Search for the Standard Model Higgs boson in the vector boson fusion-mediated diphoton decay channel using multivariate analysis techniques

by

David Di Valentino

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

Ottawa-Carleton Institute for Physics
Department of Physics
Carleton University
Ottawa, Ontario, Canada
August, 2013

© copyright
David Di Valentino, 2013
Abstract

A search for the Standard Model Higgs boson in the diphoton decay channel is presented using multivariate analysis techniques, with a focus on the vector boson fusion (VBF) production mechanism. Data events are separated into signal (VBF $H \rightarrow \gamma\gamma$) and background-like categories using a gradient boosted decision tree, with the optimized analysis applied to the full 2011—2012 dataset, corresponding to 4.8 fb$^{-1}$ of $\sqrt{s} = 7$ TeV $pp$ collisions, and 21 fb$^{-1}$ of $\sqrt{s} = 8$ TeV collisions. The best fit invariant mass for events categorized as VBF $H \rightarrow \gamma\gamma$ is found to be $m_H = 123.5$ GeV, with a local significance of 2.9$\sigma$. The best fit signal strength for the $W, Z$-mediated $H \rightarrow \gamma\gamma$ decay (VBF + associated production) is found to be $\mu_{\text{VBF+VH}} \times B/B_{\text{SM}} = 1.72^{+0.85}_{-0.77} (\text{stat})^{+0.48}_{-0.29} (\text{syst})^{+0.25}_{-0.29} (\text{theory})$ at $m_H = 126.8$ GeV, which agrees with Standard Model predictions within 2$\sigma$. 
Acknowledgements

I’d like to thank my supervisor Thomas Koffas, as well as my colleagues Dag Gillberg, Florian Bernlochner, and Jim Lacey for their guidance, advice, patience, and hard work over these past two years. Your expertise and understanding of physics are a constant source of inspiration for me, and I hope that someday I’ll be able to pass on that inspiration in the same way.

I also give my thanks and love to my parents and sisters for their unwavering love, support, and encouragement in both the good and trying times.

All my love and thanks to Jill for your constant love and support, and for putting up with all my globetrotting. Your editing skills and knowledge of language, as well, were invaluable in helping me craft a sleek and professional piece of scientific writing.

Thanks also to NSERC and Carleton University, for providing funding for this research.
# Contents

Abstract i  
Acknowledgements ii  
List of Tables vi  
List of Figures viii  

1 Introduction 1  
\hspace{2em} 1.1 The Standard Model of particle physics 1  
\hspace{2em} 1.2 The Englert-Brout-Higgs mechanism 2  
\hspace{2em} 1.3 The Large Hadron Collider 3  
\hspace{2em} 1.4 The ATLAS Detector 4  
\hspace{2em} 1.5 Higgs boson production 5  
\hspace{2em} 1.6 Higgs boson decays 6  
\hspace{2em} 1.7 Motivation of thesis topic 8  

2 The VBF $H \to \gamma\gamma$ process 11  
\hspace{2em} 2.1 What is vector boson fusion? 11  
\hspace{2em} 2.2 What is the $H \to \gamma\gamma$ decay? 12  
\hspace{2em} 2.3 NLO, NNLO QCD and electroweak corrections 13  
\hspace{2em} \hspace{2em} 2.3.1 QCD corrections to VBF production 14  
\hspace{2em} \hspace{2em} 2.3.2 Two-loop corrections to $H \to \gamma\gamma$ 15  
\hspace{2em} 2.4 Kinematics of the VBF $H \to \gamma\gamma$ process 16  
\hspace{2em} \hspace{2em} 2.4.1 VBF tree-level kinematics 17  
\hspace{2em} \hspace{2em} 2.4.2 Kinematics of the $H \to \gamma\gamma$ decay 18  
\hspace{2em} 2.5 Backgrounds 20  
\hspace{2em} \hspace{2em} 2.5.1 Dijet backgrounds 20
3 The $H \rightarrow \gamma\gamma$ analysis in ATLAS 23

3.1 Photon reconstruction ................................. 23

3.2 Jet reconstruction ........................................... 24
  3.2.1 Cluster and jet reconstruction .......................... 24
  3.2.2 Jet energy measurement and correction .................. 25
  3.2.3 Jet vertex fraction ....................................... 26

3.3 Diphoton candidate selection .............................. 27

3.4 Dijet candidate selection .................................. 29

3.5 Event categorization ....................................... 29

4 Introducing the VBF multivariate analysis 31

4.1 Boosted decision trees ..................................... 31

4.2 Signal and background modelling ........................... 33
  4.2.1 Signal modelling ......................................... 33
  4.2.2 Background modelling .................................... 33
  4.2.3 Monte Carlo event weights ............................... 34

4.3 Input sample selection ..................................... 35
  4.3.1 Samples available for the multivariate analysis ........... 35
  4.3.2 Signal, background sample configuration ................ 36
  4.3.3 Background training sample selection .................. 37

4.4 Input variable selection ................................... 39
  4.4.1 Deriving a list of discriminating variables ............... 39
  4.4.2 Optimization of input variables .......................... 42
  4.4.3 Final list of input variables ............................. 45
  4.4.4 Kolmogorov-Smirnov (overtraining) tests ................ 48

4.5 Definition of VBF MVA categories ........................ 50

4.6 Distributions of discriminating variables after VBF categorization 53

4.7 Checking for BDT sculpting in $m_{\gamma\gamma}$ .................. 57

5 Refining the VBF multivariate analysis 60

5.1 Improving the background model ........................... 60

5.2 Improving benchmark variable selection .................... 61

5.3 Optimization of the BDT configuration ...................... 64

5.4 Final definition of VBF MVA categories .................... 67
5.5 Distributions of discriminating variables after categorization . . . . 69
5.6 Checking for BDT sculpting in $m_{\gamma\gamma}$ . . . . . . . . . . . . . . . . . . 75

6 Systematic uncertainties  78
   6.1 Theoretical uncertainties . . . . . . . . . . . . . . . . . . . . . . . . . 78
      6.1.1 Higher-order perturbative uncertainty for $gg \rightarrow H + 2$ jets . 78
      6.1.2 Modelling uncertainties of $\eta_{Zep}$ . . . . . . . . . . . . . . 81
      6.1.3 Modelling uncertainties of $\Delta \phi_{jj}$ . . . . . . . . . . . . . 82
      6.1.4 Underlying event uncertainty . . . . . . . . . . . . . . . . . . . 83
   6.2 Jet and detector modelling uncertainties . . . . . . . . . . . . . . . 84
      6.2.1 Jet energy scale and resolution . . . . . . . . . . . . . . . . . . 84
      6.2.2 Jet vertex fraction selection . . . . . . . . . . . . . . . . . . . 85

7 VBF multivariate analysis results  88
   7.1 Analysis goals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88
   7.2 Training samples and BDT response . . . . . . . . . . . . . . . . . . . 91
   7.3 $m_{\gamma\gamma}$ measurement . . . . . . . . . . . . . . . . . . . . . . . . 93
   7.4 Signal strength measurement . . . . . . . . . . . . . . . . . . . . . . . 93

8 Conclusions  97

References  99

A Photon, jet reconstruction miscellanea  108
   A.1 Definitions of discriminating variables . . . . . . . . . . . . . . . . . . 108
   A.2 Comparison of background samples after dijet selection . . . . . . . . 110
   A.3 Descriptions of JES uncertainties . . . . . . . . . . . . . . . . . . . 115

B Personal contributions to the ATLAS collaboration  116
## List of Tables

4.1 Different types of diphoton + dijet backgrounds tested, and their composition. .......................................................... 38

4.2 The top ten most discriminating variables for the preliminary VBF vs. inverted isolation BDT training, ranked according to signal/background shape separation and importance. .......................... 42

4.3 Evolution of optimal significance when selecting $k$ variables only from 10-variable ‘Selection A’. The significance shown is calculated using the shape method, with statistical and fit uncertainties included as penalty terms. ......................................................... 49

4.4 Evolution of optimal significance when selecting $k$ variables only from 8-variable ‘Selection B’. The significance shown is calculated using the shape method, with statistical and fit uncertainties included as penalty terms. ......................................................... 49

4.5 Signal and background statistics for the VBF vs. inverted isolation training. .......................................................... 51

4.6 Definitions of the high and low VBF categories, along with the optimal significance and VBF purity obtained for each cut, for the BDT trained on VBF $H \rightarrow \gamma\gamma$ signal Monte Carlo and inverted isolation background. ......................................................... 53
5.1 Separation and importance for each variable used in the eight-variable VBF vs. Sherpa + inverted isolation training.

5.2 Configurable parameters for boosted decision trees in TMVA.

5.3 Possible values of each parameter tested for BDT configuration scan.

5.4 Signal and background statistics for the eight-variable VBF vs. Sherpa + inverted isolation training. The figures here are quoted after cuts on $\Delta \eta_{jj} > 2$ and $|\eta_{\gamma\gamma}| < 5$. Notably, the differences in sample statistics relative to Table 4.5 are due to the additional cut on $|\eta_{\gamma\gamma}|$, and the addition of another Sherpa sample (e1264).

5.5 Definitions of the optimal high and low VBF categories, along with the optimal significance and VBF purity obtained for each cut, for the BDT trained on VBF signal and Sherpa + inverted isolation background.

6.1 Higher-order perturbative correction uncertainties for $gg \rightarrow H$ for the high and low VBF categories.

6.2 Underlying event uncertainties for $m_H = 125$ GeV ggF and VBF Monte Carlo at $\sqrt{s} = 8$ TeV.

6.3 Largest JES uncertainties for $\sqrt{s} = 8$ TeV VBF and ggF $H \rightarrow \gamma\gamma$ Monte Carlo samples ($m_H = 125$ GeV).

6.4 Jet energy resolution uncertainties for $\sqrt{s} = 8$ TeV VBF, ggF $H \rightarrow \gamma\gamma$ Monte Carlo samples ($m_H = 125$ GeV).

6.5 Jet vertex fraction systematic uncertainties computed using $m_H = 125$ GeV $H \rightarrow \gamma\gamma$ signal Monte Carlo samples at $\sqrt{s} = 8$ TeV.

A.1 Definitions of the kinematic variables used in this analysis.

A.2 Descriptions of the largest contributors to the JES uncertainty.
List of Figures

1.1 Tree-level Feynman diagrams of the four main Higgs boson production mechanisms at the LHC. 6

1.2 Standard Model Higgs boson production cross sections for pp collisions at $\sqrt{s} = 8$ TeV. 7

1.3 Tree-level (lowest order) Feynman diagrams for notable light-to-medium Higgs boson decay channels searched for at the LHC. 8

1.4 Branching ratios for a Standard Model Higgs boson as a function of mass $m_H$. 9

2.1 Leading order diagrams for Higgs boson production through vector boson fusion. Here, $V = W^\pm$ or $Z$. 12

2.2 Lowest order contributions to the $H \rightarrow \gamma\gamma$ decay cross section. 13

2.3 Feynman diagrams for common NLO QCD corrections to the vector boson fusion vertex. 14

2.4 Feynman diagrams for NNLO QCD corrections to the vector boson fusion vertex included in the structure function approach. These are the only NNLO diagrams found to contribute non-negligibly to the VBF production cross section. 15

2.5 Two-loop electroweak corrections for the $H \rightarrow \gamma\gamma$ decay process. 15
2.6 Two-loop QCD corrections for the $H \rightarrow \gamma \gamma$ decay process. Note that all gluon loop corrections are virtual, as required by conservation of colour charge. 

2.7 Parton-level kinematics of the leading (solid line) and subleading (dashed line) VBF quarks at $m_H = 125$ GeV, as generated by Powheg with $\sqrt{s} = 8$ TeV. The plots are shown prior to any kinematic cuts, and scaled to arbitrary units.

2.8 Parton-level kinematics of the leading (solid line) and subleading (dashed line) photons resultant from VBF Higgs production at $m_H = 125$ GeV, as generated by Powheg. The plots are shown prior to any kinematic cuts, and are scaled to the expected number of events from 21 $fb^{-1}$ of $\sqrt{s} = 8$ TeV data.

2.9 Feynman diagrams of the irreducible components of the $H \rightarrow \gamma \gamma$ background.

3.1 Explanatory diagram of $\tilde{p}_T$, the magnitude of the component of the diphoton transverse momentum $\vec{p}_T^\gamma$ transverse to the diphoton thrust axis $\hat{t}$.

4.1 Schematic diagram of a binary decision tree.

4.2 The distribution of $\Delta \phi_{\gamma \gamma jj}$ in the $m_H = 125$ GeV VBF $H \rightarrow \gamma \gamma$ Monte Carlo sample (Powheg + Pythia8) after binning on $[2.94, \pi]$.

4.3 Distributions of $m_{\gamma \gamma}$ for the optimal permutations of variables (passing the BDT cut) derived from ‘Selection A’ (top) and ‘Selection B’ (bottom).

4.4 Matrices of the linear correlation coefficients for the optimized significance of ‘Selection A’ for signal (left) and background (right).
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>(a) Results of Kolmogorov-Smirnov tests performed by TMVA for the signal/background training and testing samples. (b) A comparison of the BDT score distributions of the inverted isolation and data sideband samples after BDT training and application.</td>
</tr>
<tr>
<td>4.6</td>
<td>Hierarchical structure of event selection in the VBF $H \rightarrow \gamma\gamma$ multivariate analysis.</td>
</tr>
<tr>
<td>4.7</td>
<td>Distributions of the six benchmark variables for events falling in the low VBF category ($0.18 &lt; \text{BDT score} &lt; 0.74$) based on the VBF vs. inverted isolation BDT. Shown here are data sidebands, along with VBF and ggF $H \rightarrow \gamma\gamma$ signal Monte Carlo samples.</td>
</tr>
<tr>
<td>4.8</td>
<td>Distributions of the six benchmark variables for events falling in the high VBF category ($\text{BDT score} &gt; 0.74$) based on the VBF vs. inverted isolation BDT. Shown here are data sidebands, along with VBF and ggF $H \rightarrow \gamma\gamma$ signal Monte Carlo samples.</td>
</tr>
<tr>
<td>4.9</td>
<td>Comparisons of the benchmark variable distributions of data sideband and inverted isolation events falling in the low VBF category ($0.18 &lt; \text{BDT score} &lt; 0.74$) based on the VBF vs. inverted isolation BDT.</td>
</tr>
<tr>
<td>4.10</td>
<td>Comparisons of the benchmark variable distributions of data sideband and inverted isolation events falling in the high VBF category ($\text{BDT score} &gt; 0.74$) for the VBF vs. inverted isolation BDT.</td>
</tr>
<tr>
<td>4.11</td>
<td>(Top) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of VBF signal Monte Carlo events based on the VBF vs. inverted isolation BDT. (Bottom) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of sideband data.</td>
</tr>
<tr>
<td>4.12</td>
<td>Profiles of the average BDT score with $m_{\gamma\gamma}$.</td>
</tr>
</tbody>
</table>
5.1 Comparison of data sidebands with Sherpa + inverted isolation, inverted isolation background samples for the six benchmark variables (Note: All plots have a cut of $\Delta \eta_{ij} > 2$ applied). .......................... 62

5.2 Signal (left) and background (right) correlation matrices for VBF vs. Sherpa + inverted isolation, with cuts on $\Delta \eta_{ij}$ and $\eta^{Z\gamma\gamma}_{\gamma\gamma}$. .................. 68

5.3 (a) Kolmogorov-Smirnov test results for the eight-variable training using VBF $H \rightarrow \gamma\gamma$ signal Monte Carlo events vs. Sherpa + inverted isolation background. (b) The resultant BDT score distributions for the Sherpa + inverted isolation background training sample, and data sidebands. ................................. 69

5.4 Distributions of the eight training variables for events in the optimized low VBF category ($0.51 < \text{BDT score} < 0.75$) for data sidebands, and VBF, ggF signal Monte Carlo samples. .......................... 71

5.5 Distributions of the eight benchmark variables for events in the optimized high VBF category (BDT score $> 0.75$) for data sidebands, and VBF, ggF signal Monte Carlo samples. ................................. 72

5.6 Comparisons of the benchmark variable distributions of data sideband and Sherpa + inverted isolation events in the low VBF category ($0.51 < \text{BDT score} < 0.74$) for the VBF vs. Sherpa + inverted isolation BDT. ................................. 73

5.7 Comparisons of the benchmark variable distributions of data sideband and Sherpa + inverted isolation events in the high VBF category (BDT score $> 0.74$) for the VBF vs. Sherpa + inverted isolation, eight-variable BDT. ................................. 74
5.8 (Top) Plots of the $m_{\gamma\gamma}$ sidebands for each VBF signal Monte Carlo event quartile for the VBF vs. Sherpa + inverted isolation BDT. (Bottom) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of sideband data. ................................. 76

5.9 Profiles of the average BDT score with $m_{\gamma\gamma}$. ................................. 77

6.1 Higher-order perturbative uncertainty error bands in the $gg \to H + 2$-jet decay relative to $\pi - \Delta\phi_{\gamma\gamma jj}$, derived using the Stewart-Tackmann procedure. Good agreement is observed between errors derived from MCFM and Powheg + Pythia8 Monte Carlo. ........ 80

6.2 The ratio of the distributions of $\eta^{\text{Zep}}_{\gamma\gamma}$ between Powheg + Pythia8 and MCFM (shown on the $y$-axis), along with the fitted polynomial reweighting function. ................................. 82

6.3 Weight function for truth-level $\Delta\phi_{jj}$ (as shown on the $y$-axis) derived using Sherpa and Powheg + Pythia8. ................................. 83

6.4 Reweighting functions between CT10 and non-MPI for ggF (left) and VBF (right) $H \to \gamma\gamma$ Monte Carlo generated using Powheg + Pythia8. ................................. 84

7.1 (a) BDT response of the background sample, as compared to experimental data, after diphoton + 2-jet + pre-selection cuts. (b) BDT response of the background and VBF signal samples. ........ 92

7.2 The full 2012 $m_{\gamma\gamma}$ distributions, along with expected SM signal and background fit for $m_H = 126.8$ GeV for low (left) and high (right) VBF categories. ................................. 94
7.3  (a) The observed local $p_0$ value for VBF $H \rightarrow \gamma\gamma$ events as a function of $m_H$ for the combination of $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data. (b) The best-fit values of $\mu_{ggH+tH} \times B/B_{SM}$ and $\mu_{VBF+VH} \times B/B_{SM}$ and their 68% (solid) and 95% (dashed) CL contours. . . . 96

A.1  (Left) The layout of the ATLAS detector in Cartesian coordinates, with definitions of $\theta$ and $\phi$, relative to their rotational axes. (Right) A diagram of pseudorapidity as it relates to the $z$-axis in ATLAS. . 108

A.2  A comparison of the distributions for inverted isolation, inverted ID, and Sherpa background samples, along with the data sidebands. 114
Chapter 1

Introduction

1.1 The Standard Model of particle physics

The Standard Model (SM) of particle physics is a mathematical framework that governs the generation and interactions of the known particles in the universe, along with the forces mediating these interactions. The model has stood up to continual scrutiny and testing, with its predictions verified by many experimental collaborations over the years.\(^1\) The most fundamental proposition of the Standard Model is that all visible matter in the universe originates from a finite number of particles (leptons, quarks, bosons), with their respective antimatter equivalents, to the point that any other known particle is simply a composite of one or more irreducible elements. Furthermore, the SM postulates that all particles interact via the exchange of similarly irreducible particles – namely, weak bosons, gluons, and photons. In particular, the forces described by the Standard Model include electromagnetism (mediated by photons), the weak nuclear force (mediated by W, Z bosons), and the strong nuclear force (mediated by gluons).

\(^1\) There are numerous high quality texts on the subject, with the Particle Data Group’s Review of Particle Physics [1] being a good starting point.
For all of the experimental validation the Standard Model has received, pieces of it remain unexplained or unverified to this day. For example, though the Standard Model posits symmetry between matter and antimatter, the heavy imbalance of actual matter/antimatter content in the universe appears to contradict expectations. A second notable issue is the assumption of massless neutrinos in the Standard Model, which directly contradicts the results of experiments on neutrino oscillation performed by the Sudbury Neutrino Observatory and Super-Kamiokande collaborations, among others [2, 3]. Prior to 2012, arguably the most prominent unexplored frontier in the quest to complete the Standard Model was the search for the Higgs boson, primarily because its discovery would validate (or invalidate) the Standard Model theory of how particles acquire mass, known as the Englert-Brout-Higgs (EBH) mechanism.

1.2 The Englert-Brout-Higgs mechanism

The Englert-Brout-Higgs mechanism is a theory describing the generation of massive weak bosons in the Standard Model, and has roots in Goldstone’s work on the spontaneous breaking of global symmetries. Goldstone’s paper on the subject, published in 1961, theorized that the shift of a system’s ground state from symmetric to asymmetric, e.g. through the introduction of a potential term $V(\Phi)$ in the Lagrangian of the system, induced the production of massless scalar bosons as excitations of the field, known as Goldstone bosons [4]. Roughly thirty years prior, Enrico Fermi developed the first theory of the weak interaction to explain beta decay (i.e. the emission of an $e^\pm$ from an atomic nucleus) [5]. Before Goldstone’s work, weak bosons were thought to be massive, due to the short range of the interaction. However, a gauge theory mechanism for generating their masses was unknown.
In 1964, a mechanism to introduce massive weak bosons into quantum field theory via spontaneous symmetry breaking was developed independently by three groups: Peter Higgs [6]; Brout and Englert [7]; and Guralnik, Hagen, and Kibble [8]. This framework posited the existence of a field with four degrees of freedom, permeating the entirety of space, known as the Higgs field. The introduction of the field into the Standard Model Lagrangian breaks the gauge symmetries of three out of four generators of the $\text{SU}(2) \times \text{U}(1)$ electroweak gauge group, whose field quanta are the weak bosons ($W^\pm$ and $Z$), and the photon.

Upon symmetry breaking, the Higgs field couples to the gauge fields of the weak bosons, leading to the absorption of three of its degrees of freedom (manifested as Goldstone bosons) by the $W^\pm$ and $Z$ bosons, providing them with an effective mass. The symmetries of the generators of the electromagnetic interaction, however, remain unbroken, rendering the field quantum (the photon) massless. Furthermore, the fermion fields couple to the Higgs field in the form of a Yukawa interaction, ultimately retaining the proper fermion masses. The fourth degree of freedom of the Higgs field manifests as a Higgs boson – a massive, charge neutral, scalar particle.

## 1.3 The Large Hadron Collider

The Large Hadron Collider (LHC) is a circular proton-proton collider, 27 km in circumference, located underground at CERN, along the French-Swiss border. Initial construction began in 1983, with the collider being fully functional since 2008.\(^2\) When operational, the LHC collides bunches of up to $10^{11}$ protons at rates of 20 MHz. It has a design luminosity and center-of-mass energy of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$ and $\sqrt{s} = 14 \text{ TeV}$ (or $7 \text{ TeV} / \text{ proton}$), respectively [9]; however, as of 2012,

\(^2\) A historical point: The LHC occupies the same tunnel that once held the Large Electron-Positron collider.
collisions had been occurring at $\sqrt{s} = 8 \text{ TeV} \ (7 \text{ TeV} \ \text{in} \ 2011)$. Since May 2013, the collider has been in shutdown, as it undergoes upgrades meant to increase collision energy to $\geq 6.5 \text{ GeV / beam}$. Once the upgrades are complete, it is expected that collisions and data taking will resume in 2015.

## 1.4 The ATLAS Detector

The ATLAS detector (**A** Toroidal **LHC** **A**pparatus) is a multipurpose cylindrical detector with coverage of almost $4\pi$ in solid angle, and capable of the detection, reconstruction, and proper identification of photons, charged leptons, and hadrons. The detector itself is composed of several sub-detectors which deal specifically with charged particle tracking (the Inner Detector) and energy measurement (electromagnetic, hadronic calorimeters), as well as a set of detectors that deal almost exclusively with muons (the muon spectrometer) [10].

The Inner Detector covers a pseudorapidity range of $|\eta| < 2.5$ (as defined in Equation 2.1), and encompasses a silicon pixel detector, silicon microstrip detector, and a transition radiation tracker (TRT). Using charged particle tracks, the Inner Detector is capable of accurately reconstructing the location of a given proton-proton collision, as well as the location of photon conversions to electron/positron pairs within a radius of $\sim 800 \text{ nm}$. The Inner Detector is surrounded by a solenoid producing a $2 \text{ T}$ magnetic field, which allows charged particle momenta to be determined based on the direction and degree of track bending in the field.

Immediately beyond the Inner Detector is the electromagnetic calorimeter (ECAL), a liquid argon (LAr)-based sampling detector, capable of determining the energy, mass, and direction of electromagnetic (EM) showers created by particles such as electrons or photons. It has an accordion geometry, and is divided into a barrel section, covering the region $|\eta| < 1.475$, and two end-cap sections,
covering $1.375 < |\eta| < 3.2$. The ECAL is also divided into three layers: a first layer, segmented in $|\eta|$, which provides particle identification for electrons and photons; a second layer to absorb most of the energy from EM showers; and a third layer to correct for any energy leakage beyond the ECAL. A presampler is also located in front of the first layer for $|\eta| < 1.8$, which serves to correct for particle energy losses before entering the ECAL.

The hadronic calorimeter (HCAL) surrounds the ECAL, and serves to measure the energies and directions of all manner of hadrons (baryons, mesons, etc.). The HCAL consists of steel and scintillating tiles in the range $|\eta| < 1.7$, and two copper/LAr detectors on the range $1.5 < |\eta| < 3.2$, and additional copper-tungsten/LAr calorimeters covering up to $|\eta| < 4.9$. The last detector component is the muon spectrometer, located beyond the HCAL, which provides coverage of up to $|\eta| < 2.7$. The detector is composed of three air-core superconducting toroid systems and precision tracking chambers to accurately track and measure the energy and direction of a given muon.

### 1.5 Higgs boson production

In the context of current LHC operating conditions, there are four noteworthy methods of Higgs boson production, as shown in Figure 1.1. The dominant production mechanism, as evidenced by the predicted cross sections of Figure 1.2, is gluon-gluon fusion (ggF), or $gg \rightarrow H$. This process is mediated by heavy fermion loops, with $t\bar{t}$ and $b\bar{b}$ loops being strongly preferred – a consequence of the proportionality of the couplings of the Higgs boson to particle mass [1].

The second largest production mechanism is vector boson fusion (VBF), or $qq \rightarrow H + qq$, wherein two quarks traveling antiparallel to each other emit virtual weak bosons (i.e. $W$ or $Z$), which undergo inverse pair decay to form a
1.6 Higgs boson decays

Decays of the Higgs boson are also mediated through direct couplings to weak bosons, or heavy fermion-mediated loops, as shown in Figure 1.3. A plot showing the different branching ratios of the SM Higgs boson, as a function of Higgs boson mass $m_H$, is shown in Figure 1.4. For a light Higgs boson ($m_H < 150$ GeV), decays to pairs of heavy particles are strongly preferred, including quarks ($b\bar{b}$ and $c\bar{c}$), fermions ($\tau^+\tau^-$), and $WW$, $ZZ$ pairs. These decay products can also decay...
1.6 Higgs boson decays

Further – for example, the $W$ and $Z$ bosons typically decay to $l\nu_l$ and $l^+l^-$ pairs, respectively. Because there are no direct $\gamma\gamma$ or $gg$ couplings to the Higgs boson, these decays are mediated either by loops of heavy fermions (typically $t\bar{t}$) or weak bosons ($WW$ for the $\gamma\gamma$ decay) [1]. In addition to the $\gamma\gamma$ decay, the $Z\gamma$ decay also proceeds through $W$ and heavy fermion loops.

Experimentally, the decay channels that can be fully reconstructed in the ATLAS detector are the most ideal for Higgs boson searches. In particular, the $H \to ZZ^* \to 4l$ (where $l = e, \mu$) and $H \to \gamma\gamma$ decays are both reasonably ‘clean’ channels, owing primarily to their lack of missing energy (due to neutrinos), and the excellent energy resolution and linearity of detector response for electrons, photons, and muons. In both cases, one expects a Higgs boson to be observed directly via the invariant mass distribution of each decay, manifesting as a signal peak on top of a smoothly varying background. Indeed, these are the decay channels that led to the discovery of a Higgs-like boson in 2012 [14].
1.7 Motivation of thesis topic

As of July 2012, the ATLAS experiment had already been able to claim the discovery of a Standard Model Higgs boson-like mass resonance at approximately 125 GeV [14]. A better understanding of the properties of the newly-discovered particle would require a combination of sufficient data, and more effective techniques to isolate the desired signal.

In order to judge whether the newly-discovered particle behaves like a Standard Model Higgs boson, it is necessary to measure both its couplings to massive particles, as well as the signal strength $\mu = \sigma/\sigma_{SM}$ for its various production and decay modes. The study of the $H \rightarrow \gamma\gamma$ decay, as mediated by the VBF Higgs boson production process, is an excellent vehicle for performing both of these measurements, as it is possible to isolate experimentally all of the final state physics objects necessary to observe the process. For example, comparisons with Standard Model Higgs boson couplings are made by taking ratios of the observed partial

![Feynman diagrams](image-url)
1.7 Motivation of thesis topic

widths of signal processes [15], which cancel theoretical and luminosity uncertainties. Therefore, in searching for the final state jets\(^3\) characteristic of VBF, it is possible to compare the couplings of the new boson to the weak bosons with Standard Model predictions. It is also possible to determine the signal strength for the \(H \rightarrow \gamma \gamma\) decay by studying the diphoton final state, which will indicate the consistency of the observed and expected signal rates. More precise measurements of the mass of the new boson are also possible with better discrimination of VBF \(H \rightarrow \gamma \gamma\)-like events, owing to the unique kinematics of the process (as described in Section 2.4).

Until recently, the VBF \(H \rightarrow \gamma \gamma\) analysis group used cut-based techniques to study the experimental data, defining a signal-like region of phase space by a series of quality cuts on the relevant physics objects [16], e.g. photons and hadronic jets. The use of multivariate analysis (MVA), however, allows one to

\(^3\) Narrow cones of hadrons and other particles produced by quark or gluon hadronization.
consider a larger portion of phase space, and exploits the full shapes of variables in order to discriminate signal from background. This approach has historical precedent, as well; in 2011, the $H \to \tau \tau$ group demonstrated that the usage of multivariate analysis techniques led to a better discrimination of $\tau$ leptons [17]. This success prompted several other analysis groups to evaluate the effectiveness of MVA techniques in their own studies.

This thesis presents the development of a multivariate analysis-based approach to isolate VBF $H \to \gamma\gamma$ signal, in which the unique kinematics of both the vector boson fusion and diphoton decay processes were used to form a single discriminant that judged the properties of a given hard scattering event as signal- or background-like. Also presented are the final results of the analysis, wherein the VBF $H \to \gamma\gamma$ discovery significance and production rate were computed relative to Standard Model expectations for the full 2011 – 2012 dataset.
Chapter 2

The VBF $H \rightarrow \gamma\gamma$ process

2.1 What is vector boson fusion?

Within the lexicon of particle physics, a vector boson is a boson with spin quantum number 1 – in particular, the $W$ and $Z$ bosons, photons, and gluons [18]. In the Standard Model of Particle Physics, vector boson fusion is a production mechanism wherein two incoming quarks emit virtual $W$ or $Z$ bosons which undergo inverse pair decay to form a Higgs boson [11]. As demonstrated in Figure 2.1, contributions to leading order VBF production are made in the $s$, $t$, and $u$ channels. However, at circular hadron colliders such as the LHC, the $t$ and $u$ fusion channels are heavily favored [12], as the partonic cross sections of these contributions rise logarithmically with the centre-of-mass energy of the subprocess ($\hat{\sigma} \propto \log \hat{s}/\hat{M}_W^2$) [19]. Additionally, for reasons described in Section 2.4, the $s$ channel is suppressed by the application of cuts on hadronic decay products. It is worth noting, as well, that the $WW$ fusion contribution is the dominant term in the VBF cross section – a consequence of the larger coupling of the $W$ boson to fermions [19].

VBF Higgs boson production is predicted to be the second largest contribution to the Higgs boson production cross section for the mass range $m_H \sim 100$ –
2.2 What is the $H \rightarrow \gamma\gamma$ decay?

The decay of a Higgs boson to two photons is often referred to as the “silver channel” for Higgs boson production at the LHC [19], and is cited as being one of the cleanest channels for discovery [15]. These claims are substantiated by the lower QCD backgrounds for the diphoton decay as compared to the dominant decay channels (e.g. $H \rightarrow q\bar{q}$), and the overall knowledge and control of the reducible and irreducible elements of the diphoton background (see Section 2.5.2). In addition, the reconstructed diphoton invariant mass spectrum shows direct evidence of Higgs boson production, due to the lack of neutrinos that would otherwise result in missing energy.

Figure 2.1: Leading order diagrams for Higgs boson production through vector boson fusion. Here, $V = W^\pm$ or $Z$.

200 GeV, though it is still an order of magnitude smaller than the production cross section for gluon-gluon fusion [20]. Unlike the gluon-gluon fusion process, though, the VBF production signature is highly distinctive, marked by the presence of two highly energetic final state quark jets at leading order, which tend to be found in opposite regions of cylindrical detectors such as ATLAS [12]. As such, the detection and measurement of these jets can be used to simultaneously reduce backgrounds originating in quantum chromodynamics (QCD) and identify signal-like events (see Section 2.5.1).
It is worth noting that there is no direct coupling of photons to the Higgs boson, owing to their masslessness. Instead, $H \rightarrow \gamma \gamma$ decays proceed through $W$ boson loops, or fermion loops at leading order, as shown in Figure 2.2. In the latter mechanism, decays through light fermion loops are essentially nonexistent, owing to the proportionality of Higgs boson couplings to fermion mass. Therefore, only the $W$ and top quark contribute in any significant manner to the $\gamma \gamma$ decay width [19].

2.3 NLO, NNLO QCD and electroweak corrections

At leading order, VBF Higgs boson production is purely an electroweak process (see Figure 2.1), with the differential cross section evaluated by the structure function approach used to characterize deep inelastic scattering [21]. At high energy colliders such as the LHC, though, QCD radiative corrections can become quite sizeable [22], necessitating the consideration of next-to-leading order (NLO) and next-to-next-to-leading order (NNLO) contributions to the leading order VBF cross section, as well as the $H \rightarrow \gamma \gamma$ decay width. Ultimately, the calculated higher-order electroweak and QCD corrections to the VBF $H \rightarrow \gamma \gamma$ process are small, and as a consequence, its leading order kinematics serve as a very good approximation of the true process.
2.3.1 QCD corrections to VBF production

The NLO QCD corrections to vector boson fusion constitute virtual quark self-energy, gluon exchange between \( qq \rightarrow V \) quark lines, and additional gluon emission from the initial and final states [19], as shown in Figure 2.3. The leading order VBF cross section is amended to include these diagrams by way of corrections to the structure functions that make up the tree-level \( \sigma_{VBF} \) calculation [21, 23]. Overall, NLO corrections to the VBF process only amount to about 8 – 10% of \( \sigma_{VBF} \), and thus are fairly small [23].

NLO and NNLO corrections to \( \sigma_{VBF} \) differ by the presence of gluon exchange between the first and the second incoming or outgoing quark lines. However, approximate corrections to the VBF cross section at NNLO have been computed using a similar variation on the structure function approach used at NLO [22]. In particular, diagrams of the non-negligible NNLO corrections to the cross section are shown in Figure 2.4. Similar to the NLO corrections, the overall NNLO contribution to the VBF production cross section is still on the order of percent [22]. Therefore, vector boson fusion can be considered a “rather clean Higgs production process” [19], since higher-order QCD corrections make up only a marginal fraction of the total VBF cross section.
2.3 NLO, NNLO QCD and electroweak corrections

Figure 2.4: Feynman diagrams for NNLO QCD corrections to the vector boson fusion vertex included in the structure function approach. These are the only NNLO diagrams found to contribute non-negligibly to the VBF production cross section [22].

![Feynman diagrams](image)

Figure 2.5: Two-loop electroweak corrections for the $H \rightarrow \gamma\gamma$ decay process [24].

### 2.3.2 Two-loop corrections to $H \rightarrow \gamma\gamma$

Because the $H \rightarrow \gamma\gamma$ decay proceeds not through a direct $H - \gamma$ vertex, but rather by way of virtual loops, the possibility of higher-order electroweak and QCD corrections to the one-loop, tree-level $H \rightarrow \gamma\gamma$ process must also be considered. These additions to the lowest-order $H \rightarrow \gamma\gamma$ process come in the form of secondary virtual loops within the tree-level diagram. In particular, the electroweak corrections manifest as in-loop exchanges of $W$ bosons, while the QCD corrections occur due to the exchange of gluons among elements of the loop.

Two-loop electroweak corrections can be divided into those induced by light fermions (which are assumed massless), and heavy particles ($W, t, b$) in the loop. Representative Feynman diagrams of these electroweak corrections are shown in Figure 2.5, with the last diagram demonstrating lepton and light fermion contributions to the $H \rightarrow \gamma\gamma$ process. Analytical expressions for these corrections have been derived [24, 25, 26], and for a Higgs boson mass of 125 GeV, the total
two-loop electroweak corrections amount to a fairly small $-2.07\%$ change in the tree-level decay width $\Gamma(H \to \gamma\gamma)$ [24].

With regard to gluon-based corrections, any two-loop QCD corrections made to the tree-level process are strictly virtual, as the presence of an initial/final state gluon (e.g. $H \to \gamma\gamma + g$) would violate colour charge conservation. Feynman diagrams demonstrating some generic two-loop QCD virtual corrections are shown in Figure 2.6. These corrections to the $\gamma\gamma$ decay width have been computed [27, 28], and when combined with the electroweak corrections, it is found that the overall change to $\Gamma(H \to \gamma\gamma)$ is approximately 1.5% [24]. Therefore, for a light Higgs boson ($m_H < 150$ GeV), any two-loop electroweak or QCD contributions to the decay width of the $H \to \gamma\gamma$ process are very small.

### 2.4 Kinematics of the VBF $H \to \gamma\gamma$ process

Because the higher-order corrections to both the VBF and $H \to \gamma\gamma$ processes are relatively small (as explained in Section 2.3), it is possible to describe with good accuracy the physics of the two processes in terms of their behaviour at leading order. The following sections outline the tree-level kinematics of the two processes, and motivate the usage of some relevant observables in cylindrical detectors such as ATLAS.
2.4 Kinematics of the VBF $H \to \gamma\gamma$ process

2.4.1 VBF tree-level kinematics

The presence of two energetic, well-separated jets in the VBF process leads to several distinctive characteristics that are easily observed in an experimental context. For centre-of-mass energies in the TeV range, the two final state VBF quarks tend to emerge from $pp$ collisions with high energies, and small scattering angles relative to the beam axis. In having such energetic final state quarks, there arises two important consequences used in discriminating VBF signal: 1. The intermediate vector bosons ($W$ or $Z$) will tend to have low energies, but necessarily enough to produce the Higgs boson ($\mathcal{O}(\frac{1}{2}m_H)$ each), and 2. The jets resultant from hadronization, while highly energetic, will have relatively low momentum perpendicular to the beamline, as quantified by the transverse momentum ($p_T$) [19]. In effect, there will be a ‘rapidity gap’ between the two jets (see Figure 2.7(e)), such that one jet tends to be directed forward ($\eta \gg 0$), with the other directed backward ($\eta \ll 0$), where $\eta$ is the pseudorapidity,

$$\eta = -\log\left(\tan\left(\frac{\theta}{2}\right)\right). \quad (2.1)$$

where $\theta$ is the angle between the particle momentum and the beam line / $z$-axis.$^1$

These distinct decay features allow for the use of several fairly straightforward criteria in the search for VBF signal. For example, the transverse momenta of the outgoing quark jets are determined by the scale of the weak boson masses (i.e. $p_T \sim m_V$) [20]. Therefore, a lower bound on jet $p_T$ (e.g. $p_T > 20 \text{GeV}$) has the potential to reduce QCD background contributions from low energy jets, such as those from radiated gluons. Furthermore, since the intermediate vector bosons tend to be low energy, subsequent Higgs boson decay products will often be emitted in the central region of the detector ($|\eta| \sim 0$) [19]. Therefore, one may also

$^1$See Section A.1 for details.
require the Higgs boson decay products and outgoing quark jets to be spatially well-separated, i.e. having a reasonably large $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$ between each physics object, where $\phi$ is the azimuthal angle in the $xy$-plane.

Taking advantage of the theoretically high pseudorapidity separation of the outgoing jets, a highly discriminating (and somewhat obvious) cut would be to require a large absolute dijet pseudorapidity difference, i.e. $|\Delta \eta_{jj}| \gg 0$, such that the jets are found in opposite regions of the detector. In addition, the combination of jet $p_T$ and large pseudorapidity separation also implies a larger dijet mass ($m_{jj}$) for VBF events as compared to jet background, which tends to be more centrally emitted. Therefore, placing a lower bound on $m_{jj}$ serves to reduce QCD backgrounds, as well.

The VBF production process is characterized by a lack of coloured particle exchange between quarks in the tree-level diagram, as well as the general lack of kinematic dependence of the Higgs boson on the outgoing quark jets. In contrast, the QCD background to VBF frequently proceeds through colour exchange, which tends to produce more central gluons and hadronic jets [19]. Therefore, in attempting to isolate VBF signal, one can impose a central jet veto [31] on potential signal events. In this case, a typical veto would be to exclude events containing one (or more) high $p_T$ jets in the central region, when two well-separated hadronic jets suggestive of vector boson fusion have already been detected.

### 2.4.2 Kinematics of the $H \to \gamma\gamma$ decay

Parton-level distributions of the leading and subleading photons from the $H \to \gamma\gamma$ decay, as mediated by the vector boson fusion production mechanism, are shown in Figure 2.8. Worth noting are the distinct angular distributions of the two photons, in terms of $\eta$ and $\phi$. It is evident from these plots that the $\gamma\gamma$ decay is
Figure 2.7: Parton-level kinematics of the leading (solid line) and subleading (dashed line) VBF quarks at $m_H = 125$ GeV, as generated by Powheg [29, 30] with $\sqrt{s} = 8$ TeV. The plots are shown prior to any kinematic cuts, and scaled to arbitrary units.
relatively isotropic in $\phi$, with photons strongly preferring to emerge back-to-back in the detector, i.e. $\Delta \phi_{\gamma\gamma} \sim \pi$.

Furthermore, it can be seen based on the plots of photon $\eta$ and $|\Delta \eta_{\gamma\gamma}|$ that the photon pairs tend to be emitted in the central region of the ATLAS detector, with peaks in both distributions occurring about zero. As such, requiring a high transverse momentum ($p_T$) for each photon would be an effective tool in isolating diphoton pairs in a cylindrical detector. By extension, one might also require that the invariant mass of the photon pair ($m_{\gamma\gamma}$) be sufficiently high, to eliminate the possibility of fake diphoton pairs from soft radiation, or $\pi^0$ decays. Knowing that the emitted photons tend to be spatially well-isolated [20], a secondary requirement on each photon might entail some limit on the closeness of other physics objects, either in terms of $\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}$, or as a function of the electromagnetic ‘noise’ of secondary particles overlapping the photon shower in the EM calorimeter.

\section{2.5 Backgrounds}

\subsection{2.5.1 Dijet backgrounds}

A significant contributor to VBF dijet background is Higgs boson + 2 jet production via gluon-gluon fusion. In particular, the $gg \rightarrow H$ process can mimic VBF-like signal at NNLO, due to additional gluon emission occurring in both initial and final states. Consequently, fake signal events from the $gg \rightarrow H + 2$ jet process can lead to a contamination of true VBF signal on the order of $\sim 25\%$ [32]. In theory, one could place restrictions on the amount of real emissions occurring for a given dijet signal process to reduce the contamination from ggF events. In practice, however, this treatment induces large theory uncertainties [32], necessitating the avoidance of cuts in this unsafe region of phase space, and the application of appropriately large systematic uncertainties (see Section 6.1.1).
Figure 2.8: Parton-level kinematics of the leading (solid line) and subleading (dashed line) photons resultant from VBF Higgs production at $m_H = 125$ GeV, as generated by Powheg. The plots are shown prior to any kinematic cuts, and are scaled to the expected number of events from 21 fb$^{-1}$ of $\sqrt{s} = 8$ TeV data.
2.5 Backgrounds

2.5.2 Diphoton backgrounds

The ‘silver’ nature of the $\gamma\gamma$ channel is partially due to the overall knowledge of the $\gamma\gamma$ and $\gamma j$ background components. The irreducible $\gamma\gamma$ background components constitute real photon pairs from quark-antiquark annihilation (the Born process), gluon-gluon fusion (as mediated by fermion loops, otherwise known as the box process), and gluon-(anti)quark scattering processes leading to bremsstrahlung [33]. For each case, the photons can emerge either from the hard scattering process itself, or from parton fragmentation [34]. A particularly large component of the irreducible background is the loop-induced $g g \rightarrow \gamma\gamma + X$ channel,\(^2\) which makes up approximately 50% of the irreducible background [19].

Figure 2.9: Feynman diagrams of the irreducible components of the $H \rightarrow \gamma\gamma$ background [33].

The dominant component of the reducible $\gamma\gamma$ background originates in jets faking photons, which is approximately $\mathcal{O}(10^6)$ larger than the irreducible contributions from $q\bar{q} \rightarrow \gamma\gamma + X$ and $g g \rightarrow \gamma\gamma + X$ [19]. Based on Monte Carlo driven studies (using the Pythia [35] package), it is expected that after minimizing the reducible $\gamma\gamma$ background, the dominant source of background will originate in single and multiple $\pi^0 \rightarrow \gamma\gamma$ decays. Smaller contributions to the background ($\sim 4\%$ each) originate in quark bremsstrahlung, and electrons from $b, W, Z$ decays [33]. It is predicted that $\sim 20$ fb$^{-1}$ of data would be required to be able to observe a Higgs boson in the $120 - 130$ GeV range above the irreducible diphoton background [20].

\(^2\) Here, $X$ denotes a generic hadronic final state.
Chapter 3

The $H \rightarrow \gamma \gamma$ analysis in ATLAS

3.1 Photon reconstruction

Photon reconstruction in ATLAS proceeds in three primary steps: cluster building in the EM calorimeter, track building in the Inner Detector (for photons converted to $e^+e^-$ pairs), and track/cluster matching to build a final photon measurement. Photon clusters\(^1\) are built using a *sliding window* algorithm [36], wherein the calorimeter is treated as a grid of $\Delta \eta \times \Delta \phi$ cells. Cluster reconstruction begins with the search for a calorimeter seed across each longitudinal layer of the calorimeter, with a seed required to have transverse energy $E_T = m^2 + (\vec{p}_T)^2 > 2.5$ GeV within a $3 \times 5$ projective tower (although $m = 0$ for photons). Nearby calorimeter cells are used to form the cluster by stepping through each layer and assigning all cells within $N^\eta_{\text{cluster}} \times N^\phi_{\text{cluster}}$ of the seed position to the cluster.

Inner Detector tracks for converted photons are typically built using a seed produced through the combination of $\geq 3$ silicon layer ‘hits’, which are translated to space-points. These hits form a curved track that is extended through the volume of the Inner Detector, and matched to segments in the TRT [37], known

\(^1\) A *cluster* is a group of calorimeter cells in which a particle has deposited energy.
as an inside out track. TRT segments can also form a track seed, which is extrapolated backward into the silicon detectors to form space-points, producing an outside in track. Pairs of oppositely-charged tracks are then identified as potential converted photons using a vertex fitting procedure based on the fast-Kalman filtering method [31], which seeks to match the constructed tracks to a common conversion vertex. Single tracks matched to a conversion vertex can also be stored as potential converted photons.

In the last step of the reconstruction, clusters without matching tracks are directly classified as unconverted photon candidates. Converted photons are identified after all electrons have been reconstructed, and are defined as clusters matched to tracks originating from reconstructed conversion vertices. All reconstructed photons are then subject to energy calibration [31] to account for energy loss before entering the calorimeter, as well as cluster leakage.

3.2 Jet reconstruction

3.2.1 Cluster and jet reconstruction

Within ATLAS, the “4/2/0” topological clustering scheme [36] is used to form the clusters that make up calorimeter jets. Here, the ‘4’ refers to the fact that clusters are seeded by cells with a $4\sigma$ deviation from estimated levels of electronic and pileup noise [38]. A cluster is then formed around the seed using all cells with energy $2\sigma$ above the noise threshold (the ‘2’ component). Lastly, a single layer of cells, all immediately adjacent to the $2\sigma$ region, is added to the cluster, regardless of the level above the noise threshold (the ‘0’ component).

The topological clusters built using the 4/2/0 scheme are then used to form electromagnetic jets using the “anti-$k_t$ jet clustering” algorithm [39]. Starting with
the highest $p_T$ (or ‘hardest’) object in an event defined as the seed, each other topological cluster is judged by its distance to the seed, and the beam line, by a factor inversely proportional to its transverse momentum. Well-isolated seeds (i.e. no other hard objects nearby) will combine with nearby low $p_T$ (or ‘soft’) objects to form a perfectly conical jet. Conversely, nearby hard objects will either combine with the seed into a larger jet, or individually absorb softer objects relative to their distance and $p_T$, forming two smaller, amorphous jets.

### 3.2.2 Jet energy measurement and correction

One issue that arises during jet reconstruction is the presence of pileup effects, which can complicate both cluster formation, and energy measurement. In ATLAS terminology, pileup refers to the background interactions that occur before, during, or after the collision (or crossing) of two bunches of protons. Two distinct forms of pileup are possible: those that occur at the same time as a given triggered event (in-time pileup), and those that occur in surrounding bunch crossings, and not necessarily at the same time (out-of-time pileup) [40]. Both in-time and out-of-time pileup can result in poor jet reconstruction in the ATLAS calorimeters, especially with regard to the EM calorimeter, where the time window needed to integrate the signal and produce a reading can potentially span several bunch crossings [41].

To remedy any deficiencies in response, pileup modelling and subtraction techniques are used to remove pileup contributions from jet energy measurements, and jet energy scale (JES) correction factors are used to correct jet energy and $\eta$ – a combination termed the ‘EM + JES’ scheme. The pileup subtraction method applies $E_T$ and $\eta$-based corrections to jets using the jet area $A_j$, which quantifies the susceptibility of a given jet to pileup, as well as the median $p_T$ density $\rho$, which is
3.2 Jet reconstruction

a measure of the amount of pileup in a given event [40]. The corrected transverse momentum is then calculated as \( p_T^{\text{corr}} = p_T - \rho A \). The direction of each jet then is modified to ‘point’ back to the primary interaction vertex, with JES correction factors (derived from data-Monte Carlo comparisons) applied afterward to restore the energy and direction to what is expected at hadronic energy scales [42].

Additional improvements are made using *in situ* techniques wherein the transverse momentum balance of, e.g. single jet + single Z boson events in data are used to derive correction factors not accounted for in the initial EM + JES scheme. In addition, the “\( \eta \)-intercalibration” technique uses the transverse momentum balance of dijet events to derive \( p_T \) and pseudorapidity-dependent corrections to the jet response [42]. In particular, imbalances in \( p_T \) for dijet events are compared in data and Monte Carlo samples, with the observed differences used to derive a series of scale factors that restore \( p_T \) balance in multi-jet systems.

3.2.3 Jet vertex fraction

Since several \( pp \) collisions can occur for a given bunch crossing, a reconstructed jet can originate from a number of primary vertices along the beamline, or even from pileup interactions. To aid in identifying jets originating from true hard scattering interactions, a quantity known as the jet vertex fraction (JVF) is calculated for reconstructed jets within Inner Detector acceptance (\( |\eta| < 2.37 \)). For the \( i^{\text{th}} \) jet matched to tracks emerging from vertex \( \text{vtx}_j \) in the Inner Detector, the JVF is given by [41],

\[
JVF(jet_i, \text{vtx}_j) = \frac{\sum_k p_T \left( \text{trk}_{jet}^k, \text{vtx}_j \right)}{\sum_n \sum_l p_T \left( \text{trk}_{jet}^l, \text{vtx}_n \right)}.
\]  

The JVF ranges between 0 and 1, with \( JVF = 1 \) indicating that jet \( i \) is perfectly
matched to the tracks from \(vtx_j\). Conversely, \(JVF = 0\) may imply either that a jet originates from a different hard scattering vertex, or from pileup collisions. Calorimeter jets which fall outside of the fiducial tracking region (i.e. in the forward region), or those which have not been matched to tracks, are assigned \(JVF = -1\).

### 3.3 Diphoton candidate selection

For consideration in the \(H \rightarrow \gamma\gamma\) analysis, an event must first pass the online trigger selection: a set of criteria, evaluated during data taking, that are correlated with the presence of diphoton events in the ATLAS detector. The \(H \rightarrow \gamma\gamma\) analysis uses a three-tier trigger system, with the first requirement being the presence of > 1 energy deposits in the ECAL with transverse energies of at least 35 and 25 GeV for the leading and subleading deposits, respectively. The second and third requirements examine the shape of the energy deposits to check for consistency with photon-initiated EM showering [43]. If all the criteria are met, the event is then saved as a _loose_ diphoton candidate.

As a first step for the offline analysis, the leading and subleading reconstructed photons are required to have transverse momenta of at least 40 GeV and 30 GeV, respectively, and be in the fiducial region, \(|\eta| < 2.37\) (excluding the crack region, \(1.37 < |\eta| < 1.56\)) [43]. Furthermore, it is required that both photons meet stringent EM shower shape criteria (called _tight_ identification) [9], which constitute more specific characteristics of EM showers that have been initiated by photons, thus reducing fake photons due to hadronic activity or leptons. For example, electrons tend to radiate photons before entering the calorimeter, leading the EM shower to start earlier, leaving a wider pattern of energy deposition in the presampler. Photon showers, conversely, tend to be more long and narrow, with converted
3.3 Diphoton candidate selection

photon shower shapes falling in-between in terms of shower width and length.

Beyond cuts on $p_T$ and $|\eta|$, photons are required to be well-isolated in both the calorimeter, and Inner Detector (for single and double conversions). For the calorimeter, isolation is defined in terms of $E_T$, such that the sum of the transverse energy of all topological clusters within $\Delta R < 0.4$ of the photon must be less than 6 GeV [43], excluding the energy deposits belonging to the photon itself. In the case of single- or double-track conversions, the additional requirement on track isolation is defined in terms of track $p_T$, such that the scalar sum of the transverse momenta of all tracks within $\Delta R < 0.2$ must be less than 2.6 GeV [43].

Lastly, for experimental data and any background samples, a final cut on the $\eta$-corrected diphoton mass, $m_{\gamma\gamma}$, is applied, requiring $100 < m_{\gamma\gamma} < 160$ GeV. The correction to the $\eta$ of the leading and subleading photons for each $\gamma\gamma$ event is made based on the $z$-position of the diphoton primary vertex, and $\eta$ from the first sampling layer of the ECAL. While often small, this correction is necessary to have the direction of the photon cluster properly ‘point’ back to the primary interaction vertex for an event. Using a parameterization of calorimeter radius based on the first sampling layer $\eta$, the corrected photon $\eta$ is calculated as,

$$\eta = \sinh^{-1}(\frac{z_{\text{Front}} - z_{\text{PV}}}{R_{\text{Front}}}).$$

(3.2)

Here, $z_{\text{PV}}$ is the $z$ coordinate of the primary interaction vertex, $R_{\text{Front}}$ is the radius into the first sampling layer of the ECAL, and $z_{\text{Front}}$ is the $z$ position of the cluster in the first sampling layer.
3.4 Dijet candidate selection

Since the detection of any vector boson fusion event relies heavily on the proper measurement of the outgoing quark jets, imposing quality cuts on reconstructed jets is also necessary for this analysis. As a baseline, all reconstructed jets are required to have a transverse momentum greater than 25 GeV (30 GeV) for $|\eta| < 2.5$ ($2.5 < |\eta| < 4.5$) to be considered. In addition, $|JVF| > 0.25$ is required of all jets in the central region which are matched to tracks emerging from the primary interaction vertex. Finally, in order to prevent the misidentification of photons and electrons as hadronic jets and vice versa, a cut is placed on the spatial separation of calorimeter objects, requiring $\Delta R > 0.4$ between all physics objects.\(^2\) Worth noting is that the $\eta$, $\phi$ values used in this cut are taken from the EM scale jet, negating any changes in direction from applying jet energy scale corrections. Once the selection is complete, the two highest $p_T$ jets are used for further categorization.

3.5 Event categorization

Events passing initial diphoton selection are divided into a number of exclusive categories meant to isolate events with photon, jet, and lepton properties characteristic of vector boson fusion, associated production ($VH$), or gluon-gluon fusion [43]. The categorization occurs in a hierarchical manner, with associated production-enriched categories taking top precedence. These categories exploit the presence of electrons or muons, low-mass dijet pairs, or missing transverse energy to search for Higgsstrahlung-like $H \rightarrow \gamma\gamma$ signal. Any diphoton + dijet events failing the $VH$ criteria are then considered for the secondary, VBF-enriched

\(^2\) Though not relevant to this analysis, quality cuts are also applied for electrons and muons, which are detailed in Ref. [16].
3.5 Event categorization

Any diphoton + dijet events that fail the VBF criteria, or events with < 2 good jets, are then placed in an inclusive categorization based on $p_T \hat{t}$, photon $|\eta|$, and the number of converted photons in the pair.

![Diagram of $p_T \hat{t}$](image)

Figure 3.1: Explanatory diagram of $p_T \hat{t}$, the component of the diphoton transverse momentum $\vec{p}_{\gamma\gamma}^T$ transverse to the diphoton thrust axis $\hat{t}$ [44].

Note that $p_T \hat{t}$ is the component of the diphoton transverse momentum $\vec{p}_{\gamma\gamma}^T$ transverse to the diphoton thrust axis $\hat{t}$ (see Figure 3.1), with its magnitude defined as [45, 46],

$$p_T \hat{t} \equiv |\vec{p}_{\gamma\gamma}^T| = |\vec{p}_{\gamma\gamma}^T \times \hat{t}|,$$

where

$$\hat{t} = \frac{\vec{p}_{\gamma\gamma}^T - \vec{p}_{\gamma\gamma}^T}{|\vec{p}_{\gamma\gamma}^T - \vec{p}_{\gamma\gamma}^T|}. \quad (3.3)$$

The category relevant to this analysis is that which is enriched with VBF-like events. In prior analyses, this category was defined entirely using rectangular cuts on quantities sensitive to diphoton + dijet systems, like $m_{jj}$, $\Delta \eta_{jj}$, and $\Delta \phi_{\gamma\gamma jj}$. However, for the most recent iteration of the $H \rightarrow \gamma\gamma$ analysis, the VBF category was split into two subcategories (high and low VBF-like), with the selection process motivated entirely by multivariate analysis techniques – specifically, the usage of boosted decision trees. Both the derivation of this new mode of categorization, and a study of its efficacy, are detailed in Chapters 4 and 5.
Chapter 4

Introducing the VBF multivariate analysis

4.1 Boosted decision trees

The boosted decision tree (BDT) method is a form of multivariate analysis offered in the Toolkit for Multivariate Analysis (TMVA) package [47]. A boosted decision tree uses multiple binary decision trees (as shown in Figure 4.1) to form a robust, statistically stable classifier which discriminates signal from background events in a data sample. Each constituent binary tree is built from a list of user-provided discriminating variables, and trained using distinct, non-overlapping signal and background samples. The function of each tree is to use a series of ‘yes/no’ decisions to classify individual events as signal-like or background-like.

The result from each tree (the leaf node) is then combined via weighted average into a single discriminant, known as the BDT score [47]. Typically, this score varies on [-1, +1], with background-like events assigned low (or negative) scores, and signal-like events assigned highly positive scores. The ‘boosting’ aspect derives from the fact that the weight of each tree is proportionate to its rate of
misclassification (i.e. poor signal/background discrimination yields a low weight, and vice versa).

In particular, the gradient boosted decision tree method [48] was chosen for this analysis, which treats the individual trees as a sum of terms in a function expansion approach. That is to say, given a set of input variables $\mathbf{x} = (x_1, x_2, ..., x_n)$, one can imagine a map to an output variable $y$ (in this case, the BDT score) through a function $y_i = \hat{F}(\mathbf{x}_i)$. The gradient boosting method, in essence, seeks to develop a function $F(\mathbf{x})$ that estimates $\hat{F}(\mathbf{x})$ by minimizing the expected value of a binomial log-likelihood loss function $L(y, F(\mathbf{x}))$. This method of boosting was chosen for the VBF multivariate analysis because it tends to isolate signal- and background-like events in opposite ends of the BDT score spectrum by default. Furthermore, gradient-boosted decision trees are also less susceptible to overtraining [47], wherein the BDT learns only from a narrow region of phase space, or statistical fluctuations.

There are two interrelated tasks to consider when developing a boosted decision tree-based analysis. Namely, one must select signal and background training samples that both model well the underlying physics of each system, and are sizable enough to avoid shaping the BDT response with fluctuations or overtraining. Furthermore, it is necessary to choose training variables that are known to dif-
ferentiate signal from background, and show no bias towards any value of the quantity (or quantities) of interest in the analysis. The confluence of these two elements will produce an optimized, robust BDT-based analysis.

4.2  Signal and background modelling

4.2.1  Signal modelling

The signal Monte Carlo samples used in the $H \to \gamma\gamma$ analysis are typically generated using the Powheg package [29, 30], which simulates Higgs boson signal events using exact NLO QCD matrix elements. The physical interactions that Powheg generates are deemed parton-level, as they describe only the hardest emission for a given process. All other higher-level processes (showering, parton splitting, gluon radiation, etc.) require the usage of a dedicated shower Monte Carlo (SMC) program [29]. For this reason, the parton-level output is interfaced with Pythia [35], an NLO generator that contains modules to simulate parton showering and hadronization, thus leading to truth-level information. Finally, this set of truth-level particles is run through a full simulation of the ATLAS detector using GEANT4 [49, 50], and reconstructed by the same offline ATLAS software used for data reconstruction [51].

4.2.2  Background modelling

Background estimation techniques in the $H \to \gamma\gamma$ analysis can originate both in simulated data, as well as data-driven methods [19], owing to the large reducible background from $\gamma j$ events. The data-driven methods typically form an estimate based on the sideband data,\footnote{Diphoton event candidates with $m_{\gamma\gamma} > 130\,\text{GeV}$ \& $m_{\gamma\gamma} < 120\,\text{GeV}$.} or through the reversal of a cut from the diphoton cut.
flow (as detailed in Section 3.3). Specifically, background samples can be created from a reversal of the tight identification cuts (known as inverted ID samples), or the calorimeter + track isolation cuts (known as inverted isolation).

Because any Higgs boson-like resonance will be directly observable from the $m_{\gamma\gamma}$ spectrum as a ‘bump’, one may also estimate the background distribution by fitting an analytic function to the data region(s) where the resonance is not observed. In particular, shape estimation for the $\gamma\gamma$ background is treated using an exponential distribution or Bernstein polynomial, with the estimate derived using a fit to the left and right signal region sidebands (defined as $100 < m_{\gamma\gamma} < 120$ GeV and $130 < m_{\gamma\gamma} < 160$ GeV, respectively) [34]. The ultimate goal of this sideband fit is to allow an accurate estimate of the background in the signal region, $120 < m_{\gamma\gamma} < 130$ GeV, which can be used to quantify the statistical significance of any signal-like trend in the $m_{\gamma\gamma}$ spectrum.

Additionally, several $\gamma\gamma$, $\gamma j$ process Monte Carlo generators are available within ATLAS for background estimation, which may simulate parton-level physics processes (typically at NLO), secondary interactions (such as parton showering), or perform both. In the $H \rightarrow \gamma\gamma$ analysis, Sherpa [52, 53] is typically used for background studies. Sherpa is an $O(\alpha^2)$ generator that simulates both $\gamma\gamma$ (i.e. Born, Box) and $\gamma j$ background processes. Hard scattering interactions are simulated using a matrix-element generator, which are then interfaced with modules that deal with initial- and final-state parton showering, multiple parton interactions, parton hadronization, and hadronic decay [53].

### 4.2.3 Monte Carlo event weights

Most high-level Monte Carlo generators rely on approximate or fixed-order solutions to the equations governing the hard scattering and showering interactions
4.3 Input sample selection

4.3.1 Samples available for the multivariate analysis

Any data-driven aspects of the derivation and optimization of this BDT-based analysis were based on (∼13.0 ± 3.6%) fb⁻¹ of proton-proton collision data collected at the LHC, recorded during 2012. The collisions occurred with a centre-of-mass energy of $\sqrt{s} = 8$ TeV, with the average number of primary vertices being 20.0 [56]. The standard diphoton trigger outlined in Section 3.3 was used in the analysis, with the final trigger efficiency being > 99% for the entire data set. Furthermore, for the purposes of this analysis, diphoton pairs were allowed to have a

observed in particle accelerators. So, when combined with the various detector effects and potential mismodelling of variables which are unaccounted for in full detector-simulated Monte Carlo samples, there is potential for non-trivial disagreement between experimental data and simulation.

One option to mitigate these differences is to assign data-driven weights, modifying the shapes of the Monte Carlo distributions to better match those observed in data. Most notably, a pileup weight is applied for all Monte Carlo samples, the function of which is to bring the distribution of the average bunch crossing multiplicity ($\langle \mu \rangle$) in agreement between data and Monte Carlo distributions, to account for variances in luminosity and pileup between data runs in ATLAS. A reweighting based on the hard scatter z-vertex position is also applied, which is used to correct the observed difference in z-vertex spread between data and Monte Carlo samples. For ggF samples, an interference weight is also applied, which accounts for destructive interference between the $gg \rightarrow \gamma\gamma$ background and the $gg \rightarrow H \rightarrow \gamma\gamma$ process [54, 55].
mass in the range $100 < m_{\gamma\gamma} < 170$ GeV.

Within this analysis, Monte Carlo samples were used to simulate both signal and background processes. The signal Monte Carlo samples used in the multivariate analysis, corresponding to VBF $H \rightarrow \gamma\gamma$ and ggF $H \rightarrow \gamma\gamma$ processes, were generated using Powheg + Pythia8 (as described in Section 4.2.1), with $m_H = 125$ GeV. The background Monte Carlo samples in this analysis primarily involved Sherpa $\gamma\gamma$, $\gamma j$ events (see Section 4.2.2), wherein each event contained a diphoton pair with $m_{\gamma\gamma} > 80$ GeV, and up to three parton jets generated from matrix elements. Sideband data samples were also available to use as background, along with inverted ID and inverted isolation background data, as summarized in Table 4.1.

4.3.2 Signal, background sample configuration

The first task in developing the BDT-based analysis was to select samples that could model well the VBF $H \rightarrow \gamma\gamma$ signal, as well as the diphoton and dijet components of the background. Before any specific tests were made, a series of ground rules about sample composition was decided on, in order to ensure the most unbiased analysis possible:

- **All training/testing samples must be orthogonal.** In this analysis, the requirement on sample orthogonality was two-fold. Firstly, it was required that there be no overlapping events among the signal and background samples used to build and test the BDT, which would otherwise lead to poor BDT training and ineffective signal event selection. Secondly, it was required that the training and testing samples be orthogonal, in order to prevent the introduction of biases in classifying potential VBF events.
• All diphoton + dijet data samples must have the signal region blinded. Any diphoton + dijet candidate data events with $120 < m_{\gamma\gamma} < 130$ GeV were excluded from the analysis, in order to limit any potential biases on $m_{\gamma\gamma}$. Note that this blinding does not apply to background data, or any Monte Carlo samples.

• Events in associated production-enriched categories must be excluded. In order to ensure that each sample attained the highest possible purity of VBF-like events, any diphoton + dijet events in the $VH$-enriched categories (see Section 3.5) were immediately vetoed from entering any training or testing sample.

With these rules established, it was possible to move on to select appropriate signal and background samples for training the BDT.

4.3.3 Background training sample selection

The fact that the background for the $H \rightarrow \gamma\gamma$ process encompasses pairs of real and fake photons presents a unique challenge for selecting a proper background sample. In order to successfully emulate the true diphoton background after event selection, a sample must accurately model the $\gamma\gamma$, $\gamma j$, and $jj$ components of the background. To this end, a variety of background estimation methods were tested, with the different types of models explained in Table 4.1.

The initial hope was to be able to use a data-driven estimation for training, which would remove the need to individually model the different $H \rightarrow \gamma\gamma$ background components. One option was the direct usage of a portion of the data sidebands for training, with the other being the usage of inverted isolation or inverted ID background data. It was found that after diphoton + dijet selection, $VH$ category vetoes, and division into training/testing samples, the statistics of
4.3 Input sample selection

<table>
<thead>
<tr>
<th>Background type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverted isolation</td>
<td>Reverse the calorimeter/track isolation cut in the diphoton candidate selection process (see Section 3.3), creating a sample of background $\gamma\gamma$ and $\gamma j$ candidates. Apply 2-jet selection to isolate potential dijet events.</td>
</tr>
<tr>
<td>Inverted ID</td>
<td>Reverse the tight photon ID cut from the diphoton candidate selection process, and apply dijet selection. Events where both photons fail the ID cut are denoted $jj$, with a pass-fail pair (or vice versa) denoted $\gamma j$ (or $j\gamma$).</td>
</tr>
<tr>
<td>Sherpa</td>
<td>Apply diphoton + two-jet selection to Sherpa Monte Carlo samples, requiring $100 &lt; m_{\gamma\gamma} &lt; 170 \text{ GeV}$.</td>
</tr>
<tr>
<td>Data sidebands</td>
<td>Use dijet events in the left and/or right data sidebands as a direct estimate of the irreducible $\gamma\gamma$, $\gamma j$, and $jj$ background components.</td>
</tr>
</tbody>
</table>

Table 4.1: Different types of diphoton + dijet backgrounds tested, and their composition.

The data sidebands were too meager to both train a boosted decision tree, and perform the final statistical analysis ($\sim 1200$ events per orthogonal sample for each task). Therefore, data sidebands were eliminated as a viable option for training a BDT.

The possible background models were then limited to Sherpa Monte Carlo samples, inverted isolation, and inverted ID samples. A comparison of the possible background models with the data sidebands is shown in Figure A.2 for several relevant photon and jet kinematic variables. Based on these comparisons, it was decided that inverted isolation data was the optimal choice of background model to grow the analysis, as this model provided the best agreement with sideband data, was orthogonal to all other samples by construction, and yielded adequate statistics for producing training/testing samples.
4.4 Input variable selection

4.4.1 Deriving a list of discriminating variables

Beyond selecting proper training samples, a primary concern in the development of any BDT-based analysis is the choice of discriminating variables that agree well with theoretical predictions of the process under study, provide good discrimination between the desired signal and unwanted processes, and also sidestep any undesirable biases. To this end, all potential classifiers were required to fulfill a series of preliminary criteria before training to ensure a robust, unbiased analysis. First, each variable was required to have little to no correlation to the diphoton mass, $m_{\gamma\gamma}$. This relation was determined through the calculation of a linear correlation coefficient (LCC) between $m_{\gamma\gamma}$ and a given variable, with the scale ranging from $-100\%$ (completely anticorrelated) to $+100\%$ (completely correlated). This criterion was of singular importance. If it was not met, the selection process would be immediately biased towards events with $m_{\gamma\gamma} \sim 125$ GeV.

Second, for the remaining variables uncorrelated to $m_{\gamma\gamma}$, each quantity was to be ranked according to its signal/background discriminating power, as given by the separation $\langle S^2 \rangle$, and the variable importance. The former quantity is entirely shape-based, and is calculated as \[\langle S^2 \rangle = \frac{1}{2} \sum_{i=1}^{N_{\text{bins}}} \frac{(s_i - b_i)^2}{s_i + b_i},\tag{4.1}\]
where $s$ and $b$ are signal and background probability density functions (PDFs) for a given classifier. The variable importance, by contrast, is a BDT-specific quantity, and quantifies how frequently a variable is used to split nodes (i.e. make decisions). It is computed for a given variable as the frequency of its use in node splitting, weighted by both the number of events in each node, and the separation
gain-squared it achieved [47]. If a given input variable demonstrated both high separating power in terms of \(\langle S^2 \rangle\) and importance, then it was considered for further usage in training the BDT.

A tertiary criterion for each potential classifier was a low correlation (expressed in terms of LCCs) to other candidate variables. Unlike other multivariate analysis techniques such as neural networks, uncorrelated input variables are not a fundamental requirement for BDTs [47], as the presence of two highly correlated classifiers results only in suboptimal signal discrimination, rather than a bias in the distributions of the selected events.

With a formal set of selection criteria established, the variable selection process began by training with an initially large list of variables, which would highlight the classifiers that were used most frequently and/or effectively to separate signal from background. The initial list of classifiers was composed of 23 variables, spanning a range of single jet, single photon, and dijet/diphoton quantities:

\[
p_{T}(\text{jet}1,2,3), \eta_{\text{jet}1,2,3}, \phi_{\text{jet}1,2,3}, JVF_{\text{jet}1,2,3}, \eta_{\text{Zepp jet}3}, \eta_{\text{Zepp} \gamma\gamma}, \min(\Delta R_{j\gamma}), m_{jj}, \Delta\eta_{jj}, \Delta\phi(jj), \Delta\phi_{\gamma\gamma jj}, \Delta p_{T}(jj), \Delta p_{T}(\gamma\gamma), p_{T\gamma\gamma jj}, p_{T\gamma}
\]

**Note:** Definitions of these variables are found in Table A.1.

The initial BDT training was then performed using VBF \(H \rightarrow \gamma\gamma\) Monte Carlo samples with \(m_H = 125\,\text{GeV}\) as a signal training sample, and 2012 isolated data as the background training sample. Each sample was split up into training/testing subsets according to event number, as follows:

\[
\text{Event number mod } 2 = \begin{cases} 
0 & \rightarrow \text{Testing} \\
1 & \rightarrow \text{Training}
\end{cases} \quad (4.2)
\]
By construction, the two samples are orthogonal, thereby avoiding any bias originating from training events present in the testing sample, or vice versa.

Furthermore, it was decided to bin the value of $\Delta \phi_{\gamma \gamma jj}$ on $[2.94, \pi]$, such that any value of $\Delta \phi_{\gamma \gamma jj}$ above 2.94 would be set to $\Delta \phi_{\gamma \gamma jj} = 2.95$ (see Figure 4.2). This binning was performed in order to prevent the BDT from using that particular region of phase space to make cuts, making it possible to avoid the large (and potentially unknowable) systematic uncertainties associated with restricting additional emissions in the $gg \rightarrow H \rightarrow \gamma \gamma + 2$ jet process as $\pi - \Delta \phi_{\gamma \gamma jj} \rightarrow 0$ [32], as explained in Section 6.1.1. As given by TMVA, the top ten most discriminating variables in terms of signal/background separation and importance index are shown in Table 4.2.

Somewhat unsurprisingly, the variables with consistently high rankings in terms of separation and importance are those relating to the spatial separation and energy of the jets and photons. For example, $\Delta \eta_{jj}$ and $m_{jj}$ both rank among the top three variables in both metrics, reflecting the characteristic rapidity gap and high energy quark jets in the VBF process. Furthermore, the diphoton $p_{Tt}$ is
4.4 Input variable selection

Table 4.2: The top ten most discriminating variables for the preliminary VBF vs. inverted isolation BDT training, ranked according to signal/background shape separation and importance.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th>Separation</th>
<th>Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\Delta \eta_{jj}$</td>
<td>0.386</td>
<td>$p_{Tt}$</td>
<td>0.127</td>
</tr>
<tr>
<td>2</td>
<td>$m_{jj}$</td>
<td>0.383</td>
<td>$m_{jj}$</td>
<td>0.110</td>
</tr>
<tr>
<td>3</td>
<td>$\eta_{jet2}$</td>
<td>0.173</td>
<td>$\Delta \eta_{jj}$</td>
<td>0.087</td>
</tr>
<tr>
<td>4</td>
<td>$p_{Tt}$</td>
<td>0.160</td>
<td>$\phi_{jet2}$</td>
<td>0.078</td>
</tr>
<tr>
<td>5</td>
<td>$p_{T\gamma\gamma_{jj}}$</td>
<td>0.152</td>
<td>$\Delta p_{T(jj)}$</td>
<td>0.073</td>
</tr>
<tr>
<td>6</td>
<td>$JVF_{jet2}$</td>
<td>0.113</td>
<td>$\Delta \phi(jj)$</td>
<td>0.061</td>
</tr>
<tr>
<td>7</td>
<td>$\eta_{jet1}$</td>
<td>0.109</td>
<td>$\Delta p_{T(\gamma\gamma)}$</td>
<td>0.057</td>
</tr>
<tr>
<td>8</td>
<td>$\Delta \phi_{\gamma\gamma_{jj}}$</td>
<td>0.101</td>
<td>$p_{T(jet2)}$</td>
<td>0.050</td>
</tr>
<tr>
<td>9</td>
<td>$p_{T(jet2)}$</td>
<td>0.073</td>
<td>$\eta_{jet1}$</td>
<td>0.049</td>
</tr>
<tr>
<td>10</td>
<td>$\eta_{Zep}^{\gamma\gamma}$</td>
<td>0.070</td>
<td>$p_{T(jet1)}$</td>
<td>0.048</td>
</tr>
</tbody>
</table>

also a high ranking variable, which shows the tendency of the final state photons to be well-separated (or back-to-back) in $\phi$.

An interesting outcome is that while these dijet and diphoton variables are among the top 3 – 4 classifiers in each list, the rest of the list is almost random, with few variables having a consistent ranking in terms of both separating power and importance (with the exception of $\eta_{jet1}$). Therefore, it was necessary to figure out which combination of classifiers truly drove the selection process, as opposed to those which pull rank only when in the presence of more discriminating cuts.

4.4.2 Optimization of input variables

Having derived an initial list of potential classifiers as detailed in Section 4.4.1, the next step was to define a benchmark set of variables that provided the best signal discrimination given an initial BDT configuration. Though $\langle S^2 \rangle$ and importance
calculations give an indication of how a variable performs in identifying signal and background in isolation, it is insufficient to judge a set of classifiers based on these separation criteria alone. In particular, because the final analysis is defined by the search for a resonance in the diphoton mass spectrum, it is more relevant and intuitive to use a metric that has a direct relation to the probability of particle discovery. For this reason, it was chosen to evaluate the list of benchmark variables as the set of classifiers optimizing the approximate VBF significance for \( s \ll b \),

\[
S \approx \frac{s_{\text{VBF}}}{\sqrt{s_{\text{VBF}} + s_{\text{ggF}} + b + \sum (\text{syst. uncert.})^2}},
\]

where \( s_{\text{VBF}}, s_{\text{ggF}}, \) and \( b \) are integrals of the \( m_{\gamma\gamma} \) distributions of the VBF and ggF \( H \rightarrow \gamma\gamma \) signal Monte Carlo and background samples (respectively) over a given mass range. Here, \( \sum (\text{syst. uncert.})^2 \) is the sum in quadrature of the different systematic uncertainties of the analysis (see Chapter 6 for details). This term is unique to the multivariate analysis, and serves as a ‘penalty’ for the choice of variables with high associated theoretical and/or experimental uncertainties.

Two forms of the approximate significance were calculated, based on different \( m_{\gamma\gamma} \) regions. The window significance was defined using integrals of the signal and background distributions on \([123, 127]\) GeV, while the shape significance was obtained using integrals on the entire mass range. Owing to the narrow width of the Higgs boson resonance in the VBF signal Monte Carlo sample, the variable set maximizing shape significance should also maximize the window significance. With a metric for judging performance determined, the following approach was taken to derive a benchmark list of variables:

1. Based on the results of Section 4.4.1, a list of \( 8 - 10 \) variables was created, with which a new training was performed using VBF \( H \rightarrow \gamma\gamma \) signal Monte Carlo and inverted isolation background samples. Classifiers for this training
were required to have high separating power, and be frequently/effectively used within the initial BDT.

2. BDTs were trained using every permutation of \( k \) variables in the \( n \)-variable list (where \( 2 \leq k \leq n \)), with each trained BDT applied to each testing sample (constituting VBF, ggF signal Monte Carlo samples, and sideband data) separately. In total, for each \( k \) variables chosen, \( \binom{n}{k} \) BDT trainings were performed.

3. For each permutation, using the resultant BDT score distributions for each sample, a scan was made over the range of possible cuts on the discriminant. At each step of the scan, shape and window significance calculations were performed. The combination of variables maximizing the approximate significance values was taken to be the benchmark.

Of particular note is that the amount of VBF signal rejected by each permutation of variables and BDT score cut was also considered during the optimization process, with the signal rejection obtained by translating the BDT score distribution to a flat distribution of 5\% VBF rejection quantiles.\(^2\) That is to say, instead of placing a cut on the BDT score itself by requiring BDT score > \( X \), the \( n^{th} \) cut is more aptly described as requiring,

\[
\text{BDT score} > X', \text{ such that } \int_{-1}^{X'} \frac{dN_{\text{VBF}}}{dX} \, dX = n \times 0.05N_{\text{VBF}}, \text{ where } n \in [0, 20].
\]

The shape and window significance values for the cases of BDT score > \( X' \) and BDT score < \( X' \) were added in quadrature to form the final significance, although the contribution from the latter case is heavily suppressed by background, such

\(^2\) Scanning over the range of BDT scores to produce a cut should yield the same optimal significance, hence why we don’t need or use the BDT score directly.
that $S_n \approx S(\text{BDT score} > X')$. Ultimately, it was decided to test two different sets of variables, with each having a mixture of single jet, dijet/diphoton, and dijet + diphoton system quantities:

**Selection A** (10 total)
- $\eta_{\text{jet}1}, \eta_{\text{jet}2}, m_{jj}$
- $\Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj}, p_{Tt}$
- $p_{T\gamma\gamma jj}, \Delta p_T(jj)$
- $\Delta \phi(jj), \Delta p_T(\gamma\gamma)$

**Selection B** (8 total)
- $\eta_{\text{jet}1}, \eta_{\text{jet}2}$
- $m_{jj}, \Delta \eta_{jj}$
- $\Delta \phi_{\gamma\gamma jj}, p_{Tt}$
- $\eta_{\gamma\gamma}^{\text{Zapp}}, \min(\Delta R_{j\gamma})$

### 4.4.3 Final list of input variables

After comparing the shape and window VBF significance values for each permutation of variables, as summarized in Tables 4.3 and 4.4, it was found that the following six variables consistently performed the best at discriminating signal from background:

- $m_{jj}, \Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj}, \eta_{\text{jet}1}, \eta_{\text{jet}2}, p_{Tt}$

Each of these variables relates directly to the spatial separation of either the dijet or diphoton systems. In particular, as shown in Table 4.3, when restricted to selecting $2 \rightarrow 4$ variables, the dijet mass ($m_{jj}$) and diphoton thrust $p_{Tt}$ are continuously the best variables for discriminating VBF $H \rightarrow \gamma\gamma$ signal from background. Furthermore, these results reinforce the hierarchy of variables shown in Table 4.2, which also point to $m_{jj}$ and $p_{Tt}$ as being highly discriminating both in terms of shape separation and importance.

One interesting outcome is the discriminating power of $\Delta \phi_{\gamma\gamma jj}$, which initially did not rank very highly in terms of separating power/importance on its own. However, when used in confluence with several other variables, this dijet + dipho-
4.4 Input variable selection

Figure 4.3: Distributions of $m_{\gamma\gamma}$ for the optimal permutations of variables (passing the BDT cut) derived from ‘Selection A’ (top) and ‘Selection B’ (bottom).
4.4 Input variable selection

Figure 4.4: Matrices of the linear correlation coefficients for the optimized significance of ‘Selection A’ for signal (left) and background (right).

The correlation matrices for signal and background in ‘Selection A’ are shown in Figure 4.4, where ±100% represents total linear correlation (+), or anti-correlation (−). Since the optimized list of variables for ‘Selection B’ is a subset of those in ‘Selection A’, the same correlation coefficients apply. Aside from the relations between $m_{jj}$ and $\Delta \eta_{jj}$, and the leading/subleading jet pseudorapidity, the correlation between each variable is reasonably small. This result suggests that the regions of phase space being probed by each variable have fairly little overlap, leading to more efficient signal/background discrimination.

After the variable optimization procedure, other VBF MVA group members
4.4 Input variable selection

found that adding a preliminary cut (or *precut*) on the minimum pseudorapidity difference between the leading and subleading jets ($\Delta \eta_{jj} > 2$) increased the discriminating power of the variables selected using the optimization process [56]. Therefore, an additional selection requirement was placed on the dijet + diphoton samples, such that each event must have $\Delta \eta_{jj} > 2$ to be included in any training/testing sample.

### 4.4.4 Kolmogorov-Smirnov (overtraining) tests

![Figure 4.5: (a) Results of Kolmogorov-Smirnov tests performed by TMVA for the signal/background training and testing samples. (b) A comparison of the BDT score distributions of the inverted isolation and data sideband samples after BDT training and application.](image)

Kolmogorov-Smirnov tests are performed automatically by TMVA, and indicate the probability that the training and testing samples used to model signal (or background) emerged from the same parent population. In general, high K-S probabilities imply less overtraining, and consequently, more effective signal discrimination. For the purposes of this analysis, a K-S test probability of $> 0.1$ was required of a given training [56]. Figure 4.5 shows the results of Kolmogorov-Smirnov tests performed on the signal and background samples, showing good...
### 4.4 Input variable selection

<table>
<thead>
<tr>
<th>k</th>
<th>Optimal variables</th>
<th>Optimal significance</th>
<th>VBF rejection (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( m_{jj}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.842</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>( m_{jj}, p_{T1}, p_{T\gamma jj} )</td>
<td>0.904</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>( m_{jj}, \Delta \eta_{jj}, p_{T1}, p_{T\gamma jj} )</td>
<td>0.913</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>( \eta_{jet1}, m_{jj}, \Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.9293</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>( \eta_{jet1}, \eta_{jet2}, \Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.915</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>( \eta_{jet1}, \eta_{jet2}, \eta_{\gamma\gamma}, \Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj}, p_{T1}, \min(\Delta R_{j\gamma}) )</td>
<td>0.887</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 4.3: Evolution of optimal significance when selecting \( k \) variables only from 10-variable ‘Selection A’. The significance shown is calculated using the shape method, with statistical and fit uncertainties included as penalty terms.

<table>
<thead>
<tr>
<th>k</th>
<th>Optimal variables</th>
<th>Optimal significance</th>
<th>VBF rejection (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( m_{jj}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.842</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>( m_{jj}, p_{T1}, p_{T\gamma jj} )</td>
<td>0.904</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>( m_{jj}, \Delta \eta_{jj}, p_{T1}, p_{T\gamma jj} )</td>
<td>0.913</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>( \eta_{jet1}, m_{jj}, \Delta \eta_{jj}, p_{T1}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.9293</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>( \eta_{jet1}, \eta_{jet2}, m_{jj}, \Delta \eta_{jj}, p_{T1}, \Delta \phi_{\gamma\gamma jj} )</td>
<td>0.9289</td>
<td>70</td>
</tr>
<tr>
<td>7</td>
<td>( \eta_{jet1}, \eta_{jet2}, \eta_{\gamma\gamma}, \Delta \eta_{jj}, \Delta \phi_{\gamma\gamma jj}, p_{T1}, \min(\Delta R_{j\gamma}) )</td>
<td>0.915</td>
<td>80</td>
</tr>
<tr>
<td>8</td>
<td>( \text{All variables} )</td>
<td>0.915</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 4.4: Evolution of optimal significance when selecting \( k \) variables only from 8-variable ‘Selection B’. The significance shown is calculated using the shape method, with statistical and fit uncertainties included as penalty terms.
agreement between the training and testing samples, as evidenced by the reasonably high K-S test probabilities. Consequently, the amount of overtraining on signal and background is small, ensuring a more robust analysis.

4.5 Definition of VBF MVA categories

![Hierarchical structure of event selection in the VBF $H \rightarrow \gamma\gamma$ multivariate analysis](image)

Figure 4.6: Hierarchical structure of event selection in the VBF $H \rightarrow \gamma\gamma$ multivariate analysis [43].

Once the input sample and variable selection had been optimized, the next goal was to establish the method by which VBF-like events would be classified. It was decided that a two-level categorization, using high and low VBF categories, would be used (see Figure 4.6), with the categories defined via cuts on the BDT discriminant itself. The basis for this choice was two-fold: 1. The gradient boosted decision tree approach isolates signal-like and background-like events in opposite ends of the BDT score spectrum, by construction, and 2. The BDT score for each event is always normalized to a range of [-1, +1] by TMVA. With the function of the cut defined, it was then possible to define the procedure and criteria for identifying the optimal categorization, in terms of VBF purity and significance. The statistics of the VBF and inverted isolation samples used to define the VBF categories are given in Table 4.5.
4.5 Definition of VBF MVA categories

<table>
<thead>
<tr>
<th>Type</th>
<th>Sample</th>
<th>Events</th>
<th>Yield for 13 fb$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>VBF</td>
<td>177727</td>
<td>8.80</td>
</tr>
<tr>
<td>Background</td>
<td>Inverted isolation</td>
<td>4429</td>
<td>3377.0</td>
</tr>
</tbody>
</table>

Table 4.5: Signal and background statistics for the VBF vs. inverted isolation training.

Similar to the previous variable optimization study, the approximate VBF significance was chosen to quantify improvement in VBF event selection and background rejection. In particular, the approximate Asimov significance,\(^3\) denoted median$[Z_0|s+b]$, was chosen as the final metric, which is calculated as \([57]\),

$$\text{median}[Z_0|s+b] \approx \sqrt{2 \times (s_{\text{VBF}} + s_{\text{ggF}} + b) \times \ln \left(1 + \frac{s_{\text{VBF}}}{s_{\text{ggF}} + b}\right) - s_{\text{VBF}}}, \quad (4.4)$$

such that it behaves the same as the significance of Equation 4.3 for $s \ll b$. After a method of quantifying improvement between trials was established, an iterative scan was then performed over the entire range of possible scores, with a BDT score step size of 0.01 chosen. Therefore, 201 potential categorizations were tested (including no cut, i.e. BDT score $>-1$), with the optimization procedure of the scan as follows \([56]\):

1. A cut on the BDT score for the $n^{\text{th}}$ iteration of the scan was defined as $X = -1 + 0.01n$. An event was required to have ‘BDT score $> X$’ in order to pass VBF categorization.

2. The categorization requirement for the $n^{\text{th}}$ iteration was applied to the two-jet VBF signal Monte Carlo testing sample, as well as the two-jet ggF signal Monte Carlo and sideband data samples. Events passing the cut on the BDT

---

\(^3\) The Asimov data set is an idealized data set, defined in such a way that “when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values” \([57]\).
score formed two VBF-enriched subsets (VBF signal, and VBF background, containing dijet background and ggF signal). Events failing this cut were re-injected into the main $H \rightarrow \gamma\gamma$ analysis, and placed in one of the nine inclusive categories.

3. Using the VBF-enriched subsets, the median significance $\text{median}[Z_0|s + b]$ was computed in the $m_{\gamma\gamma}$ mass window $[122.5, 127.5]$ GeV. Here, the effective number of VBF and ggF events ($s_{\text{VBF}}$ and $s_{\text{ggF}}$, respectively) were derived from integrals of the effective number of events from each optimization sample in this mass range, i.e. Monte Carlo signal, data sidebands as background. The number of background events ($b$) was derived from an exponential fit to the background model, which was integrated on the same mass range.

The above procedure was used to define the high VBF category, i.e. a subset of events demonstrating a strong resemblance to VBF kinematics. Furthermore, the distribution of median$[Z_0|s + b]$ against BDT score cut $X$ was checked to ensure that the optimal cut value (henceforth referred to as $X_{\text{High}}$) was the result of a smoothly varying distribution, rather than a statistical fluctuation. Note that the following configuration options were used in TMVA to train the BDT:

```
"NTrees = 500 : nEventsMin = 150 : BoostType = Grad : Shrinkage = 0.05 :
UseBaggedGrad : GradBaggingFraction = 0.5 : nCuts = 30 : MaxDepth = 3"
```

With the high category defined, a second iteration of the scan was performed, in order to define a ‘low’ category of moderately VBF-like events. However, in order to remain orthogonal with the high category, a scan over the BDT score range $[-1, X_{\text{High}})$ was used instead. Therefore, events already classified as high
Table 4.6: Definitions of the high and low VBF categories, along with the optimal significance and VBF purity obtained for each cut [56], for the BDT trained on VBF $H \rightarrow \gamma\gamma$ signal Monte Carlo and inverted isolation background.

VBF-like are excluded from low categorization. The optimal cuts for the two VBF categories, along with the optimized VBF significances, are shown in Table 4.6.

4.6 Distributions of discriminating variables after VBF categorization

Figures 4.7 and 4.8 show the distributions of the benchmark variables derived in Section 4.4.1 for data sideband events, as well as VBF and ggF signal Monte Carlo events after categorization in the low and high VBF categories (respectively). There are several differences worth noting not just between the two categories, but between the samples themselves. Of primary interest are the differences in the dijet mass $m_{jj}$, and the pseudorapidity separation of the jets, $\Delta\eta_{jj}$. The peaks of both distributions shift to larger values (i.e. larger invariant mass, greater separation) with harder cuts on the BDT score – an unsurprising result, considering the overall hardness of the quark jets in true VBF processes (see Section 2.4). Likewise, the pseudorapcities of the leading and subleading jets tend to peak more sharply in the forward/backward regions with tighter cuts on the BDT score.

The peak of the VBF diphoton $p_Tt$ spectrum also shifts from approximately 60 to 80 GeV between the low to high VBF categories, although this behaviour is
not unique to VBF.\textsuperscript{4} However, this shift is indicative that more central diphoton pairs are being isolated by these cuts, which is reflective of the low-momentum Higgs boson produced in the VBF process. Lastly, $\Delta \phi_{\gamma\gamma jj}$ also tends towards higher values between the low and high categories, indicating a momentum balance between the dijet and diphoton systems, as expected of the tree-level VBF $H \to \gamma\gamma$ process.

Figures 4.9 and 4.10 show comparisons of the inverted isolation and data sideband events in the low and high VBF categories, respectively. While agreement in the former category is fairly good between the two samples, the low statistics of the samples after the tighter BDT score cut lead to disagreement in the shapes of some of the benchmark variables (e.g. $\eta_{\text{jet2}}$). Otherwise, the two samples generally agree within statistical errors, so it can be said that the inverted isolation background is a fair descriptor of the diphoton background.

\textsuperscript{4} This behaviour is also observed in the ggF $p_{Tt}$ spectrum.
4.6 Distributions of discriminating variables after VBF categorization

Figure 4.7: Distributions of the six benchmark variables for events falling in the low VBF category (0.18 < BDT score < 0.74) based on the VBF vs. inverted isolation BDT. Shown here are data sidebands, along with VBF and ggF $H \rightarrow \gamma\gamma$ signal Monte Carlo samples.

Figure 4.8: Distributions of the six benchmark variables for events falling in the high VBF category (BDT score > 0.74) based on the VBF vs. inverted isolation BDT. Shown here are data sidebands, along with VBF and ggF $H \rightarrow \gamma\gamma$ signal Monte Carlo samples.
4.6 Distributions of discriminating variables after VBF categorization

Figure 4.9: Comparisons of the benchmark variable distributions of data sideband and inverted isolation events falling in the low VBF category (0.18 < BDT score < 0.74) based on the VBF vs. inverted isolation BDT.

Figure 4.10: Comparisons of the benchmark variable distributions of data sideband and inverted isolation events falling in the high VBF category (BDT score > 0.74) for the VBF vs. inverted isolation BDT.
4.7 Checking for BDT sculpting in $m_{\gamma\gamma}$

One major concern in the usage of a BDT-based analysis is the sculpting of kinematic distributions resultant from a hard cut on the BDT score. Here, the term connotes a distortion of the shape of a given variable to the point where the initial trend (i.e. prior to any BDT score cut) has been changed non-trivially, or a bias is introduced to its expected value. The presence of this effect would be particularly worrisome if occurring in the diphoton invariant mass distribution $m_{\gamma\gamma}$, as the final analysis – in which discovery probability and signal strength are evaluated for the VBF $H \rightarrow \gamma\gamma$ process – is built around the expectation of a Gaussian-like $\gamma\gamma$ signal ‘bump’ on top of a monotonically decreasing background. Therefore, the presence of sculpting-like effects such as new peaks or non-monotony in the $m_{\gamma\gamma}$ sidebands (outside of statistical error) would cause significant difficulty in merging the results of the VBF analysis with those of the other ATLAS $H \rightarrow \gamma\gamma$ groups.

For these reasons, it was necessary to check that no sculpting-like effects were induced in the $m_{\gamma\gamma}$ sidebands after successively tighter cuts on the BDT score. A procedure to perform this check was developed and performed as follows. The $m_{\gamma\gamma}$ sidebands were divided up into four subsets, with each subset defined by a BDT score range $X_1 < \text{BDT score} < X_2$, such that the range $[X_1, X_2]$ contained 25% of the diphoton + dijet VBF signal Monte Carlo events – otherwise known as ‘quartiles’ of events. An exponential fit was then performed on each sideband data subset, with the resultant distributions checked for sculpting-like effects, such as new peaks in the $m_{\gamma\gamma}$ distribution, or a monotonically increasing (or flat) fit. This same procedure was then repeated using sideband data quartiles in lieu of VBF Monte Carlo event quartiles.
4.7 Checking for BDT sculpting in $m_{\gamma\gamma}$

Figure 4.11: (Top) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of VBF signal Monte Carlo events based on the VBF vs. inverted isolation BDT. (Bottom) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of sideband data (Figure used with permission of D. Gillberg, F. Bernlochner) [56].
The resultant sideband data distributions are shown in Figure 4.11. The top plot shows the sideband distributions for each quartile of VBF signal Monte Carlo events, with the bottom plot showing the sidebands for each sideband data quartile. Although the slope of the exponential fit is dependent on the cut on the BDT discriminant, the overall trend for each quartile agrees with a monotonically decreasing exponential, within statistical uncertainty. Therefore, it can be said that no significant sculpting effects were introduced to the $m_{\gamma\gamma}$ sideband distribution as a consequence of cutting on the BDT discriminant.

![Figure 4.12: Profiles of the average BDT score with $m_{\gamma\gamma}$](image)

A secondary check of BDT stability is shown in Figure 4.12, where the average BDT score with respect to $m_{\gamma\gamma}$ is calculated for the inverted isolation background and data sidebands. In both cases, the average BDT response is roughly the same, with no distinct correlation between BDT score and $m_{\gamma\gamma}$ observed.
Chapter 5

Refining the VBF multivariate analysis

Once the “rough” BDT-based analysis was established using inverted isolation background, the next step was to refine the methodology to improve the efficacy and efficiency of signal/background discrimination. To this end, many of the same approaches detailed in Chapter 4 were enhanced with new choices of background samples and variables, with the goal of improving the final VBF significance and selection purity.

5.1 Improving the background model

A second option in deriving a background model, outside of using pure samples, is to mix two (or more) of the samples described in Table 4.1, with each element representing a unique component of the background. For example, one could use Sherpa Monte Carlo events to represent the irreducible $\gamma\gamma$ background contribution, with another component handling the $\gamma j + j \gamma + jj$ elements. The primary motivation for using such a mixture is the large sample statistics, which makes the
5.2 Improving benchmark variable selection

BDT less beholden to fluctuations in variable correlations and shape information that may occur with small training samples. Furthermore, the more accurately one is able to model the background, the more efficient and effective the BDT discrimination becomes.

It was decided that any mixture of background components should be weighted according to the approximate $H \rightarrow \gamma\gamma$ background composition measured using the $2 \times 1$D fit method [58], which estimated a diphoton background composition of 79.5% $\gamma\gamma$ events, and 20.5% $\gamma j + j\gamma + jj$ events. Therefore, a new background model was tested according to these guidelines, which mixed 79.5% Sherpa $\gamma\gamma$ Monte Carlo events with 20.5% inverted isolation data, reweighted to the statistics of the data sidebands. That is, the Sherpa component was weighted to have $0.795 \times N_{\text{Sidebands}}$ effective events, while the inverted isolation component was weighted to have $0.205 \times N_{\text{Sidebands}}$ effective events.

Figure 5.1 shows a comparison of the resultant kinematic distributions of the Sherpa + inverted isolation mixture, as well as a pure inverted isolation sample, with the data sidebands. While both background samples model the data sidebands fairly well in terms of the dijet variables, these plots demonstrate that the mixed sample improves greatly the agreement with sideband data for the diphoton and dijet + diphoton variables. It was decided, then, to train another BDT using this Sherpa + inverted isolation mixture to test its modelling capabilities.

5.2 Improving benchmark variable selection

It was initially decided that the original six benchmark variables should be included in the refined BDT, having demonstrated their discriminating power in Section 4.4.2. However, in an effort to garner the best performance out of the BDT, additional variables and post-dijet selection cuts were also considered.
5.2 Improving benchmark variable selection

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

η ∆
0 1 2 3 4 5 6 7 8 9

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

γ γ
(φ ∆
0 0.5 1 1.5 2 2.5 3

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

jj
m [GeV]
0 200 400 600 800 1000 1200 1400

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

Tt
P
0 20 40 60 80 100 120 140 160 180 200 220

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

jet1 η
-4 -3 -2 -1 0 1 2 3 4

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

jet2 η
-4 -3 -2 -1 0 1 2 3 4

MC / Data

Counts normalized to unity

Sherpa+InvIso
InvIso
Data

Figure 5.1: Comparison of data sidebands with Sherpa + inverted isolation, inverted isolation background samples for the six benchmark variables (Note: All plots have a cut of $\Delta \eta_{jj} > 2$ applied).
Evidence from the cut-based VBF analysis [16] (developed simultaneously with the multivariate analysis) suggested that additional discriminating power was provided by the minimum jet-photon spatial difference (min(ΔR_{jγ})) and diphoton Zeppenfeld variable (∆η_{γγ}). While the study detailed in Section 4.4.2 did not show much initial discriminating power in either variable, the cut-based analysis also required a minimum pseudorapidity separation between the leading and subleading jets (specifically, ∆η_{jj} > 2.8), which was not required during the initial variable selection process. However, it was found that the addition of these two variables did hold more discriminating power when used in tandem with the precut of ∆η_{jj} > 2 mentioned in Section 4.4.3, as shown in Table 5.1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th>Separation</th>
<th>Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m_{jj}</td>
<td>0.256</td>
<td>∆η_{jj}</td>
<td>0.172</td>
</tr>
<tr>
<td>2</td>
<td>∆η_{jj}</td>
<td>0.228</td>
<td>m_{jj}</td>
<td>0.170</td>
</tr>
<tr>
<td>3</td>
<td>min(ΔR_{jγ})</td>
<td>0.186</td>
<td>∆φ_{γγjj}</td>
<td>0.146</td>
</tr>
<tr>
<td>4</td>
<td>η_{Zepp}γγ</td>
<td>0.124</td>
<td>η_{Zepp}γγ</td>
<td>0.146</td>
</tr>
<tr>
<td>5</td>
<td>p_{Tt}</td>
<td>0.098</td>
<td>min(ΔR_{jγ})</td>
<td>0.145</td>
</tr>
<tr>
<td>6</td>
<td>∆φ_{γγjj}</td>
<td>0.091</td>
<td>p_{Tt}</td>
<td>0.097</td>
</tr>
<tr>
<td>7</td>
<td>η_{jet2}</td>
<td>0.078</td>
<td>η_{jet1}</td>
<td>0.070</td>
</tr>
<tr>
<td>8</td>
<td>η_{jet1}</td>
<td>0.047</td>
<td>η_{jet2}</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Table 5.1: Separation and importance for each variable used in the eight-variable VBF vs. Sherpa + inverted isolation training [56].

Disagreement between the data and signal/background Monte Carlo distributions was also observed in the region |η_{Zepp}γγ| > 5, prompting a second precut on η_{Zepp}γγ to ensure no outlying dijet events were represented as signal in the BDT. The two-jet selection criteria were then amended, such that any event used to train the BDT must have:
5.3 Optimization of the BDT configuration

In using the boosted decision tree method within TMVA, one has access to several configuration options for the tree itself, beyond selecting the type of samples and classifiers to use. A list of BDT parameters that can be configured using TMVA is given in Table 5.2. While most parameters are straightforward in name and definition, a few terms deserve further explanation:

- **GradBaggingFraction** – When building a BDT, a bagging procedure may also be employed, which has the potential to improve the efficacy of the gradient boosting. Bagging (or stochastic gradient boosting) is a procedure akin to re-sampling, in which random subsets of training events are used to build each binary tree [48]. Therefore, one will produce decision trees using samples that resemble the full data set in terms of their distributions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>nTrees</td>
<td>Number of trees to be built.</td>
</tr>
<tr>
<td>nEventsMin</td>
<td>Minimum number of events required in a tree node.</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>Controls the weight of the individual trees.</td>
</tr>
<tr>
<td>GradBaggingFraction</td>
<td>Controls the fraction of events used when employing a bagging procedure to build the BDT.</td>
</tr>
<tr>
<td>nCuts</td>
<td>Number of cuts used to define a leaf node.</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>Maximum allowed depth of a decision tree.</td>
</tr>
</tbody>
</table>

Table 5.2: Configurable parameters for boosted decision trees in TMVA [47].

1. $|\Delta \eta_{jj}| > 2$ between the leading, subleading jets.

2. $|\eta^{Zep}_{\gamma\gamma}| < 5$ for the leading/subleading diphoton and dijet pairs.

5.3 Optimization of the BDT configuration

In using the boosted decision tree method within TMVA, one has access to several configuration options for the tree itself, beyond selecting the type of samples and classifiers to use. A list of BDT parameters that can be configured using TMVA is given in Table 5.2. While most parameters are straightforward in name and definition, a few terms deserve further explanation:

- **GradBaggingFraction** – When building a BDT, a bagging procedure may also be employed, which has the potential to improve the efficacy of the gradient boosting. Bagging (or stochastic gradient boosting) is a procedure akin to re-sampling, in which random subsets of training events are used to build each binary tree [48]. Therefore, one will produce decision trees using samples that resemble the full data set in terms of their distributions.
5.3 Optimization of the BDT configuration

and variance, but are effectively unique. In theory, the usage of this procedure will lead to more statistical stability in the classifier, depending on the fraction of events used in the bagging (a fraction between 0.5 and 0.8 is recommended [47]).

- **Shrinkage** – The robustness of a BDT is controlled by the shrinkage, or nominal weight, of the individual trees. This factor can be compared to the learning rate of the BDT, and its size should be relative to the number of trees being grown ($nTrees$), i.e. A smaller shrinkage (e.g. 0.1 - 0.3) requires more trees to be grown [47].

- **MaxDepth** – This parameter controls the maximum number of decisions that can be made by each tree. After **MaxDepth** number of decisions has been reached by a given branch, a node is reached, with a signal/background decision made.

Because there are several parameters with which the structure of the BDT can be modified, it became necessary to ensure that the BDT-based analysis was actually using the most optimal configuration possible. In theory, the best configuration of the BDT would provide statistical stability, not detract from the established discriminating power, and, if possible, improve signal/background identification for the particular samples chosen.

Similar to the input variable optimization method described in Section 4.4.2, a multivariable scan was performed for the BDT input parameters. A list of the parameters tested, along with the values for the scan, is given in Table 5.3, which amounts to a total of 2880 combinations of input parameters. For each configuration, a BDT was trained using a VBF $H \to \gamma\gamma$ signal Monte Carlo sample and Sherpa + inverted isolation background, as described in Section 5.1, divided using
Table 5.3: Possible values of each parameter tested for BDT configuration scan.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>nTrees</td>
<td>100, 200, 700, 1000, 1500</td>
</tr>
<tr>
<td>nEventsMin</td>
<td>25, 50, 100, 150</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>0.05, 0.1, 0.2, 0.4</td>
</tr>
<tr>
<td>GradBaggingFraction</td>
<td>0.5, 0.6, 0.7, 0.8</td>
</tr>
<tr>
<td>nCuts</td>
<td>20, 40, 60</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>3, 5, 10</td>
</tr>
</tbody>
</table>

the same scheme as defined in Equation 4.2. After training and application, the approximate significance was computed as the median significance for the Asimov data set, as in Equation 4.4. Here, the background was approximated in the signal region using an exponential fit to the sideband data, while the VBF and ggF signal estimates were derived from integrals of their respective $m_{\gamma\gamma}$ distributions, about the peak at $m_H = 125$ GeV. The configuration deemed to be the optimal choice was that which maximized the approximate VBF significance.

This choice of configuration was validated using a series of stability tests for each parameter, in order to confirm that the maximal significance was not the result of a statistical fluctuation. Individual parameters were allowed to vary, while fixing the other values to their optimal choices, ensuring that each optimal value was at the peak of a smooth distribution, rather than a fluctuation. It was initially found that the optimal BDT configuration was given by,

"NTrees = 700 : nEventsMin = 25 : BoostType = Grad : Shrinkage = 0.4 : 
UseBaggedGrad : GradBaggingFraction = 0.8 : nCuts = 60 : MaxDepth = 3"

This method of parameter optimization was repeated with higher granularity about the same approximate range of values by other members of the VBF MVA
group, with the final, optimized BDT configuration found to be [56],

\[ \text{"NTrees} = 700 : \text{nEventsMin} = 100 : \text{BoostType} = \text{Grad} : \text{Shrinkage} = 0.05 : \text{UseBaggedGrad} : \text{GradBaggingFraction} = 0.5 : \text{nCuts} = 30 : \text{MaxDepth} = 3" \]

5.4 Final definition of VBF MVA categories

Having decided on a new background sample for BDT training, with an updated list of discriminating variables, it was then necessary to reoptimize the BDT score cuts used to define the high and low VBF categories. To this end, the category derivation process outlined in Section 4.5 was repeated for the eight-variable BDT, using the mixed Sherpa + inverted isolation background. The background was trained against the same VBF $H \to \gamma\gamma$ signal Monte Carlo sample described in Chapter 4, with the signal and background sample statistics given in Table 5.4. The optimal BDT score cuts defining the new set of categories, along with the resultant VBF significance and purity values, are shown in Table 5.5. It is evident that the combination of Sherpa + inverted isolation background, in tandem with the addition of $\eta^{\text{Zepp}}_{\gamma\gamma}$ and $\text{min}(\Delta R_{j\gamma})$ as discriminating variables, provides greater signal discrimination (by way of a higher VBF significance) and similar VBF selection purity, as compared to the “rough” BDT analysis documented in Chapter 4.

Correlation matrices for the eight variables used to build the BDT are shown in Figure 5.2 for signal and background, demonstrating that the additional variables ($\eta^{\text{Zepp}}_{\gamma\gamma}$ and $\text{min}(\Delta R_{j\gamma})$) do not introduce any problematic correlations, validating their addition to the training process. Furthermore, the resultant distributions of the BDT discriminant for the signal and background training/testing samples, along with the resultant K-S test probabilities for each, are given in Figure 5.3(a).
5.4 Final definition of VBF MVA categories

The reasonably large probabilities (Signal: 0.421, Background: 0.405) indicate that there is sufficient overlap in the BDT response – hence, reasonably good agreement in the kinematics – of the training and testing samples for each model.

<table>
<thead>
<tr>
<th>Type</th>
<th>Sample</th>
<th>Events</th>
<th>Yield for 13 fb$^{-1}$</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>VBF</td>
<td>161000</td>
<td>7.79</td>
<td>-</td>
</tr>
<tr>
<td>Backgrounds</td>
<td>MC Sherpa $\gamma\gamma$ (e1264)</td>
<td>3582</td>
<td>2668.0</td>
<td>79.5</td>
</tr>
<tr>
<td></td>
<td>MC Sherpa $\gamma\gamma$ (e1434)</td>
<td>3594</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\gamma j + j\gamma + jj$ (inv. iso.)</td>
<td>4410</td>
<td>688.0</td>
<td>20.5</td>
</tr>
<tr>
<td></td>
<td>$\gamma\gamma + \gamma j + j\gamma + jj$</td>
<td>8040</td>
<td>3356.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 5.4: Signal and background statistics for the eight-variable VBF vs. Sherpa + inverted isolation training. The figures here are quoted after cuts on $\Delta\eta_{ij} > 2$ and $|\eta_{ij}^{\gamma\gamma}| < 5$. Notably, the differences in sample statistics relative to Table 4.5 are due to the additional cut on $|\eta_{ij}^{\gamma\gamma}|$, and the addition of another Sherpa sample (e1264).
5.5 **Distributions of discriminating variables after categorization**

After re-optimizing the VBF categories as described in Section 5.4, it was possible to look at the kinematic distributions of events in the low and high VBF categories. As expected, a higher BDT score implied more VBF $H \to \gamma\gamma$-like behaviour, as shown in Figures 5.4 and 5.5. The trends shown in the training variables are consistent with those remarked upon in Section 4.6 – in particular, a tighter cut on the BDT score tends to result in larger $m_{jj}$ and $\Delta\eta_{jj}$, both positive indicators of VBF-mediated signal.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Low VBF</th>
<th>High VBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal cut</td>
<td>$0.44 &lt; \text{BDT score} &lt; 0.74$</td>
<td>$\text{BDT score} \geq 0.74$</td>
</tr>
<tr>
<td>Optimal median $[Z_0</td>
<td>s + b]$</td>
<td>0.4</td>
</tr>
<tr>
<td>VBF purity (%)</td>
<td>48.0</td>
<td>77.1</td>
</tr>
</tbody>
</table>

Table 5.5: Definitions of the optimal high and low VBF categories, along with the optimal significance and VBF purity obtained for each cut [56], for the BDT trained on VBF signal and Sherpa + inverted isolation background.

![Figure 5.3](image_url)  

(a) Kolmogorov-Smirnov test results for the eight-variable training using VBF $H \to \gamma\gamma$ signal Monte Carlo events vs. Sherpa + inverted isolation background [56]. (b) The resultant BDT score distributions for the Sherpa + inverted isolation background training sample, and data sidebands.
5.5 Distributions of discriminating variables after categorization

Greater agreement was also observed in $p_T$ between the VBF $H \to \gamma\gamma$ Monte Carlo sample and the data sidebands, suggesting that the diphoton system tends towards more back-to-back diphoton pairs with a higher BDT score. Similarly, the value of $\Delta\phi_{\gamma\gamma jj}$ tends to be much closer to $\pi$ with successively tighter cuts on the BDT score, which is expected of a system where the momenta of the two final state jets are balanced with the photons. Although similar trends were observed using the BDT trained on inverted isolation background in Section 4.6, the higher calculated VBF significance when using the eight-variable VBF vs. Sherpa + inverted isolation BDT implies that events with VBF $H \to \gamma\gamma$-like kinematics are selected with a higher purity and efficiency with the optimized BDT.

Good agreement was observed between the data sidebands and the Sherpa + inverted isolation testing sample both in terms of the final BDT score distributions (c.f. Figure 5.3(b)) and the agreement of the distributions of the training variables. In particular, as shown in Figures 5.6 and 5.7, reasonable agreement was maintained for background events in the low and high VBF categories, demonstrating that the mixture of Sherpa $\gamma\gamma$ Monte Carlo events and inverted isolation background was an effective model of the $\gamma\gamma + \gamma j + j \gamma + jj$ components of the VBF $H \to \gamma\gamma$ background. For all of the above reasons, the eight-variable VBF vs. Sherpa + inverted isolation BDT was taken to be the optimal choice for categorizing VBF $H \to \gamma\gamma$-like events, to be used in the final $H \to \gamma\gamma$ statistical analysis, reported in Chapter 7.
Figure 5.4: Distributions of the eight training variables for events in the optimized low VBF category ($0.51 < \text{BDT score} < 0.75$) for data sidebands, and VBF, ggF signal Monte Carlo samples.
5.5 Distributions of discriminating variables after categorization

Figure 5.5: Distributions of the eight benchmark variables for events in the optimized high VBF category (BDT score > 0.75) for data sidebands, and VBF, ggF signal Monte Carlo samples.
Figure 5.6: Comparisons of the benchmark variable distributions of data sideband and Sherpa + inverted isolation events in the low VBF category (0.51 < BDT score < 0.74) for the VBF vs. Sherpa + inverted isolation BDT.
Figure 5.7: Comparisons of the benchmark variable distributions of data sideband and Sherpa + inverted isolation events in the high VBF category (BDT score > 0.74) for the VBF vs. Sherpa + inverted isolation, eight-variable BDT.
5.6 Checking for BDT sculpting in $m_{\gamma\gamma}$

As in the previous training based on inverted isolation background, it was necessary to check whether the optimized BDT induced any sculpting-like effects in $m_{\gamma\gamma}$ through cuts on the BDT score. To this end, the procedure to check for sculpting, as defined in Section 4.7, was repeated for the eight-variable VBF $H \rightarrow \gamma\gamma$ vs. Sherpa + inverted isolation BDT, with the resultant $m_{\gamma\gamma}$ quartile distributions shown in Figure 5.8. Although the fourth quartile regions in each plot show fluctuations in $m_{\gamma\gamma}$ that fall outside of the exponential fit, the overall trends for each quartile are still in sufficient agreement with the fitted distributions. Therefore, it can be said that the eight-variable VBF vs. Sherpa + inverted isolation BDT does not induce any sculpting effects in $m_{\gamma\gamma}$.
5.6 Checking for BDT sculpting in $m_{\gamma\gamma}$

Figure 5.8: (Top) Plots of the $m_{\gamma\gamma}$ sidebands for each VBF signal Monte Carlo event quartile for the VBF vs. Sherpa + inverted isolation BDT. (Bottom) Plots of the $m_{\gamma\gamma}$ sidebands for each quartile of sideband data (Figure used with permission of D. Gillberg, F. Bernlochner) [56].
A secondary check of BDT stability is shown in Figure 5.9, where the average BDT score with respect to $m_{\gamma\gamma}$ is calculated for each component of the Sherpa + inverted isolation background, as well as the data sidebands. In each sample, the average BDT response is roughly the same, with no distinct correlation between BDT score and $m_{\gamma\gamma}$ observed.
Chapter 6

Systematic uncertainties

The careful evaluation of systematic uncertainties is crucial for any physics analysis, as any measurement or scientific claim – especially as it relates to the agreement with (or deviation from) a physical model – must be made relative to the bounds of systematic and statistical error. In ATLAS, systematic uncertainties are defined relative to Standard Model yield predictions, and are propagated through to the final statistical analysis as nuisance parameters (see Chapter 7). The systematic uncertainties in the VBF $H \rightarrow \gamma\gamma$ analysis are both theoretical, derived from variable modelling and perturbative uncertainties, and experimental, emerging from detector effects, object reconstruction, and energy calibration.

6.1 Theoretical uncertainties

6.1.1 Higher-order perturbative uncertainty for $gg \rightarrow H + 2$ jets

As alluded to in Section 2.5.1, the usage of dijet + diphoton variables such as $\Delta \phi_{\gamma\gamma jj}$ must be carefully considered with regard to the $gg \rightarrow H + 2$ jet cross section. In applying cuts to isolate VBF-like signal using such a variable, one
implicitly places restrictions on additional emissions from the gluon-gluon fusion process, since $\pi - \Delta \phi_{\gamma\gamma jj} \rightarrow 0$ strongly implies a VBF-like dijet + diphoton system. Consequently, with a tight enough cut, fixed order perturbation theory breaks down, resulting in unphysical predictions, such as a negative $gg \rightarrow H + 2$ jet cross section. Furthermore, the traditional method of computing uncertainties due to missing higher-order perturbative corrections (i.e. varying the renormalization ($\mu_R$) or factorization ($\mu_F$) scales) can lead to unrealistically small errors in this regime [32]. It was for these reasons that $\Delta \phi_{\gamma\gamma jj}$ was binned above 2.94, thereby preventing the BDT from using phase space information in the problematic region.

After performing the binning, it was possible to compute a systematic uncertainty for the missing higher-order $gg \rightarrow H$ perturbative corrections using the Stewart-Tackmann (ST) procedure [59]. This method deals specifically with cases where an inclusive $N$-jet sample is divided into exclusive $N$-jet and inclusive $(N+1)$-jet bins by a cut on a variable $p_{(N+1)}$ that characterizes additional emissions (denoted an ‘infrared sensitive’ variable) [32]. The overall goal of the method is to compute an accurate uncertainty on the yield of the exclusive $N$-jet bin, using a combination of jet bin boundary conditions, and perturbative uncertainties derived from the standard $\mu_R$, $\mu_F$ scale variations.

For a rectangular cut on $p_{(N+1)}$, the sample is divided into an exclusive $N$-jet bin, and one inclusive $(N+1)$-jet bin, introducing two related sources of uncertainty: the absolute yield uncertainty $\Delta^y$, which is correlated between jet bins, and the migration uncertainty $\Delta_{Cut}$, which is anticorrelated between bins, and goes to zero when the bins are merged (thereby retrieving the original $\geq N$-jet sample). The exclusive $N$-jet bin perturbative uncertainty (denoted $\Delta_N$) is found by solving a $2 \times 2$ covariance matrix to obtain $\Delta_N^2 = (\Delta_N^y)^2 + (\Delta_{Cut})^2$. Both $\Delta_N^y$ and $\Delta_{Cut}$ are expressed in terms of the inclusive $N$-jet and $(N+1)$-jet
6.1 Theoretical uncertainties

Figure 6.1: Higher-order perturbative uncertainty error bands in the \( gg \to H + 2 \)-jet decay relative to \( \pi - \Delta \phi_{\gamma jj} \), derived using the Stewart-Tackmann procedure (Figure used with permission of D. Gillberg, F. Bernlochner) [56].

<table>
<thead>
<tr>
<th>Sample</th>
<th>Low VBF (%)</th>
<th>High VBF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ggF + 2 )-jets</td>
<td>20.1</td>
<td>20.0</td>
</tr>
<tr>
<td>( ggF + 3 )-jets</td>
<td>20.1</td>
<td>44.1</td>
</tr>
</tbody>
</table>

Table 6.1: Higher-order perturbative correction uncertainties for \( gg \to H \) for the high and low VBF categories.

bin perturbative uncertainties, which are approximated using the standard scale variations [32].

The generalized ST procedure allows the inclusive \((N + 1)\)-jet phase space to be further subdivided into an arbitrary number of bins,\(^1\) which is necessary for the derivation of perturbative uncertainties when the exclusive 2-jet phase space is defined by simultaneous cuts on multiple infrared sensitive variables. In this framework, the \( 2 \times 2 \) covariance matrix instead becomes an \( n \times n \) matrix (where \( n \) is the total number of bins), with the covariance of bins \( i \) and \( j \) treated using a simple linear correlation model [60].

In the VBF multivariate analysis, a hard cut on the BDT score translates to

\(^1\) This formulation is defined rigorously in Refs. [32] and [60].
a nonlinear, tight selection in $\Delta \phi_{\gamma\gamma jj}$ phase space [60]. Consequently, the $gg \rightarrow H + 2$ jets sample is divided into exclusive 2-jet and inclusive 3-jet bins, with the former bin composed of events passing the score threshold, and all others relegated to the inclusive 3-jet bin. Because $\Delta \phi_{\gamma\gamma jj}$ was the only infrared sensitive variable used in the BDT, it was possible to apply the ST method for $\Delta \phi_{\gamma\gamma jj}$ to compute the perturbative uncertainties for dijet $gg \rightarrow H$ events in the VBF MVA categories. Specifically, after applying the categorization via BDT score cuts (hence, creating 2- and $\geq 3$-jet bins), the modelled ST covariance matrix was built using scale variation uncertainties from the $\geq 2$- and $\geq 3$-jet bins, thereby allowing for the derivation of $\Delta_2$ and $\Delta_{\geq 3}$. This procedure was performed for both the high and low VBF categories, with the exclusive 2-jet bin error bands in $\Delta \phi_{\gamma\gamma jj}$ appearing similar to those in Figure 6.1. The resultant exclusive 2-jet and inclusive 3-jet perturbative uncertainties for each VBF category, expressed as a fraction of the nominal dijet yield, are given in Table 6.1.

6.1.2 Modelling uncertainties of $\eta_{\gamma\gamma}^{Zepp}$

One particular issue that arises when using Monte Carlo samples to form an analysis is the potential for disagreement in the shapes of discriminating variables at parton- or truth-level between Monte Carlo generators. These disparities may arise due to the usage of different approximations, or calculations being made at different orders (e.g. leading order vs. NLO). Therefore, beyond just looking at reconstruction-level comparisons between Monte Carlo and data, lower-level differences between generators must also be considered, which can be mitigated by way of event weights and/or systematic uncertainties.

In this analysis, it was found that the distribution of the diphoton Zeppenfeld variable ($\eta_{\gamma\gamma}^{Zepp}$) for $gg \rightarrow H + 2$ jets events in Powheg disagreed with MCFM [61]
6.1 Theoretical uncertainties

Figure 6.2: The ratio of the distributions of $\eta_{\gamma\gamma}^{\text{Zepp}}$ between Powheg + Pythia8 and MCFM (shown on the $y$-axis), along with the fitted polynomial reweighting function (Figure used with permission of D. Gillberg, F. Bernlochner) [56].

(Another parton-level generator used in ATLAS) outside of the bounds of QCD scale variation uncertainties, likely due to better modelling of the higher-order $ggF$ perturbative corrections in MCFM. In order to account for this difference, the distribution of $\eta_{\gamma\gamma}^{\text{Zepp}}$ from Powheg + Pythia8 was reweighted to match that of MCFM, with the result being a polynomial weight function, as seen in Figure 6.2. The corresponding systematic uncertainties from the reweighting were found to be 6.2% for the low VBF category, and 7.6% for the high VBF category [56].

6.1.3 Modelling uncertainties of $\Delta\phi_{jj}$

A second modelling disparity was found in the simulation of the dijet $gg \rightarrow H$ process by Powheg + Pythia8 (see Section 4.2.1). Rather than originating from the parton-level process in Powheg, the jets instead originate from the subsequent parton showering performed by Pythia8. These two steps are effectively independent, therefore eliminating the angular correlations present in the simulation of two additional jets using QCD matrix elements. Consequently, there is a noticeable disagreement in the distribution of $\Delta\phi_{jj}$ between Powheg + Pythia8 and other NLO Monte Carlo generators.
6.1 Theoretical uncertainties

Figure 6.3: Weight function for truth-level $\Delta \phi_{jj}$ (as shown on the $y$-axis) derived using Sherpa and Powheg + Pythia8 (Figure used with permission of D. Gillberg, F. Bernlochner) [56].

A systematic uncertainty was then introduced to account for the disparity in the shape of $\Delta \phi_{jj}$. In particular, a polynomial weight function was derived to correct the Powheg + Pythia8 distribution of $\Delta \phi_{jj}$ in $gg \rightarrow H + 2$ jets at truth-level using the Sherpa generator, which does simulate the additional jets using QCD matrix elements. The resultant reweighting function is shown in Figure 6.3, with the corresponding uncertainties being 8.5% and 12.1% for the low and high VBF categories, respectively [56].

6.1.4 Underlying event uncertainty

The multi-parton interaction (MPI) model in a Monte Carlo generator simulates the QCD processes underlying each hard scattering interaction, and as such, may also be a source of systematic error. Following the recommendations of the Higgs cross section working group, the uncertainties due to the modelling of the underlying event were estimated for both ggF and VBF $H \rightarrow \gamma\gamma$ Monte Carlo in $\Delta \phi_{\gamma\gamma jj}$ using two Powheg + Pythia8 samples: one with the AU2-CT10 tune, which is a more recent parameterization of the underlying QCD effects of LHC $pp$ interactions [62, 63], and one with MPI entirely turned off [56].

It is then possible to reweight the CT10 sample to correspond to the truth-level
6.2 Jet and detector modelling uncertainties

6.2.1 Jet energy scale and resolution

As described in Section 3.2.2, the EM + JES + \textit{in situ} scheme takes into consideration several reconstruction-level (e.g. in-time/out-of-time pileup) and truth-level quantities (e.g. flavour composition) in the correction of jet energy and direction. Consequently, the $i^{th}$ element in the correction has its own associated systematic uncertainty (denoted $\sigma^{\text{JES}}_i$) which may also vary with jet $p_T$, $|\eta|$, flavour,
6.2 Jet and detector modelling uncertainties

etc. Therefore, it is necessary to propagate the largest of these JES uncertainties through the rest of the statistical analysis. One propagates a given uncertainty $\sigma_{i}^{\text{JES}}$ by producing two separate samples: a sample with all jet energies scaled by a factor $(1 + \sigma_{i}^{\text{JES}})$, and another with all jet energies scaled by $(1 - \sigma_{i}^{\text{JES}})$. The average difference in dijet event yield between each shifted sample and the nominal sample (in % of the nominal yield) becomes the JES uncertainty for that component.

The average, non-negligible contributions to the JES uncertainty for both VBF and $\text{ggF} \rightarrow \gamma\gamma$ Monte Carlo ($m_{H} = 125$ GeV), as divided into low and high VBF, and inclusive categories, are found in Table 6.3. Worth noting is that the sign of each uncertainty is based on the size of the $(1 + \sigma)$ dijet yield relative to the nominal sample (i.e. ‘+’ means a greater yield, ‘−’ a smaller yield).

Another uncertainty related to jet and detector modelling is derived from the observed differences in experimental and simulated jet energy resolution (JER) [64]. Similar to the evaluation of JES uncertainties, one propagates the JER uncertainty by applying a smearing factor $(1 + \sigma_{\text{JER}})$ to the energy of every jet, with the uncertainty derived from the average difference in dijet event yield between the nominal and smeared samples. Table 6.4 shows the resultant JER uncertainties for VBF, $\text{ggF} \rightarrow \gamma\gamma$ Monte Carlo samples at $\sqrt{s} = 8$ TeV.

6.2.2 Jet vertex fraction selection

The jet vertex fraction, as described in Section 3.2.3, has its own associated uncertainty, owing to the fact that a single cut is used across the entire $p_{T}$ and $|\eta|$ range. Since the $JVF$ is a measure of the probability that a jet emerged from a hard scattering interaction in a given event, one can evaluate a systematic uncertainty based on $JVF$ by varying the cut (nominally $|JVF| > 0.25$) by some factor,
Table 6.3: Largest JES uncertainties for $\sqrt{s} = 8$ TeV VBF and ggF $H \rightarrow \gamma\gamma$ Monte Carlo samples ($m_H = 125$ GeV).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Low VBF (%)</th>
<th>High VBF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBF</td>
<td>-0.66</td>
<td>-1.29</td>
</tr>
<tr>
<td>ggF</td>
<td>3.49</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Table 6.4: Jet energy resolution uncertainties for $\sqrt{s} = 8$ TeV VBF, ggF $H \rightarrow \gamma\gamma$ Monte Carlo samples ($m_H = 125$ GeV).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Low VBF (%)</th>
<th>High VBF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBF</td>
<td>-0.31</td>
<td>-0.03</td>
</tr>
<tr>
<td>ggF</td>
<td>-1.2</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 6.5: Jet vertex fraction systematic uncertainties computed using $m_H = 125$ GeV $H \rightarrow \gamma\gamma$ signal Monte Carlo samples at $\sqrt{s} = 8$ TeV.
and examining the difference in the dijet yield relative to the nominal sample.

The systematic shift is applied only for jets with $p_T < 50$ GeV and $|\eta| < 2.4$, the region which is known to be most affected by pileup jets. Jets are classified as ‘pileup’ or ‘hard scatter’ by way of matching reconstructed jets in $(y, \phi)$ with a corresponding truth jet,\(^2\) such that shifts are calculated for the pileup and hard scatter jet subsets separately. The uncertainty is then derived by comparing the yields in Monte Carlo and data for some control region using the nominal $JVF$ cut (denoted $JVF^0$). One then attempts to find looser and tighter cuts on the $JVF$ in Monte Carlo (denoted $JVF^-$ and $JVF^+$, respectively) that produce the same change in yield between the Monte Carlo sample using the nominal cut ($JVF^0$), and the shifted Monte Carlo samples with cuts on $JVF^-$ and $JVF^+$ [65].

For example, if a nominal cut of $|JVF| > 0.25 \equiv JVF^0$ shows a 5% difference in yield between the Monte Carlo and data samples, one would look for cuts $|JVF| > JVF^+$ and $|JVF| > JVF^-$ that produce a $\sim 5\%$ change in yield between the nominal Monte Carlo sample, and the Monte Carlo samples with the shifted cuts. By running the analysis thrice – alternately requiring $|JVF| > JVF^0$, then $|JVF| > JVF^+$, and finally $|JVF| > JVF^-$ – one obtains the uncertainty from the average difference in dijet yield between the nominal and shifted samples. For the VBF multivariate analysis, the $JVF$-based systematic uncertainties were evaluated for VBF and ggF $H \rightarrow \gamma\gamma$ Monte Carlo ($m_H = 125$ GeV), as shown in Table 6.5.

\(^2\) Here, $y$ is the rapidity, $y = \frac{1}{2} \ln \left( \frac{E + p_L}{E - p_L} \right)$. Each truth jet is also required to have $p_T^{\text{Truth}} > 10$ GeV.
Chapter 7

VBF multivariate analysis results

After the BDT-based VBF $H \rightarrow \gamma\gamma$ categorization was built and optimized, the final test of the analysis was the application to the full, unblinded 2011 and 2012 data sets, constituting $\sim 7 \text{ fb}^{-1}$ of $\sqrt{s} = 7 \text{ TeV}$ $pp$ collisions, and $\sim 21 \text{ fb}^{-1}$ of $\sqrt{s} = 8 \text{ TeV}$ $pp$ collisions, respectively. The events selected using the BDT categorization underwent statistical tests to search for evidence (or perhaps the discovery) of a VBF-mediated $H \rightarrow \gamma\gamma$-like process, and quantify the observed amount of signal relative to Standard Model predictions.

7.1 Analysis goals

With the discovery of a Standard Model Higgs boson-like particle claimed midway through 2012, the goal of the final analysis was to update the results of the statistical tests that led to the initial discovery. In doing so, it was possible to re-examine and validate the previous results from the VBF $H \rightarrow \gamma\gamma$ channel using both greater statistics, and improved signal discrimination by way of the new BDT-based categorization. These tasks were performed by calculating two statistical quantities:
• **Signal strength and error,** \( \mu \pm \delta \mu \), where \( \mu = \sigma / \sigma_{SM} \) – For a given signal process, \( \mu \) is defined as the ratio between the observed cross section and the Standard Model prediction at a Higgs boson mass \( m_H \). The signal strength effectively quantifies how the signal rate behaves relative to what is expected of the Standard Model under the same conditions \[66\]. In this framework, \( \mu = 0 \) corresponds to no signal (i.e. background only), while \( \mu = 1 \) implies perfect agreement with Standard Model predictions. Given that the cross sectional measurements are fixed to what is observed using the ATLAS detector, the main goal is the minimization of \( \delta \mu \), which will put a hard limit on whether the observed signal rate agrees with the Standard Model, or indicates new physics, i.e. \( \mu \) deviating from 1, beyond the limits of statistical and systematic error.

• **p-value for background-only hypothesis,** \( p_0 \) – The quantity \( p_0 \) is representative of the level of disagreement between recorded data and the background-only hypothesis, and is found by integrating the PDF of the maximum likelihood test statistic \( q_0 \) (see Equation 7.4). This p-value is used to quantify particle discovery, such that a smaller \( p_0 \) implies a greater probability that a signal-like resonance has been observed.

In practice, the best fit signal strengths for each production channel are found by way of a profile likelihood ratio which depends on the Higgs boson mass \( m_H \), and a series of nuisance parameters \( \theta \) (i.e. systematic uncertainties) \[57\],

\[
\lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}, \tag{7.1}
\]

The likelihood function \( L(\mu, \theta) \) for a given channel originates in the idea of experimental data being represented as an N-bin histogram,\(^1\) defined as \( n = (n_1, ..., n_N) \),

\(^1\) Typically, the likelihood is formed using a binned distribution of \( m_{\gamma\gamma} \).
7.1 Analysis goals

with another $M$-bin histogram used to constrain nuisance parameters. The resultant function is then given by the product of the Poisson probabilities for each bin [57],

$$L(\mu, \theta) = \prod_{j=1}^{N} \frac{\left(\mu s_j + b_j\right)^{n_j}}{n_j!} e^{-\left(\mu s_j + b_j\right)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}. \quad (7.2)$$

where $s_j$ and $b_j$ are the mean number of entries in bin $n_j$ for signal and background (respectively), and $u_i(\theta) = E[m_i]$ is the expectation value of bin $m_i$ from the $M$-bin histogram. Here, the single circumflex denotes the unconditional maximum likelihood estimate of a parameter, and the double circumflex denotes the conditional maximum likelihood estimate for a fixed $\mu$ [66]. For a given Higgs boson production process observable through ATLAS, the cross section is given by [15],

$$\sigma_H = \frac{N_{S+B} - \langle N_B \rangle}{\epsilon \int \mathcal{L} \ dt}. \quad (7.3)$$

Here, $N_{S+B}$ is the combined signal + background rate, $\langle N_B \rangle$ is the expected number of background events derived from fitted data, $\epsilon$ is the product of various efficiency factors (such as detector efficiency), and $\int \mathcal{L} \ dt$ is the integrated luminosity (in units of fb$^{-1}$).

We may use the profile likelihood ratio $\lambda(\mu)$ to calculate the probability of signal discovery for a given $m_H$ under the assumption that $\hat{\mu} \geq 0$ (i.e. a signal process can only add to the existing background statistics). A test statistic $q_0$ is derived to rule out the background-only hypothesis ($\mu = 0$) [57] as,
Accordingly, the level of disagreement between data and the background-only hypothesis is quantified as,

\[
p_0 = \int_{q_0(\text{obs})}^{\infty} f(q_0|0) \, dq_0,
\]

where \( \lambda(0) \) is the profile likelihood ratio for \( \mu = 0 \). When evaluating the background-only hypothesis, the function \( f(q_0|0) \) can be approximated analytically as [57],

\[
f(q_0|0) = \frac{1}{2} \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} e^{-q_0/2}.
\]

This expression can be integrated easily to derive the value of \( p_0 \),

\[
p_0 = 1 - \Phi(\sqrt{q_0}),
\]

with the corresponding significance being \( Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0} \), with \( \Phi \) being the cumulative PDF of the Gaussian distribution. If \( Z_0 > 5\sigma \) (or, equivalently, if \( p_0 < 2.87 \times 10^{-7} \)), then a particle discovery is claimed for the observed signal.

### 7.2 Training samples and BDT response

In Figure 7.1, distributions of the BDT score are shown for the various testing samples used in the analysis. Specifically, in Figure 7.1(a), good agreement is observed in the BDT response of the Sherpa + inverted isolation background testing sample, and the full 2012 dataset, within experimental uncertainty. Furthermore, in Figure 7.1(b), a sharp contrast is observed between the BDT response of the VBF Monte Carlo testing sample, and the ggF Monte Carlo and background samples.
Figure 7.1: (a) BDT response of the background sample, as compared to experimental data, after diphoton + 2-jet + pre-selection cuts. (b) BDT response of the background and VBF signal samples [43].
7.3 $m_{\gamma\gamma}$ measurement

The invariant mass spectra for events in the low and high VBF categories are shown in Figure 7.2, along with fits for the background, and expected signal for a Standard Model Higgs boson with $m_H = 126.5$ GeV. A plot of the local $p_0$ for for events in the low and high VBF categories is shown in Figure 7.3. Here, the largest local significance (as given by $Z_0 = \Phi^{-1}(1 - p_0)$) was found to be $2.9\sigma$ at $m_H = 123.5$ GeV. This local significance value is just under the threshold of being able to claim evidence for a Higgs boson-like particle in the VBF $H \rightarrow \gamma\gamma$ channel alone, which would correspond to a significance of $3\sigma$.

7.4 Signal strength measurement

Several Higgs boson production modes are available at the LHC, therefore a simplification is made to group the signal strengths for $W$, $Z$-mediated Higgs boson production modes (VBF, associated production) and $t\bar{t}$-mediated modes (ggF, top fusion) into two quantities, $\mu_{\text{VBF}+\text{VH}}$ and $\mu_{\text{ggF}+\text{ttH}}$. Both signal strength values are then computed by way of a simultaneous likelihood fit, which allows for a 2D representation of $\mu_{\text{VBF}+\text{VH}}$ vs. $\mu_{\text{ggF}+\text{ttH}}$. Confidence limit contours for this representation can also be computed by using the asymptotic test statistic $-2\ln \Lambda(\mu)$, which behaves as a $\chi^2$ distribution with $n$ degrees of freedom, where $n$ is the dimension of the vector $\mu$. For $n = 2$, the 68% and 95% confidence limits are given by $-2\ln \Lambda(\mu) = 2.3$ and 6.0, respectively [67].

Since the signal strength is derived from Higgs boson production cross sections, in order to obtain a value specific to the $H \rightarrow \gamma\gamma$ decay, it is necessary to multiply $\mu_{\text{VBF}+\text{VH}}$ and $\mu_{\text{ggF}+\text{ttH}}$ by $\mu_{H \rightarrow \gamma\gamma} = B_{H \rightarrow \gamma\gamma}/B_{H \rightarrow \gamma\gamma(\text{SM})}$, where $B_{H \rightarrow \gamma\gamma}$ is the observed $H \rightarrow \gamma\gamma$ branching ratio. This extra term is also a signal strength factor, therefore $\mu_{\text{VBF}+\text{VH}} \times \mu_{H \rightarrow \gamma\gamma} = 1$ still implies perfect agreement with the
7.4 Signal strength measurement

Figure 7.2: The full 2012 $m_{\gamma\gamma}$ distributions, along with expected SM signal and background fit for $m_H = 126.8$ GeV for low (left) and high (right) VBF categories [43].
Standard Model. After performing the profile likelihood ratio maximization, the best fit signal strength measurement for $W, Z$-mediated $H \rightarrow \gamma \gamma$ signal was found at a Higgs boson mass of $m_H = 126.8 \text{ GeV}$ to be \([43]\),

$$
\mu_{\text{VBF}+\text{VH}} \times \mu_{H\rightarrow\gamma\gamma} = 1.72^{+0.85}_{-0.77} \text{ (stat)}^{+0.48}_{-0.29} \text{ (syst)}^{+0.25}_{-0.29} \text{ (theory)},
$$

which agrees with Standard Model predictions (given as $\times$ in Figure 7.3(b)) within $2\sigma$, as represented by the 95% confidence limit contour.
7.4 Signal strength measurement

Figure 7.3: (a) The observed local $p_0$ value for VBF $H \rightarrow \gamma\gamma$ events as a function of $m_H$ for the combination of $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data. (b) The best-fit values of $\mu_{ggH+ttH} \times B/B_{SM}$ and $\mu_{VBF+VH} \times B/B_{SM}$ and their 68% (solid) and 95% (dashed) CL contours [43].
Chapter 8

Conclusions

The kinematically unique properties of the vector boson fusion production mechanism allow for a multitude of methods to isolate Higgs boson-like signal, with varying degrees of success. When coupled with the distinct kinematics and heavily reducible backgrounds of the $H \rightarrow \gamma\gamma$ process, it becomes possible to search for both direct evidence of Higgs boson production, as well as quantify its rate of production relative to Standard Model predictions. It has been shown that the isolation of signal resembling VBF $H \rightarrow \gamma\gamma$ is possible (and indeed, quite successful) using a multivariate analysis approach – specifically, through the use of a boosted decision tree, developed using orthogonal signal and background samples.

A primary iteration of the VBF multivariate analysis was derived using a BDT trained on inverted isolation background and VBF $H \rightarrow \gamma\gamma$ signal Monte Carlo samples. It was found that the variables contributing the most discriminating power related to the dijet mass, jet pseudorapidity separation, and balance of the dijet + diphoton system. A second iteration of the analysis was then performed, with the goal of improving signal discrimination through better background modelling and variable selection. It was found that the irreducible $\gamma\gamma$ and $\gamma j + j\gamma + jj$ components of the $H \rightarrow \gamma\gamma$ background were modeled reasonably well by the combination of a Sherpa $\gamma\gamma$ Monte Carlo sample and inverted isolation diphoton data,
weighted to the approximate proportions of true and fake diphoton components, respectively, of the $H \rightarrow \gamma\gamma$ background. Evidence of this similarity was found in the agreement of the BDT score distributions of the $\gamma\gamma$ data sidebands and background model, as well as the agreement in the resultant kinematic distributions after performing VBF categorization via cuts on the event-by-event BDT score.

A final statistical analysis was performed on the full 2011 – 2012 dataset to quantify the discovery significance and signal strength of signal resembling the VBF-mediated $H \rightarrow \gamma\gamma$ decay. The resultant discovery significance was found to be $Z_0 = 2.9\sigma$ at an invariant mass of $m_H = 123.5$ GeV. In combining the results of the VBF categorization with those of the associated production-enriched categories, the best-fit signal strength for $W, Z$-mediated Higgs boson production was found to be $\mu_{VBF+VH} \times B/B_{SM} = 1.72$ at an invariant mass of $m_H = 126.8$ GeV, which agrees with Standard Model production rates within $2\sigma$.

Ongoing work involving multivariate analysis techniques in the $H \rightarrow \gamma\gamma$ channel includes the implementation of two-dimensional BDTs, the refinement of signal and background samples, and the application of the boosted decision tree method to other facets of the $H \rightarrow \gamma\gamma$ analysis (e.g. spin studies). The same type of machinery is also being extended to other VBF-mediated Higgs boson decays, such as $H \rightarrow ZZ^* \rightarrow 4l$. One practical consideration for the extension of this method is the rate of VBF production for a given decay process. For example, VBF-mediated $H \rightarrow ZZ^* \rightarrow 4l$ is an exceptionally rare process, necessitating much larger volumes of data to ensure the practicality of a multivariate analysis. However, with the planned resumption of collisions and data taking in 2015 at full design luminosity and centre-of-mass energy, the prospect of discovering and studying additional VBF-mediated Higgs boson decays using multivariate analysis techniques looks increasingly bright.
References


REFERENCES


Appendix A

Photon, jet reconstruction miscellanea

A.1 Definitions of discriminating variables

Figure A.1: (Left) The layout of the ATLAS detector in Cartesian coordinates, with definitions of \( \theta \) and \( \phi \), relative to their rotational axes. (Right) A diagram of pseudorapidity as it relates to the z-axis in ATLAS.

Descriptions of the jet- and photon-related kinematic variables used in this analysis are given in Table A.1. Furthermore, a representation of the ATLAS coordinate system (as it relates to these quantities) is shown in Figure A.1. Of particular note is the pseudorapidity \( (\eta) \), which is commonly used in place of \( \theta \), owing to the boost invariance of \( \Delta \eta \) between objects. Here, the subscripts ‘1, 2, 3’ refer to the leading, subleading, and third-highest \( p_T \) objects in a set (e.g. photons or jets).
### A.1 Definitions of discriminating variables

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_T(j_{1,2,3})$</td>
<td>Leading (subleading, third) jet transverse momentum.</td>
</tr>
<tr>
<td>$\eta_{j_{1,2,3}}$</td>
<td>Leading (subleading, third) jet transverse pseudorapidity.</td>
</tr>
<tr>
<td>$\phi_{j_{1,2,3}}$</td>
<td>Leading (subleading, third) jet $\phi$.</td>
</tr>
<tr>
<td>$JVF_{j_{1,2,3}}$</td>
<td>Jet vertex fraction for leading (subleading, third) jet.</td>
</tr>
<tr>
<td>$\eta_{Zepp}^{\text{jet3}}$</td>
<td>Zeppenfeld variable for third jet, $\eta_{\text{jet3}} - \langle \eta_{\text{jet1}}, \eta_{\text{jet2}} \rangle$</td>
</tr>
<tr>
<td>$\eta_{\gamma\gamma}$</td>
<td>Zeppenfeld variable for diphoton system, $\eta_{\gamma\gamma} - \langle \eta_{\text{jet1}}, \eta_{\text{jet2}} \rangle$</td>
</tr>
<tr>
<td>$\text{min}(\Delta R_{\gamma j})$</td>
<td>Minimum $\Delta R(\gamma_i, j_j)$ where $i,j = 1$ or 2</td>
</tr>
<tr>
<td>$m_{jj}$</td>
<td>Invariant mass of dijet system</td>
</tr>
<tr>
<td>$\Delta \eta_{jj}$</td>
<td>$\Delta \eta$ between leading, subleading jets</td>
</tr>
<tr>
<td>$\Delta \phi(jj)$</td>
<td>$\Delta \phi$ between leading, subleading jets</td>
</tr>
<tr>
<td>$\Delta \phi_{\gamma\gamma jj}$</td>
<td>$\Delta \phi$ between diphoton, dijet systems</td>
</tr>
<tr>
<td>$\Delta p_T(jj)$</td>
<td>Difference in leading, subleading jet transverse momenta ($p_T$)</td>
</tr>
<tr>
<td>$\Delta p_T(\gamma\gamma)$</td>
<td>Difference in leading, subleading photon transverse momenta ($p_T$)</td>
</tr>
<tr>
<td>$p_{T,\gamma\gamma jj}$</td>
<td>Transverse momentum of diphoton + dijet system</td>
</tr>
<tr>
<td>$p_Tt$</td>
<td>Transverse momentum perpendicular to the diphoton thrust axis</td>
</tr>
</tbody>
</table>

Table A.1: Definitions of the kinematic variables used in this analysis.
A.2 Comparison of background samples after dijet selection

An extensive comparison of jet- and photon-related quantities for different background samples is given in Figure A.2. Samples under consideration include inverted isolation data, Sherpa and Pythia $\gamma\gamma$ Monte Carlo, as well as data sidebands, with each sample restricted to an invariant mass range $100 < m_{\gamma\gamma} < 160$ GeV. Below each plot, the ratio of each background model distribution to the sideband data is shown. In terms of general agreement between jet, photon, and dijet/diphoton variables, inverted isolation background data is the best choice to model the data sidebands using a single, unmixed sample.
A.2 Comparison of background samples after dijet selection

Counts normalized to unity

0.0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16

Sherpa
InvID
InvIso
Data

\( \eta \Delta \)

0 1 2 3 4 5 6 7

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0.0 0.05 0.1 0.15 0.2 0.25 0.3 0.35

Sherpa
InvID
InvIso
Data

\( \phi \Delta \)

0 0.5 1 1.5 2 2.5 3

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0.0 0.02 0.04 0.06 0.08 0.1 0.12 0.14

Sherpa
InvID
InvIso
Data

\( m_j \) [GeV]

0 100 200 300 400 500 600 700 800 900 1000

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0 0.05 0.1 0.15 0.2 0.25 0.3

Sherpa
InvID
InvIso
Data

\( \gamma \gamma \)

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0.0 0.02 0.04 0.06 0.08 0.1 0.12 0.14

Sherpa
InvID
InvIso
Data

\( \gamma \gamma \)

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0.0 0.05 0.1 0.15 0.2 0.25 0.3

Sherpa
InvID
InvIso
Data

\( T_t \) [GeV]

0 20 40 60 80 100 120 140 160 180 200 220

MC / Data

0 0.5 1 1.5 2

Counts normalized to unity

0 0.05 0.1 0.15 0.2 0.25 0.3

Sherpa
InvID
InvIso
Data

| Zepp(\( T_t \)) |
A.2 Comparison of background samples after dijet selection

Counts normalized to unity

MC / Data

Counts normalized to unity

MC / Data

Counts normalized to unity

MC / Data

Counts normalized to unity

MC / Data

Counts normalized to unity

MC / Data
A.2 Comparison of background samples after dijet selection
Figure A.2: A comparison of the distributions for inverted isolation, inverted ID, and Sherpa background samples, along with the data sidebands.
A.3 Descriptions of JES uncertainties

The following are brief descriptions of the largest contributors to the overall JES systematic uncertainty. A more detailed explanation of each component can be found in Ref. [42].

<table>
<thead>
<tr>
<th>Component</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EffectiveNP_Modeling1</td>
<td>This uncertainty is derived from studying disparities in the modelling of jet fragmentation between different MC generators, e.g. Alpgen + Herwig + Jimmy vs. Pythia [68].</td>
</tr>
<tr>
<td>EtaIntercalibration_Modeling</td>
<td>Monte Carlo-based correction factors are used in the calibration to ensure uniformity in jet response across the entire detector. This uncertainty is derived from studying the effect these correction factors on events with one reconstructed jet in the central region (</td>
</tr>
<tr>
<td>Pileup_RhoTopology</td>
<td>Pileup-related JES correction factors are derived with reference to mean values of the number of primary vertices and bunch crossings (N_{PV}^{ref} and &lt;μ^{ref}&gt;, respectively). The uncertainty in using the pileup-related quantity ρ (see Section 3.2.2) is formed by studying the correlation in μ, N_{PV}, and jet p_T with reference to N_{PV}^{ref} and &lt;μ^{ref)}.</td>
</tr>
<tr>
<td>Closeby</td>
<td>This uncertainty emerges from studying the differences in calibration between isolated (non-overlapping) and non-isolated jets in data and Monte Carlo.</td>
</tr>
<tr>
<td>FlavourComp</td>
<td>This nuisance parameter is derived from the difference in calibration response between light quark jets and gluons, relative to the response of the entire sample. Worth noting is that this uncertainty is not computed for b-jets, which have their own specific uncertainty terms.</td>
</tr>
<tr>
<td>FlavourResponse</td>
<td>Similar to the FlavourComp uncertainty, this uncertainty is also related to light quark and gluon jet response. However, here, the differences between light quark, gluon jet response in Pythia and Herwig are evaluated to compute the nuisance parameter [69].</td>
</tr>
</tbody>
</table>

Table A.2: Descriptions of the largest contributors to the JES uncertainty.
Appendix B

Personal contributions to the 
ATLAS collaboration

Several collaborators worked together to develop the VBF $H \rightarrow \gamma\gamma$ multivariate analysis, which superseded the former cut-based categorization, thereby becoming an integral part of the ATLAS $H \rightarrow \gamma\gamma$ analysis. Furthermore, many others were involved in performing the final statistical analysis. Therefore, this section details my most significant contributions to the analysis, and the relevant chapters in which the work is found, along with my contributions to the larger ATLAS collaboration.

1. Development and testing of $H \rightarrow \gamma\gamma$ analysis software for the Canadian VBF $H \rightarrow \gamma\gamma$ group

   ▶ Relevant chapter: Chapter 3

   ▶ Summary of contribution: In preparation for the new VBF analysis, I assumed the lead developer role for the $H \rightarrow \gamma\gamma$ analysis software that had previously been developed and maintained by Canadian VBF $H \rightarrow \gamma\gamma$ group members D. Gillberg and F. Bernlochmer. This task involved continually updating the event selection to meet the standards laid out by the lead analyzers, incorporating new tools and event weights,
restructuring and optimizing C++ code, and testing the code to ensure agreement with other analyzers. I also contributed a significant portion of the code used by the Canadian group to perform the final BDT analysis.

2. Production of data and Monte Carlo samples for the Canadian VBF $H \rightarrow \gamma\gamma$ group

▷ **Relevant chapters:** Chapter 3, Chapter 6

▷ **Summary of contribution:** Once the $H \rightarrow \gamma\gamma$ analysis software was stable, it was necessary to produce skims of both ATLAS data, and the signal, background Monte Carlo samples used in the multivariate analysis. I was wholly responsible for producing and validating these skims for the Canadian VBF $H \rightarrow \gamma\gamma$ group, as well as putting together a centralized archive for their storage and access. I also produced skims of the differently-tuned ggF samples for Section 6.1.4.

3. Background model studies for Monte Carlo and data-driven samples

▷ **Relevant chapters:** Chapter 4, Chapter 5

▷ **Summary of contribution:** Much of the original ideas and derivation of the “rough” BDT analysis came from the Canadian VBF $H \rightarrow \gamma\gamma$ group. I contributed to discussions within the MVA group on possible background models, and performed studies comparing the different models in Table 4.1 with regard to their accuracy in modelling the sideband data, the results of which are shown in Figure A.2. Though proposed by other members of the MVA group, I also contributed a study of the accuracy of the mixed background model discussed in Section 5.1.

4. Derivation of the benchmark list of BDT classifiers

▷ **Relevant chapter:** Chapter 4

▷ **Summary of contribution:** The input variable selection and optimization processes detailed in Sections 4.4.2 and 4.4.1 were devised by
myself and other Canadian group members. I procured the initial list of variables, performed the initial training, and ranked variables based on separation and importance, with the subsequent BDTs trained and generated by F. Bernlochner. I then performed the significance calculations, rankings, etc. and generated the resultant benchmark variable list.

5. Derivation of the optimal BDT configuration

▷ **Relevant chapter:** Chapter 5

▷ **Summary of contribution:** The BDT configuration optimization process was also developed by myself and F. Bernlochner, and subsequently adopted by the MVA group. As in the benchmark variable list derivation, I devised a series of parameters to test, and developed the final analysis and ranking, while F. Bernlochner performed the BDT training.

6. Calculation of JES and JER systematic uncertainties

▷ **Relevant chapters:** Chapter 6, Chapter 7

▷ **Summary of contribution:** I was responsible for the evaluation of both the JES and JER systematic uncertainties, as detailed in Chapter 6, which are large contributors to the detector-related uncertainties in the VBF $H \rightarrow \gamma\gamma$ analysis. I was also responsible for presenting these results to the larger $H \rightarrow \gamma\gamma$ group in a format usable for the statistical analysis of Chapter 7.

7. Liquid argon data quality & inspection shift work

▷ **Summary of contribution:** During my shift work, I was responsible for the offline evaluation of the quality of recently recorded calorimeter data. Tasks include scanning data for high voltage trips, noise bursts, and instances of non-trivial data rejection during data acquisition. Noisy calorimeter cells were flagged for vetoing, and lists of defects in luminosity blocks were uploaded to a database for further event cleaning. Presentations were also prepared for daily and weekly
meetings which summarized the status of each run inspected, and relayed any systematic problems with calorimeter behaviour. For this shift work, I received Class 3 OTP credit for Carleton University.

8. Authorship qualification task in electron-photon group

▷ **Summary of contribution:** My authorship qualification task involved the revival and development of a method to combine electron track and cluster four momentum measurements into a single, optimal four momentum (an $E - p$ combination). Initial studies were performed to determine the optimal circumstances for combination, in terms of observable quantities that were correlated with truth-level information, and the quality of the track and cluster reconstruction, such as the approximate amount of bremsstrahlung losses in the Inner Detector.

Validation was then performed on a variety of data and Monte Carlo samples: $J/\psi$ and $Z \rightarrow ee$ decays; single electron samples; and $H \rightarrow ZZ \rightarrow 4l$ decays. In addition to developing the method for the combination, I rewrote existing $E - p$ combination software to implement the new method, optimize the C++ code, and enhance code modularity and readability. For this task, I received my authorship qualification. Furthermore, a public conference note on the tool’s development is in progress, and will make up a component of a future paper on electron calibration by the ATLAS electron-photon group.