the measurement (left) and the observed measurement (right) of $\mu$ for the mass 125 GeV Higgs boson is shown with both statistical uncertainties only and full systematics.

Figure 7.8. The expected (left) and observed (right) measurements of signal strengths relative to the SM prediction for the four production modes. The error bars corresponding to one standard deviation are shown for each point and the green band displays the combined uncertainty.
Figure 7.9. The observed signal strengths relative to the SM expectation for each category. The ZH Leptonic Tag has no events in the signal region and is therefore grayed out.
Figure 7.10. The expected (left) and observed (right) signal strength for vector boson ($\mu_{VBF,VH}$) vs fermion ($\mu_{ggH,tH}$) Higgs boson production modes relative to the SM prediction. One and two $\sigma$ uncertainty contours are shown.

Figure 7.11. The observed vector boson vs fermion (left) coupling to the Higgs boson relative to the SM prediction. The observed effective coupling of gluons vs photons to the Higgs boson relative to Standard Model prediction (right). One and two $\sigma$ uncertainty contours are shown.
The work presented in this thesis was carried out within the large group of CMS collaborators working on measuring the properties of the Higgs boson as it decays to two photons. My main contributions to the analysis were the design and testing of the analysis trigger, as well as characterizing the $p_T^{\text{miss}}$, and defining the VH categories (focusing on the VH $E_T^{\text{miss}}$ category).

A measurement of the properties of the Higgs boson decaying into two photons has been presented. The analysis covers 35.9 fb$^{-1}$ of 13 TeV data collected by the CMS detector at the LHC. The best fit signal strength was found to be consistent with the Standard Model: $1.16^{+0.15}_{-0.14}$ times the Standard Model expectation. The signal strengths are also consistent with the Standard Model when looking at each production mode of the Higgs boson individually. And finally, the relative coupling of the Higgs boson to fermions and vector boson as well as the relative coupling to photons and gluons are shown to be consistent with the Standard Model.


