Study of Higgs Boson Production Through Vector Boson Fusion at the CMS Experiment Using a Dense Convolutional Neural Network

Jack Charles Wright

Imperial College London
Department of Physics

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

Abstract

Measurements of the Higgs boson using the $H \to \gamma\gamma$ Higgs boson decay mode and two different methods for identifying Higgs bosons produced via vector boson fusion are presented. These analyses use proton-proton collision data collected by the CMS collaboration during the 2016 running period and constitute $35.9\,\text{fb}^{-1}$ of integrated luminosity at $\sqrt{s} = 13\,\text{TeV}$. One vector boson fusion identification method is based on boosted decision trees, and the other is based on jets formulated as images and a dense convolutional neural network. The categorisations produced by both methods are subjected to the overall $H \to \gamma\gamma$ statistical analysis and their results compared. The neural network itself is also subjected to analysis to determine what features it has learned to extract from the jet images.

The main objectives of this new approach are to reduce contamination from gluon fusion in the vector boson fusion categories and to improve their statistical significance. This is indeed observed in the expected yields measured from simulation. The vector boson fusion signal strength relative to the Standard Model is measured to be $0.8^{+0.6}_{-0.5}$ in the boosted decision tree variant and $1.5^{+0.5}_{-0.5}$ in the neural network variant. The neural network is also observed to give a reduced uncertainty on many of the other measurements, especially those more directly impacted by vector boson fusion production.

To Mum and Dad.
I'm sorry about the electricity bill.

Declaration

I declare that this thesis is my own work. It has been built upon the work of others and this is stated in detail below. When the work of others is used in the text it is referenced appropriately.

Chapter 1 introduces the work in this thesis referencing prior results in the fields of particle physics and machine learning in my own words.

Chapter 2 describes particle physics theory that has been entirely developed by others, but in my own words.

Chapter 3 describes the Large Hadron Collider and Compact Muon Solenoid in my own words, but these again were developed and studied by many experimental physicists before me.

Chapter 4 describes machine learning theory and practise. This is covers the work of various individuals in the field with my own words.

Chapter 5 describes physics objects at CMS. I had a role in part of the calibration of the ECAL for energy scales and smearing. The systems and studies are the work of my other colleagues at CMS.

Chapter 6 describes event categorisation. The vector boson fusion tagging is the focus of my work and is based on the official analysis approach. I developed the current version of the boosted decision tree based vector boson fusion tag along with Dr Yacine Haddad. The neural network based vector boson fusion tag is my own work. The other tags are the work of other members the CMS $H \rightarrow \gamma\gamma$ analysis group.

Chapter 7 describes the final statistical analysis. The official results are the work of the entire $H \rightarrow \gamma\gamma$ analysis group. For the neural network based results I developed a framework to produce information in a format that could be consumed by the existing final fits machinery. The final fits over the neural network variant categories were run by Ed Scott.

Chapter 8 summarises and draws conclusions. This is my own writing, and the future developments are my own suggestions.

Jack Charles Wright

Acknowledgements

This thesis depends on the contributions of many more people than I can name here. I will always be thankful for the last four years and all the good people I've had the privilege to meet and work with. Therefore I'd like to express my sincere gratitude to everyone I'm about to neglect.

To begin with I'd like to thank Imperial College London and the HEP group for giving me this opportunity in the first place, and STFC for providing funding. I'd also like to thank CERN, the LHC and the CMS collaboration for building and running such remarkable machines and for providing a great environment for research. I am particularly grateful to the Max Planck Institute for Intelligent Systems for accepting me into MLSS 2017 and providing such an enriching two weeks that taught me so much.

A special thank you must go to my supervisors Prof Paul Dauncey and Dr Chris Seez for their support and deep expertise. Thank you Chris for welcoming me to CERN and your uncompromising honesty, and thank you to Paul for making so much time for me and being so supportive even though you're head of group. I couldn't have asked for anyone better.

I am grateful for my colleagues Dr Seth Zenz, Dr Yacine Haddad and Ed Scott. Seth, without your tireless hard work and dedication I and many others would be up the proverbial creek with no paddle. Yacine, your encouragement and permanently sunny disposition helped me so much. Both of you have taught me a lot. Finally, special thanks to Ed for your help over the past couple of years but especially for all your work with the final fits. I wish you the best of luck as you start your thesis, I sincerely hope it goes a lot smoother than mine.

Thank you to Daniel Saunders and his partner Elliot for helping me at an especially difficult time. I wish you both the very best for your future together.

I am also grateful for the deep kindness and generosity of the Paslay family. I'll never forget what you've done for me over the years.

I am eternally thankful to my amazing partner Lucie Altenburg for all the love and support she has shown me over the last year. You've helped me back up when it all seemed to be going to bits, you've proofread for me and more. I can't begin to

thank you enough for how you've looked after me.

Finally, none of this would have been possible without the love and warmth of my family: you're all the bedrock of my life. Above all I'd like to thank my wonderful parents Maureen and Kevin Wright who have shown me unwavering support and fed my curiosity from a the start. I fondly remember childhood visits to the old Birmingham science museum where we'd look all the machines, at the wave motion and light spectrum displays. I especially remember the boxes with table top physics experiments we used to do with prisms, bridge building and other things. You gave me all the books I could ever want on all the subjects I was interested in and then some. You've done so much more for me than I could ever express in here, and this thesis is a culmination of all of that. I feel this is your achievement as well as mine.

Contents

1 Introduction **1**

2 Theory **5**

2.1 Introduction 5

2.2 Yang-Mills Theories 5

2.2.1 From Geometry to Gauge Fields 5

2.2.2 Constructing a Lagrangian 7

2.2.3 Phenomenology 7

2.3 Spontaneous Symmetry Breaking 8

2.3.1 Gauge Symmetry Breaking 10

2.4 The Standard Model of Particle Physics 11

2.4.1 Electroweak Theory 12

2.4.2 Quantum Chromodynamics 17

2.4.3 Higgs Boson Phenomenology 18

3 Apparatus **21**

3.1 Introduction 21

3.2 The Large Hadron Collider 21

3.2.1 LHC Accelerator Chain 21

3.2.2 LHC Structure and Operation 22

3.3 The Compact Muon Solenoid 23

3.3.1 Design Overview 24

3.3.2 Solenoid and Return Yoke 26

3.3.3 Inner Tracking 27

3.3.4 Electromagnetic Calorimetry 29

3.3.5 Hadron Calorimetry 31

3.3.6 Muon Detection 32

3.3.7 Trigger System and Storage 34

4 Machine Learning **35**
4.1 Fundamentals . . . 35
4.1.1 The Learning Process . . . 35
4.1.2 Model Capacity and Generalisation . . . 38
4.1.3 Ensembles . . . 41
4.1.4 Algorithm Design, Evaluation and Optimisation . . . 44
4.2 Deep Learning . . . 46
4.2.1 Artificial Neural Networks . . . 47
4.2.2 Images and Convolutional Neural Networks . . . 52
4.2.3 Dense Convolutional Neural Networks . . . 55

5 Object Reconstruction and Selection **59**
5.1 Introduction . . . 59
5.2 Tracks, Clusters, and Physics Objects . . . 59
5.3 Samples . . . 61
5.3.1 Trigger . . . 61
5.3.2 Data . . . 62
5.3.3 Simulation . . . 62
5.4 Photon Reconstruction . . . 62
5.4.1 Common Variables . . . 63
5.4.2 Photon Energy . . . 64
5.4.3 Photon Identification . . . 65
5.4.4 Photon Preselection . . . 66
5.5 Vertex Reconstruction . . . 67
5.5.1 Vertex Selection . . . 67
5.5.2 Vertex Probability . . . 68
5.5.3 Performance . . . 68
5.6 Other Objects . . . 70
5.6.1 Leptons . . . 70
5.6.2 Jets . . . 70

6 Event Categorisation **73**
6.1 Overview and Objectives . . . 73
6.1.1 The Diphoton BDT . . . 73
6.1.2 Tagging Scheme . . . 75
6.2 Top Fusion Tagging . . . 75
6.2.1 $t\bar{t}H$ Leptonic . . . 76
6.2.2 $t\bar{t}H$ Hadronic . . . 77
6.3 VH Tagging . . . 78
6.3.1 ZH Leptonic . . . 78

6.3.2 WH Leptonic . . . 78
6.3.3 VH Leptonic Loose . . . 79
6.3.4 VH MET . . . 79
6.3.5 VH Hadronic . . . 79
6.4 Untagged . . . 79
6.5 VBF Tagging . . . 80
6.5.1 Selections . . . 81
6.6 VBF Tag with BDTs . . . 82
6.6.1 Dijet BDT . . . 83
6.6.2 Combined BDT . . . 83
6.6.3 Model Interpretation . . . 85
6.6.4 Categorisation and Tag Performance . . . 87
6.6.5 Validation . . . 88
6.6.6 Single BDT Model . . . 90
6.7 DCNN VBF Tag . . . 90
6.7.1 Jet Images . . . 91
6.7.2 Dense CNN Model . . . 94
6.7.3 Model Performance . . . 101
6.7.4 Categorisation and Tag Performance . . . 102
6.7.5 Model Interpretation . . . 103
6.7.6 Validation . . . 109
6.7.7 Conclusions . . . 112

7 Statistical Analysis and Results 115
7.1 Introduction . . . 115
7.2 Statistical Models . . . 115
7.2.1 Signal Modelling . . . 115
7.2.2 Background Modelling . . . 116
7.3 Systematic Uncertainties . . . 116
7.3.1 Theoretical Uncertainties . . . 117
7.3.2 Experimental Uncertainties . . . 119
7.4 Results . . . 122
7.4.1 Best Fit of Model to Data . . . 123
7.4.2 Signal Strength Likelihood Scans . . . 126
7.4.3 Couplings Measurements . . . 129
7.4.4 Conclusions . . . 130

8 Conclusions **133**
8.1 Summary of Results 133
8.2 Future Development 134
8.3 Conclusions 136

A VBF Tag Plots with Loose Preselection **137**

B VBF Tag $Z \to e^+e^-$ Validation Plots **143**

C Feature Visualisation of Different Network Layers **147**

D Per-Category Mass Plots **153**

List of Figures

2.1 A fibre bundle with $\mathcal{V} = \mathrm{U}(1)$. A section is shown (grey line) choosing $g(x) \in \mathrm{U}(1)$ for a set of points in $\mathcal{M}$. 6
2.2 The three types of vertex in Yang-Mills theories. 8
2.3 A ferromagnet above (left) and below (right) the Curie temperature. The example below the Curie temperature has magnetised along the z-direction breaking the SO(3) symmetry to just SO(2) about the z-axis. 9
2.4 Perturbations around the vacuum state at $\theta_0 = 0$ of a symmetry breaking potential $V(\phi)$. The family of degenerate minima are shown by the black circle. 9
2.5 The four main production modes of the Higgs boson at the LHC. Clockwise from top left: ggH, VBF, $\mathrm{t\bar{t}H}$ and VH. 19
2.6 Main contributing diagrams to the Higgs boson to diphoton decay mode. 20

3.1 A schematic view of the LHC accelerator chain for proton-proton operation. 22
3.2 Left: total integrated luminosity over the 2016 proton-proton running period delivered to (blue) and recorded by (orange) the CMS experiment [31]. Right: the 2016 pileup distribution [31]. 23
3.3 Coordinate systems used at CMS. Example values for the pseudorapidity η are shown by the red arrows. 24
3.4 The CMS experiment separated into barrel (top) and endcap (bottom), both have an azimuthal section removed to show structure of the detector subsystems. Rendering was built with the model in [35]. 25
3.5 The CMS solenoid (white) within the steel return yoke (red). Rendering uses [35]. 27
3.6 Left: the tracker subsystem showing the the central pixel detector in yellow, TIB in orange, TID in red, TOB in purple and TECs in blue [35]. Right: the material before the ECAL in radiation lengths (X_0) [32]. 28
3.7 The CMS ECAL with a section removed to show structure [32]. 30

3.8 The CMS HCAL with the barrel and endcap sections (left) with part removed to show structure and the forward hadronic calorimeter (right) [35]. 32
3.9 Left: the CMS muon detector subsystems (white) within the structure of the steel return yoke (red) [35]. Right: a diagram showing a quarter-view of of the muon system with detector types labelled. 33

4.1 Training a linear regressor with SGD (blue) and SGD plus momentum (magenta). Top row: loss histories over training (left), the trajectory of the model parameters in parameter space during training (centre), and the final result with the result in red and the true value in black. An example minibatch is also shown by the black points (right). Lower plot: how the optimisation descends the 'loss landscape' during training. The surface shows the loss calculated over the entire dataset at once for each parameter value. Each step during training computes an estimation of this surface using the sampled minibatch. 39
4.2 Linear regressors with different order polynomials fitted to the same data. 40
4.3 Fits for different regularisation strengths with L_2 (top) and L_1 (bottom) regularisation data drawn from a uniformly sample of $y = 1 + x^2$ plus noise. The red curve is the unregularised fit, the orange curve is the result with the lowest loss with respect to the validation set (triangles). The bar charts show the parameter values of the overfitted result and optimal regularised result. 42
4.4 Range of the side length to cover a fraction of the volume of a unit cube in up to nine dimensions. The grey lines show the fraction required to cover 25% of the volume. 45
4.5 ROC curve construction. On the left is the definition of the True Positive Rate (TPR) and the False Positive Rate (FPR) where TP is true positive, FP is false positive, TN is true negative and FN is false negative. In the centre and right are the distributions that are thresholded with the coloured lines in the central plot being cuts that correspond to the same coloured point on the right plot, a ROC curve. 45
4.6 Schematic of an artificial neuron(left) and a plot of three commonly-used activation functions (right). 47
4.7 Decision boundaries for a two-input (x_0, x_1), two-class neural network classifier with no hidden layer (left) and one hidden layer (right). The outputs of the networks are mapped to probabilities with the softmax function and are shown by the background contour plot. 49

4.8 Left: a multi-layer perceptron with no dropout. Right: the same network with dropout. Dropped neurons are shown greyed-out. 50
4.9 Convolution layer: a single neuron connected and its connection to a 4×4 patch of input (left) and an image patch with neurons in context with an input image [62] (right) 53
4.10 An example input to a pooling layer is shown on the left with two outputs on the right from max pooling (above) and average pooling (below). 54
4.11 A typical CNN architecture with three convolutional layers (grey) containing neurons with a restricted FOV (green), pooling layers for downsampling (orange), a flattening of the final feature map (purple) and a set of fully-connected layers. 55
4.12 The separate components of a composite layer. 56
4.13 A dense block with depth 5 and growth rate 4. Input feature volume is shown by the stack of white squares, each composite layer is shown as a grey square and the output feature volume of the layer is shown by the coloured layered stack. Coloured arrows show the which layers each feature volume is input to. The final concatenated output of the dense layer is shown by the white and coloured stack on the right. . . 56
4.14 A transition layer with a reduction factor of 0.5. 57

5.1 A comparison between data and simulation of dielectron invariant mass. 65
5.2 Photon ID BDT performance and validation. (Left) Photon ID BDT output score of the lower-scoring photon of each diphoton passing the photon preselection. Signal photons from simulated Higgs events are shown in red and simulated background events are shown in blue, data is shown by the back dots. (Right) Validation on $Z \to e^+e^-$ events. . . 66
5.3 Vertex ID efficiency of dimuon events reconstructed as diphotons as a function of p_T in simulation and data. 69
5.4 Vertex ID efficiency (dots) and average vertex probability (shaded band) as a function of diphoton p_T (left) and number of event vertices (right). 69

6.1 Stacked diphoton BDT score distributions for simulated signal and background, with data shown superimposed (left). Diphoton BDT score in the $Z \to e^+e^-$ control region (right). The same transformation has been applied to the score distribution in both plots such that the total signal distribution is flat. 75
6.2 Top quark decay modes: a fully-hadronic decay (left) and a semi-leptonic decay (right). 76

6.3 Score distribution of the hadronic $t\bar{t}H$ BDT. The blue lined histogram shows the distribution for the control region, the red filled histogram shows the score distribution for simulated signal, and the points show the score distribution of the data sideband regions ($m_{\gamma\gamma} < 115\,\text{GeV}$ or $m_{\gamma\gamma} > 135\,\text{GeV}$). 77

6.4 Dijet BDT feature distributions with the full VBF preselection. Distributions are all normalised to unity with the solid red line corresponding to VBF, blue line to ggH, and black line to SM background. The SM background distribution is shown as a stacked histogram. 84

6.5 Dijet BDT performance. On the left are the output score distributions for VBF (red), ggH (blue) and SM background (black). The SM background distribution is shown as a stacked histogram. On the right are the ROC curves for the dijet BDT split into the different samples. The performance against ggH is noticeably lower than the other backgrounds. 85

6.6 Combined BDT feature distributions with the full VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH. The SM background is shown as a stacked histogram. 85

6.7 Combined BDT score distribution with the full VBF preselection are shown on the left. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM backgrounds. The SM background distribution is shown as a stacked histogram. ROC curves broken down by background sample are shown in the same colours on the right. 86

6.8 Coloured regions correspond to mean values for top percentile (red) and bottom percentile (black) combined score events. The top and bottom five scoring events are also shown by lines and dots. 86

6.9 Mass fits for estimating AMS. 87

6.10 Data/Simulation comparison for dijet and combined BDT output scores. 88

6.11 Joint distribution study with BDT on the $Z \to e^+e^-$ control region data-simulation test set. 89

6.12 ROC curves for parton shower and underlying event variations. The nominal performance is shown in black, and the magenta lines show the upper and lower bounds of the envelope covering all the curves. . . 90

6.13 Single BDT performance and comparison to the two step approach. Distributions of the single BDT score are shown on the left and are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM backgrounds. The SM background distribution is shown as a stacked histogram. Corresponding ROC curves for the single BDT (centre) and original approach (right) are shown with colours denoting the same samples as the histogram. 91

6.14 Construction of single 12×12-pixel three-channel jet image (top). Arrows correspond to individual jet constituents where red arrows are charged, green are neutral and the opacity of the arrows corresponds to candidate p_T. The multiplicity channel is drawn separately, and black pixels lightened so the charged and neutral channels can be seen clearly. The final image (bottom) shown in both $(\Delta\eta, \Delta\phi)$ coordinates (left) and the $(\Delta R, \varphi)$ coordinates seen by the network (right). 93

6.15 Mean dijet images for VBF events (top), gluon fusion events (bottom left) and Standard Model background processes (bottom right). In each dijet image the left hand image corresponds to the leading jet and the right corresponds to the subleading jet. 94

6.16 A schematic view of the dense CNN model architecture. The convolutional section is indicated by blue, the merge section by red, and the main discriminant by purple. Grey squares and rectangles show layers of neurons: a depthwise convolution layer in the spread layer, composite layers in the dense blocks, and fully-connected layers in the merge section and main discriminant. 96

6.17 VBF/ggH discrimination performance for the image-only model with the full VBF preselection. The score distribution for VBF (red) and ggH (blue) is shown on the left. The associated ROC curve measuring VBF/ggH discrimination power is shown on the right. 101

6.18 Discrimination performance for all of the background samples with the full model and the full VBF preselection. Score distributions (left) are shown as stacked histograms for the SM background, a blue line histogram for ggH, and a red line histogram for VBF. Associated ROC curves (left) are shown in the same colours for each sample, with all background together shown in black. 102

6.19 Mass distributions and fits for the three optimised DCNN-based tag categories. 103

6.20 Generated images for feature visualisation which maximally activate the output neuron: VBF (top) and ggH (bottom). 104

6.21 Real images that maximally activate the class logits. The top image is the maximally-activating VBF image, and the bottom is the maximally-activating ggH image. 106
6.22 Front filters of the network grouped by the six image channels. Positive weight values are darker red and more negative values are darker blue, weights close to zero are shown as white. The vertical direction in the filters corresponds to φ and the horizontal direction corresponds to ΔR. 107
6.23 The effect of selected filters. Each subplot shows the effect of the filter and subsequent neural activation (left), the original preprocessed image (centre) and the filter (right). Clockwise starting from top left the filters are: radial gap detector, angular band smearer, general smearer, angular smear with gap in front, radial and angular smearer with shifts, double shifter. 108
6.24 $Z \to e^+e^-$ validation plots for the output scores of the image-only network (left) and the full network (right). 109
6.25 $Z \to e^+e^-$ Simulation/data discriminant performance. The score distribution for the simulation and data classes is shown on the left, and the associated ROC curve is shown on the right. 110
6.26 Sim/Data discriminant feature visualisation. The top dijet image is optimised for the simulation output neuron, and the bottom is optimised for the data neuron. 111
6.27 Mean images for top and bottom 5% maximally activating events. Score selection for simulation-like is shown at the top, selection for data-like is at the bottom. 112
6.28 Variation in the score distributions and ROC curves with parton shower and underlying event variation. The top pair corresponds to the image-only model (step 1 of training) and the bottom plots correspond to the full model used in the tag. 113

7.1 VBF category mass fits for the BDT-based VBF tag (left) and the DCNN-based VBF tag (right). Categories are shown in order from the most stringent to least: VBF 0 at the top to VBF 2 at the bottom. . . 124
7.2 Diphoton mass distribution plots for all categories combined using the BDT-based VBF tag. The unweighted combined distribution is shown on the left, and the sensitivity weighted combination is shown on the right. 126
7.3 Likelihood scan of the global signal strength modifier μ with a 2ΔNLL test statistic for analysis with the BDT-based VBF tag (top) and the DCNN-based VBF tag (bottom). 127

7.4 Likelihood scan results of the production mode signal strength modifiers μ with a 2ΔNLL test statistic. Analysis with the BDT-based VBF tag is shown at the top and the DCNN-based variant is at the bottom. 128

7.5 SM prediction to measured cross section ratios in the STXS Stage 0 framework. Analysis with the BDT-based VBF tag is shown on top and the DCNN-based variant is on the bottom. 130

7.6 Two-dimensional likelihood scan of signal strength modifiers for bosonic (VBF, VH) and fermionic (ggH, $t\bar{t}H$) production modes. Analysis with the BDT-based VBF tag is shown on the left and the DCNN-based variant is on the right. 131

7.7 Two-dimensional likelihood scan of κ values for bosonic versus fermionic production modes (left) and effective gluon coupling versus effective photon coupling (right). The BDT-based VBF tag is shown on the top, and the DCNN-based tag is shown at the bottom. 132

8.1 Fake celebrity faces generated by a progressive GAN [101]. 135

A.1 Dijet BDT performance and combined BDT performance evaluated with the loose preselection. 137

A.2 Dijet BDT feature distributions with the loose VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM background. 138

A.3 Dijet BDT feature distributions with the loose VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM background. 139

A.4 Single BDT performance with the loose VBF preselection. 139

A.5 Mean images in the loose VBF selection. From top to bottom: VBF, ggH and SM background processes. 140

A.6 Dense CNN model performance for images only (top) and the full model (bottom) in the loose VBF preselection. 141

B.1 $Z \to e^+e^-$ validation plots of pseudorapidity distributions for leading jet in p_T (left) and subleading jet (right). 143

B.2 $Z \to e^+e^-$ validation plots for kinematic features used by the VBF tag. Clockwise from top left: dijet mass, dijet pseudorapidity gap, subleading jet p_T, minimum ΔR between either photon and either jet, centrality, and leading jet p_T. 144

B.3 $Z \to e^+e^-$ validation plots for kinematic features used by the VBF tag. Clockwise from top left: leading photon p_T scaled by the diphoton mass, subleading photon p_T scaled by the diphoton mass, azimuthal angular difference between dijet and diphoton, total diphoton p_T scaled by diphoton mass, diphoton BDT score, and azimuthal angular difference between the dijet jets. 145

C.1 Feature visualisation of the spread layer features. Red is the charged p_T channel, green is the neutral p_T channel and blue is the PF candidate multiplicity channel. This layer only constructs features in individual channels. Optimisation objective is the mean of the values over a whole feature map. 148
C.2 Feature visualisation of the output of TU1. Here the low level features have been combined together to compare structure across channels, directly opposite around the jet axis and between the jets of the dijet. Optimisation objective is the mean of the values over a whole feature map. 149
C.3 Feature visualisation of the output of TU2. Here the features of TU1 are combined to make more complex features, but they are also reused (this is facilitated by the skip connections and is a capability of dense CNNs). Optimisation objective is the mean of the values over a whole feature map. 150
C.4 Feature visualisations of individual neuron values after TU3. These constitute the learned features used in the main discriminant. These images are optimised to maximally activate a single neuron rather than the mean of the neurons of one feature map. 151

D.1 Mass plots of the $t\bar{t}H$ tags. BDT-based VBF analysis is on the left and DCNN-based is on the right. 154
D.2 VH leptonic tags. BDT-based VBF analysis is on the left and DCNN-based is on the right. 155
D.3 VBF tag categories. BDT-based VBF analysis is on the left and DCNN-based is on the right. 156
D.4 VH MET and VH hadronic tags. BDT-based VBF analysis is on the left and DCNN-based is on the right. 157
D.5 Untagged categories 0 and 1. BDT-based VBF analysis is on the left and DCNN-based is on the right. 158
D.6 Untagged categories 2 and 3. BDT-based VBF analysis is on the left and DCNN-based is on the right. 159

List of Tables

2.1 Electroweak quantum numbers of the electroweak gauge bosons and the Higgs boson. 14
2.2 Electroweak quantum numbers of the leptons. 15
2.3 Electroweak quantum numbers of the quarks. 16
2.4 Main branching ratios of the Higgs boson. 19

5.1 Additional photon preselection requirements specific to different $|\eta|$ and R_9 regions. 67
5.2 Photon preselection efficiencies measured in four different bins. 67

6.1 The H $\rightarrow \gamma\gamma$ tag sequence in order of tag priority from highest (top) to lowest (bottom). 76
6.2 Pileup jet ID cuts of the tight working point. 82
6.3 Estimated category attributes for the BDT-based VBF tag. 88
6.4 Comparison table for BDT/DCNN AUROCs with full and loose preselections broken down by background sample. 102
6.5 Estimated category attributes for the DCNN-based VBF tag. 103

7.1 Expected signal yields per category for the BDT-based VBF tag (top) and the DCNN-based VBF tag (bottom). Only the downstream tags are shown for the DCNN-based tag as the others are unaffected. The width values σ_{eff} and σ_{HM} correspond to the smallest interval containing 68.3% of the $m_{\gamma\gamma}$ distribution and the width at half maximum of the signal peak respectively. 125
7.2 VBF tag category expected signal significances comparing the BDT-based VBF tag to the DCNN-based VBF tag. 126

8.1 Measurement results. 133

"ALL THIS IS A DREAM. Still, examine it by a few experiments."

Michael Faraday
Laboratory journal entry #10040

Chapter 1

Introduction

The Standard Model (SM) of particle physics has had its remarkable predictive power demonstrated through many experimental confirmations. At the core of this success is the Higgs field and the way its behaviour radically alters the phenomenology of a pristine, symmetric and massless theory.

The SM consists of a collection of fields whose quanta constitute fundamental matter particles and force mediators, as well as the couplings between them. These couplings determine the interactions in particle physics processes: the signals we observe in experiment. All of these particles, and many of their interactions, have been discovered by generations of high-energy particle physics experiments. This culminated in the completion of the field content of the SM with discovery of the Higgs boson itself in 2012 by the ATLAS and CMS collaborations [1, 2].

It is also known that the SM gives an incomplete description of nature. There is no dark matter candidate to explain experimental observations such as the bullet cluster [3], there is no description of the force of gravity, and neutrinos are considered massless when they are known not to be [4]. Various extensions to the SM have been suggested [5], and these can manifest as entirely new particles and as deviations in SM-expected rates for some processes.

This raises the question of precisely what sort of Higgs boson has been discovered. How does the Higgs field grant mass to the fermions? How does it self-interact and what is the shape of the potential it experiences? Are there any unexpected couplings that alter the Higgs boson's production and decay? As we enter the precision measurement era of Higgs physics we aim to answer these questions, and hopefully shed light on physics beyond the SM.

Precision measurement depends on high-quality data and superior signal extraction. The field of machine learning (ML) has produced many algorithms that are

used throughout experimental particle physics in both detector operation and data analysis. The discovery of the Higgs boson decaying to a diphoton system ($H \to \gamma\gamma$) in particular used these techniques in concert with its characteristically clean signal. Here signal extraction is enhanced using information from extra objects characteristic of certain predicted Higgs production modes and an algorithm called a boosted decision tree (BDT).

BDTs have been a reliable workhorse of particle physics for over a decade [6], but ML has made exceptional leaps in recent years thanks to deep learning (DL). In particular, DL applied to image recognition has resulted in powerful approaches that achieve super-human performance [7]. This thesis explores how to use these techniques to improve the extraction of $H \to \gamma\gamma$ events produced via Vector Boson Fusion (VBF). Specifically, VBF signal extraction is reformulated in part as an image classification problem where VBF's characteristic jets of particles are treated as images.

This thesis is based on the 2016 $H \to \gamma\gamma$ analysis [8], and is structured as follows: Chapter 2 begins by describing the theory underlying the SM, then the SM itself with emphasis on the Higgs sector and its phenomenology.

Chapter 3 describes the experimental apparatus used to produce and record the proton collision dataset used in this thesis: the Large Hadron Collider (LHC) and the Compact Muon Solenoid (CMS). CMS is described in detail with each detector subsystem's structure and operation explained with emphasis on the electromagnetic calorimeter.

Chapter 4 presents an introduction to machine learning covering basic theory, how to control model capacity for generalisation performance, ensembles (BDTs), plus how to design and tune an ML algorithm. The chapter then introduces neural networks and deep learning, culminating in the more advanced dense convolutional neural network (DCNN) models used in this thesis.

Chapter 5 describes how physics objects are reconstructed at CMS with emphasis on photons and how they are formed into $H \to \gamma\gamma$ diphoton candidates.

Chapter 6 describes how candidate $H \to \gamma\gamma$ events are categorised by different tags in the analysis for signal enhancement. Each of the tags is described in turn, but VBF will be described and validated in fine detail in both the BDT-based and DCNN-based variants. The DCNN will also be examined to determine what features it has learned to detect using a collection of network interpretation techniques.

Chapter 7 describes the final statistical analysis of the categorised $H \to \gamma\gamma$ candidates and the resulting measurements. The construction of the statistical models for signal and background are described, and a full description of all of the systematic uncertainties is given. Final results of yields and likelihood scans of signal strength and coupling modifiers performed for analyses with the DCNN-based VBF tags and

compared to the BDT-based results.

Finally, Chapter 8 discusses the conclusions we may draw as well as possible avenues for future development and research.

Chapter 2

Theory

2.1 Introduction

Modern particle physics theory is built upon the twin pillars of Yang-Mills theories and spontaneous symmetry breaking. Our best current model, the SM, is built from two such theories: electroweak theory and quantum chromodynamics. The former of these has its gauge symmetry spontaneously broken. In this chapter we will explore these two ideas before moving on to how they are used to construct the SM with particular emphasis on the mechanism of symmetry breaking and one of its phenomenological consequences: the Higgs boson.

2.2 Yang-Mills Theories

2.2.1 From Geometry to Gauge Fields

The gauge covariant derivative D_μ and the field strength tensor $F_{\mu\nu}$ are two vital mathematical objects when one wants to construct the Lagrangian of a Yang-Mills theory [9]. Far from simply being an ansatz, they have a deep origin in the fundamental geometry of field theory [10]. Their origin is outlined in this subsection: we start by describing the concept of a fibre bundle, its relationship to the internal symmetries of a field, and how the 'warping' of a fibre bundle is related to the covariant derivative and the field strength tensor.

A fibre bundle $\mathcal{B}$ is a space which can be considered to consist of two parts: the base space $\mathcal{M}$ and the fibre $\mathcal{V}$. For each point p in the base space there is an associated copy of the fibre space and these fibres do not intersect. In the context of a field one can consider these to be the external and internal spaces respectively. A special case is an ordinary product space where $\mathcal{B}$ is simply the Cartesian product of $\mathcal{M}$ and $\mathcal{V}$,

generally one has more warped examples with curvature and less trivial topology. A visual example is given in Figure 2.1. These warped examples are of interest in gauge

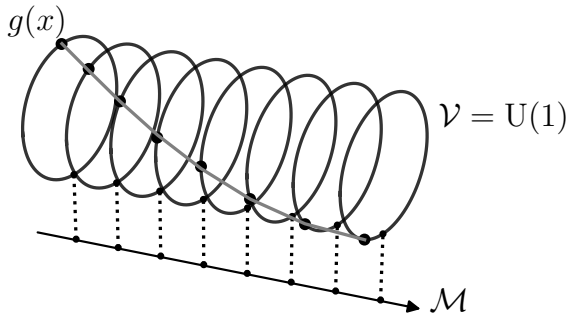

Figure 2.1: A fibre bundle with $\mathcal{V} = \mathrm{U}(1)$. A section is shown (grey line) choosing $g(x) \in \mathrm{U}(1)$ for a set of points in $\mathcal{M}$.

theory, specifically when there is curvature in the fibre space with no torsion.

Particularly, we are interested in the cases when $\mathcal{V}$ is symmetric under some Lie group $\mathcal{G}$ (in our case SU(N)). These symmetries allow for the warping of the fibre bundle and correspond to the internal symmetries of a field. Furthermore, one can model these examples by taking the fibre to be $\mathcal{G}$ with the identity element not at a fixed location. One can then 'lift' the base space into the bundle: for each base space point we get a point within the associated fibre. In the gauge theory context choosing a section of the fibre bundle means choosing a particular $g(x) \in \mathcal{G}$; this is picking a gauge.

To understand the warping of the fibre bundle we need the notion of a connection just like with the warped spaces of General Relativity. This will allow for the introduction of warping to the internal space, and construction of invariants such as curvature and torsion tensors. We can do this by constructing a differential operator D_μ, and in our case of a fibre bundle with $\mathcal{V} = \mathcal{G} = \mathrm{SU}(N)$ where the internal space is simply stretched with no torsion we have,

$$D_\mu = \partial_\mu - igA^a_\mu T^a, \tag{2.1}$$

where A^a_μ are generally complex-valued functions that depend on x_μ and operate by multiplying the input, and T^a are the generators of the Lie group SU(N) which provide a basis in the fibre space with $a = 0, \ldots, N^2 - 1$. We recognise this as having the familiar form of the gauge covariant derivative and the A^a_μ as the gauge potential.

Now we have the connection we can begin to construct invariants of the geometry

of the internal space. In particular we can construct the curvature tensor as follows

$$\frac{i}{g}[D_\mu, D_\nu] = F_{\mu\nu} = \partial_\mu A^b_\nu T^b - \partial_\nu A^a_\mu T^a - ig[A^a_\mu T^a, A^b_\nu T^b]. \tag{2.2}$$

We recognise this form as the field strength tensor.

One can now see what occurs when a global symmetry is promoted to a gauge symmetry: we have induced some non-trivial warping of the field's internal space that gives rise to the A^a_μ gauge fields and their kinematics through the curvature $F_{\mu\nu}$.

2.2.2 Constructing a Lagrangian

With these ingredients we can construct a generic Yang-Mills Lagrangian with a straightforward procedure: we begin with a global symmetry of the fields that we promote to a gauge symmetry, we construct the gauge covariant derivative, replace $\partial_\mu \to D_\mu$ in the free theory, and add an interaction term based on the field strength tensor [9]. As a concrete example, consider the collection of massive free Dirac fermions which we will turn into an interacting gauge theory with $\mathcal{G} = \mathrm{SU}(N)$. We first construct the gauge covariant derivative,

$$D_\mu = \partial_\mu - igA^a_\mu T^a, \tag{2.3}$$

and replace $\partial_\mu \to D_\mu$ in the free Lagrangian

$$\mathcal{L} = \sum_\alpha \bar{\Psi}^\alpha [i\gamma^\mu (D_\mu \Psi)^\alpha - m\Psi^\alpha]. \tag{2.4}$$

We must also introduce a kinematic term for the gauge fields, but the contraction of the general non-Abelian field strength tensor with itself is not gauge invariant, only its trace over the generator indices is. Therefore we use this as the gauge-invariant kinetic term for our final Yang-Mills Lagrangian [9],

$$\mathcal{L}_{YM} = \sum_\alpha \bar{\Psi}^\alpha [i\gamma^\mu (D_\mu \Psi)^\alpha - m\Psi^\alpha] - \frac{1}{2}\mathrm{Tr} F_{\mu\nu} F^{\mu\nu}. \tag{2.5}$$

2.2.3 Phenomenology

To analyse what sort of particle interactions occur in this theory we 'unpack' equation 2.5 and isolate the fields and interaction terms. Firstly, in the spectrum of this theory we have $N^2 - 1$ gauge fields (one for each of the generators of $\mathrm{SU}(N)$) that are all massless. These fields couple to the massive fermionic fields via a trilinear interaction term proportional to g introduced by the gauge covariant derivative.

$$\mathcal{L}_{A\Psi} = gA^a_\mu \bar{\Psi}^\alpha \gamma^\mu (T^a)_{\alpha\beta} \Psi^\beta \tag{2.6}$$

Now consider the gauge field kinetic term: one can reformulate this as $-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}$ using $\mathrm{Tr}T^aT^b = \frac{1}{2}\delta^a_b$ and $F_{\mu\nu} = F^a_{\mu\nu}T^a$. Once the product has been evaluated one finds the following forms of interaction terms

$$\mathcal{L}_{3A} \propto g f^{abc}(\partial_\mu A_{\nu\lambda}A^{\lambda a})A^{b\mu}A^{c\nu} \tag{2.7}$$

$$\mathcal{L}_{4A} \propto g^2 f^{abc}f^{ade}A^b_\mu A^c_\nu A^d_\lambda A^e_\sigma \tag{2.8}$$

that correspond to interactions between three and four gauge bosons respectively. We now have the three types of interaction vertices which allow for the construction of Feynman diagrams for a generic Yang-Mills theory (Figure 2.2). Their strengths are all set in terms of a single parameter: the gauge coupling g. One should note that the three and four-gauge boson interactions come from the commutator in the gauge field kinematic term and are not present in the Abelian case.

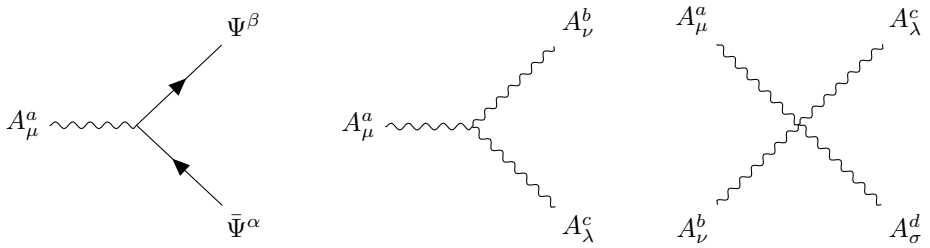

Figure 2.2: The three types of vertex in Yang-Mills theories.

2.3 Spontaneous Symmetry Breaking

Spontaneous symmetry breaking (SSB) occurs when the lowest energy solutions to a theory do not respect the symmetries of the Lagrangian that describes it. A straightforward example [11] is that of a three-dimensional ferromagnetic material cooling down from above its Curie temperature. Above this threshold there is no magnetisation and solutions obey the SO(3) symmetry of the Lagranian. Below this threshold the ferromagnet becomes magnetised and must 'choose' one of a degenerate family of lowest-energy solutions. This picks out a direction of magnetisation. The SO(3) symmetry of the ferromagnet has now been broken to SO(2) (Figure 2.3).

In the context of a field theory, symmetry can be spontaneously broken in the following way: a field experiences a potential whose minima are a family of degenerate states transforming under the symmetry group. Consider the Lagrangian of a complex

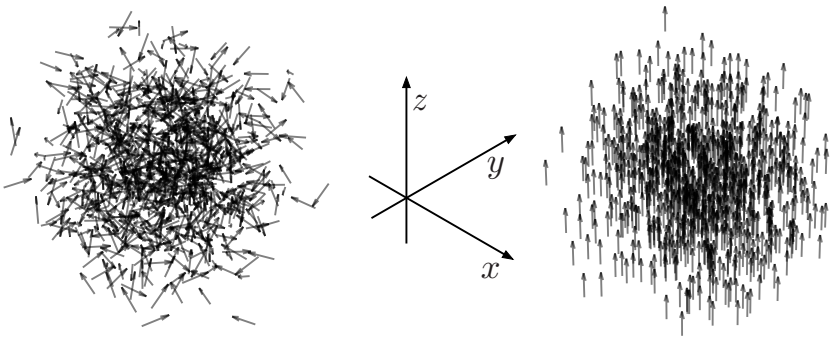

Figure 2.3: A ferromagnet above (left) and below (right) the Curie temperature. The example below the Curie temperature has magnetised along the z-direction breaking the SO(3) symmetry to just SO(2) about the z-axis.

scalar field ϕ experiencing a potential $V(\phi)$,

$$\mathcal{L} = (\partial_\mu \phi)^\dagger (\partial^\mu \phi) - V(\phi) \tag{2.9}$$

where the potential has the form

$$V(\phi) = -\mu^2(\phi^\dagger \phi) + \lambda(\phi^\dagger \phi)^2. \tag{2.10}$$

This has a global U(1) symmetry, $\phi \to e^{i\theta}\phi$, and the potential has a circle of degenerate minima at $|\phi| = \mu/\sqrt{2\lambda} = v$. The vacuum expectation value (VEV) of ϕ, $\langle\phi\rangle$ is now non-zero and will pick a state in this circle parameterised by $\langle\theta\rangle$ which can take any value θ_0. The global symmetry has been spontaneously broken. To see the

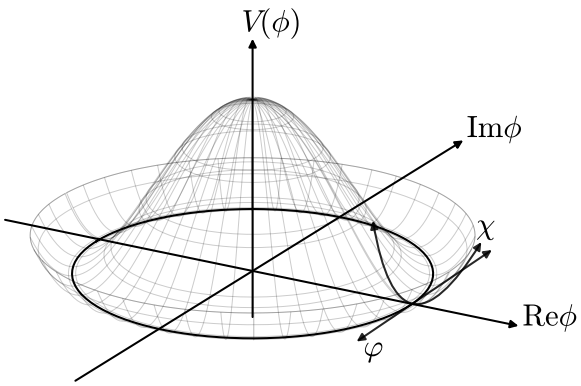

Figure 2.4: Perturbations around the vacuum state at $\theta_0 = 0$ of a symmetry breaking potential $V(\phi)$. The family of degenerate minima are shown by the black circle.

effects of this SSB consider a small perturbation around the vacuum with $\theta_0 = 0$ as shown in Figure 2.4. We describe ϕ in terms of two real scalar fields: one along the imaginary direction of ϕ (along the family of vacuum states) and one along the real (against the potential gradient),

$$\phi(x) = v + \frac{1}{\sqrt{2}}(\chi(x) + i\varphi(x)). \tag{2.11}$$

If one substitutes this into equation 2.9 and evaluates the non-kinematic part we find that the field χ is granted a mass term of the form $\frac{1}{2}m^2\chi^2$,

$$\frac{1}{2}m_\chi^2\chi^2 = \frac{1}{2}\lambda v^2\chi^2 \tag{2.12}$$

and there is no equivalent term for φ. In the spectrum of this theory we now have a massive and a massless scalar boson upon quantisation [12, 13]. This massless boson is known as a Nambu-Goldstone boson and is a general result of breaking a global symmetry: for each broken symmetry generator there is a massless Nambu-Goldstone boson.

To see this more clearly consider the case of a global SU(N) symmetry where the Lagrangian has the same form as before (equation 2.10) but ϕ is now a complex scalar N-tuple. There is a global symmetry $\phi \to e^{i\theta^a T^a}\phi$, and a family of minima at $\phi^\dagger\phi = \frac{\mu^2}{2\lambda}$ that form an (N^2-1)-dimensional surface instead of a circle. There are (N^2-1)-many ways to move on this surface and the one remaining direction is away from the centre. The former are the fields associated with the (N^2-1) Nambu-Goldstone bosons and the latter is the single massive scalar boson as before.

2.3.1 Gauge Symmetry Breaking

In the case where we have a gauge symmetry that is spontaneously broken the behaviour is rather different: there are no Nambu-Goldstone bosons and the gauge bosons are granted mass [14–17]. To see this take the example of equation 2.9 and consider a local SU(N) symmetry: we construct the gauge-covariant derivative

$$D_\mu = \partial_\mu + igA_\mu^a T^a, \tag{2.13}$$

replace the partial derivative, and introduce a gauge field kinetic term to get the gauge-invariant Lagrangian

$$\mathcal{L} = (D_\mu\phi)^\dagger(D^\mu\phi) - V(\phi) - \frac{1}{2}\mathrm{Tr}F_{\mu\nu}F^{\mu\nu}. \tag{2.14}$$

We can consider the field ϕ in its ground state in terms of its norm and a local SU(N) transformation, and then expand around v,

$$\phi(x) = e^{i\theta^a(x)T^a} \begin{pmatrix} 0 \\ \vdots \\ v + \frac{1}{\sqrt{2}}H(x) \end{pmatrix} \tag{2.15}$$

where H is a real scalar field corresponding to the direction orthogonal to the family of vacua and the θ^a correspond to the directions along its surface. The fields $\theta^a(x)$ now completely parameterise the vacua in contrast to the global case where it was an infinitesimal perturbation around a vacuum state. As a result of this we recognise that the SU(N) transformations can always be removed by some gauge transformation $\exp(-i\theta^a(x)T^a)$, so we can freely set it to zero. We have removed $2N-1$ degrees of freedom and we only have only one real scalar left: the gauge freedom has eliminated the Nambu-Goldstone bosons from the spectrum of the theory.

When we substitute equation 2.15 with $\theta^a(x) = 0$ into the Lagrangian 2.14 and then collect the terms that contain A_μ we get the following Lagrangian for the gauge fields (neglecting interaction terms)

$$\mathcal{L}_A = -\frac{1}{2}\mathrm{Tr}F_{\mu\nu}F^{\mu\nu} + g^2v^2A^a_\mu A^{a\mu} \tag{2.16}$$

This contains mass terms of the form $\frac{1}{2}m_A^2A_\mu A^\mu$, so we conclude that the fields $\theta^a(x)$ have indeed been eliminated and that these degrees of freedom have been absorbed into the longitudinal components of the gauge fields A^a_μ which have been granted mass $m_A^2 = 2g^2v^2$.

Collecting the scalar field H terms in the same way we have

$$\mathcal{L}_H = \frac{1}{2}(\partial_\mu H)(\partial^\mu H) - \lambda^2v^2H^2 \tag{2.17}$$

and we conclude that the theory contains a massive scalar field with $m_H^2 = 2v^2\lambda^2$ as in the global case. Upon quantisation fields such as H give rise to particles [16] called Higgs bosons. These fields have far-reaching consequences for theories of fundamental physics, playing a crucial role in the SM by granting mass to all the fundamental field quanta such as electrons and quarks and by breaking part of the gauge symmetry group of the SM.

2.4 The Standard Model of Particle Physics

The SM is a phenomenologically-motivated theory of fundamental particle interactions consisting of two Yang-Mills theories: one of the unified weak and electromag-

netic interaction (electroweak theory) and one of the strong interaction (quantum chromodynamics). This section will treat these in turn using the theoretical machinery presented in previous sections. At the end of this section the resulting Higgs boson and its behaviour will then be discussed.

2.4.1 Electroweak Theory

The electroweak unification model of Glashow, Weinberg and Salam [18–20] marks the birth of the SM and our modern understanding of fundamental physics. In this subsection we will begin by constructing the interaction itself as a massless gauge theory and then break its gauge symmetry via the Brout-Englert-Higgs mechanism. We will then move on to introduce the leptonic sector and then the quark sector discussing their dynamical mass generation and properties.

Gauge Fields and the Higgs Field

We begin with a Yang-Mills theory consisting of a complex scalar SU(2) doublet ϕ

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.18}$$

and a global symmetry group $\mathcal{G} = \mathrm{SU}(2) \times \mathrm{U}(1)$ experiencing a potential $V(\phi)$ of the same form as equation 2.10. We build the gauge-covariant derivative

$$D_\mu = \partial_\mu \mathbb{1} + igW_\mu^a T^a + \frac{ig'}{2} y B_\mu \mathbb{1} \tag{2.19}$$

where W_μ^a are the gauge fields corresponding to each of the generators of the SU(2) subgroup of $\mathcal{G}$, the τ^a are the SU(2) generators (Pauli Matrices), B_μ is the gauge field corresponding to the Abelian subgroup U(1), and g,g' are the gauge couplings corresponding to the SU(2) and U(1) respectively. The internal field space here has two complex dimensions and the internal geometry corresponds to a unit circle in the 2D complex space (SU(2)) warped by a position-dependent complex phase (U(1)).

We construct the following Lagrangian for the complex scalar theory

$$\mathcal{L} = (D_\mu \phi)^\dagger (D_\mu \phi) - V(\phi) - \frac{1}{2}\mathrm{Tr} F_{\mu\nu} F^{\mu\nu} - \frac{1}{4} G_{\mu\nu} G^{\mu\nu} \tag{2.20}$$

where $G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ is the Abelian field strength tensor. When $\mu^2 > 0$, ϕ adopts a ground state from the family of minima, gains a non-zero vacuum expectation value (VEV) and breaks the SU(2) subgroup. As described previously the SU(2)-associated weak gauge fields will gain mass terms, but there are extra complications. Consider

the gauge fields in terms of their physical states $W_\mu^\pm$

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2) \tag{2.21}$$

and the isospin structure of D_μ is shown explicitly by writing it in matrix form

$$D_\mu = \begin{pmatrix} \partial_\mu & 0 \\ 0 & \partial_\mu \end{pmatrix} + \frac{ig}{\sqrt{2}} \begin{pmatrix} 0 & W_\mu^+ \\ W_\mu^- & 0 \end{pmatrix} + \frac{i}{2} \begin{pmatrix} gW_\mu^3 + g'yB_\mu & 0 \\ 0 & -gW_\mu^3 + g'yB_\mu \end{pmatrix}. \tag{2.22}$$

Note that both the third component of the SU(2) gauge field and B_μ both multiply a diagonal matrix in the internal isospin space, as a result of this the symmetry breaking pattern is more complex and the two fields will later need to be unmixed. If we substitute this expression into the Lagrangian (equation 2.20) and gather terms quadratic in the fields we have,

$$\begin{aligned} \mathcal{L} =& (\partial_\mu H)(\partial^\mu H) - 4\lambda v^2 H^2 \\ &+ \frac{1}{2} g^2 v^2 W_\mu^+ W^{\mu -} \\ &+ \frac{1}{4} v^2 (g W_\mu^3 - g' y B_\mu)(g W^{\mu 3} - g' y B^\mu) \\ &- \frac{1}{2} \mathrm{Tr} F_{\mu\nu} F^{\mu\nu} - \frac{1}{4} G_{\mu\nu} G^{\mu\nu} \end{aligned}. \tag{2.23}$$

We observe that there is a mass term present for the Higgs field H and the charged weak bosons $W^\pm$, however, the quadratic terms for W^3 and B are 'mixed' and we do not have a simple mass term for W^3 and a massless B field. These fields must be unmixed by performing a rotation in the internal field space

$$\begin{aligned} A_\mu &= \cos\theta_W B_\mu + \sin\theta_W W_\mu^3 \\ Z_\mu &= -\sin\theta_W B_\mu + \cos\theta_W W_\mu^3 \end{aligned} \tag{2.24}$$

where θ_W is called the weak mixing angle and is defined as $\tan\theta_W = g'/g$. The field Z_μ now picks up a mass term,

$$\frac{1}{2} m_Z^2 Z_\mu Z^\mu = \frac{1}{4} v^2 (g^2 + g'^2)^2 Z_\mu Z^\mu \tag{2.25}$$

and the field A_μ does not have a mass term. Upon quantisation these are the neutral weak boson, Z, and the photon of electromagnetism. We can also examine the Abelian part of the gauge covariant derivative with the unmixed fields,

$$D_\mu^{\mathrm{Abel}} = \partial_\mu + ig \sin\theta_W (T^3 + \frac{1}{2} y) A_\mu \tag{2.26}$$

this leads to the interpretation of $T^3 + \frac{1}{2}y$ as the electromagnetic charge operator where T^3 is the third component of weak isospin and y is the hypercharge.

We now have four vector bosons and one scalar boson: the three weak bosons of the weak interaction ($W^{\pm}_{\mu}, Z_{\mu}$), the photon (A_{μ}) of the electromagnetic interaction and the Higgs boson (H) with the quantum numbers shown in table 2.1.

Particle	T^3	y	$Q = T^3 + \frac{y}{2}$
$W^{\pm}$	± 1	0	± 1
Z	0	0	0
γ	0	0	0
H	$-\frac{1}{2}$	1	0

Table 2.1: Electroweak quantum numbers of the electroweak gauge bosons and the Higgs boson.

Leptons

Leptons, fermionic constituents of the SM that interact only via electroweak interactions, must be introduced in a more careful fashion than in equation 2.5. Firstly, neutrinos are assumed massless in the SM (but this not the case in nature [4]). Experiment also observes that neutrinos have left-handed chirality [21], and that there are processes involving the decay of $W^- \to e^- + \bar{\nu_e}$. Therefore we begin by assigning each lepton and their counterpart neutrino to a weak isodoublet with $T_3 = \pm\frac{1}{2}$

$$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}, \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}, \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}. \tag{2.27}$$

However, the fact that there are no right-handed neutrino interactions necessitates a different structure: we need to split the leptonic isodoublets into left and right-handed versions with the projection operator

$$\ell_e = \begin{pmatrix} \nu_e \\ e^-_L \end{pmatrix}, e^-_L = \frac{1-\gamma^5}{2} e^- \tag{2.28}$$

and we note that the SU(2) gauge symmetry is actually $\mathrm{SU}(2)_L$ which denotes operation with the chirality operator along with the elements of the group, and the L subscript has been omitted from the neutrino field as it is assumed that they are only left-handed. We can also write the right-handed lepton doublets as

$$\begin{pmatrix} \frac{1+\gamma^5}{2}\nu_e \\ \frac{1+\gamma^5}{2} e^- \end{pmatrix} = \begin{pmatrix} 0 \\ e^-_R \end{pmatrix}. \tag{2.29}$$

This transforms as a singlet under $\mathrm{SU}(2)_L$ due to the properties of the projection operator. The electroweak quantum numbers of the leptonic families are shown in table 2.2

Particle	T^3	y	$Q = T^3 + \frac{y}{2}$
ν_e, ν_μ, ν_τ	$\frac{1}{2}$	-1	0
$\bar{e}_L, \bar{\mu}_L, \bar{\tau}_L$	$-\frac{1}{2}$	-1	-1
$\bar{e}_R, \bar{\mu}_R, \bar{\tau}_R$	0	-2	-1

Table 2.2: Electroweak quantum numbers of the leptons.

To preserve gauge invariance (due to the singlet nature of the right-handed leptons), and to grant mass only to the lower component of the lepton weak isodoublets, mass is granted dynamically via Yukawa couplings [11]. For each isodoublet there is a coupling to the Higgs field ϕ of the form

$$g_f(\bar{e}_R\phi^\dagger\ell_e + \bar{\ell}_e\phi e_R), \tag{2.30}$$

which is invariant under $\mathrm{SU}(2)_L$. Upon spontaneous symmetry breaking the Higgs field vacuum expectation value generates mass terms and interactions of the form

$$g_f v(\bar{e}_R e_L + \bar{e}_L e_R) + g_f(\bar{e}_R e_L H + \bar{e}_L e_R H) \tag{2.31}$$

where we recognise the left hand part of the expression as a fermionic mass term with $m_f = g_f v$, where g_f is the Yukawa coupling strength. The Lagrangian of the leptonic sector of the SM is then

$$\begin{aligned}\mathcal{L} = &-\frac{1}{2}\mathrm{Tr}F_{\mu\nu}F^{\mu\nu} - \frac{1}{4}G_{\mu\nu}G^{\mu\nu} \\ &+ (D_\mu\phi)^\dagger(D_\mu\phi) + \mu^2(\phi^\dagger\phi) - \lambda(\phi^\dagger\phi)^2 \\ &+ i\sum_{f=e,\nu,\tau}(\bar{\ell}_f\gamma^\mu D_\mu\ell_f + g_f(\bar{f}_R\phi^\dagger\ell_f + \bar{\ell}_f\phi f_R)) \\ &+ i\sum_{f=e,\nu,\tau}(\bar{f}_R\gamma^\mu D^Y_\mu f_R),\end{aligned} \tag{2.32}$$

where D^Y_μ denotes the part of the covariant derivative that corresponds to the hypercharge, and f labels lepton generation.

Quarks

To complete the fermionic content of the SM we must include quarks: fermions with fractional electric charge that transform non-trivially under the full SM gauge group

[22]. In analogy with the leptons we begin by grouping the quarks into three generations of $\mathrm{SU}(2)_L$ isodoublets with $T_3 = \pm\frac{1}{2}$

$$\begin{pmatrix} u \\ d \end{pmatrix}, \begin{pmatrix} c \\ s \end{pmatrix}, \begin{pmatrix} t \\ b \end{pmatrix}, \tag{2.33}$$

where from left to right and top to bottom we have the up, down, charm, strange, top and bottom quarks. There are also right-handed singlet fields for each flavour of quark.

The introduction of the electroweak interaction is performed in the same way as before with the introduction of the covariant derivative and the breaking of the $\mathrm{SU}(2)_L$ subgroup by the Higgs mechanism. The quantum numbers of the quarks are shown in table 2.3.

Particle	T^3	y	$Q = T^3 + \frac{y}{2}$
u, c, t	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{3}$
d, s, b	$-\frac{1}{2}$	$\frac{1}{3}$	$-\frac{1}{3}$
u_R, c_R, t_R	0	$\frac{4}{3}$	$\frac{2}{3}$
d_R, s_R, b_R	0	$-\frac{2}{3}$	$-\frac{1}{3}$

Table 2.3: Electroweak quantum numbers of the quarks.

However, the mechanism for the generating quark masses is slightly different. One still uses couplings of the same form as before, but a few modifications are required [22] to generate masses for the up-type quarks which would remain massless if we proceeded in the exact same way as for leptons. Firstly, note that the following also transforms as an $\mathrm{SU}(2)_L$ doublet

$$\phi_C = i\tau_2\phi^* = \begin{pmatrix} \phi^{0*} \\ -\phi^{+*} \end{pmatrix}, \langle\phi_C\rangle = \begin{pmatrix} v \\ 0 \end{pmatrix}, \tag{2.34}$$

where τ_2 is the second Pauli matrix. Now, when the Higgs field isodoublet gains a vacuum expectation value this transformed version has the value in the upper part of the isodoublet and one can use this to construct gauge-invariant masses for the quarks of the following form

$$\sum_{i=1,2,3} g_i(\bar{u}_{iR}\phi^\dagger q_{iL} + \bar{d}_{iR}\phi_C^\dagger q_{iL} + \text{h.c.}) \tag{2.35}$$

where h.c. denotes Hermitian conjugate. This grants equal masses to the up and down-type quarks in disagreement with experiment. We therefore have to introduce

matrix-valued couplings which mix the flavours

$$\sum_{i,j=1,2,3} g(\alpha_{ij}\bar{u}_{iR}\phi^{\dagger}q_{jL} + \beta_{ij}\bar{d}_{iR}\phi_C^{\dagger}q_{jL} + \text{h.c.}). \tag{2.36}$$

Finally, in the physical flavour-changing currents of the SM it is combinations of down type quarks that appear. Each type of quark is measured to have a preference for their own generation but can also decay through the weak interaction to others [22]. We can consider the down-type quarks to be 'rotated' in flavour space such that the lower component of the quark isodoublets are actually mixed between d, s, b. We therefore replace the down part of each with

$$\begin{pmatrix} u \\ d' \end{pmatrix}, \begin{pmatrix} c \\ s' \end{pmatrix}, \begin{pmatrix} t \\ b' \end{pmatrix}, q'_f = \sum_{f'=d,s,b} V_{ff'}q_{f'} \tag{2.37}$$

where the $V_{ff'}$ are elements of the Cabbibo-Kobayashi-Maskawa (CKM) matrix [23, 24], a unitary matrix that performs the required rotation in flavour space.

2.4.2 Quantum Chromodynamics

As mentioned previously quarks are the only fermions of the SM to transform non-trivially under the full SM gauge group. This means that they experience an extra interaction from the gauging of the SU(3) subgroup, the strong interaction, and carry colour charge. This is described by quantum chromodynamics (QCD) [22], the other Yang-Mills theory that constitutes the SM.

To construct the Lagrangian of QCD we begin by defining the SU(3) covariant derivative

$$D_\mu = \partial_\mu + ig_s\lambda^a A^a_\mu, \tag{2.38}$$

where λ^a are the generators of SU(3), and g_s is the QCD gauge coupling. We then replace ∂_μ in a Dirac-type Lagrangian that has the corresponding non-Abelian field strength tensor $F^a_{\mu\nu}$ and whose fermions are the six flavours of quarks. Each of these are isodoublets that also carry the colour charge q_f^C, $C = R, G, B$ (red, green, blue) and are structured in colour triplets transforming under SU(3)

$$\psi_f = \begin{pmatrix} q_f^R \\ q_f^G \\ q_f^B \end{pmatrix}. \tag{2.39}$$

The Lagrangian is then

$$\mathcal{L}_{\text{QCD}} = \sum_f i\bar{\psi}_f \gamma^\mu D_\mu \psi_f - \frac{1}{2}\text{Tr}F_{\mu\nu}F^{\mu\nu}, \tag{2.40}$$

where the index f denotes quark flavour. Upon quantisation we will have eight massless gauge bosons called gluons that couple to quark pairs of the same flavour.

The strong interaction behaves differently to the weak or electromagnetic interactions: instead of the weakening with distance the strong force increases in strength. An important result of this phenomenon is that colour-carrying particles such as quarks and gluons are confined and can only exist within composite particles called hadrons [22]. This is responsible for the phenomenon of jets in high-energy particle collisions.

Jets are collimated cone-shaped sprays of particles that result from quark or gluon (parton) production [25]. When produced these will radiate other partons in a similar process to an electromagnetic particle shower (parton showering). Collinear gluon emission is much more common during this process giving the conical shape of the jet. However, due to confinement the partons can only exist bound within composite particles and therefore must hadronize to form colourless hadrons. These hadrons may further fragment or decay to daughter particles.

We now have the full fundamental particle content of the SM and the couplings between them. Next we will consider how the above is manifested as the phenomenology of the Higgs boson.

2.4.3 Higgs Boson Phenomenology

The Higgs field couples directly to every massive particle in the SM [22]: either through the gauge-covariant derivative that gives interaction between the Higgs boson and the weak gauge bosons, or the Yukawa couplings which cause interaction with the non-neutrino fermions. In this section we will look at how Higgs bosons are created in proton-proton collisions at the Large Hadron Collider, and their subsequent decays.

Higgs Boson Production in Proton Collisions

There are four main ways that a Higgs boson can be produced in proton collisions: gluon fusion (ggH), vector boson fusion (VBF), associated production (VH), and top fusion ($t\bar{t}H$) (Figure 2.5). At $\sqrt{s} = 13\,\text{TeV}$ gluon fusion dominates over the other processes for a Higgs boson of mass $m_H = 125\,\text{GeV}$ from proton collisions with a cross section of 49 pb. VBF is the second largest with 3.8 pb, VH is third with 2.3 pb and $t\bar{t}H$ is last with 0.5 pb [26]. Although ggH has by far the largest crossection, the other production modes contain extra objects in their final state aiding the separation

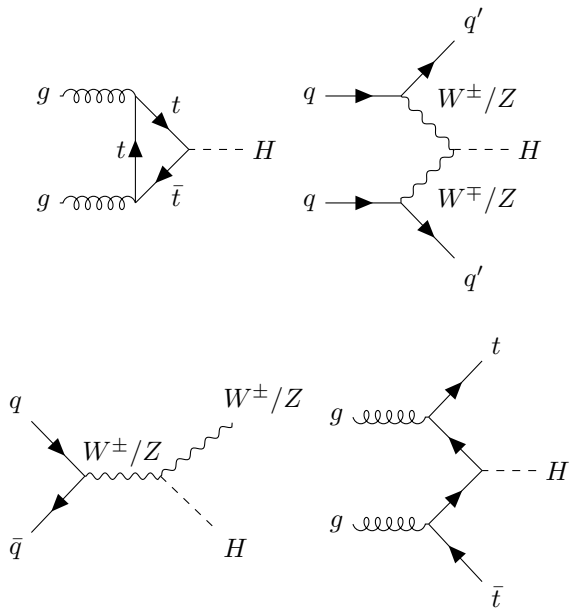

Figure 2.5: The four main production modes of the Higgs boson at the LHC. Clockwise from top left: ggH, VBF, $t\bar{t}$H and VH.

of Higgs bosons from events that resemble them. In particular VBF gives rise to two highly energetic jets from the final-state quarks with a large angular separation.

Higgs Boson Decays

Once produced, the Higgs boson is predicted to decay very quickly [27] to pairs of particles. At tree-level it will decay into massive particles in proportion to their masses in two different ways: in the case of particles granted mass through the Yukawa couplings (fermions) the branching ratio will be proportional to the square of the mass, in the case of particles granted mass through the gauge-covariant derivative the branching ratio will be proportional to the fourth power of the mass (weak gauge bosons). The Higgs boson can also decay via loop diagrams that have a reduced branching ratio. The prevalences of the main decay modes [28] are summarised in table 2.4.

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

Table 2.4: Main branching ratios of the Higgs boson.

The H $\rightarrow \gamma\gamma$ decay (Figure 2.6) is of particular interest in experimental searches

despite its relatively small branching ratio. It has a simple, fully-reconstructed final state with no composite objects such a jets or missing momentum that cause difficulty in the high-multiplicity hadronic environment of the LHC. This decay mode's clean signal led to it being one of the two channels in which the Higgs boson was discovered in 2012.

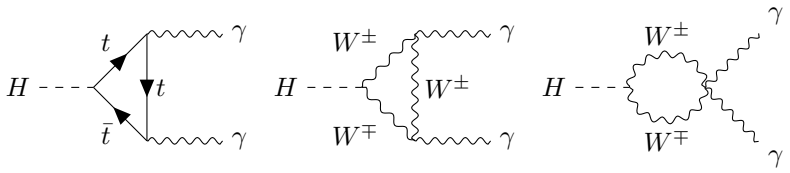

Figure 2.6: Main contributing diagrams to the Higgs boson to diphoton decay mode.

Chapter 3

Apparatus

3.1 Introduction

This chapter describes the experimental apparatus used to produce the 2016 proton-proton collision dataset used in this thesis. A description of the means of collision production, the Large Hadron Collider (LHC), will be given and their measurement with the Compact Muon Solenoid (CMS) will be described in particular detail. The CMS design and operation described here will correspond to the 2016 running period.

3.2 The Large Hadron Collider

The LHC [29] is a large synchrotron-type particle accelerator whose purpose is to provide collisions to survey electroweak-scale physics, particularly the mechanism of electroweak symmetry breaking, in addition to a broad physics program ranging from dark matter searches to flavour physics to studies of quark-gluon plasma. These studies are afforded by the production of high-energy and high-luminosity proton and lead ion collisions in counter-circulating beams. These achieve a long reach for the production of heavy particles and superior statistical power for the study of rare processes. The remainder of this section will be solely concerned with the LHC's proton-proton operation.

3.2.1 LHC Accelerator Chain

Before collisions occur at the LHC interaction points the beams must be produced and conditioned with a collection of accelerators [30] before injection into the LHC (Figure 3.1). This process begins with the acquisition of protons from a small quantity of hydrogen gas: hydrogen is injected into a chamber and the atomic electrons are

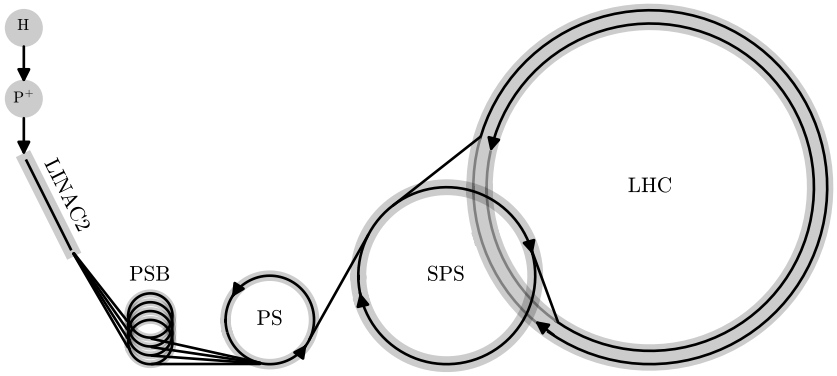

Figure 3.1: A schematic view of the LHC accelerator chain for proton-proton operation.

stripped off using a strong electric field. The resulting bare protons are then injected into a linear accelerator (LINAC 2) and accelerated by radio-frequency (RF) cavities to an energy of 50 MeV. The protons then enter the Proton Synchrotron Booster (PSB), consisting of four synchrotron rings stacked on top of each other with a radius of 25 m, the protons are further accelerated up to an energy of 1.4 GeV allowing for more protons to be injected into the next part of the accelerator chain and therefore higher-intensity beams. The protons from each ring of the PSB enter the Proton Synchrotron (PS) in sequence with 25 ns spacing to form the bunch structure. The PS is another synchrotron with a radius of 72 m where they are accelerated to 25 GeV. When they have reached this energy the protons are then injected into the Super Proton Synchrotron (SPS) which has a circumference of nearly 7 km and accelerates protons to 450 GeV before their injection into the LHC in two opposing directions.

3.2.2 LHC Structure and Operation

The LHC itself is a 27 km ring consisting of 1232 8 T superconducting dipole magnets that force the protons into a circular path so they can be repeatedly accelerated by 16 superconducting RF cavities oscillating at 400 MHz. As on-time protons with correct energy come in to the cavity they do not experience any acceleration, if they arrive slightly later they experience an accelerating potential, slightly early and they experience a deceleration. This maintains the energy of the protons and the bunch structure as they circulate in the LHC ring. The beams are then further adjusted by 392 quadrupole magnets to maintain stable beam conditions and stronger quadrupole magnets are used to focus the beams at the four LHC collision points.

Bunches of protons are brought together to collide at each of the LHC interaction points every 25 ns during normal operation. This is referred to as a bunch crossing and usually produces multiple superimposed proton collisions (pileup) and a large dose

of radiation for any equipment situated nearby. These conditions pose challenges for the design and operation of the LHC's experiments.

The LHC is designed to operate with a centre of mass energy of $\sqrt{s} = 14\,\text{TeV}$ and instantaneous luminosity of $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ with two beams of 2880 bunches. During the 2016 period the LHC operated at $\sqrt{s} = 13\,\text{TeV}$ and an instantaneous luminosity above design specification at $1.4 \times 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$. This culminated in $40.82\,\text{fb}^{-1}$ of integrated luminosity delivered to the CMS experiment in the 2016 period with an average pileup rate of 27 [31] (Figure 3.2).

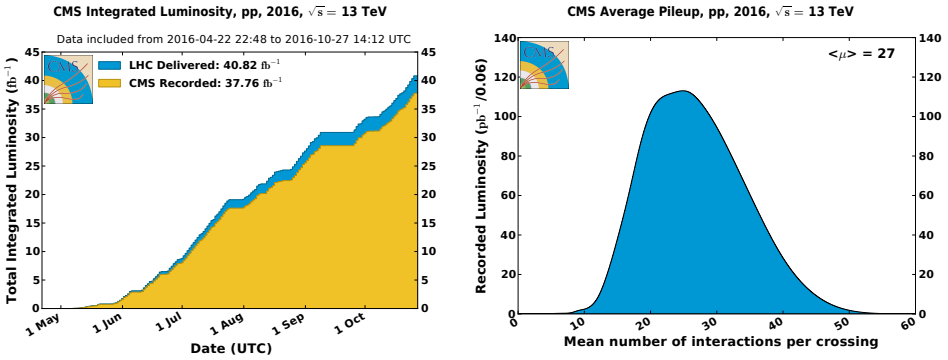

Figure 3.2: Left: total integrated luminosity over the 2016 proton-proton running period delivered to (blue) and recorded by (orange) the CMS experiment [31]. Right: the 2016 pileup distribution [31].

3.3 The Compact Muon Solenoid

The CMS experiment [32] is a general-purpose detector situated at Point 5 on the LHC directly opposite its counterpart, ATLAS [33]. CMS uses a right-handed coordinate system with the x-axis pointing horizontally towards the centre of the ring, the y-axis pointing vertically, and the z-axis pointing along the beamline in the anti-clockwise direction. An angular coordinate system is commonly used in physics analyses which consists of the coordinates (ϕ, η, z), where ϕ is the azimuthal angle in the x-y plane and η is the pseudorapidity defined from the polar angle θ as

$$\eta = -\ln\tan\frac{\theta}{2}. \tag{3.1}$$

Generally, the high $|\eta|$ regions closer to the beamline are referred to as forward, and the low $|\eta|$ region is referred to as central. In addition, the radial distance in the x, y plane r is sometimes used. The different coordinates used at CMS are summarised in Figure 3.3. This thesis will use the (ϕ, η, z) coordinate system.

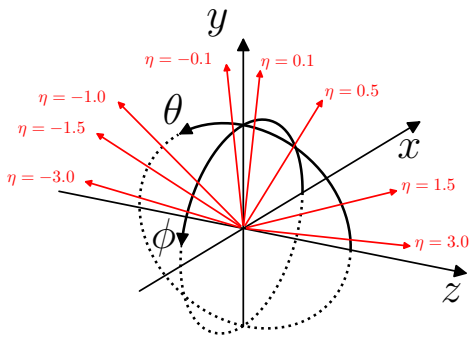

Figure 3.3: Coordinate systems used at CMS. Example values for the pseudorapidity η are shown by the red arrows.

3.3.1 Design Overview

The design of CMS is driven by the challenges of operating in the LHC collision environment and by the broad range of its physics goals [34]. The LHC produces proton-proton bunch crossings at a high rate (40 MHz) and this requires CMS to be very responsive to facilitate a short decision time on accepting an event. This high collision rate also means that the components of CMS operate in a high-radiation environment, so their performance must be robust to large doses of radiation. Finally, the pileup in each bunch crossing puts particular requirements on the CMS design to achieve isolation of different kinds of particles in a complex, high-multiplicity environment: it requires fine granularity, both spatial and temporal.

The physics goals of CMS include the discovery and measurement of the Higgs boson and searches for supersymmetry amongst other topics like extra gauge bosons, extra dimensions and heavy ion collisions. The CMS Higgs physics programme has prioritised searches for the Higgs boson in the leptonic final states as well as the diphoton final state as these have superior signal separation potential and mass resolution in the LHC collision environment when compared with hadronic searches. For supersymmetry searches one expects events with a significant degree of missing-transverse energy (E_T^{miss}) and this, along with maximising the acceptance of other analyses, motivates the hermetic design of CMS. Therefore, the main CMS performance goals were decided to be [32]:

- Good identification of muons and good muon momentum resolution,
- Good charged particle momentum resolution and reconstruction efficiency,
- Efficient triggering and offline tagging for τ leptons and b quark jets,
- Good energy resolution for electromagnetically interacting particles over a large geometric area,

- Good π^0 rejection,
- Good missing-transverse energy and dijet mass resolution.

The design of CMS, shown in Figure 3.4, is driven by these goals.

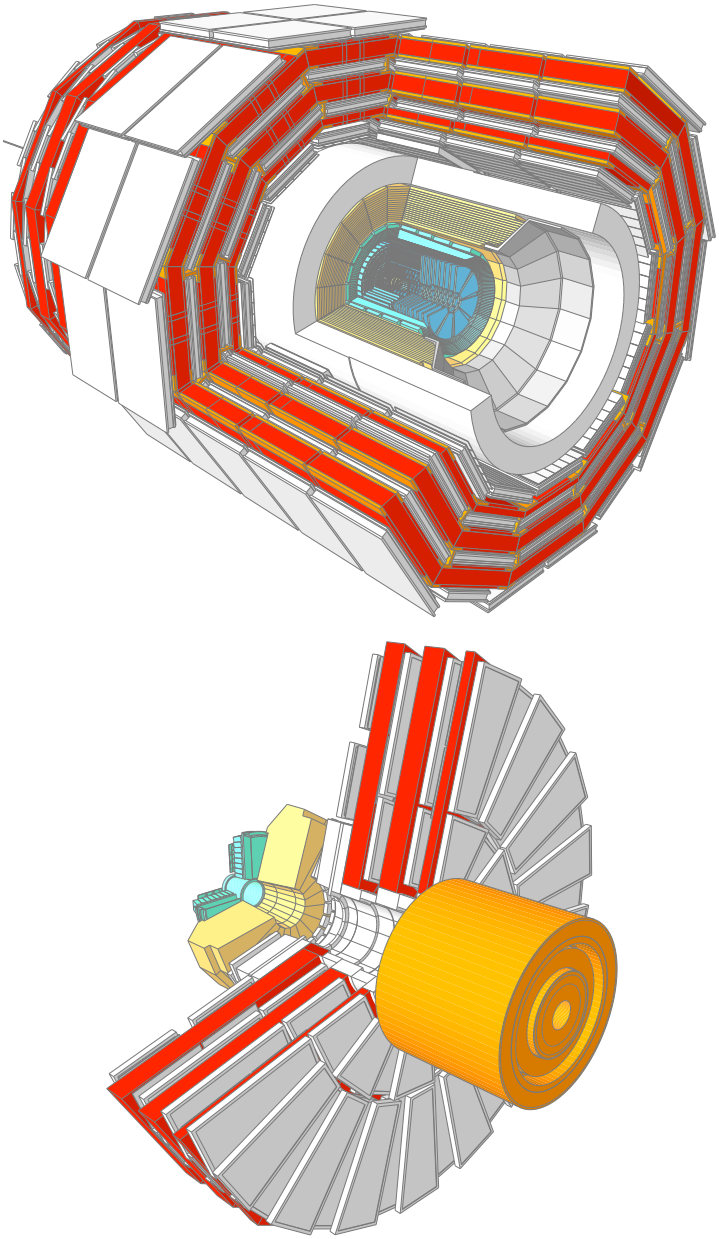

Figure 3.4: The CMS experiment separated into barrel (top) and endcap (bottom), both have an azimuthal section removed to show structure of the detector subsystems. Rendering was built with the model in [35].

The general structure of CMS is a classic hermetic design with a main cylindrical section centred around the interaction point called the 'barrel' that is then sealed by two 'endcaps'. This gives a detector with a pseudorapidity range from -5 to 5 that almost covers the entire 4π solid angle. Both barrel and endcaps consist of multiple concentric detector subsystems with different functionality, all of this is based around the main feature of the CMS detector: its large superconducting solenoid. The CMS solenoid, along with the steel return yoke it is supported by, achieves a high-strength and homogenous magnetic field over a large volume. This field bends the trajectories of charged particles into a helix, and when this bend is measured accurately one can achieve a precise momentum measurement. This meets the performance requirements for momentum resolution of charged particles and especially muons. Within the bore of the solenoid there are three subsystems: the tracker, the electromagnetic calorimeter (ECAL) and the hadron calorimeter (HCAL). The tracker consists entirely of silicon-based sensors and performs precise measurements of charged particle trajectories, at the centre are pixel detectors which allow CMS to meet its τ lepton and b-quark jet tagging goals by reconstructing tracks and secondary vertices with great precision. After this, the ECAL measures the energy of electromagnetically interacting particles with good resolution. This energy resolution allows for excellent mass resolution for dilepton and diphoton objects and is crucial to the measurement of Higgs boson decays to $\gamma\gamma$ and $ZZ^{(*)}$. Situated around the ECAL, the HCAL measures the energies of neutral hadrons and covers a large area with fine granularity to achieve hermeticity and to meet the objective of good E_T^{miss} measurement. Finally, the muon detectors are sited around the outside of the solenoid and cover a large area. These systems are interleaved with the steel return yoke and measure muon trajectories and energy precisely to achieve good muon momentum resolution and particle identification with a fast response. Each of these subdetector systems, with the exception of the tracker, provide fast measurements for the CMS trigger system. This allows CMS to cope with the very high data rate by making fast decisions about whether to keep events. Each of these systems will be described in detail in the following subsections. Particular attention will be given to the ECAL due to its importance to the measurement of the Higgs diphoton decay mode.

3.3.2 Solenoid and Return Yoke

The CMS magnet is a liquid helium-cooled superconducting solenoid, 12.5 m in length with a bore of diameter 6 m and a mass of 220 t (including all systems operating at cryogenic temperature) [34]. This is situated within a steel return yoke [36] that weighs 12000 T extends outwards to a diameter of 14 m and is made of five barrel 'wheels' with three layers, and then completed by three disks in each endcap (Figure 3.5). The niobium-titanium coils of the solenoid carry 18160 A of current that pro-

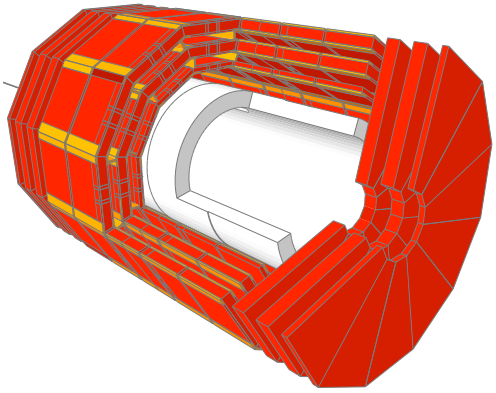

Figure 3.5: The CMS solenoid (white) within the steel return yoke (red). Rendering uses [35].

duces a magnetic field strength within the bore of 3.8 T and a stored energy of 2.3 GJ. This field strength is below the design capability of 4 T to maximise the operating lifetime of the solenoid. The return yoke then conditions this field, increasing the strength in the bore by a small amount (8%) [36] and improving the field homogeneity in the barrel and the muons systems.

3.3.3 Inner Tracking

The CMS tracker [37] is the innermost detector layer and is designed to measure the trajectories and the secondary vertices of charged particles with high precision. To meet the requirements of the LHC operating environment the tracker must have fine granularity to handle the high-multiplicity environment, precise timing to match the particle tracks to the correct bunch crossing, and it must be robust to large doses of radiation. It must also introduce minimal material in front of the other detector subsystems to avoid photon conversion, bremsstrahlung and other interactions. This would particularly degrade the quality of reconstructed photons and electrons. These requirements motivated the choice of silicon sensor technology for the tracker that uses two different types: pixel detectors and microstrip detectors. Pixel detectors are made up of many small pixels and measure a position on a trajectory in two dimensions. Microstrips consist of small parallel strips separated by a distance called the 'pitch' that detects the ionisation from an incident charged particle in one dimension.

The general structure of the tracker (Figure 3.6) has a length of 5.8 m, a diameter of 2.5 m and covers a pseudorapidity range of $|\eta| < 2.5$. Its active surface consists of 1440 pixel sensors, 15148 microstrip detectors and covers an area of approximately 200 m^2. These are assembled into several subsystems: the pixel, the tracker inner barrel (TIB), the tracker inner disks (TID), the tracker outer barrel (TOB) and the

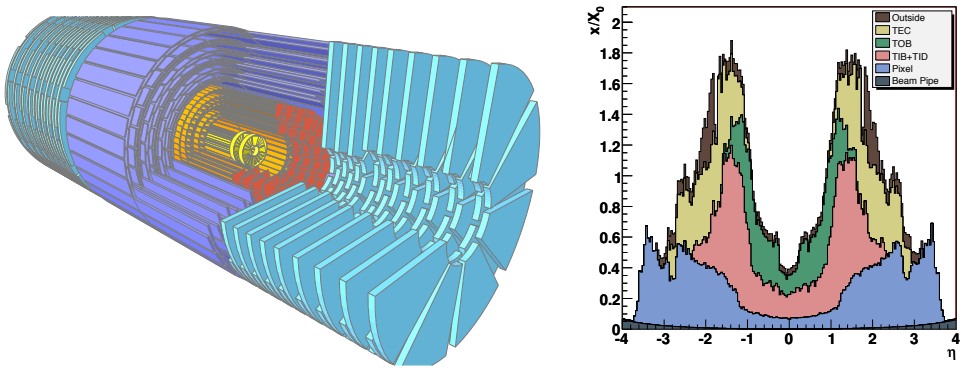

Figure 3.6: Left: the tracker subsystem showing the the central pixel detector in yellow, TIB in orange, TID in red, TOB in purple and TECs in blue [35]. Right: the material before the ECAL in radiation lengths (X_0) [32].

tracker endcaps (TEC).

The innermost subsystem of the tracker is the pixel detector which provides precise measurements in ϕ, r and z of particle trajectories and achieves a high transverse and longitudinal position resolution. The pixel consists of pixel sensors arranged in cylinders of 4.4 cm, 7.3 cm, and 10.2 cm with two disks of pixels sensors at either end to give a total of 66 million pixels and an active area of 1 m^2. This gives a pseudorapidity coverage of $|\eta|< 2.5$ and at least three position determinations along each track with a transverse resolution of 10 μm and longitudinal resolution of 20-40 μm. Further out from the pixel detector all of the tracker subsystems use silicon microstrip sensors.

Situated around the pixel detector are the TIB and TID subsystems that extend in radius out to 55 cm and consist of four cylinders of sensors in the TIB and three disks of sensors in the TID. These deliver up to four r-ϕ measurements using sensors oriented parallel to the beam axis in the barrel and radially on the disks. These sensors have a pitch of 80 μm in layers one and two of the TIB and 120 μm further out giving a position resolution of 23 μm and 35 μm respectively. Surrounding the TIB and TID is the TOB subsystem with an outer radius of 116 cm, z between $\pm$118 cm and consists of six cylindrical layers with pitches of 183 μm in the first four layers and 122 μm in the fifth and sixth. This gives six r-ϕ measurements with single point resolutions of 53 μm and 35 μm respectively. Finally, at either end, are the TEC tracker subsystems that extend in radius from 22.5 cm to 113.5 cm and extend in $|z|$ between 124 cm and 282 cm. Each of the two TECs consist of nine disks of up to seven rings of radially-oriented microstrip sensors. This gives up to nine measurements of ϕ for each trajectory. Some of the microstrip modules are in pairs to provide a second coordinate measurement: z in the cylindrical layers, r in the discs. They are mounted back to back and rotated by 100 mrad with respect to each other. These modules

make up the first two layers of the TIB and TOB, the first two rings of the TID and the first, second and fifth rings of the TECs. All of this ensures ≈9 precision measurements of trajectories in the silicon microstrip part of the tracker in the range $|\eta|< 2.4$ with ≈4 of them being two-dimensional.

The entire tracker material budget manages to remain under two radiation lengths: it ranges from $0.4\,X_0$ at $|\eta|\approx 0$ to about $1.8\,X_0$ at $|\eta|\approx 1.4$ back to about $1\,X_0$ at $|\eta|\approx 2.5$. This is shown in Figure 3.6.

3.3.4 Electromagnetic Calorimetry

The CMS ECAL [38] is a calorimeter designed to reconstruct the energy of electromagnetically interacting particles such as photons with good resolution. In particular the ECAL is aimed at the detection and reconstruction of leptonic and diphoton Higgs final states with good mass resolution within the confines of the CMS solenoid and in LHC operating conditions. To meet these requirements the ECAL must have fine spatial granularity, a large spatial acceptance, a fast response time, and it must capture maximal information from the showers in the restricted space available within the solenoid.

The ECAL as a whole has the following geometry (Figure 3.7): the barrel region (EB) covers a pseudorapidity range of $|\eta|< 1.442$, there is then a gap between $1.442 < |\eta|< 1.566$, and finally the endcaps (EE) cover the range $1.566 < |\eta|< 3$. Due to prohibitive radiation and pileup conditions, electrons and photons are only measured with precision up to $|\eta|< 2.5$. In addition to these subsystems there is the preshower detector (ES) mounted in front of the endcaps that occupies the range $1.54 < |\eta|< 2.61$. The ES consist of two disks of lead absorber followed by two planes of silicon strip detectors with pitch 1.9 mm. This is used for π^0 rejection.

The EB and the EE regions are constructed out of lead tungstate ($PbWO_4$) crystals, 61200 in the former and 7324 in the latter. The choice of $PbWO_4$ was made due to its short radiation length (0.89 cm), and small Molière radius. The short radiation length ensures that particle showers are shorter in extent and can be contained in as small a depth as possible. The short Molière radius (a measure of showers spread transverse to their direction in a material) ensures that the showers are more contained in the η,ϕ directions.

In the EB each crystal has a front face of 22x22 mm^2 corresponding to the Molière radius of 21.9 mm and a segmentation of $(\Delta\eta, \Delta\phi) = (0.0174, 0.0174)$. The crystals are also tapered in η with $25.8\,X_0$ length (230 mm) and are oriented at a 3° offset from the average primary vertex position in η and ϕ. This improves the hermeticity. In the EE the crystals have a front face of 28.6×28.6 mm and are $24.7\,X_0$ (220 mm) in length.

The individual crystals are grouped together in both the EB and EE. In the EB

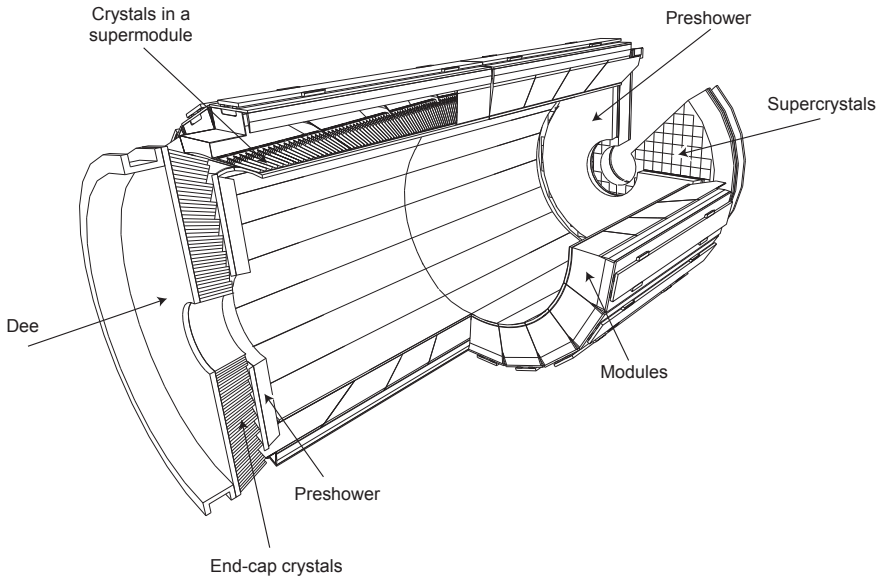

Figure 3.7: The CMS ECAL with a section removed to show structure [32].

they are grouped into 36 'supermodules' that cover $\Delta\phi = 20°$ and extend half the barrel length in z. In each EE identically-shaped crystals are grouped into 5×5 'supercrystals' arranged in a rectangular x-y grid with angular offsets of 2-8°.

When a particle enters one of the crystals and then showers, scintillation light will be produced and collected by a sensor at the opposite end. This sensor will produce a pulse that is amplified and then converted into a digital signal. The height of this digitised pulse is then used to determine the energy deposition within the crystal. Two different types of sensor are used: in the barrel region avalanche photodiodes are used, while in the endcaps vacuum phototriodes are used due to the different magnetic field properties and higher radiation levels. Crystals are read out as 5×5 trigger towers whose digital signal sum constitutes the fast, coarse information provided to the trigger system with every bunch crossing.

The energy resolution of a single $PbWO_4$ crystal is modelled with the following equation [34]

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2, \tag{3.2}$$

where S is the stochastic term, N is a noise term, and C is a constant term. The crystal performance was measured in a test beam and the above parameters determined by fitting a Gaussian function to the reconstructed energy distributions. Their values were measured to be $S = 2.8\,\mathrm{GeV}^{\frac{1}{2}}$, $N = 0.12\,\mathrm{GeV}$, and $C = 0.3\%$.

To reconstruct the energy of a photon or electron, multiple crystals are typically used as the shower spreads in the ECAL, or even starts before it reaches it. Clustering

algorithms [39] are used to reconstruct energy from these crystals and assemble them into a supercluster (SC). The energy associated to the supercluster is then calculated as

$$E_{SC} = F_{SC} G \sum_{i=0}^{N_c} A_i C_i S_i(t) \tag{3.3}$$

where N_c is the number of crystals in the SC, A_i is the amplitude of the pulse of crystal i, $S_i(t)$ corrects crystal transparency loss due to radiation, C_i is a factor that adjusts the response of the crystal, G is a conversion factor from the digital signal to GeV (global energy scale) and F_{SC} is a correction to the SC energy sum due to second order effects.

To calibrate the ECAL [40] one must use a variety of measurements to determine the factors G, C_i, and $S_i(t)$ corresponding to calibration of the overall energy scale, uniformity of measurement in space and uniformity over time respectively. Corrections over time due to radiation-induced transparency change in the crystals ($S_i(t)$) are derived by injecting laser light at 440 nm every 40 minutes.

Several methods are used to derive the factors for an even crystal response over the ECAL spatial extent using the symmetries of CMS. Firstly one uses ϕ symmetry to find factors in rings of η that should all have the same response. Other methods reconstruct particles of known mass decaying to diphotons and use this as a standard candle. The mass should be the same in each part of the detector which allows for the determination of regional differences in response. These different methods are combined to give a collection of per-crystal corrections.

The final factor, the global energy scale, is derived by reconstructing Z bosons decaying to an e^+e^- pair and comparing the measured dielectron mass to the known Z boson value.

3.3.5 Hadron Calorimetry

The CMS HCAL [41] is situated around the ECAL and its function is to identify neutral hadrons, measure their energies and positions, and to determine E_T^{miss} with good resolution over a large acceptance. It is a sampling calorimeter that uses material to produce particle showers (absorber) distinct from the active material measuring deposited energy, unlike the ECAL which is a homogeneous calorimeter where one material (lead tungstate) performs both functions.

The structure of the HCAL is shown in Figure 3.8. It consists of a barrel region (HB), two endcaps (HE), a region outside the solenoid (HO), and two forward calorimeters (HF) that take the HCAL acceptance up to $|\eta| < 5$. The HB region covers a pseudorapidity range of $|\eta| < 1.3$ and consists of two half-barrel sections that slot in to either end of the solenoid bore and each consist of 36 identical wedges in the

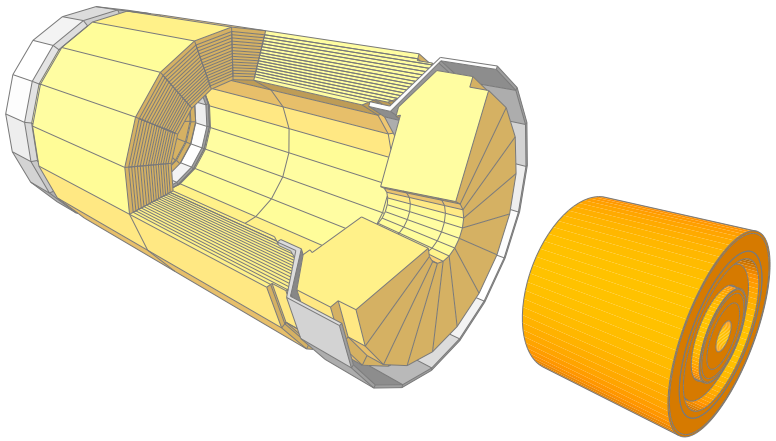

Figure 3.8: The CMS HCAL with the barrel and endcap sections (left) with part removed to show structure and the forward hadronic calorimeter (right) [35].

azimuthal angle ϕ. These wedges are constructed from 2 steel and 14 brass absorber plates with a plastic scintillator active material in alternating layers. Brass is used because it is non-magnetic, while steel absorber is only used in the innermost and outermost plates to provide structural support. Each wedge is segmented into four azimuthal regions and the plastic scintillator is divided into 16 pseudorapidty regions giving a granularity of $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$. The HE regions cover a pseudorapidity range $1.3 < |\eta| < 3$ and each is divided into 36 azimuthal wedges. It uses brass absorber plates and achieves a granularity between $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$ and $(0.017, 0.017)$ The HF region covers a pseudorapidity range from $3 < |\eta| < 5$ and must deal with extremely high levels of radiation. This motivates a different construction with quartz fibres chosen for the active material and steel for the absorber. It is cylindrical in structure with 5 mm thick grooved plates where the fibres fit into the grooves and operate by detecting Cherenkov light produced by incident particles. The fibres are bundled to form $(\Delta\eta, \Delta\phi) = (0.175, 0.175)$ towers. Finally, the HO covers the barrel region around the solenoid and consists of plastic scintillator tiles matching the granularity of the HB. It uses the solenoid itself as the absorber and is designed to operate as a shower 'tail catcher' that compensates for hadron showers that begin later in the HCAL and may not be properly measured. This leakage has a direct effect on the measurement of E_T^{miss}.

3.3.6 Muon Detection

The CMS muon system [42] is situated around the outside of the solenoid and consists of detectors interleaved with the steel return yoke assembled into barrel and endcap regions (Figure 3.9). The muon system has three objectives: to identify muons,

to measure their momentum with precision, and to trigger on them over a large spatial acceptance. The muon detectors are all gas-based and belong to three different

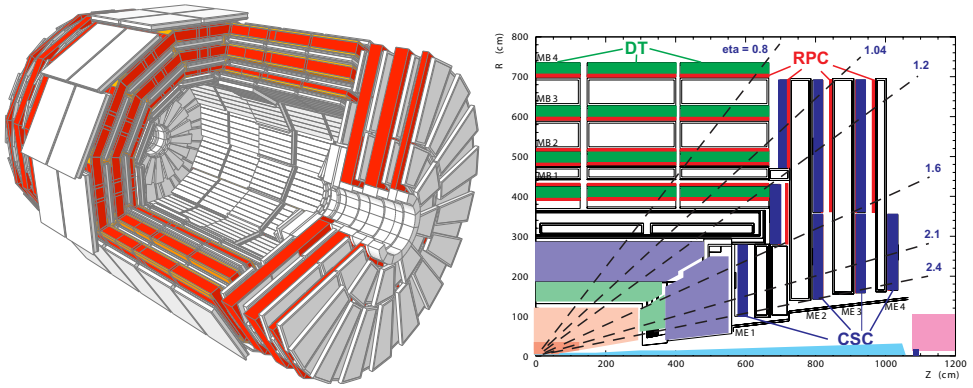

Figure 3.9: Left: the CMS muon detector subsystems (white) within the structure of the steel return yoke (red) [35]. Right: a diagram showing a quarter-view of of the muon system with detector types labelled.

types: drift tubes (DTs), cathode strip chambers (CSCs) and reactive plate chambers (RPCs). They all operate the same way, with incident particles ionising gas, producing free electrons which drift towards the anode and produce an electrical signal.

In the barrel region DTs are used as the magnetic field is more uniform, the neutron flux is small and the muon rate is lower. They are arranged in four layers covering the pseudrorapidity $|\eta|< 1.2$ where the detectors in the first three layers have a different construction to the fourth. The first three layers' detectors consist of eight chambers in two groups of four: the first half measure in the r-ϕ plane, the second half measure in z. The outermost detectors do not have the z determination. All of these detectors use aluminium wires with an Ar plus CO_2 gas mix and achieve a position resolution of 100 μm.

In the endcap regions the magnetic field is less uniform, the neutron-induced background is high and so is the muon rate. This led to the adoption of CSC technology due to their fast timing ability, granularity, and robustness to high radiation. These detectors cover a pseudorapidity region of $0.9 < |\eta|< 2.4$ and are arranged in four layers. Each detector consists of 6 gas gaps with cathode strips running radially away from the beamline and anode wires running perpendicularly to the cathode strips. The cathode strips give a precise but relatively slow measurement in the r-ϕ plane and the anodes measure in η with fast timing for triggering and bunch crossing attribution. They achieve a position resolution of around 200 μm.

Finally, in addition to the DTs and CSCs, the muon system uses RPCs placed in the barrel and the endcaps (up to $|\eta|< 1.6$) as an independent and complimentary way of triggering. The RPCs are double gap chambers operating in avalanche mode

to give a fast response time and good timing resolution. In the barrel region there are six layers of RPCs: two layers in each of the first two layers of drift tubes and one each in the last two. This redundancy helps with triggering on muons with low-p_T. In the endcaps there are planes of RPCs in each of the layers that, in addition to triggering, help to resolve ambiguities in the CSCs when there are multiple tracks in a chamber.

3.3.7 Trigger System and Storage

The LHC delivered bunch crossings to CMS at a rate of 40 MHz in the 2016 period. With each of these events requiring up to 1MB of memory this would amount 40 TB per second of readout and storage which is not feasible. However, most of these events are not physically interesting: they will mostly be low-energy interactions where the protons only glance off each other rather than collide head-on.

To filter these events a fast measurement and decision must be made whether to store or discard; this is achieved with the CMS trigger system [43]. The trigger system operates as a two-step process: first the hardware-based level-1 trigger (L1T) makes fast decisions on whether to keep events using coarse information from some of the subsystems; and then the software-based high-level trigger (HLT) cuts the rate further by using all detector subsystems and a basic physics object reconstruction.

The L1T achieves a rate reduction of 40 MHz to 100 kHz by performing fast calculations using custom, reprogrammable hardware called field-programmable gate arrays (FPGAs). To achieve this the L1T must make an accept decision within 3.2 μs, including the time of transmission from the detector and decision return. Coarse information, due to speed and bandwidth limitations, is received from the ECAL, HCAL, and the muon system to be stored in a buffer that contains information from multiple bunch crossings. Furthermore there is insufficient time for using correlations between subdetectors and also insufficient time and other resources to use information from the tracker. Once information is received, a collection of algorithms are used to pick out relevant events and if the event is accepted the entire detector readout is passed on to the HLT.

At the HLT the data rate is still too high and must be cut further to 1 kHz. This is achieved with a computing farm a short distance away from the CMS detector which runs a basic reconstruction of physics objects from the full CMS readout including the tracker. Here more sophisticated algorithms may be run to pick out collections of objects in events and make the final accept decision. After this, events are recorded to permanent storage, put through the full CMS reconstruction, evaluated for data quality and then made available to physics analyses.

Chapter 4

Machine Learning

4.1 Fundamentals

The field of Machine Learning is concerned with algorithms with the capacity to 'learn' from experience; this may be contrasted with algorithms that achieve some task with a set of statically-defined steps [44]. The ability to learn allows these algorithms to solve problems which may be too complex for a collection of explicitly defined instructions. This chapter will give an overview of machine learning, and deep learning in particular, as pertinent to the field of high-energy physics. We begin with the fundamentals of machine learning, then move on to ensembling and decision trees, and finally neural networks and deep learning.

4.1.1 The Learning Process

Problem Formulation

The data in a machine learning problem are often formulated in terms of a vector space $X = \mathbb{R}^n$, where each dimension is an observable quantity referred to as a 'feature' and a particular datum corresponds to a single feature vector $\vec{x} \in X$ [45]. A dataset is a set of feature vectors $\vec{x}_i$ sampled from some underlying probability distribution $P(\vec{x})$: the data-generating distribution. A machine learning algorithm can then be considered [45] to consist of a model f,

$$f(\vec{x}, \vec{w}) \to Y \tag{4.1}$$

a function that maps from a feature vector $\vec{x}$ to an outcome Y given a vector of parameters $\vec{w}$; a loss function L

$$L(f, \vec{x}) \to \mathbb{R} \tag{4.2}$$

that measures a notion of performance given the model, a set of feature vectors and sometimes the desired outcome; and finally, an optimisation scheme that tunes the parameters of the model with respect to the loss to drive the learning process.

Learning is said to occur when the model's performance at some class of tasks T, as measured by some performance measure P, improves given experience E [46]. There are many types of task that depend on the dataset and the desired outcome, but the two main tasks of interest are classification and regression [44]:

- Classification tasks aim to predict one of k-many classes given a feature vector, $f(\vec{x}) \to y$ where $y \in \{1, \ldots, k\}$. This is often an integer class label but can be a probability distribution over classes. An example of a classification problem from physics would be signal-background event discrimination where we attempt to classify events into background-like or signal-like classes.
- Regression tasks aim to predict a continuous value given the input features, $f(\vec{x}) \to y$ where $y \in \mathbb{R}$. An example of a regression task in physics would be detector calibration where we attempt to predict the true value from the measured value.

In addition to these main tasks there are others such as structured prediction that attempt to predict more complicated structures such as trees and lists.

The experience that the model receives depends on the data that the model is exposed to during optimisation, and can be split into two broad categories [44]:

- Supervised machine learning algorithms experience target values y as well as the input features x and learn properties of the conditional probability distribution $P(\vec{x}|y)$. An example would be a classifier trained on simulated data where we know the true signal-background class label.
- Unsupervised algorithms do not have access to target values and will attempt to learn properties of the data-generating distribution itself such as clusters. An example of this would be a Gaussian mixture model used in calorimetric clustering.

Training and Evaluation

The learning process is also referred to as 'training' and has a different objective to a typical optimisation problem. Rather than just finding the parameters giving the optimal loss over the training dataset, we require the model to find useful properties that generalise to new data [44]. To estimate the generalisation power of a model, we evaluate performance over another unseen dataset, the test set, which should be chosen such that it is representative of the distribution of the whole dataset.

During training most machine learning algorithms will use some form of gradient-based optimisation where one descends the gradient of L with respect to $\vec{w}$ to find

the minimum of L,

$$\vec{\nabla}_{\vec{w}} L = 0. \tag{4.3}$$

The most conceptually simple approach is to evaluate this expression over the entire training set. However, this is often impractical for large datasets with a large population or high dimensionality. An alternative is to use stochastic gradient descent (SGD) [44]. In SGD the optimiser evaluates the gradient with small batches of training data (minibatches). The parameters of the model are then updated as

$$\vec{w} \to \vec{w} - \eta \vec{\nabla}_{\vec{w}} L \tag{4.4}$$

where η is the learning rate, a non-learned parameter that controls the size of the change at each parameter update. As we iterate the model parameters should ideally converge to a global optimum. This is not always guaranteed, as there are sometimes local optima that the optimisation can get stuck in.

More intuitively, the loss as a function of the model parameters is like a mountainous landscape. Each optimiser iteration during SGD is like a hiker evaluating the gradient at their location and taking a step in the direction of the negative gradient.

SGD is often extended with 'momentum' where the update depends on accumulated steps over time: the gradient changes the parameters indirectly through a 'velocity'. In practise this often gives better results more quickly [47]. This is implemented mathematically as,

$$\begin{aligned} \vec{v} &\to \mu \vec{v} - \eta \vec{\nabla}_{\vec{w}} L \\ \vec{w} &\to \vec{w} + \vec{v} \end{aligned} \tag{4.5}$$

where $\vec{v}$ is the velocity and μ is a non-learned parameter referred to as momentum but actually behaves more like a coefficient of friction. Typical values for μ are between 0.5 and 0.9 so it decays the accumulated velocity and has a damping effect on oscillation during training. In contrast to the hiker example, SGD with momentum is more like a skiier: the skiier begins with zero velocity but accumulates velocity over time as they descend the mountain in the direction of negative gradient.

There are also more advanced optimisation algorithms and an alternative formulation of momentum. Nesterov momentum [48] allows the velocity to carry the evaluation point forward and then the gradient is calculated at this new point. This results in a scheme that 'looks forwards' and then corrects. Optimisation with Nesterov momentum has better theoretical guarantees for convergence and is often superior to ordinary momentum. Adaptive optimisers have a learning rate for each parameter that adapts over training depending on how often the parameter is updated [47]. An example of this is Adam [49] which also includes classical momentum. A variant on

this that uses Nesterov momentum (Nadam [50]) is used to train the DCNNs in this thesis.

Example: Linear Regressor

Now we have the ingredients of a machine learning algorithm, we can consider the simple example of a linear regressor. The task of linear regression is to predict the value y given a collection of features. In this formulation we will consider a single feature x, and the algorithm will have access to the y during training, making this a supervised learning problem.

The formula of the model is

$$\hat{y} = w_1 + w_2 x, \tag{4.6}$$

and the term linear refers to the model parameters, powers of x are considered features. The dataset will consist of n-many points of single features, and the value to predict. These are distributed as a linear function of x like the model (Equation 4.6) with $\vec{w} = (-2, 2)$, plus Gaussian noise. The loss will be the mean squared error,

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{4.7}$$

and a single SGD step will be

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \to \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \eta\frac{1}{m}\sum_{i=0}^{m}\begin{pmatrix} 2(w_1 + w_2 x_i - y_i) \\ 2x_i(w_1 + w_2 x_i - y_i), \end{pmatrix} \tag{4.8}$$

where m is the minibatch size.

Training is performed for 500 minibatches with a learning rate of 0.0001, parameters initialised to $\vec{w} = (1, 1)$, and $\mu = 0.9$ in the training with momentum. A small learning rate was chosen deliberately to show the progress of the model during training. The training process and final outcome of this algorithm are shown in Figure 4.1. The training with SGD plus momentum shows a much faster convergence to the true model parameter values.

4.1.2 Model Capacity and Generalisation

The space of functions that a model can draw upon to describe observed data is referred to as the model's hypothesis space. Taking the example of a linear regressor,

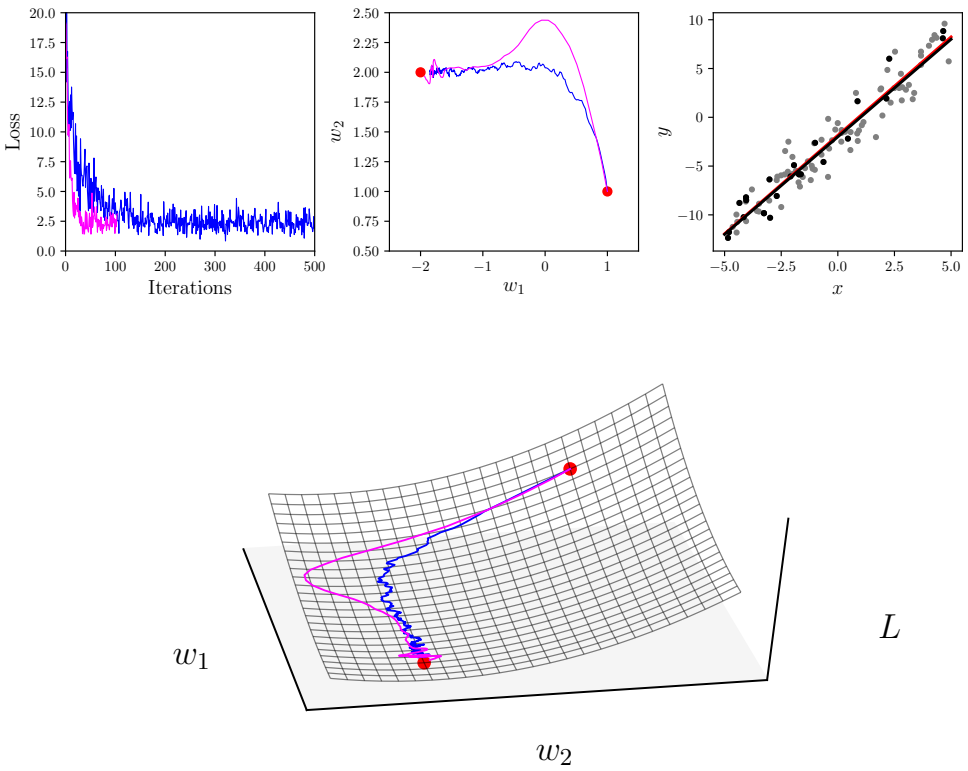

Figure 4.1: Training a linear regressor with SGD (blue) and SGD plus momentum (magenta). Top row: loss histories over training (left), the trajectory of the model parameters in parameter space during training (centre), and the final result with the result in red and the true value in black. An example minibatch is also shown by the black points (right). Lower plot: how the optimisation descends the 'loss landscape' during training. The surface shows the loss calculated over the entire dataset at once for each parameter value. Each step during training computes an estimation of this surface using the sampled minibatch.

the hypothesis space can be expanded by using higher-order polynomials

$$\hat{y} = \sum_{i=0}^{N} w_i x^i. \tag{4.9}$$

When we increase or decrease the size of this space we are increasing or decreasing the model's descriptive power, known as its 'capacity' [44]. If it is inappropriately large or small, the model can experience problems with generalisation. Specifically, over or under-capacity can lead to generalisation error from two sources which often need to be traded against each other: bias and variance. Bias is the error that comes from the model approximating the underlying function. Variance is how much the

trained model estimate will change if the training dataset is changed [44].

If the capacity is too small, this leads to 'underfitting'. Here the model does not have enough descriptive power to fit the data and we get generalisation error due to bias. If there is too much capacity, the model will find a function fit to the training set arbitrarily well and have large variance. This will cause another sort of generalisation error called 'overfitting'. Examples of how inappropriate capacity can lead to generalisation error are shown in Figure 4.2 where a linear regressor is trained with different order polynomials. In this example, we note that the high-capacity

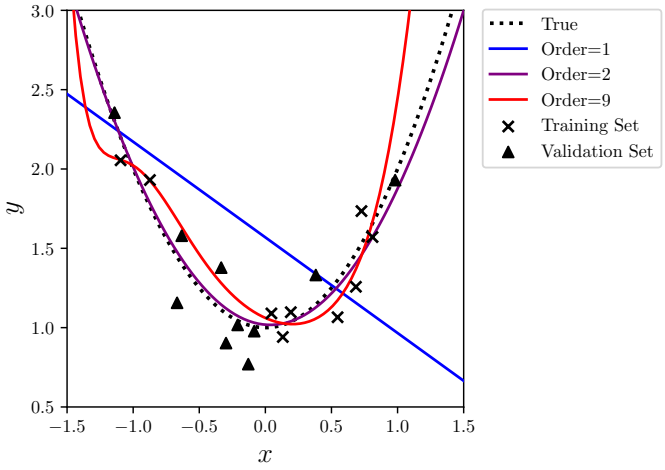

Figure 4.2: Linear regressors with different order polynomials fitted to the same data.

model has been fitted to noise and fails to generalise to unseen data.

An alternative and more general approach to controlling model capacity is to penalise parts of the hypothesis space rather than remove them. Removing them is equivalent to infinite penalisation. This way, if we have two functions performing equally well, we can express some sort of preference for which one to choose by adding a term to the loss. This penalty term is called a 'regulariser' term [44] and has the form

$$\lambda\Omega(\vec{w}), \tag{4.10}$$

where λ is a value scaling the strength of the regulariser's effect. The variable λ is an example of a 'hyperparameter', defined as any unlearned parameter of a machine learning algorithm. Furthermore, adding regulariser terms to the loss is an example of regularisation, a broad class of techniques that aim to improve an algorithm's generalisation error.

Two common forms of regularisation are penalty terms based on the squared-L_2

and L_1 norm of the model's weight vector

$$\begin{aligned} \lambda\Omega(\vec{w}) &= \lambda||\vec{w}||_2^2 = \lambda \sum_{i=0}^{n} w_i^2 \\ \lambda\Omega(\vec{w}) &= \lambda||\vec{w}||_1 = \lambda \sum_{i=0}^{n} |w_i|. \end{aligned} \tag{4.11}$$

These two regularisers have a different effect on the model parameters: L_2 regularisation expresses a preference for using each parameter a small amount, L_1 regularisation will prefer sparsity where a few parameters are large, and the rest close to zero. A comparison of a range of values for both regularisers applied to a linear regressor with a 9^{th}-degree polynomial is shown in Figure 4.3.

These models have been evaluated for generalisation error on another unseen dataset: the 'validation' set. The reason to use this rather than the test set, is that when we choose a hyperparameter value we are essentially doing another fit to data. If we do this on the test set we may fit to unrepresentative patterns in the test data and overfit again. Evaluation on the test set is then no longer a good measurement of generalisation.

4.1.3 Ensembles

It is sometimes useful to train multiple models (base learners) and combine them in some way, for example by some weighted sum of their outputs, chaining them together, or some other approach. This technique is called ensembling, and one of the most popular machine learning algorithms in particle physics is an example of this: boosted decision trees (BDTs) [6]. This subsection will give only a narrow overview of this area as relevant to the CMS Higgs diphoton analysis: we will focus on a single ensembling method, gradient boosting, and a single base learner: the decision tree (DT).

Decision Trees

Decision trees [51] are binary tree structures that recursively partition the feature space into non-overlapping regions. Each node of the tree corresponds to a region, and each additional child node corresponds to further splits into subregions. We eventually reach a node with no child, this last node is a leaf of the tree and assigns a value to the corresponding region. Decision trees are trained by calculating a collection of possible splits and then choosing one optimising a measure of purity in classification case, or a loss function such as mean-squared error in the regression case. This process is usually stopped once the leaf nodes have reached some optimal value, or a maximum

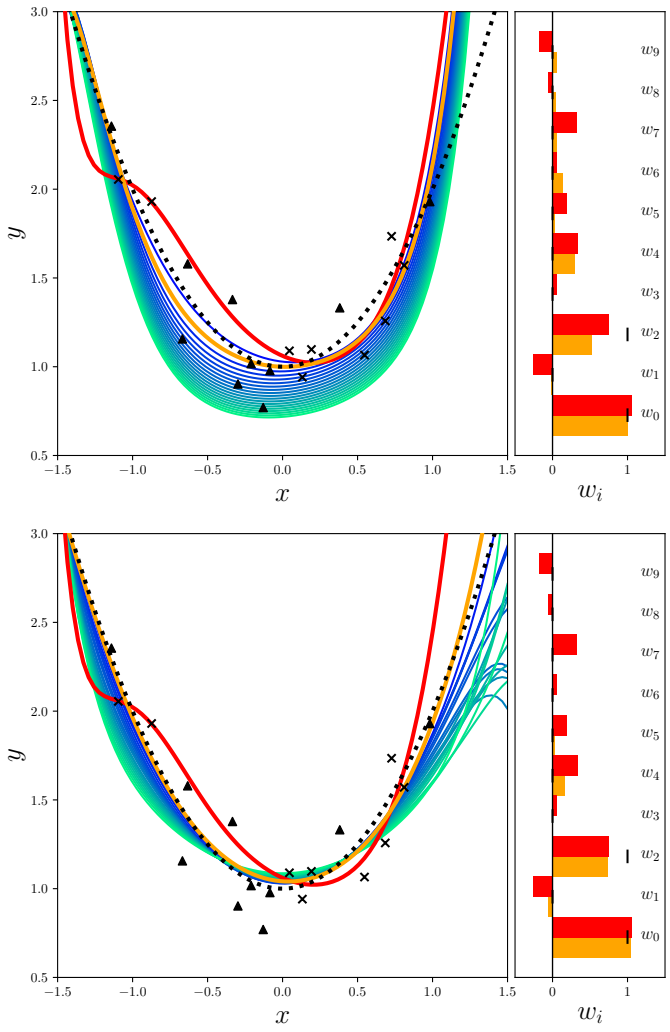

Figure 4.3: Fits for different regularisation strengths with L_2 (top) and L_1 (bottom) regularisation data drawn from a uniformly sample of $y = 1 + x^2$ plus noise. The red curve is the unregularised fit, the orange curve is the result with the lowest loss with respect to the validation set (triangles). The bar charts show the parameter values of the overfitted result and optimal regularised result.

depth has been reached. Decision trees are also regularised by pruning which removes branches that use unimportant features and give no overall performance improvement.

Decision trees have advantages such as their simplicity, interpretability, and their ability to handle different types of data. However, they also have various disadvantages such as their cuts being aligned with the dimensions of the feature space (so diagonal decision boundaries need to be constructed out of many orthogonal cuts), their tendency to get stuck in local minima, and their training variance leading to

overfitting. These disadvantages can be mitigated by using decision trees as base learners in an ensemble producing an algorithm stronger than its constituent parts.

Gradient Boosting of Decision Trees

Generally, boosting algorithms [52] construct a strong learner from base learners by iteratively training base learners, compensating for earlier weakness in some way. Each base learner is then added together in a weighted fashion to produce the final ensemble. Gradient boosting [53] is a particular boosting algorithm that fits to the errors of prior base learners in a way that is equivalent to gradient descent. Gradient boosting assumes that at each iteration n, $1 < n \leq N$ there is a base model $f_n(\vec{x})$ that can be improved by addition of another estimator (in our case another decision tree) $h(\vec{x})$

$$f_{n+1}(\vec{x}) = f_n(\vec{x}) + h(\vec{x}). \tag{4.12}$$

If $h(\vec{x})$ perfectly corrects $f_n(\vec{x})$ this implies that,

$$\begin{aligned} f_{n+1}(\vec{x}) = f_n(\vec{x}) + h(\vec{x}) = y \\ h(\vec{x}) = y - f_n(\vec{x}) \end{aligned} \tag{4.13}$$

where $y - f_n(\vec{x})$ is referred to as the 'residual'. A key insight was that this process is analogous to gradient descent as the residuals are the negative gradients with respect to $F(\vec{x})$ of the squared error loss function

$$\frac{1}{2}(y - F(\vec{x}))^2. \tag{4.14}$$

This can then be generalised to other differentiable loss functions.

When the base learners are decision trees, the process is as follows: at each iteration n we train a DT on the residual

$$h_n(\vec{x}) = \sum_{j=1}^{J_n} b_{jn} \mathbf{1}_{R_{jn}}(\vec{x}) \tag{4.15}$$

where J_n is the number of regions of h_n, $R_{1n}, \ldots, R_{J_n n}$ are the regions themselves, b_{jn} is the value predicted in region R_{jn}, and $\mathbf{1}_{R_{jn}}$ returns 1 for $\vec{x}$ in region R_{jn} and zero otherwise. The output of this tree is multiplied by a value γ_n minimising the loss chosen by line search,

$$\gamma_n = \text{argmin}_\gamma \sum_{i=1}^{m} L(y_i, f_{n-1}(\vec{x}) + \gamma h_n(\vec{x}_i)), \tag{4.16}$$

and then the ensemble is added to as

$$f_n(\vec{x}) = f_{n-1}(\vec{x}) + \eta\gamma_n h_n(\vec{x}_i). \tag{4.17}$$

where η is a learning rate parameter and controls steps size just like in normal gradient descent.

4.1.4 Algorithm Design, Evaluation and Optimisation

The No Free Lunch theorem of machine learning [54] states that, averaged over all possible data-generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points. This result essentially means that no machine learning algorithm is universally superior, but it does not mean that they are all equally powerful for a particular task. The theorem only holds averaged over all distributions, and some algorithms will indeed perform better given specific focus. We must make assumptions given prior knowledge and build our algorithms accordingly. This will inform how we choose the model, how we measure performance, and how we optimise the hyperparameters.

A particularly important phenomenon is the 'curse of dimensionality' [45] where machine learning algorithms can under-perform given a dataset with a large number of features (high-dimensionality). For such a dataset the number of possible configurations of the features are far larger than the size of the training set. This can also be formulated in terms of coverage of a hypervolume (Figure 4.4): if one considers a unit cube of dimension D, the portion of the sides required to cover a given volume increases rapidly with D. This issue is a primary motivator for the development of deep learning, and is also something that needs to be considered during hyperparameter optimisation.

Choice of model, and input features, will depend on a number of practical constraints such as time and computational resources, but also constraints that avoid biases particular to physics analyses. In training a classifier to separate Higgs boson signals on simulation we do not want the algorithm to reconstruct the mass and bias itself to the simulation value. This can happen if the algorithm is given this value explicitly or if it is capable of reconstructing it from the other features. Furthermore one must use assumptions from prior knowledge of dataset size, dimensionality and other properties such as linear separability and class balance to choose the model. One can then evaluate candidate algorithms using appropriate performance measures.

Once the model is chosen, hyperparameters can be selected with a variety of optimisation approaches. However, because the underlying function mapping from hyperparameters to performance values is unknown we cannot use gradient-based methods. Here we use derivative-free optimisation methods, two of which will be

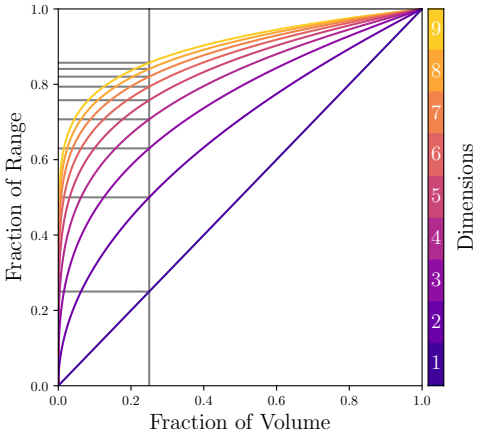

Figure 4.4: Range of the side length to cover a fraction of the volume of a unit cube in up to nine dimensions. The grey lines show the fraction required to cover 25% of the volume.

presented here and used later in this thesis: grid search and Bayesian optimisation.

Before we can optimise we need to define a performance measure to optimise with respect to. A common measure for binary classifiers, and the one used in this thesis, is the area under the 'Receiver Operating Characteristic' curve (AUROC). ROC curves are constructed by cutting on the output score of a machine learning model and plotting the false positive rate versus the true positive rate. The area under this curve will be between 1 and 0.5 where 1 indicates a perfect classifier and 0.5 is equivalent to random guessing. This is demonstrated in Figure 4.5.

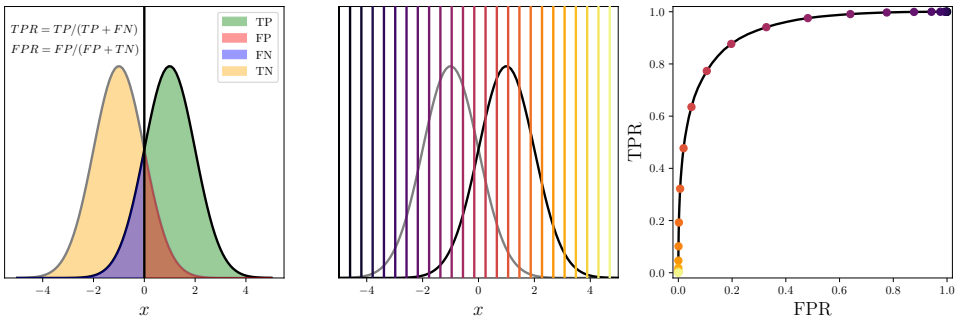

Figure 4.5: ROC curve construction. On the left is the definition of the True Positive Rate (TPR) and the False Positive Rate (FPR) where TP is true positive, FP is false positive, TN is true negative and FN is false negative. In the centre and right are the distributions that are thresholded with the coloured lines in the central plot being cuts that correspond to the same coloured point on the right plot, a ROC curve.

Grid search is a simple method that samples a set of evenly-spaced values over a given region of the hyperparameter space. This method can be used when there are fewer hyperparameters to optimise and the time cost of sampling each point is not too great. As the number of hyperparameters goes up, dimensionality increases and the sampling becomes extremely sparse as each point represents a smaller portion of the space.

Bayesian optimisation [55] belongs to a class of optimisation algorithms that use previous observations of the performance to determine the next point to sample. The method consists of two main steps:

1. using evaluated points in the hyperparameter space, calculate a posterior expectation of the performance as a function of the hyperparameters
2. evaluate the performance at a new point maximising an 'acquisition function'. This is a function that trades off exploration versus exploitation in choosing the next optimal point to sample given the posterior expectation.

Bayesian optimisation makes efficient use of sampling and is more appropriate when evaluating a single point in the hyperparameter space is expensive. The difficulty of the optimisation will still increase rapidly with the dimensionality so one should consider, where possible, the optimisation as a set of orthogonal problems. These two methods can be combined where an initial grid search is performed, and the set of evaluated points are used in the first iteration of the Bayesian optimisation. This step is called a 'warm start'.

Often we do not know the form of the data-generating distribution a priori. Therefore a good approach to choosing and tuning a model is to have as much capacity as design constraints allow and then restrict this capacity with regularisation using an optimisation over the validation set as described earlier.

4.2 Deep Learning

Deep learning is a powerful approach to machine learning problems based on artificial neural networks (ANNs), especially with a large quantity of input features. The name of the field refers to the depth of the ANNs: as depth increases they can model ever-more complex functions of the input features. This section will give a description of ANNs, how they are trained and regularised, the challenges that come with increasing network depth, and finally convolutional neural networks including dense convolutional neural networks.

4.2.1 Artificial Neural Networks

The Single Neuron

Artificial neurons receive a weighted collection of input signals and then 'fire' depending on their sum [47]. Mathematically, they consists of an input feature vector x_i, a weight vector w_i, a bias b and a nonlinear activation function f that produces an output o via the following computation,

$$o = f(w_i x_i + b). \tag{4.18}$$

A schematic of an artificial neuron is shown in Figure 4.6 along with three commonly used activation functions:

$$\begin{aligned} f(z) &= \frac{e^z - e^{-z}}{e^z + e^{-z}}, \\ f(z) &= \begin{cases} 0 & z \leq 0 \\ z & z > 0 \end{cases}, \\ f(z) &= \begin{cases} \alpha z & z \leq 0 \\ z & z > 0 \end{cases}. \end{aligned} \tag{4.19}$$

These are tanh, the rectified linear unit (ReLU) and the leaky ReLU respectively. The value α in the leaky ReLU is a hyperparameter and is typically set to a value of 0.2 [47].

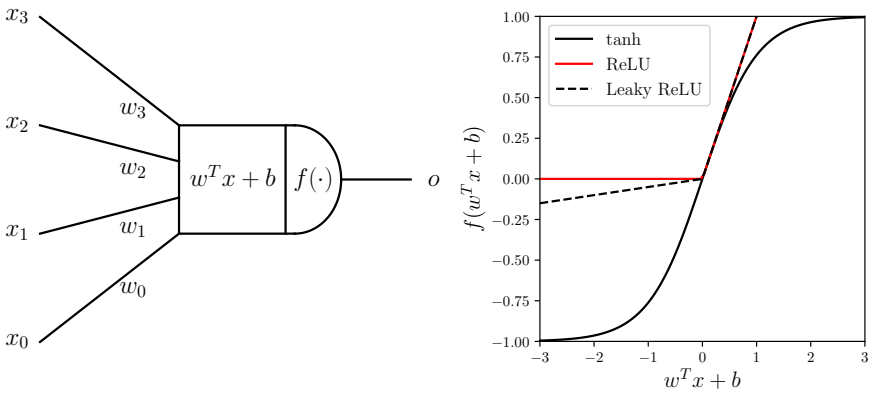

Figure 4.6: Schematic of an artificial neuron(left) and a plot of three commonly-used activation functions (right).

The weights vector and the bias constitute the learnable parameters of this model. With the correct loss and activation, this structure is equivalent to a linear classifier:

we can consider each neuron to be attempting to place an optimal linear decision boundary in the space of its input features. If the data are linearly separable this will be achievable, but if they are not, this simple classifier will struggle. To help the neuron we could construct a function $\phi(x_i)$ on the feature space to produce transformed feature vectors in which the data are now linearly separable [44]. This could be constructed explicitly, or it could be learned from data.

Feedforward Neural Networks

ANNs come in many different architectures, but the ones we will consider here will be exclusively feedforward networks. These networks are constructed from layers of neurons where each layer feeds into the ones after it, starting with the input layer, then often multiple hidden layers, with the final output layer giving the prediction. The most common layer type used is the 'fully-connected' layer where each neuron in layer l is connected to every neuron in layer $l+1$.

A classic feedforward architecture is the multi-layer perceptron (MLP) consisting of a series of fully-connected layers. Often the outputs $\vec{o}$ of these networks (referred to as 'logits') are transformed with the 'softmax' function,

$$\sigma(\vec{o})_i = \frac{e^{o_i}}{\sum_{k=1}^{N} e^{o_k}}, \tag{4.20}$$

where N is the number of outputs. This maps the vector of network logits to a vector of probabilities that sum to one.

When we connect multiple layers together we are constructing a model capable of performing a chain of feature space transformations $\phi(x_i)$ where each layer produces features for the one that follows it, and the final layer can place a linear decision boundary on this transformed feature space [44]. The effect of composing layers together can be seen by comparing a model with no hidden layers versus a model with a single hidden layer on data that are not linearly separable (Figure 4.7).

With increasing depth neural networks are able to construct ever-more complex functions of their input. Mathematically, this corresponds to a chain of matrix multiplications of feature vectors plus biases interleaved with non-linear activation functions,

$$o_k = f_n(f_{n-1}(\dots(f_1(w_{ij}^1 x_i + b^1)))). \tag{4.21}$$

Training Neural Networks

ANN classifiers may use a variety of different loss functions. Let o_j indicate the j^{th} element of the output class vector of the neural network and y_i indicate the true class label of datapoint i. Two popular choices of loss function [47] are the hinge loss and

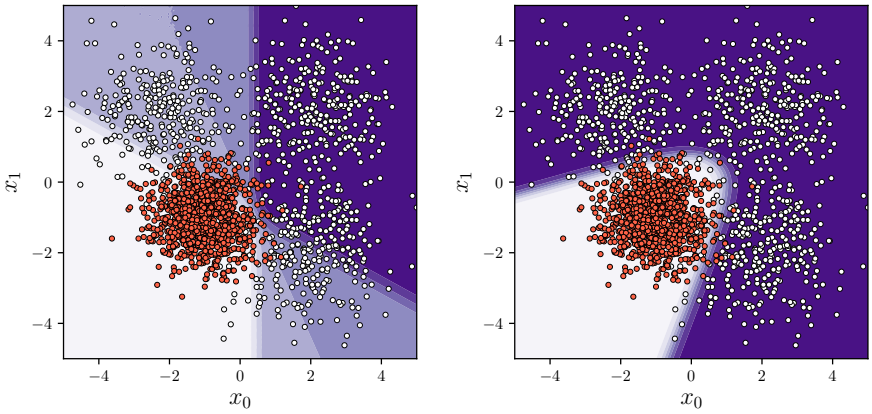

Figure 4.7: Decision boundaries for a two-input (x_0, x_1), two-class neural network classifier with no hidden layer (left) and one hidden layer (right). The outputs of the networks are mapped to probabilities with the softmax function and are shown by the background contour plot.

cross entropy loss with softmax function. The hinge loss has the form

$$L_i = \sum_{j \neq y_i} \max(0, o_j - o_{y_i} + 1) \tag{4.22}$$

and is a 'maximum margin' loss that attempts to strongly penalise misclassified examples. The cross entropy loss has the form

$$L_i = -\log\left(\frac{e^{o_{y_i}}}{\sum_j e^{o_j}}\right) \tag{4.23}$$

and is used in conjunction with a softmax function. This loss can be interpreted as minimising the negative log likelihood of the correct class. Each of these will cause the network to behave in a different way: the hinge loss will in effect prioritise accurate classification at the cost of modelling probability, whereas the cross entropy will model $p(y|x_i)$ with less priority on accuracy [47].

Once the loss has been defined, neural networks are usually trained via SGD with gradients computed using an algorithm called backpropagation [56]. A single iteration works as follows:

1. Forward pass: inputs are repeatedly transformed from the first layer to the last with product sums dependent on w_{ij}^k and then by the activation functions. The final output $\hat{y}$, the input to each activation, and the outputs from each activation are stored for the next step.
2. Backward pass: this works like the forward pass but from the output layer

backwards to the input calculating $\partial L/\partial w_{ij}^k$ as it goes. The 'input' is the output layer's loss terms and these are repeatedly transformed by product sums with the weights w_{ij}^k and then multiplied by the values of the derivatives of the activation function with inputs from the forward pass.

3. Weight update: after the backward pass, we now have the gradients required to update the weights by SGD.

Regularising Neural Networks

Just like other machine learning models neural networks will overfit when they have too much capacity and therefore regularisation is crucial. The L_1 and squared-L_2 regularisers are commonly used and L_2 is often preferred [47]. Another regularisation method, max-norm clipping [47], restricts the norm of the gradient vectors during weight update and stops them from getting above a certain size whilst preserving their direction in parameter space. Finally, a highly-effective and complimentary method is 'dropout' [57] (Figure 4.8) where neurons are switched off at random. This aids model generalisation by stopping the network from over-using neurons and also acts as an effective ensembling where many random subnetworks are trained and then combined together at inference time.

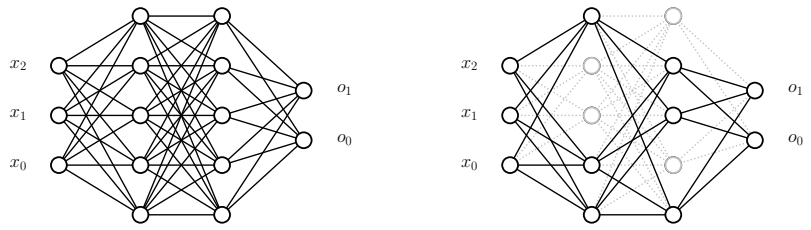

Figure 4.8: Left: a multi-layer perceptron with no dropout. Right: the same network with dropout. Dropped neurons are shown greyed-out.

Additionally, although they are not strictly regularisations, data preprocessing and augmentation can greatly help with model generalisation. Augmentation is when we apply random transformations to the input features such as random rotations and reflections of an image. Standardisation is when each of the features are mean-subtracted and divided by their standard deviation so that each feature has zero mean and unit variance. This helps greatly with training time and convergence.

Network Depth and its Problems

As one simply increases the depth of a neural network, the performance does not always increase. Depth brings with it a collection of problems each of which have their solutions and therefore an impact on network design.

A major problem that occurs with deep ANNs using sigmoidal activation functions is the vanishing/exploding gradient problem [58]. During training, these functions can have gradients in the range $[0, 1]$ that are multiplied n-many times given how many layers from the output layer the backward pass is. This causes the gradients to become small, the weight updates in turn become small, and the training slows down heavily or stops altogether. If the gradients are greater than one, the opposite problem can happen where the gradient becomes very large, this will cause inputs to neurons to become large and push the activation functions into the saturation region where the gradient is again small. This can be mitigated by using a non-sigmoidal activation such as ReLU, and choosing the correct weight initialisation. The recommended initialisation [47, 59] is to draw values for each neuron weight vector from a Gaussian with $\mu = 0$ and $\sigma = \sqrt{2/n}$ where n is the number of inputs to the neuron.

Using ReLU activations can pose its own problems as well. With these activations there can be large negative corrections to the weights that then cause the inputs to a neuron to become strongly negative. Due to the ReLU having zero gradient for negative input, all future weight updates for this neuron will be zero and it will never activate again. This is the 'dying ReLU' problem. The solution to this is to use leaky ReLU activations [60] that have a small non-zero gradient below zero and are therefore able to recover from large negative corrections.

Another problem is internal covariate shift [61]: because the input to each layer depends on all the ones before it, small changes can be amplified and cause the distributions in later layers to shift. This makes learning in later layers harder. To solve this issue, we introduce 'batch normalisation' layers [61]. These layers normalise the input into each layer on a per-minibatch basis during training and also have a learnable scale γ and shift β,

$$x' = \gamma \frac{x - \mu}{\sigma} - \beta. \tag{4.24}$$

During inference, the population statistics of the layer inputs are used, so this must be calculated from the training set during the training process. These layers also help with the vanishing gradient problem.

Finally, even with the above measures in place the accuracy of a deep network can saturate as depth increases and then drop. This is not caused by overfitting but is actually caused by the network's failure to reproduce identity transformations leading to the degradation of information as it passes to deeper layers [7]. This has been solved by using different sorts of bypass connections where outputs from earlier

layers are directly connected to later ones to facilitate the flow of information. This was the key insight that drove the development of the ResNet architecture [7], and the dense convolutional network architecture used in this thesis.

4.2.2 Images and Convolutional Neural Networks

Introduction

This thesis is concerned with treating jets within CMS as images, and then using these images to enhance the classification of VBF Higgs events with their characteristic dijets. A particular class of ANN architectures, convolutional neural networks (CNNs), are used to solve such image-based machine learning problems. This section will describe how images are formulated, classic CNNs with their constituent parts, and finally the more advanced architecture used in this thesis.

Images are formulated as three-dimensional volumes of values with height, width, and then a depth usually corresponding to the RGB channels. Each of these values is an individual feature that measures a particular type of information at a transverse location (such as 'redness' in a non-transformed image). These features can be more complex, such as an image after a local edge-detection transform where each feature now corresponds to whether an edge is present at that location. Therefore the features located at a particular depth are said to form 2D feature maps, and the depth-wise stack of feature maps are said to form a feature volume. Image processing problems are concerned with manipulating and extracting information from these feature volumes.

Images often have certain properties allowing two assumptions to be made which inform the design of CNNs. Firstly, the statistical properties of an image dataset are uniform in the transverse directions and therefore local feature detectors will be useful over the whole image extent. Secondly, that features near to each other in the transverse directions are highly correlated and we can downsample the image without losing much information.

If we try to apply an ordinary MLP to an image processing problem the large number of input features will give rise to a very large number of model parameters (proportional to the square of the inputs). For example, if the input is a $100 \times 100 \times 3$ RGB image, each neuron in the next fully-connected layer would receive 30000 inputs, leading to a very large number of parameters even in a shallow network. A network with such a large number of parameters will be difficult to train and suffer from severe overfitting.

CNNs are feedforward-type ANNs constructed from successive 3D volumes of neurons using local connectivity and parameter sharing to avoid the issues with MLPs in this area [47]. Specifically, CNNs introduce two new layer types: convolution layers and pooling layers corresponding to the two image assumptions described above.

Convolution Layers

Convolution layers (Figure 4.9) learn to find local feature transformations that pick out features in the image which are then transformed into ever-more abstract and complex features by later convolution layers. They receive a feature volume as input, and consist of a convolution operation that produces a transformed (convolved) feature volume and then a volume of activation functions. Each activation function receives one feature in the convolved feature volume.

This convolution operation consists of a collection of local transformations that multiply the values of a local patch of the input by learned weights and then sums them. These weights are shared over the transverse extent of the image: each patch will be multiplied by the same set of weights, and all the patches together will produce a feature map. This parameter sharing design is inspired by the feature translation assumption. Each convolution layer will have multiple filters that will each make a different feature map constituting the transformed feature volume given to the activations.

This can be interpreted as a volume of neurons each of whom only see a small portion of the input feature volume (its field of view, FOV). All of the neurons at the same depth in the volume will share the same weights and can be considered to be looking for the same feature at different locations. The volume of activation values constitutes the convolution layer output volume and each value will correspond to the possible presence of some higher-level feature constructed from the lower-level input features.

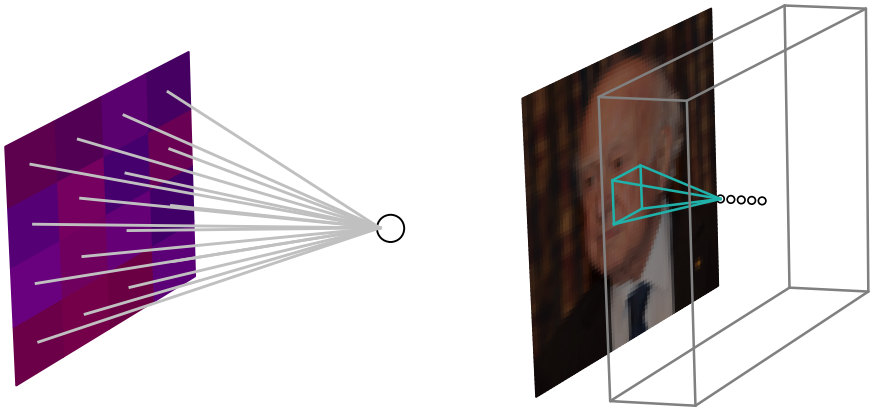

Figure 4.9: Convolution layer: a single neuron connected and its connection to a 4×4 patch of input (left) and an image patch with neurons in context with an input image [62] (right)

Sometimes the individual convolutions only see a single feature (1×1 patch), but still the full extent of the input depth. These transformations operate only on differences between the feature maps and are used to reduce the depth of the input feature volume. This corresponds to feature reduction, where we combine or remove features to produce a small set of performant features.

Pooling Layers

Pooling layers are inspired by the local correlation assumption and reduce the size of their input by mapping sections of neuron outputs to a single value. For example: a 2×2 patch of each feature map is mapped to a single value whilst keeping the depth of the feature volume the same. This has the dual function of reducing model complexity and increasing the local field of view of later neurons as each later neuron input corresponds to multiple neuron outputs from prior layers. The two main types of mapping are the maximum value of the patch and the average value, referred to as max pooling and average pooling respectively (Figure 4.10).

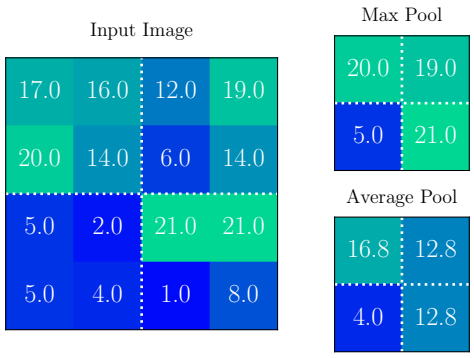

Figure 4.10: An example input to a pooling layer is shown on the left with two outputs on the right from max pooling (above) and average pooling (below).

Max pooling leads to faster trainings and to translation invariance of feature detection, but loses information from the non-maximal values and will cause the network to be less aware of spatial arrangement. Average pooling takes the average in the region of interest and does not lose as much information, however it can be slower and does not have the same translation invariance properties.

Classic Architecture

The classic example of a CNN [47] consists of interleaved convolutional and pooling layers. At the end of these, the feature map is flattened to a 1D array of neurons and the rest of the network has the structure of an MLP (Figure 4.11).

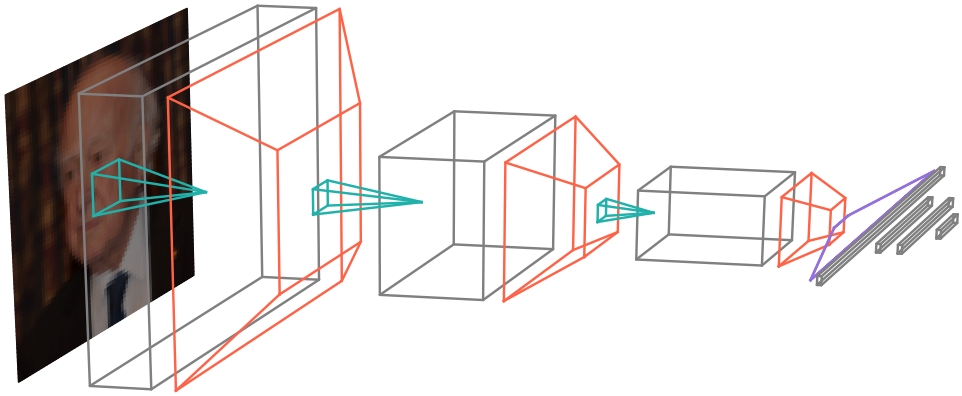

Figure 4.11: A typical CNN architecture with three convolutional layers (grey) containing neurons with a restricted FOV (green), pooling layers for downsampling (orange), a flattening of the final feature map (purple) and a set of fully-connected layers.

4.2.3 Dense Convolutional Neural Networks

Dense convolutional neural networks [63] are inspired by the bypass layers in models such as ResNet, but here every layer is connected to the layers situated after it. This dense connectivity gives superior gradient flow during backpropagation and allows for increased depth (due to the mitigation of depth degradation) and therefore much more sophisticated features. It also encourages feature reuse, and allows the network to achieve high performance with fewer parameters.

Dense CNNs have a similar general structure to ordinary CNNs: they are feed-forward and make use of convolution and pooling, but their structure is much more complicated. Instead of interleaved convolution and pooling layers dense CNNs have dense blocks made of multiple composite layers of convolutions, these are interleaved with transition layers which play the role of pooling but can also perform feature reduction. After this the feature volume is flattened and input to a MLP classifier structure as normal.

Composite Layers

These layers are the basic unit constituting the dense blocks. They consist of (Figure 4.12) a batch normalisation, a ReLU activation function, a 1×1 convolution for compressing the depth of the input volume, a second batch normalisation, a second ReLU activation and then an ordinary convolution that outputs a feature volume of a set length k called the growth rate. The 1×1 convolution is called a bottleneck and serves to reduce model complexity and perform feature reduction. These layers, as formulated in this thesis, have two hyperparameters: the depth of the output feature

volume, the 'growth rate' and the size (width and height) of the filter in the second convolution.

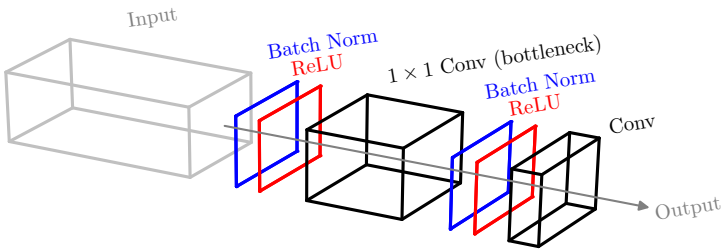

Figure 4.12: The separate components of a composite layer.

Dense Blocks

Dense blocks consist of d-many composite layers where there is direct connection from each layer to all those after it (Figure 4.13). In other words each layer receives all of the feature volumes produced before it, concatenated along the depth axis. Each

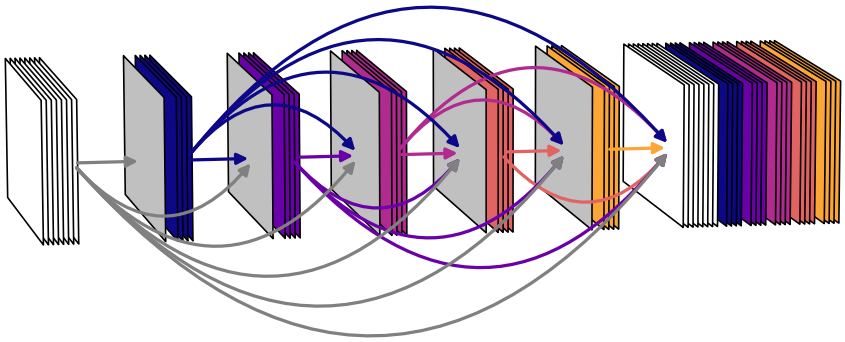

Figure 4.13: A dense block with depth 5 and growth rate 4. Input feature volume is shown by the stack of white squares, each composite layer is shown as a grey square and the output feature volume of the layer is shown by the coloured layered stack. Coloured arrows show the which layers each feature volume is input to. The final concatenated output of the dense layer is shown by the white and coloured stack on the right.

dense block has a number of hyperparameters. In the formulation used in this thesis they are the following: depth is the number of composite layers in the dense block, filter size is the size of the filter of the second convolution in the composite layers, and the growth rate of the composite layers.

Transition Layers

Transition layers are pooling layers with average pooling, but with batch normalisation applied to the input and a 1×1 convolution for compressing the input feature volume before the pooling operation (Figure 4.14). This compression of the feature volume performs feature reduction where less useful features are removed. It will also reduce model complexity and overfitting. This reduction is controlled by another hyperparameter: the reduction factor.

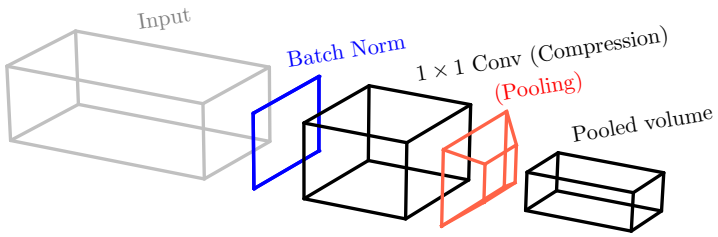

Figure 4.14: A transition layer with a reduction factor of 0.5.

Chapter 5

Object Reconstruction and Selection

5.1 Introduction

The CMS H $\rightarrow \gamma\gamma$ analysis works by searching for excess production in the distribution of diphoton invariant masses. A Higgs boson signal will manifest itself as a small bump on top of a continuous distribution due to background processes from Standard Model diphoton production.

The invariant mass of a diphoton system is calculated with the expression

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

where E_{γ_1} and E_{γ_2} are the energies of the leading energy photon and subleading energy photon respectively, and α is the opening angle between them. To determine the value of α we require the locations of the photons in the ECAL and the correct originating vertex. Good identification and measurement of photons and their verices are therefore crucial to the analysis, and the reconstruction of these will be described in detail in this chapter. Other objects such as jets and leptons provide extra information on the production mode and allow for improved signal isolation. The reconstruction of these objects will also be described.

5.2 Tracks, Clusters, and Physics Objects

Tracks are reconstructed from hits in the CMS tracker using a standard iterative procedure based on Kalman filters [64]. Each iteration is as follows: first, track candidates are seeded from two or three hits, next a Kalman filter extrapolates these

trajectories and looks for more hits to associate to the track, then another Kalman filter and a smoother is used to produce estimates for parameters of each candidate track, finally a selection rejects low-quality candidates.

Calorimetry clusters are used in all of the objects of interest in this chapter, and ECAL clusters are especially important for photons. The clustering for ECAL deposition is described here, the HCAL clustering proceeds in similar fashion. Energy deposition from photon and electron electromagnetic showers is often spread out over multiple ECAL crystals due to the magnetic field and interaction with tracker material. The objective of the ECAL clustering is to gather these energy deposits into 'superclusters' (SCs) to achieve good energy containment, pileup robustness and to take variation in the ECAL structure into account.

The process [38] begins with the identification of 'seed' crystals. These are crystals over a threshold energy (greater than 230 MeV in the barrel and 600 MeV in the endcaps) that alse have more energy than all their neighbours. The seeds are then grown into 'topological clusters' by iteratively including crystals that neighbour with crystals already in the cluster beginning with the seed. For a crystal to be included, its energy must must be over another threshold equal to twice the noise level in the associated ECAL region (greater than 80 MeV in the barrel and 150 MeV in the endcaps). There is also an extra requirement for endcap clusters: because the noise increases with η, seeds have an additional criterion of $E_T > 150\,\text{MeV}$. Topological clusters are then assembled into superclusters with a dynamic clustering algorithm [38].

Physics objects are reconstructed with the CMS global event description known as particle flow (PF) [65]. PF uses information from all of the subdetectors to identify and reconstruct individual particles produced within CMS, and to achieve good energy resolution. The information used as inputs are tracks from the tracker, tracks from the muon systems, and energy clusters from the ECAL and HCAL. Depending on which of these are present, PF will output 'PF candidates' corresponding to different types of semi-stable particles:

- **Photons**: ECAL supercluster is present with no associated track in the tracker. The energy of the photons is obtained from the ECAL deposition.
- **Electrons**: ECAL supercluster is present with associated track in the tracker. Energy is determined from the electron momentum at the primary vertex, the ECAL deposition, and the energy of associated bremsstrahlung photons.
- **Muons**: compatible tracks in the tracker and muon system. Energy is determined from the curvature of the tracks.
- **Charged Hadrons**: a compatible track in the tracker, ECAL supercluster and associated HCAL cluster. Energy is determined from the track curvature, and

the matching ECAL/HCAL deposition.

- **Neutral Hadrons**: HCAL and ECAL deposition with no associated track in the tracker. Energy is measured with the deposition from the ECAL and HCAL.

These PF candidates are then used to construct jets, and to determine missing transverse momentum. This process is applied in the same way to data collected with the CMS detector and data from simulation.

5.3 Samples

5.3.1 Trigger

The analysis uses events selected with the two-step CMS triggering system (L1T and HLT). The objective of this system is to keep the event rate below an acceptable level due to limited bandwidth resources, whilst keeping the signal efficiency as high as possible. Requirements at L1T are looser because it uses fast coarse measurements, HLT uses more stringent requirements to compensate for any false positives due to this.

At L1T we require one or two energy deposits in the ECAL with energy thresholds that varied over the 2016 running period. For the single deposit, energy requirements are tighter, and were 25 GeV during low luminosity and going up to 40 GeV at high-luminosity periods to keep the trigger rate to an acceptable level. For two deposits at high-luminosity, 22 GeV for the leading energy deposit and 15 GeV for the subleading were required.

At HLT, events were selected with E_T thresholds of 30 GeV and 18 GeV for the leading and subleading photon respectively. Furthermore, the selection had loose requirements on the shape of the electromagnetic showers, isolation variables, and the ratio of deposition in the ECAL compared to the HCAL.

These selections have their efficiencies measured with the 'tag-and-probe' technique [66]. This uses the resonant production and decay to pairs of well-understood particles near their mass peak to ensure a pure and well-understood sample. In the $H \to \gamma\gamma$ analysis $Z \to e^+e^-$ is used as both electrons and photons are reconstructed with the ECAL clustering, so one can use dielectron decays as a proxy for diphotons. A strict ID requirement is placed on one of the decay products (the tag) and a looser requirement is placed on the other (the probe). The requirement on the probe should be loose enough that it does not affect the selection being measured. The selection efficiency may then be measured as the proportion of the probes which satisfy the selection.

5.3.2 Data

The data used in the analysis corresponds to the $35.9\,\text{fb}^{-1}$ of proton-proton collision data recorded by the CMS experiment in the 2016 run period with a centre of mass energy of $\sqrt{s} = 13\,\text{TeV}$ and selected with the trigger requirements described above.

5.3.3 Simulation

Simulated samples are used for a variety of tasks such as to train the ML models of the analysis, to optimise cuts and categorisations, producing signal models, and to perform validations.

Signal events are simulated for a range of mass points from 120 GeV to 130 GeV using cross sections and branching ratios recommended by the LHC cross section working group. The signal events are generated at next-to-leading order in perturbative QCD with `MadGraph5_aMC@NLO`, with parton showers and hadronization modelled with `pythia8`. The `pythia` tune parameter set `CUETP8M1` is used.

The background simulations are generated in different ways. For the main irreducible background from prompt diphotons, `Sherpa` is used which includes Born processes with up to three jets as well as box diagram processes at leading order. For the γ-jet and jet-jet reducible backgrounds where jets are mistakenly reconstructed as photons we use `pythia8` with a filter applied to enhance the electromagnetic energy content of the jets. Finally, $W\gamma$ and $Z\gamma$ samples used in validation studies are simulated with `Madgraph` and Drell-Yan (DY) is simulated with `Madgraph_aMC@NLO`.

The CMS detector itself is simulated in detail with `GEANT4`. This includes the simulation of both in-time and out-of-time pileup. Simulated events are then weighted such that they reproduce the pileup distribution observed in data from CMS.

5.4 Photon Reconstruction

Candidate photons are reconstructed from calibrated ECAL superclusters. However, these constitute an imperfect measurement of the underlying object and must be corrected. This section will describe these corrections in both simulation and data: first the SC energy is corrected using a trained regressor model and a collection of features; then the photon energies are scaled or smeared depending on whether they are from simulation or data. Once the energy is finalised the photons are evaluated by a BDT classifier that attempts to identify fake photons from jet fragments. This is implemented as a preselection based on the output score and some other features. All of these steps, including their validation, use a common collection of variables detailed in the next subsection for later reference.

5.4.1 Common Variables

The set of common variables can be divided into two main types: shower shape variables and isolation variables, plus some other miscellaneous variables. Shower shape variables describe properties of the electromagnetic showers within the ECAL which will allow us to infer information about the object, for example, whether a shower is from a converted or unconverted photon. The set of shower shape variables consists of the following:

- $E_{2\times2}/E_{5\times5}$: the ratio of energy in the 2×2 grid containing the most energetic crystals in the SC to the energy in the 5×5 grid around the SC seed crystal.
- $cov_{i\eta i\phi}$: the covariance of the crystal η and ϕ locations within the 5×5 grid around the SC seed crystal.
- $\sigma_{i\eta i\eta}$: pseudorapidity width of the shower in terms of crystals.
- R_9: the ratio $E_{3\times3}/E_{SC}$, where $E_{3\times3}$ is the energy in the 3×3 grid around the SC seed crystal and E_{SC} is the energy of the SC.
- $\sigma_{\eta\eta}$: The energy-weighted η width. Computed as the standard deviation of the logarithmic energy-weighted crystal positions in η of a SC.
- $\sigma_{\phi\phi}$: The energy-weighted ϕ width. Computed as the standard deviation of the logarithmic energy-weighted crystal positions in ϕ of a SC.
- σ_{rr}: the standard deviation of the shower width in the $x - y$ plane as measured by the preshower subsystem (only for photons measured in the endcaps).

Isolation variables measure how well-separated an object, in this case a photon, is from other objects in the event such as electrons or charged hadrons that could imitate the true signal. The set of isolation variables consists of the following:

- $\mathcal{I}_\gamma$: photon isolation, the sum of the transverse energy of the particles identified as photons in a cone of $R = 0.3$ around the candidate photon.
- $\mathcal{I}^V_{\mathrm{CH}}$: charged hadron isolation, the sum of transverse momenta of charged particles in a $R = 0.3$ cone around the candidate photon associated with vertex V.
- $\mathcal{I}_{\mathrm{T}}$: track isolation, the sum of transverse momenta of tracks in a hollow cone between $R = 0.3$ and $R = 0.04$ around the candidate photon.
- H/E: the ratio of energy measured in the HCAL to the energy measured in the ECAL in a cone of $R = 0.15$ around the candidate photon.

Finally there are other miscellaneous variables used throughout the selection for different purposes,

- Electron veto: true or false if there is a track associated with the candidate SC or not.
- ρ: the event median energy density per unit area.

- η_{SC}: the pseudorapidity of the candidate SC.
- E_{SC}: the energy of the candidate SC.

5.4.2 Photon Energy

The deviation between the true photon energy (E_{true}) and the SC energy (E_{SC}) occurs due to loss of energy prior to the ECAL and mismeasurement of the energy that is actually deposited. Photons can interact with the pre-ECAL material and begin to shower early. This leads to a more spread-out shower in the ECAL and an overall loss of energy in the pre-ECAL material. Once at the ECAL, the resulting electromagnetic showers can be mismeasured when energy is lost due to improper containment. This can occur via leakage into the gaps between crystals of the ECAL and even through the back of the ECAL if the photon begins to shower deep in the crystal. These effects on the energy are corrected using the photon energy regression.

There are also extra corrections applied due to changing detector conditions that depend on whether the photon is from simulation or data. A time-dependent energy scale correction is applied in data, and in simulation a smearing is applied to match the energy resolution to data. More detail on these processes can be found in [67].

Photon Energy Regression

The objective of the photon energy regression is to predict the correction factor E_{true}/E_{SC}. The regression problem is formulated to target the parameters of the E_{true}/E_{SC} probability distribution function on a per-photon basis. This is taken to have a modified Crystal Ball form with a Gaussian core and two power law tails. The correction is then the peak value of the distribution and a resolution energy estimate may be computed from the width of the Gaussian core. The training itself uses a large set of variables formed from positions within the ECAL, shower shape variables, and region specific information such as from the preshower detector [68].

Energy Scale Correction

Once we have the corrected energies, the overall energy scale needs to be corrected to account for detector effects. During operation the CMS ECAL receives large doses of radiation that can degrade its performance over time. This will lead to drifts and jumps in the detector response as conditions change, and as a result the measured energy will also drift and jump. Scale factors to account for this effect in data are calculated using the $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ decay as a standard candle where the electrons are reconstructed as photons. Using comparison to simulation, and the well-known value of the Z boson mass, scales are derived for different times and detector regions to bring the measured value of real Z bosons back to the true value.

Energy Resolution Correction

Simulation also needs to be corrected by comparison to data to make it more realistic. The photon energy resolution in simulated events has a Gaussian smearing added to it that is derived from comparing the width of the $Z \to e^+e^-$ mass distribution in different categories depending on $|\eta|$-location within the detector (two in the barrel either side of $|\eta|= 1$ and two in the endcaps either side of $|\eta|= 2$), and the R_9 variable that measures photon quality (above or below $R_9 = 0.94$). The mass peak in two of these bins is shown in Figure 5.1.

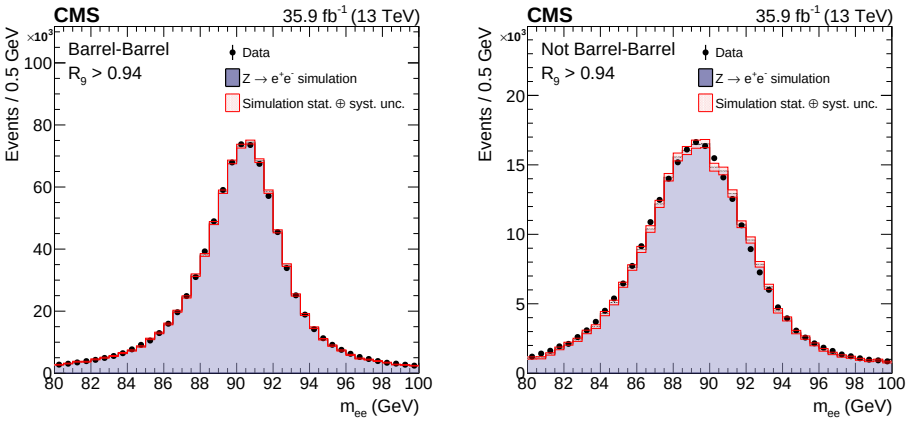

Figure 5.1: A comparison between data and simulation of dielectron invariant mass.

5.4.3 Photon Identification

The photon identification BDT is a classifier whose task is to discriminate between real prompt photons and photon-like jet fragments which satisfy the preselection criteria [67]. The BDT is trained using simulated γ+jet events where the reconstructed photons are matched to a generator-level particle; if there is no match it is considered to be in the non-prompt class. To avoid the BDT introducing a dependence on photon kinematics, the signal photons are re-weighted such that their distribution in p_T and η is flat. The classifier then receives the following input features:

- Shower shape features: $\sigma_{i\eta i\eta}$, $cov_{i\eta i\phi}$, $E_{2\times2}/E_{5\times5}$, R_9, $\sigma_{\eta\eta}$, $\sigma_{\phi\phi}$, and σ_{rr}.
- Isolation features: $\mathcal{I}_\gamma$, $\mathcal{I}_{CH}^{SV}$, and $\mathcal{I}_{CH}^{WV}$, where SV and WV refer to the selected vertex and worst vertex in terms of the vertex probability BDT respectively.
- Other features: ρ, η_{SC}, E_{SC}^{RAW}, and E_{ES}/E_{SC}^{RAW} where E_{ES} is the energy measured by the ECAL preshower (endcaps only).

The performance of this classifier is shown in Figure 5.2. The systematic uncertainty on the BDT output is shown by the shaded region of the right hand plot. This

is estimated so that it covers the largest disagreement between data and simulation of $Z \to e^+e^-$ reconstructed as photons in the endcap regions where agreement is worst.

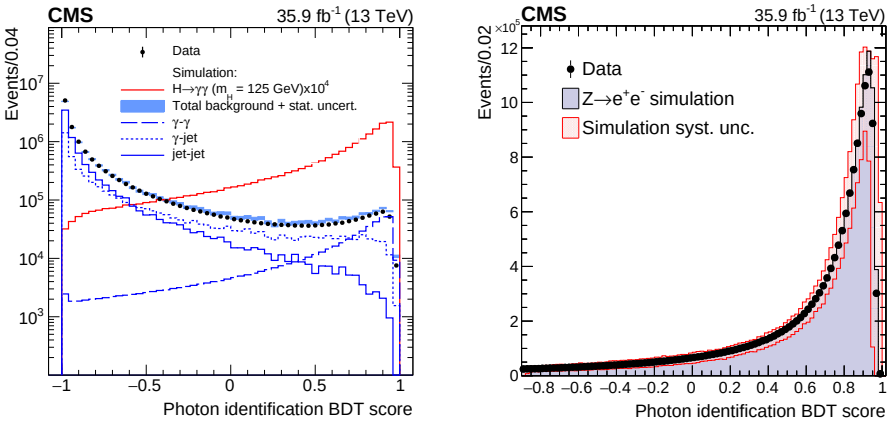

Figure 5.2: Photon ID BDT performance and validation. (Left) Photon ID BDT output score of the lower-scoring photon of each diphoton passing the photon preselection. Signal photons from simulated Higgs events are shown in red and simulated background events are shown in blue, data is shown by the back dots. (Right) Validation on $Z \to e^+e^-$ events.

5.4.4 Photon Preselection

Before a photon can enter the analysis it must pass a set of selection criteria, the photon preselection, which are tighter than the trigger and vary by location. First, photons are grouped into candidate diphotons by considering all possible pairs in the event, then the criteria are applied (with the exception of p_T cuts) on a per-photon basis. The criteria are:

- Electron veto: rejection if there is a track associated to the supercluster.
- Photon p_T: $p_T^{\gamma_1} > 30\,\text{GeV}$ and $p_T^{\gamma_2} > 20\,\text{GeV}$.
- Photon ID: $\hat{y} > -0.9$.

Both photons must then also satisfy either of two requirements:

- $R_9 > 0.8$ and $\mathcal{I}_{CH} < 20\,\text{GeV}$.
- $\mathcal{I}_{CH}/p_T^{\gamma} < 0.3$.

Finally, additional requirements are applied depending on the $|\eta|$ and R_9 of the photon (Table 5.1).

The efficiencies of these criteria are measured using $Z \to e^+e^-$ and the tag-and-probe method, with the exception of the electron veto which uses $Z \to \mu^+\mu^-\gamma$. The preselection efficiencies are summarised in Table 5.2.

Preselection Category	H/E	$\sigma_{\eta\eta}$	R_9	$\mathcal{I}_\gamma$	$\mathcal{I}_T$
Barrel, $R_9 > 0.85$	< 0.08	N/A	> 0.5	N/A	N/A
Barrel, $R_9 < 0.85$	< 0.08	< 0.015	> 0.5	< 4.0	< 6.0
Endcap, $R_9 > 0.90$	< 0.08	N/A	> 0.8	N/A	N/A
Endcap, $R_9 < 0.90$	< 0.08	< 0.035	> 0.8	< 4.0	< 6.0

Table 5.1: Additional photon preselection requirements specific to different $|\eta|$ and R_9 regions.

Preselection Category	$\epsilon_{\text{data}}(\%)$	$\epsilon_{\text{sim}}(\%)$	$\epsilon_{\text{data}}/\epsilon_{\text{sim}}$
Barrel, $R_9 > 0.85$	94.2 ± 0.9	94.7 ± 0.9	0.995 ± 0.001
Barrel, $R_9 < 0.85$	82.5 ± 0.7	82.5 ± 0.7	1.000 ± 0.003
Endcap, $R_9 > 0.90$	90.1 ± 0.2	91.3 ± 0.1	0.987 ± 0.005
Endcap, $R_9 < 0.90$	49.7 ± 1.4	53.8 ± 1.5	0.923 ± 0.010

Table 5.2: Photon preselection efficiencies measured in four different bins.

5.5 Vertex Reconstruction

If the selected vertex is within 1 cm of the correct vertex the contribution of spatial uncertainty to the mass resolution is negligible and is dominated by the energy resolution of the CMS ECAL [8]. The ECAL gives a good determination of the photon location in z and ϕ, but it does not provide any pointing information: to determine α precisely we need to determine the correct vertex by other means. When diphotons are produced in proton collisions there are often charged tracks present from jets or from the proton remnants that are associated to the same vertex. One can exploit this information to choose the correct vertex.

The process begins by gathering the tracks in the central tracker and grouping them by their common points of origin. These are the candidate vertices. The next step will be to choose the vertex most compatible with the candidate diphoton under consideration.

5.5.1 Vertex Selection

Vertex selection is performed with a BDT classifier which takes a set of input features formed from the transverse momenta of tracks associated to the candidate vertex and

the candidate diphoton. The features are

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

where i enumerates the tracks of the candidate vertex. In the case of converted photons there are two additional features: the number of conversion tracks and the pull $|z_{vtx} - z_{conv}|/\sigma_z$ where z_{vtx} is the z position from the vertex, z_{conv} is the position estimated from the conversion tracks, and σ_z is the uncertainty on z_{conv}. The BDT is trained on vertices from simulated Higgs boson diphoton events with vertices that correspond to the true Higgs vertex considered the signal class and all others background. The selected vertex is then the candidate with the highest BDT score.

5.5.2 Vertex Probability

Once a candidate vertex is selected, another BDT is used to score the probability that it is within 1 cm of the true vertex location in z. This is also trained on simulated Higgs diphoton events and is given the following input features:

- The number of vertices in the event.
- The three highest vertex ID scores.
- p_T of the candidate diphoton.
- Δz between the highest scoring vertex and the second highest.
- Δz between the highest scoring vertex and the third highest.
- The number of converted photon tracks.

5.5.3 Performance

The performance of the vertex selection BDT is validated with both simulated and real $Z \to \mu^+\mu^-$ events where the muon tracks have been removed and the event re-reconstructed as a diphoton system. In the converted-photon case a similar procedure uses γ+jet events where the vertex is found using the tracks of the jet. The tracks of the jet are then removed and the event is re-reconstructed as a diphoton system. Validation of the BDT for unconverted photons is shown in Figure 5.3.

The selection efficiency for selecting a vertex within 1 cm of the true position is evaluated using simulated Higgs diphoton decay events. This efficiency is shown as a function of the number of vertices in the event and the diphoton p_T in Figure 5.4. The efficiency over all events is approximately 81%.

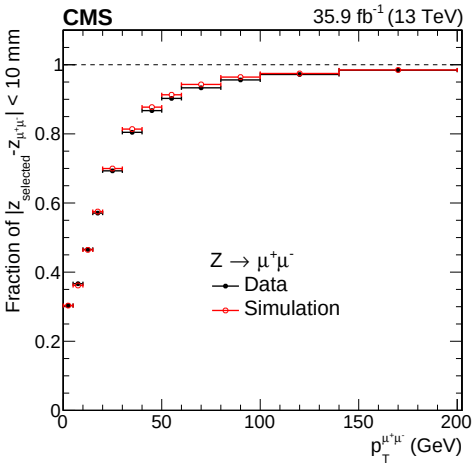

Figure 5.3: Vertex ID efficiency of dimuon events reconstructed as diphotons as a function of p_T in simulation and data.

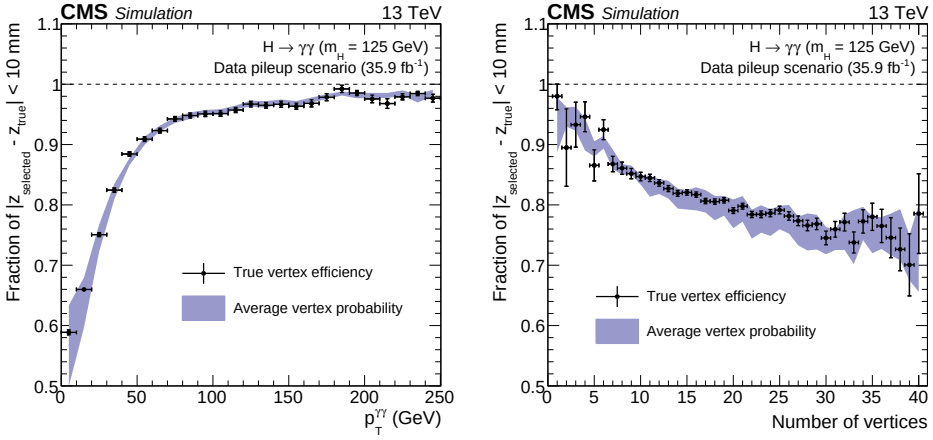

Figure 5.4: Vertex ID efficiency (dots) and average vertex probability (shaded band) as a function of diphoton p_T (left) and number of event vertices (right).

Corrections are applied to account for discrepancies between data and simulation and an associated systematic uncertainty is assigned. This is estimated by varying the ratio of data to simulation within their uncertainties.

The width of the distribution of vertex positions in z is a factor of 1.5 wider in simulation than that measured in data. This is corrected by weighting simulated events with selected vertices more than 0.1 cm from the generator-level Higgs vertex to reproduce the data distribution [8].

5.6 Other Objects

5.6.1 Leptons

Leptons are used for validation purposes throughout the analysis, and also event categorisation where there are leptons in the final state such as production with the $t\bar{t}H$ or VH modes.

Electrons

Electrons are reconstructed in a similar way to photons, but with the extra requirement of an associated track. This association can be achieved in two ways depending on the properties of the candidate:

- ECAL-based: starting with an energetic and well-isolated ECAL SC, match the track that is closest to the energy-weighted position of the SC and also within a window of $\Delta\eta = \pm 0.02$, $\Delta\phi = \pm 0.15$ of it.
- Tracker-based: starting with candidate tracks, associate them to a geometrically compatible SC. This is used for low p_T electrons ($p_T < 10\,\text{GeV}$).

Candidates from both methods are combined to produce the set of PF candidate electrons of the event. There are also corrections applied to the tracks to correct for bremsstrahlung effects, and an energy correction from a BDT regressor analogous to the photon reconstruction. More information on electron reconstruction can be found in [69].

Muons

Hits in the muon system are formed into track segments and these are assembled via a clustering algorithm into muon system tracks called standalone muons. These are then associated with tracks from the inner tracker to produce global muons. A third approach is also used where inner tracks with $p_T < 0.5\,\text{GeV}$ and total $p < 2.5\,\text{GeV}$ are extrapolated into the muon system. If a compatible muon segment is found this makes a 'tracker muon'. Global muons and tracker muons that share the same inner track are merged into a single candidate. Standalone muons have worse momentum resolution and are more contaminated by cosmic rays. More information on muon reconstruction can be found in [70].

5.6.2 Jets

Jets are composite objects reconstructed from PF candidates using the anti-k_T algorithm [71] with $R = 0.4$. The dedicated calibration for each PF candidate type, as well

as 90% of the jet energy being in the form of photons and charged hadrons, allows for high-resolution measurement with the inner tracker and ECAL. The remaining 10% consists of neutral hadrons which are measured at lower energy resolution in the HCAL [72]. The tracks also allow for the identification and rejection of particles originating from pileup vertices. Pileup gives extra energy to signal jets, and the soft pileup jets of these interactions can also be clustered into so-called fake jets of relatively large p_T when they overlap. This latter effect rises quadratically with the number of pileup interactions.

The jet objects receive corrections to their energy that relate the energy of the reconstructed jets to the energy at particle-level. A factorised approach is used [72]:

- Subtract an offset in p_T due to pileup contamination. This is an average correction derived from the global per-event density ρ
- Correct p_T and η dependence of the average jet detector response: this is due to non-linearities in the calorimetry, and differences in detector construction as a function of η, and p_T thresholds.
- Jet energy scale: corrects for average residual discrepancies between data and simulation.

Finally, to reject pileup jets, charged hadrons from vertices other than the chosen vertex are ignored within the tracker acceptance where this information is available. Outside the tracker a selection criterion is placed on the width of the jet expressed as

$$\sigma_{RMS} = \frac{\sum_i p_T^{i\,2} \Delta R^i}{\sum_j p_T^{j\,2}} \tag{5.2}$$

where ΔR^i is the distance between the constituent and the jet axis. The threshold value used is $\sigma_{RMS} > 0.03$.

Chapter 6

Event Categorisation

6.1 Overview and Objectives

Once a set of candidate photons is assembled they are tagged and categorised using extra final-state objects characteristic of particular Higgs production modes. The objective of this tagging procedure is to enhance overall significance, to construct categories of events with superior mass resolution, and to separate out the Higgs production modes for individual measurement.

The event categorisation begins with a selection on the photon candidates with $p_T^{\gamma 1}/m_{\gamma\gamma} > 1/3$, $p_T^{\gamma 2}/m_{\gamma\gamma} > 1/4$, and $100 < m_{\gamma\gamma} < 180\,\text{GeV}$. The use of mass-scaled p_T here and in later machine learning models serves a dual purpose: firstly it avoids distortion of lower values in the $m_{\gamma\gamma}$ spectrum, secondly it avoids introducing mass bias from the simulated data during model trainings. There are then further requirements on the photons' supercluster pseudorapidities: both must have $|\eta|<$ 2.5 to keep them in the fiducial region of the ECAL, and also must not be in the barrel-endcap transition region $1.44 < |\eta|< 1.57$ to ensure full containment of the electromagnetic showers.

6.1.1 The Diphoton BDT

Selected diphoton candidates are then evaluated for signal-like kinematics and mass resolution by a BDT, the diphoton BDT, whose output score is used as a discriminating variable by the tags. The input features of this BDT are the following:

- the mass-scaled transverse momentum $p_T^{\gamma}/m_{\gamma\gamma}$ for the leading and subleading photons;
- the pseudorapidity η for the leading and subleading photons;
- the cosine of the azimuthal angle $\Delta\phi$ between the photons;

- the score from the photon identification BDT for both photons;
- the mass resolution estimate given the assumption that the correct vertex is selected, $\sigma_{\gamma\gamma}^{RV}/m_{\gamma\gamma}$;
- the mass resolution estimate given the assumption that the incorrect vertex is selected, $\sigma_{\gamma\gamma}^{WV}/m_{\gamma\gamma}$;
- the probability that the correct diphoton vertex has been selected p^{RV},, estimated with the vertex probability BDT.

The mass resolution in the right vertex case, $\sigma_{\gamma\gamma}^{RV}/m_{\gamma\gamma}$, is assumed to be completely dominated by the ECAL photon energy resolution; one can therefore neglect vertex uncertainty. The energy resolution for each photon can be approximated by a Gaussian distribution and combined in quadrature to give the following expression for mass resolution,

$$\sigma_{\gamma\gamma}^{RV} = \frac{1}{2}\sqrt{(\sigma_{\gamma1}^{E}/E_{\gamma1})^2 + (\sigma_{\gamma2}^{E}/E_{\gamma2})^2} \tag{6.1}$$

where $\sigma_{\gamma1}^{E}/E_{\gamma1}, \sigma_{\gamma2}^{E}/E_{\gamma2}$ are the relative uncertainties on the photon energies for the leading and subleading photons respectively. In the wrong vertex case, $\sigma_{\gamma\gamma}^{WV}/m_{\gamma\gamma}$, the extra contribution to the mass resolution is modelled with an extra term. This term is assumed to be Gaussian in form, with a width equal to the extent in z of the beam spot multiplied by $\sqrt{2}$. This extra term is then summed in quadrature with the mass resolution for the right vertex case,

$$\sigma_{\gamma\gamma}^{WV} = \frac{1}{2}\sqrt{(\sigma_{\gamma\gamma}^{RV}/m_{\gamma\gamma})^2 + (\sigma_{\gamma\gamma}^{V}/m_{\gamma\gamma})^2}. \tag{6.2}$$

The diphoton BDT is trained on all four signal samples and the QCD, GJet and diphoton background samples. Each training event is weighted in proportion to its cross section, its event weight and its expected mass resolution. When events are weighted during training like this it can be considered to be a way of defining the 'cost' of misclassifying a particular event. Higher weight events will have a higher associated misclassification cost and will therefore be prioritised over lower weight events. Specifically, signal weight events are weighted as follows,

$$w^{sig} = \frac{p^{RV}}{\sigma_{\gamma\gamma}^{RV}/m_{\gamma\gamma}} + \frac{1-p^{RV}}{\sigma_{\gamma\gamma}^{WV}/m_{\gamma\gamma}}. \tag{6.3}$$

This scheme helps ensure that the diphoton BDT will assign a relatively high score to events with good expected mass resolution. The signal-flattened score distribution for all simulated signal and background samples, as well as data, is shown in the lefthand plot in Figure 6.1.

The performance of the diphoton BDT is validated in a $Z \to e^+e^-$ control region

where the normal diphoton selection has been applied, but the electron veto is inverted (Figure 6.1 right). The diphoton BDT output score has good agreement between data and simulation in the score region used by the event tagging.

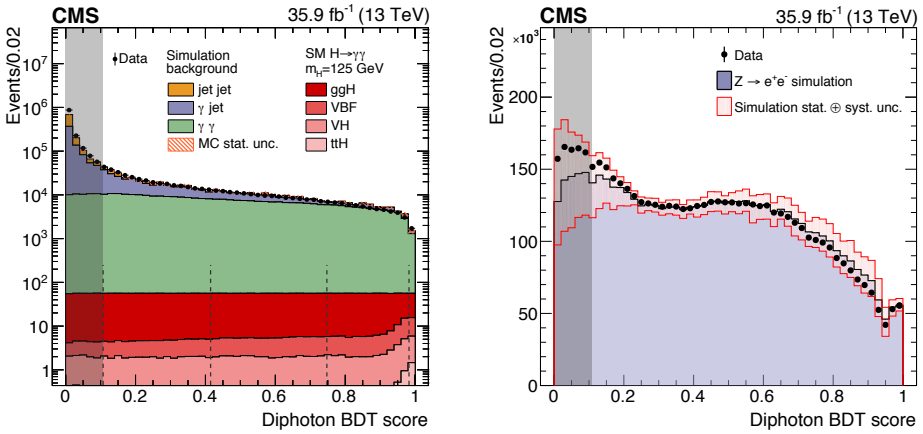

Figure 6.1: Stacked diphoton BDT score distributions for simulated signal and background, with data shown superimposed (left). Diphoton BDT score in the $Z \to e^+e^-$ control region (right). The same transformation has been applied to the score distribution in both plots such that the total signal distribution is flat.

6.1.2 Tagging Scheme

Tagging is implemented as a fall-through sequence where diphoton events are offered to each tag in order of priority (Table 6.1). If a diphoton event is not accepted by a tag it then passes to the next tag for consideration until the final 'Untagged' tag category. If the event does not meet the criteria for this last tag it is discarded. In the case of multiple tagged candidate diphotons in an event the one with the highest priority tag and category is selected, if they are in the same category the diphoton with the highest diphoton p_T is chosen. The criteria for each tag will be specified in following sections in this chapter.

6.2 Top Fusion Tagging

In the $t\bar{t}H$ production mode, a top-antitop pair is produced in association with the Higgs boson. The top quark immediately decays to a b quark and a W boson which will subsequently decay leptonically or hadronically. In the former (semi-leptonic) case there will be a bottom quark jet plus an associated lepton with E_T^{miss} from the W decay. In the latter (fully-hadronic) case there will be a bottom quark jet plus two quark jets from the W decay to quarks (Figure 6.2).

Tag	Target Process	Structure
$t\bar{t}H$ Leptonic	$t\bar{t}H$ with semi-leptonic top decays	Single category
$t\bar{t}H$ Hadronic	$t\bar{t}H$ with fully-hadronic top decays	Single category
ZH Leptonic	VH with leptonically-decaying Z boson	Single category
WH Leptonic	VH with leptonically-decaying W boson	Single category
VH Leptonic Loose	VH with leptonically-decaying W or Z boson	Single category
VBF	VBF with dijet in the final state	Three categories
VH MET	VH with significant amount of E_T^{miss}	Single category
VH Hadronic	VH with hadronically-decaying W or Z boson	Single category
Untagged	Inclusive	Four categories

Table 6.1: The $H \to \gamma\gamma$ tag sequence in order of tag priority from highest (top) to lowest (bottom).

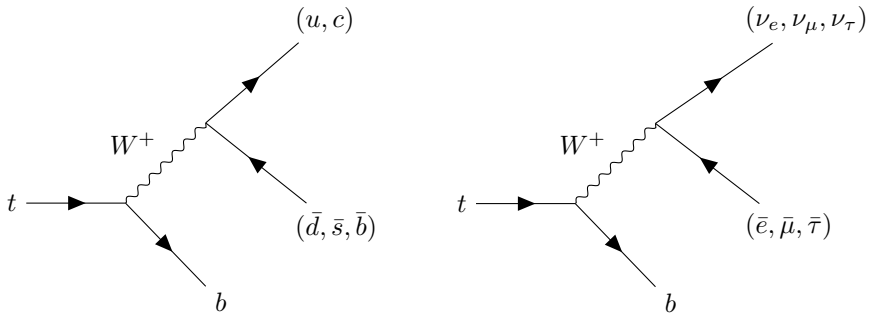

Figure 6.2: Top quark decay modes: a fully-hadronic decay (left) and a semi-leptonic decay (right).

The top tags target these two decay modes: the leptonic tag searches for $t\bar{t}H$ events where at least one top quark decays semi-leptonically, and the hadronic tag searches for $t\bar{t}H$ events where both top quarks decay fully-hadronically.

6.2.1 $t\bar{t}H$ Leptonic

This tag uses a set of selections on kinematic properties of leptons and jets in the event. Leptons are required to pass selection requirements depending on their flavour:

- diphoton BDT score > 0.11;
- at least one selected lepton with $p_T > 20\,\text{GeV}$;
- all selected leptons are required to have an angular separation from a signal photon of $R(\ell, \gamma) > 0.35$;
- $|m_{e\gamma} - m_Z| > 5\,\text{GeV}$ (electrons only);
- a minimum of two jets in the event with $p_T > 25\,\text{GeV}$, $|\eta| < 2.4$, $R(j, \gamma) > 0.4$

and $R(j, \ell) > 0.4$;

- at least one jet is tagged as a b jet by the CSV tagger (medium requirement).

6.2.2 $t\bar{t}H$ Hadronic

This tag uses a set of selections on kinematic properties of the jets in the event, as well as a dedicated BDT. The $t\bar{t}H$ hadronic BDT is trained on the following input features:

- the number of jets with $p_T > 25\,\text{GeV}$;
- the p_T of the leading jet;
- the two highest scores of the CSV b-tagger.

A selection on the BDT output score (Figure 6.3) is optimised simultaneously on simulation with a selection on the diphoton BDT score to maximise expected precision on the signal strength of the $t\bar{t}H$ production channel.

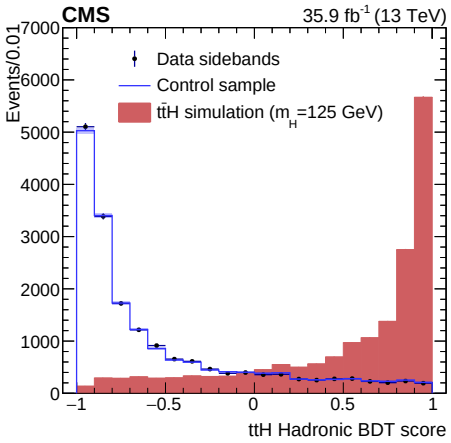

Figure 6.3: Score distribution of the hadronic $t\bar{t}H$ BDT. The blue lined histogram shows the distribution for the control region, the red filled histogram shows the score distribution for simulated signal, and the points show the score distribution of the data sideband regions ($m_{\gamma\gamma} < 115\,\text{GeV}$ or $m_{\gamma\gamma} > 135\,\text{GeV}$).

A control region is constructed by selecting photon pairs where one passes the preselection and photon ID requirements, whilst the other has no preselection requirement and the photon ID is inverted. These events are then weighted in η and p_T to reproduce the kinematic properties of the photons in the signal region.

The selection requirements of the $t\bar{t}H$ Hadronic tag are as follows:

- $p_T/m_{\gamma\gamma} > 1/3$ and $1/4$ for leading and subleading photons respectively;
- diphoton BDT score > 0.4;

- no leptons that meet the criteria of the $t\bar{t}H$ Leptonic tag;
- a minimum of three jets in the event with $p_T > 25\,\text{GeV}$ and $|\eta| < 2.4$;
- at least one jet is tagged as a b jet by the CSV tagger (medium requirement);
- a $t\bar{t}H$ Hadronic BDT score above 0.75.

6.3 Associated Production Tagging

In the associated production (VH) mode a $W^{\pm}$ or Z boson is produced in association with the Higgs boson. The VH tags target different vector bosons decaying in different ways which can manifest as leptons, jets or E_T^{miss} in the event. All of the leptonic VH tags are selection-based and have various isolation requirements to avoid contamination from Drell-Yan background processes.

6.3.1 ZH Leptonic

This tag targets Higgs production in association with a Z boson that subsequently decays leptonically with stringent requirements. The selection criteria are as follows:

- $p_T^{\gamma}/m_{\gamma\gamma} > 3/8$ for leading photon;
- diphoton BDT score > 0.11;
- two same-flavour leptons with $p_T > 20\,\text{GeV}$ and satisfying the same requirements as in the $t\bar{t}H$ Leptonic tag;
- $70 < m_{\ell\ell} < 110\,\text{GeV}$;
- $R(\gamma, e) > 1.0$, or $R(\gamma, \mu) > 0.5$;
- conversion electron veto: if an electron and a photon share a supercluster, the electron track must be well-separated from the supercluster centre ($R(SC, e) > 0.4$).

6.3.2 WH Leptonic

Targets Higgs production in association with a $W^{\pm}$ boson that subsequently decays leptonically with stringent requirements. The selection criteria are as follows:

- $p_T^{\gamma}/m_{\gamma\gamma} > 3/8$ for leading photon;
- diphoton BDT score > 0.28;
- at minimum one lepton with $p_T > 20\,\text{GeV}$ and satisfying the same requirements as in the $t\bar{t}H$ Leptonic tag;
- $R(\gamma, \ell) > 1.0$;
- $E_T^{\text{miss}} > 45\,\text{GeV}$;
- a maximum of two jets each satisfying $p_T > 20\,\text{GeV}$, $|\eta| < 2.4$, $R(j, \ell) > 0.4$, $R(j, \gamma) > 0.4$;
- electron conversion veto as in the ZH Leptonic tag.

6.3.3 VH Leptonic Loose

This tag targets Higgs production in association with either $W^{\pm}$ or Z which then decay leptonically. This tag uses an orthogonal E_T^{miss} selection of $E_T^{\text{miss}} < 45$ GeV, with the rest of the selection being the same as WH Leptonic.

6.3.4 VH MET

Targets Higgs associated production with E_T^{miss} from at least one missing lepton. The selection criteria are as follows:

- $p_T^{\gamma}/m_{\gamma\gamma} > 3/8$ for leading photon;
- diphoton BDT score > 0.79;
- $E_T^{\text{miss}} > 85\,\text{GeV}$;
- $|\Delta\phi(\gamma\gamma, E_T^{miss})| > 2.4$.

6.3.5 VH Hadronic

This tag targets Higgs production in association with a W or Z boson that decays hadronically. The selection criteria are as follows:

- $p_T^{\gamma}/m_{\gamma\gamma} > 1/2$ for leading photon;
- diphoton BDT score > 0.79;
- a minimum of two jets with $p_T > 40\,\text{GeV}$ and $|\eta| < 2.4$, $R(j,\gamma) > 0.4$;
- dijet invariant mass $60 < m_{jj} < 120\,\text{GeV}$;
- $|\cos\theta^*| < 0.5$, where θ^* is the difference in diphoton polar angles $\theta_{\gamma\gamma}$ in the diphoton-dijet centre-of-mass frame, and the lab frame.

6.4 Untagged

If the candidate is not assigned to any of the other tags it is considered for inclusion in a final inclusive tag: Untagged. The Untagged tag consists of four categories defined as exclusive selections on the diphoton BDT score and consists mostly of gluon fusion events. If an event does not meet the lowest requirement it is discarded from the analysis.

These categories are optimised simultaneously in a similar way to the VBF tag, but with a few differences. The procedure begins with the boundaries spaced equally along the score distribution. Each category is evaluated in simulation by fitting an exponential for background plus two Gaussian distributions for the signal. Significance is extracted from each category based on a fit to an Azimov dataset derived from the earlier exponential and Gaussian function fits. The boundaries are optimised to maximise overall significance of the Untagged tag.

This procedure is repeated for increasing numbers of Untagged categories until there is no significant increase in performance. This is then the number of Untagged categories to be used in the tag.

6.5 VBF Tagging

The VBF production mode is characterised by its distinctive event topology and kinematics: two high-p_T jets with large pseudorapidity separation and high invariant mass. Furthermore, the dijet substructure will also be distinctive with both jets originating from quarks, having colour connections to the proton remnants and possibly having other correlations in structure between the two jets.

Other production modes can also produce a Higgs boson in association with jets to produce a VBF-like final state. In particular, ggH can be a significant source of background due to its larger cross section and capacity to produce jets at next-to-leading order or from initial-state radiation. These dijets will mostly be from gluons, therefore targeting the jet substructure will be important in discriminating these production modes.

The VBF tag targets the VBF production mode by exploiting the distinctive properties of VBF dijets. At the core of the VBF tag is a machine learning model which takes these distinctive properties as input features. The selection and category assignment of the tag is then based on the output of this model. This chapter explores two approaches.

- A tag based on two BDTs with engineered kinematic features using `Scikit-learn` [73]. This is the approach used in the 2016 $H \to \gamma\gamma$ analysis.
- A tag based on a single dense convolutional neural network built in `TensorFlow` [74] that receives jet structure information in the form of images in addition to engineered kinematic features.

Both tags use the same event preselection, and produce scores used to define event categories that enhance the expected significance of the VBF channel and will be evaluated in the same way. The only difference will be the machine learning model that the tags are based around, and the extra image-based information in the dense CNN tag.

The problem formulation is the following: to separate VBF from SM background and ggH events using simulated data. Constraints are that the model must generalise to real data and must not introduce a bias to the diphoton mass. There are a few challenges associated with this problem.

- There is a severe class imbalance where there are approximately seven times more examples of the background class than the signal.

- The events are weighted, some events can be equivalent to multiple others.
- Some events have negative weight and are needed for correct distribution shapes.
- The total weight difference between the classes is very large and make the class imbalance problem even worse.
- The QCD background sample has very large weights and very few events. This causes the background distribution shapes to become very jagged.

The model is evaluated using the area under the ROC curve (AUROC), a performance measure of a binary classifier. This measure is chosen because it is robust to class imbalance and can easily be evaluated with weighted events.

When developing the tag itself and its categories the approximate mean significance (AMS) [75] is used as the figure of merit. This is defined as

$$\text{AMS} = \sqrt{2\left((s+b+b_{\text{reg}})\log\left(1+\frac{s}{b+b_{\text{reg}}}\right)-s\right)}, \tag{6.4}$$

where s is the total number of signal events, b is the total number of background events, and b_{reg} is a regularisation term that reduces sensitivity to local optima. The value of b_{reg} is chosen to be 5. AMS is estimated by simultaneously fitting an exponential plus a double Gaussian function to the diphoton mass distribution. The background and signal event weights from an interval of two effective standard deviations around the peak are summed to produce s and b, and to estimate AMS.

6.5.1 Selections

When a candidate diphoton is considered for VBF selection additional requirements are applied based on the jet content of the event. First requirements are applied on a per-jet basis, if there are more than two jets which meet these requirements the top two in p_T are selected to form a dijet. Finally, a preselection based on dijet kinematics is applied to the candidate dijet. If it does not pass, the event falls through to the lower-priority categories (VH MET, VH Hadronic, and Untagged).

Jet Selection

Jets are required to meet the criteria detailed in the previous chapter plus some additional requirements specific to the VBF tag. Pileup jet ID (PUJID) uses a BDT classifier [76] that takes a collection of jet shape variables and produces a score for each jet. A collection of selections on this score are then applied for bins in p_T and η (Table 6.2). In this analysis the tight working point is used as this gives the highest expected significance for the VBF tag, and also leads to marked improvement in data/simulation agreement in η in the $\text{Z} \to \text{e}^+\text{e}^-$ plus jets control region. Furthermore

	$\|\eta\|< 2.5$	$2.5 \leq \|\eta\|< 2.75$	$2.75 \leq \|\eta\|< 3.0$	$3.0 \leq \|\eta\|< 5.0$
$20 < p_T \leq 30\,\text{GeV}$	0.69	−0.35	−0.26	−0.21
$30 < p_T \leq 50\,\text{GeV}$	0.86	−0.10	−0.05	−0.01
$50 < p_T \leq 100\,\text{GeV}$	0.95	0.28	0.31	0.28

Table 6.2: Pileup jet ID cuts of the tight working point.

there is a photon-jet isolation criterion that requires the jet to have $\Delta R(\gamma, j) > 0.4$ with both of the photons of the candidate diphoton and a jet pseudorapidity requirement of $|\eta_j|< 4.7$.

Dijet Preselection

Dijets are formed by selecting the two highest-p_T jets in the event that pass the jet selection requirements. The highest-p_T jet in the pair is referred to as the leading jet, and the other jet as the subleading jet. If there are fewer than two jets the event is rejected by the VBF tag and falls through to Untagged. Candidate dijets are required to meet the following selection criteria before being presented to the machine learning model:

- $p_T^\gamma / m_{\gamma\gamma} > 1/3$ and $1/4$ for leading and subleading photon respectively;
- photon ID BDT score > -0.2 for both photons;
- dijet invariant mass $m_{jj} > 250\,\text{GeV}$;
- jet $p_T > 40\,\text{GeV}$ and $> 30\,\text{GeV}$ for the leading and subleading jets respectively.

The criterion on the photon ID score is motivated by under-performance of the diphoton BDT in the VBF phase space. The diphoton BDT was trained over all signal, the bulk of which will consist of ggH where the diphoton is produced with no extra objects such as jets. In this phase space the transverse momentum of the diphoton system is highly discriminating. This leads the diphoton BDT to assign a high score just on high values of diphoton p_T, and with lax requirements on photon ID. This in turn leads to under-performance in the VBF phase space where the p_T spectrum is harder and low photon ID background events are given a high score.

6.6 VBF Tag with BDTs

Once a candidate event passes the preselection it is presented to a machine learning model consisting of two BDTs in sequence: the dijet BDT and the combined BDT. The output classification score of this model will be used to define the categories of the tag selection.

6.6.1 Dijet BDT

The purpose of the dijet BDT is to evaluate how VBF-like events are based on kinematic information from the dijet and the diphoton, and in particular to handle the rejection of ggH. The BDT receives the following features which are chosen to minimise correlation with the diphoton mass:

- $p_T^\gamma/m_{\gamma\gamma}$ for the leading and subleading photons;
- p_T^{j1} and p_T^{j2}, the transverse momenta of the leading and subleading jets respectively;
- m_{jj} the invariant mass of the dijet;
- $\Delta\eta$ the pseudorapidity gap between the two jets;
- $\min\Delta R(\gamma, j)$ the smallest angular separation between either of the diphoton photons and either of the jets;
- $|\Delta\phi_{\gamma\gamma jj}|$ the absolute azimuthal angular difference between the diphoton and dijet;
- $|\Delta\phi_{jj}|$ the absolute azimuthal angular difference between the jets of the dijet;
- $C_{\gamma\gamma}$ the diphoton centrality expressed as:

$$C_{\gamma\gamma} = \exp\left(-\frac{4}{(\eta_{j1}-\eta_{j2})^2}\left(\eta_{\gamma\gamma} - \frac{\eta_{j1}+\eta_{j2}}{2}\right)^2\right) \tag{6.5}$$

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

Their distributions with the VBF preselection applied are shown in Figure 6.4.

This dijet BDT is trained on all simulated SM background samples (with ggH included) versus VBF. To increase the number of training examples the training uses a loosened dijet preselection requirement where the $p_T^\gamma/m_{\gamma\gamma}$ are reduced to 1/4 and 1/5, the jet p_T cuts are reduced by 10 GeV, the dijet invariant mass cut is reduced to 100 GeV and the photon ID cuts are not applied. The normalised score distributions for the classes and the ROC curves for each individual sample are shown in Figure 6.5. These scores are then used as an input feature in the next BDT in the VBF tag: the combined BDT.

6.6.2 Combined BDT

The purpose of the combined BDT is to combine information from the diphoton BDT and dijet BDT to produce the final discriminant score for defining VBF tag categories. Specifically, it takes the following input features:

- diphoton BDT score;
- dijet BDT score;

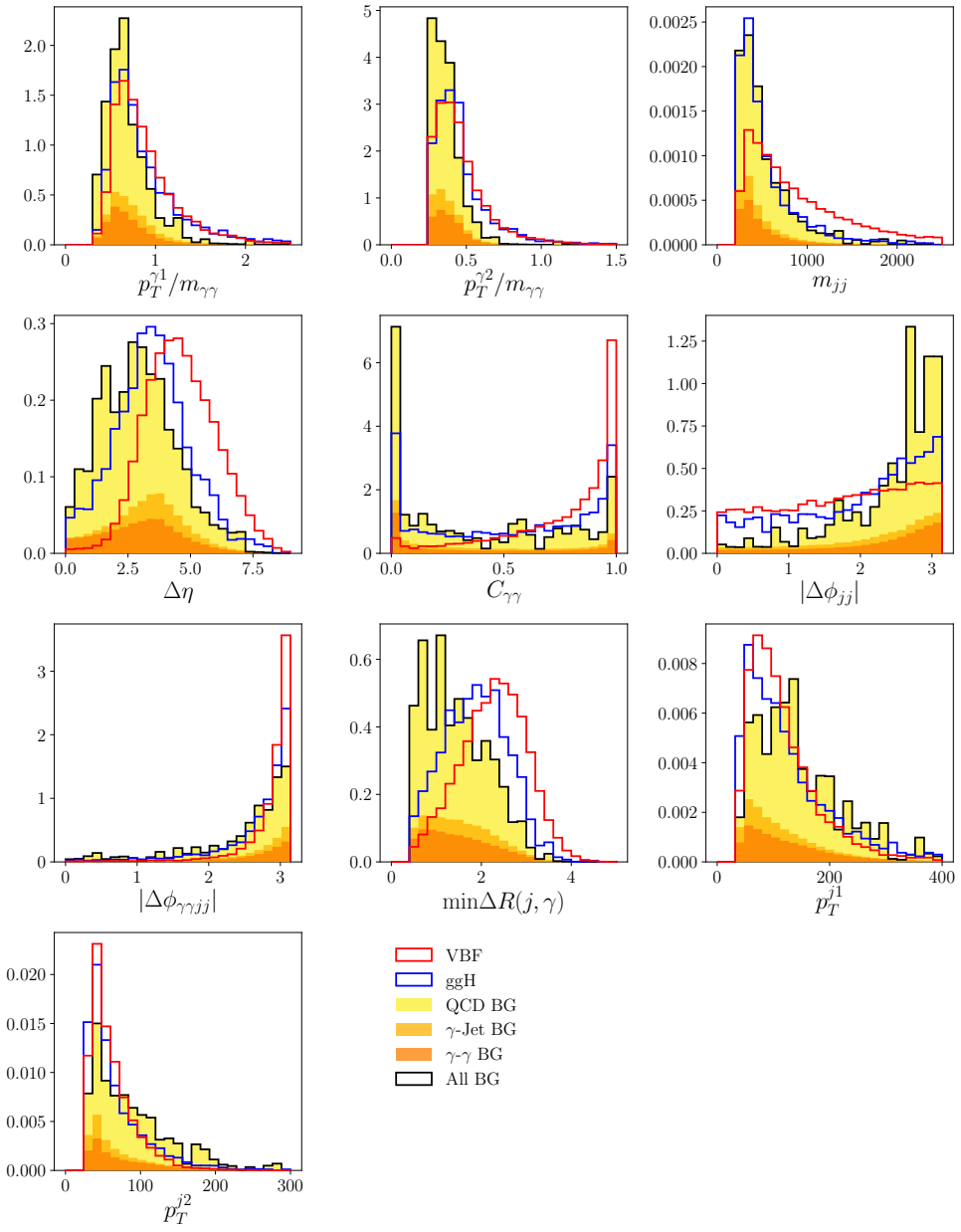

Figure 6.4: Dijet BDT feature distributions with the full VBF preselection. Distributions are all normalised to unity with the solid red line corresponding to VBF, blue line to ggH, and black line to SM background. The SM background distribution is shown as a stacked histogram.

- $p_T^{\gamma\gamma}/m_{\gamma\gamma}$, the mass-scaled transverse momentum of the diphoton.

The distributions of these features with the full VBF preselection applied are shown in Figure 6.6.

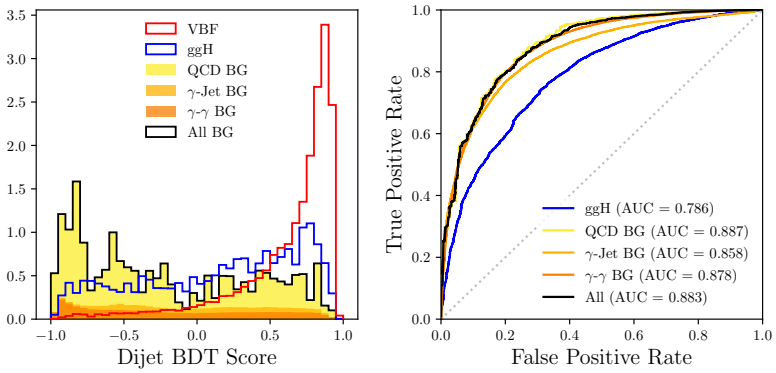

Figure 6.5: Dijet BDT performance. On the left are the output score distributions for VBF (red), ggH (blue) and SM background (black). The SM background distribution is shown as a stacked histogram. On the right are the ROC curves for the dijet BDT split into the different samples. The performance against ggH is noticeably lower than the other backgrounds.

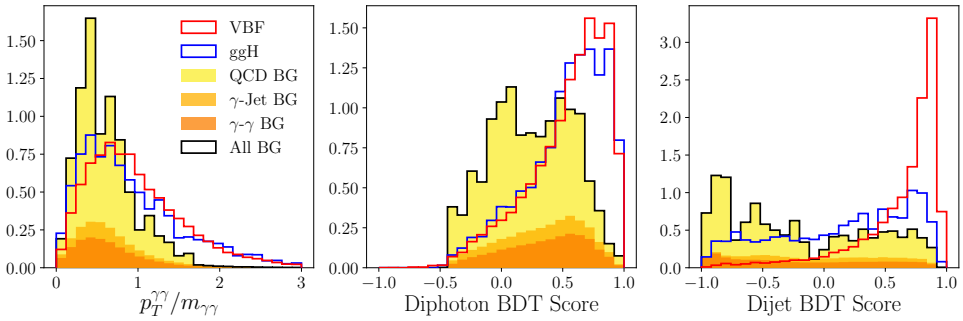

Figure 6.6: Combined BDT feature distributions with the full VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH. The SM background is shown as a stacked histogram.

The combined BDT is then trained with the SM background samples vs VBF. Gluon fusion is not included in this training as it is found to reduce the ability of this BDT to reject SM background. This is considered to be a higher priority than ggH rejection because rejection of SM background has the largest impact on statistical significance. The normalised combined score distributions for the classes and the ROC curves for each individual sample are shown in Figure 6.7.

6.6.3 Model Interpretation

The model can be interpreted by observing how features are used together in high and low-scoring events. This is achieved by examining their joint distribution for highest and lowest percentile scoring events. The result is shown in Figure 6.8 where the solid

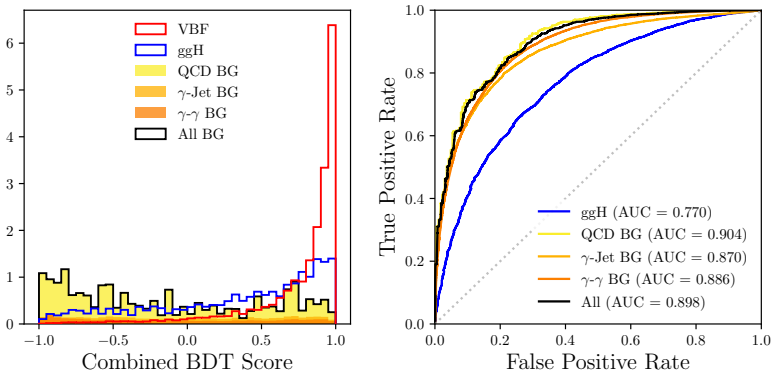

Figure 6.7: Combined BDT score distribution with the full VBF preselection are shown on the left. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM backgrounds. The SM background distribution is shown as a stacked histogram. ROC curves broken down by background sample are shown in the same colours on the right.

colour shows the values averaged over each percentile, and the lines show the top five highest (or lowest) scoring events.

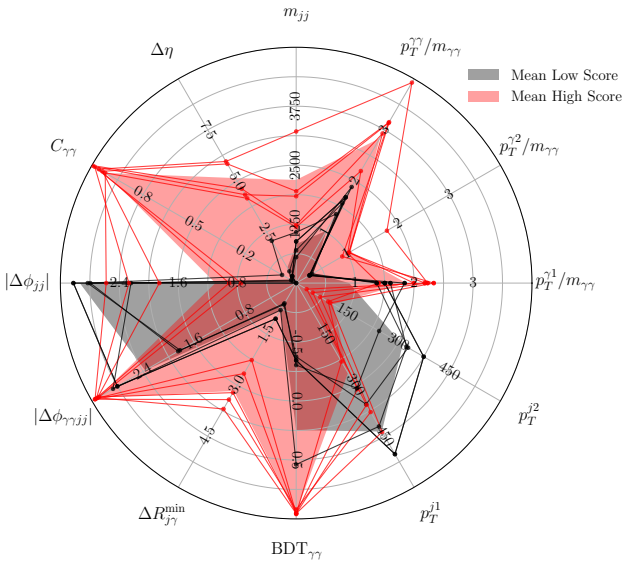

Figure 6.8: Coloured regions correspond to mean values for top percentile (red) and bottom percentile (black) combined score events. The top and bottom five scoring events are also shown by lines and dots.

The discrimination power of the model is driven by the dijet angular variables, with the exception of $|\Delta\phi_{\gamma\gamma jj}|$ and the dijet mass. High jet p_T is actually more of an indicator of background. Diphoton BDT score shows substantial overlap, likely from

ggH but there could still be residual high-p_T SM background events and underperformance in the VBF phase space.

6.6.4 Categorisation and Tag Performance

Once a candidate event has been preselected and evaluated by the model it is considered for inclusion in the VBF categories. These are defined as exclusive selections on the combined BDT output score and are chosen to maximise the AMS over all categories. If the combined score is not high enough for inclusion in the lowest category it is rejected and is passed for consideration by lower-priority tags.

The AMS is estimated for each category by constructing a diphoton mass histogram of all events in the category score range, an exponential function is then fitted to the mass sidebands, and a double Gaussian in the signal region to the background-subtracted mass histogram. The parameters of these fits are then used as initial values for a simultaneous signal-background fit. The values of s and b are evaluated in a region around the peak with width equal to two effective sigma either side estimated from the the simultaneous fit. Overall significance is then calculated as the sum in quadrature of the significance of each category.

The boundaries are optimised simultaneously for overall significance with a random search algorithm. The number of categories is chosen considering an increasingly larger number and performing a category optimisation for each one. The procedure stops when the improvement is less than one percent. This was found to happen with three categories.

Approximate studies (in lieu of the full final fits machinery) of the three VBF categories using the procedure described above are carried out to study tag performance. Their boundaries and their estimated performance is shown in Figure 6.9 and Table 6.3.

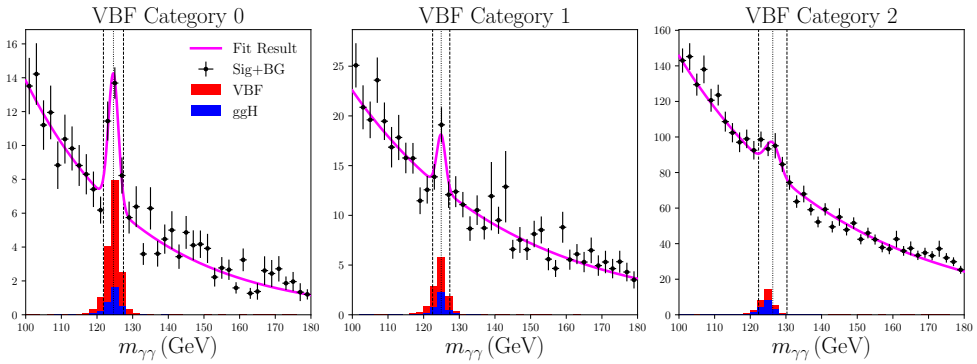

Figure 6.9: Mass fits for estimating AMS.

Category	Score Range	σ_{eff}	AMS	$B_{\text{ggH}}/(S+B_{\text{ggH}})$	$S/(S+B)$
VBF 0	$[1.00, 0.957)$	1.4	2.16	0.20	0.37
VBF 1	$[0.957, 0.902)$	1.2	1.00	0.34	0.17
VBF 2	$[0.902, 0.553)$	2.0	0.69	0.53	0.04

Table 6.3: Estimated category attributes for the BDT-based VBF tag.

6.6.5 Validation

$Z \to e^+e^-$ Control Region

Validation of the VBF tag uses the same $Z \to e^+e^-$ control region as the other tags, but with the extra requirements of the VBF preselection. This control region is used for simulation/data comparison of both the features input to the BDTs and the BDT scores themselves (Figure 6.10). There is good data-simulation agreement in the

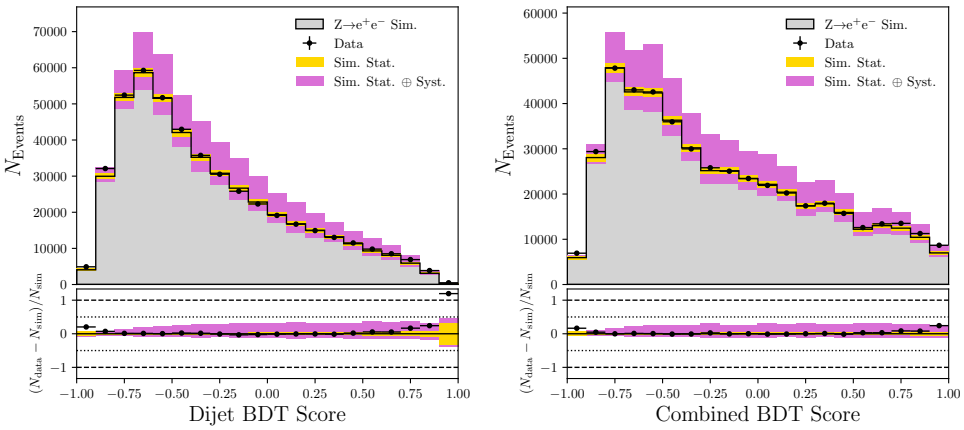

Figure 6.10: Data/Simulation comparison for dijet and combined BDT output scores.

output scores of the BDTs and in the kinematic features (Appendix B).

These plots only show the marginal distributions of these variables, they do not show their joint behaviour. To examine data-simulation agreement of the joint distribution a BDT is trained to discriminate between simulation and data events. If the discrimination power of the resulting model is high then the agreement is bad, if it is equivalent to guessing this suggests that the joint distributions match closely. Selections on the BDT score can then be used to try to isolate regions of the joint distribution where there is disagreement.

The results shown in Figure 6.11 show that the score distributions are similar, discrimination is very low, and therefore the joint feature distribution has little disagreement.

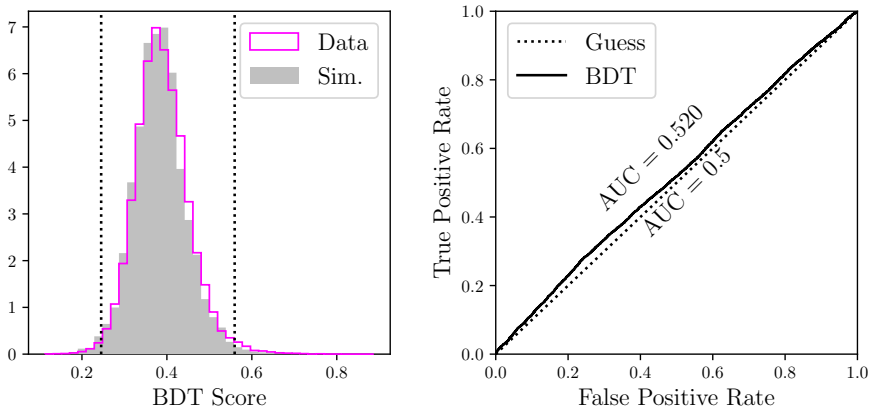

Figure 6.11: Joint distribution study with BDT on the $Z \to e^+e^-$ control region data-simulation test set.

QCD Modelling Variations

The accuracy of QCD process modelling in hadronic collisions is a significant source of theoretical uncertainty. This will have an impact on categorisation with hadronic objects such as jets, and is especially pertinent for any features that are based on jet substructure.

To test for how such mismodelling affects the VBF tag and VBF/ggH discrimination, samples are evaluated with up and down variations on aspects of jet production:

- **Underlying event**: interactions between parton pairs in the proton collision that are not part of the hard scatter;
- **Parton shower**: simulates a succession of parton emissions from the incoming and outgoing partons of the interaction.

Both of these will affect the cross section for jet production, the configuration of the jet substructure and particularly the degree of colour connection to the proton remnant.

To evaluate how the performance of the model changes depending on the QCD modelling, a ROC curve is constructed for each variant, and an envelope is drawn around the curves to estimate performance bounds. This is shown in figure 6.12

The change in performance is modest. This is expected as the VBF model only uses kinematic variables, and the main impact of these variations may be through pileup mitigation and selecting the incorrect jet rather than substructure mismodelling.

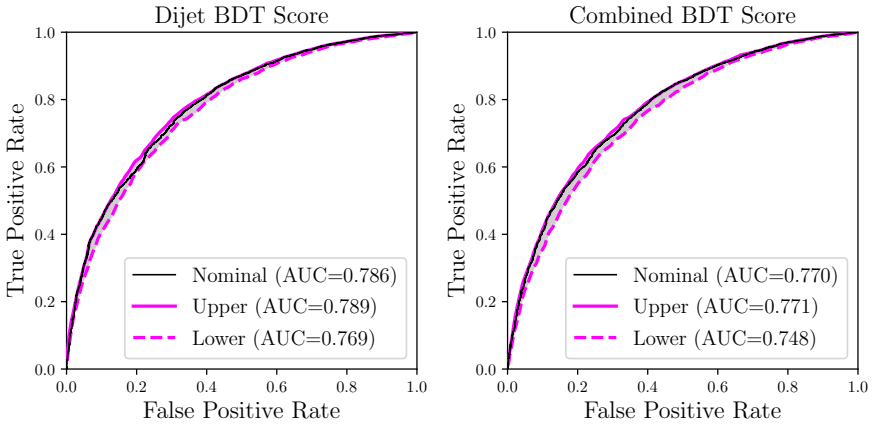

Figure 6.12: ROC curves for parton shower and underlying event variations. The nominal performance is shown in black, and the magenta lines show the upper and lower bounds of the envelope covering all the curves.

6.6.6 Single BDT Model

The two-step structure of the VBF tag was first developed for the Higgs boson Run 1 analysis on less performant software and different selections. In particular, the photon ID cut of the preselection removes much of the background that the combined BDT targets. When trained over events with this cut applied, the combined BDT adds little to the performance of the VBF Tag. Using a single BDT equivalent to simply adding the diphoton BDT score to the dijet BDT performs at the same level as the two step tag (Figure 6.13).

Furthermore, the original train-test split of the two-step BDT was unavailable. Evaluating over the entire sample will include training events and exaggerate the performance of the BDT-based tag. A single-step BDT was trained and found to be equal in performance. A small increase in the two-step tag is possibly from the inclusion of training events. It is a fairer test to compare this model to the dense CNN because the exact same training and test sets can be used.

6.7 VBF Tag with a Dense Convolutional Neural Network

In the BDT-based tag the ggH separation power is much lower than the SM background samples. This is a challenging problem to solve when equipped with only kinematic variables. Jet structure variables should offer important extra information.

Rather than using hand-engineered jet structure features, this problem is ap-

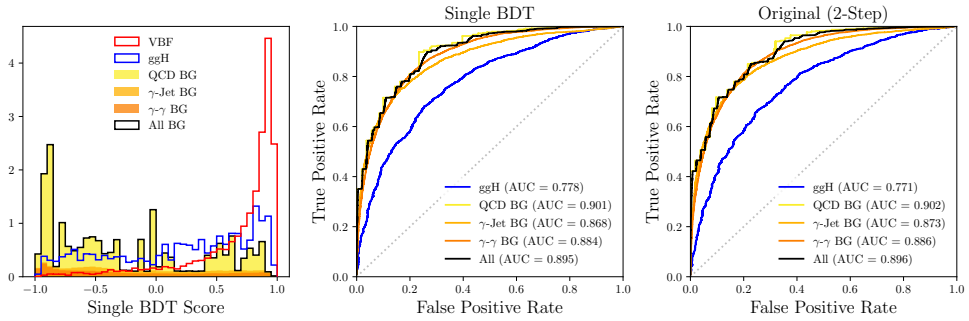

Figure 6.13: Single BDT performance and comparison to the two step approach. Distributions of the single BDT score are shown on the left and are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM backgrounds. The SM background distribution is shown as a stacked histogram. Corresponding ROC curves for the single BDT (centre) and original approach (right) are shown with colours denoting the same samples as the histogram.

proached by formulating jet structure as an image and then training a dense convolutional neural network. The model will form many discriminating features as part of the learning process, some hopefully more sophisticated and more powerful than common engineered features.

This section will describe the construction of these images, and the model built to process them in detail. The optimisation techniques for selecting the model structure are described, and the features the model has learned will be explored using a collection of techniques. After this the model will be validated with particular emphasis on how it is affected by the quality of QCD event modelling.

6.7.1 Jet Images

The properties of a jet's constituent particles can provide important information about the originating parton. An image is a natural way of representing this information, with the spatial distribution represented by the arrangement of pixel values and the channels of the image representing properties such as charged particle p_T deposition in the pixel region.

Formulation

The image formulation used in this thesis (Figure 6.14) is inspired by [77], and uses two three-channel jet images (corresponding to the leading and subleading jets) stacked in the channels' dimension to produce a $n \times n \times 6$ dijet image. The three channels are the following: p_T deposition of charged PF candidates, p_T deposition of neutral PF candidates, and PF candidate multiplicity. The space that the pixels correspond

to is the space of particle displacements in pseudorapidity and azimuthal angle from the jet axis $(\Delta\eta, \Delta\phi)$,

$$\begin{aligned} \Delta\eta =& \eta_p - \eta_j \\ \Delta\phi =& \phi_p - \phi_j \end{aligned} \tag{6.6}$$

where subscript p denotes a constituent particle and j denotes the jet. The pixels themselves are not a rectilinear grid in $(\Delta\eta, \Delta\phi)$, but are evenly-spaced in the polar coordinates

$$\begin{aligned} \Delta R =& \sqrt{\Delta\eta^2 + \Delta\phi^2} \\ \varphi =& \text{atan2}(\Delta\phi, \Delta\eta) \end{aligned} \tag{6.7}$$

These have been rotated by half a pixel in φ so that the $(\Delta\eta, \Delta\phi)$ axes line up with the centres of a row of pixels rather than the boundary between them. Finally, the images are normalised such that the sum of the p_T-based channels equal one, and the sum of the individual multiplicity channels equal one. All of the images in this thesis will have the same form and are always centred on the jet axis.

This stacked dijet image formulation is used to facilitate finding correlations in structure between the two dijet jets. In this formulation it will happen at a lower level rather than constructing complex features on a per-jet basis and then comparing them. For example, this is needed for detecting the characteristic colour connection of VBF: if the jets are in opposite hemispheres the jet image will show the candidates to be pulled in opposite $\Delta\eta$ directions.

Polar coordinate pixels are used to give finer segmentation at the centre of the jet and to make it easier for a DCNN to construct translation-invariant filters. Features will depend more on $(\Delta R, \varphi)$ than $(\Delta\eta, \Delta\phi)$.

Image Dataset and Preprocessing

These images are different in their formulation and behaviour compared to a typical image one finds in computer vision problems. Firstly, they are sparse with only a fraction of the pixels ever non-zero in any one image. Secondly, assumptions about local correlations between pixels do not apply: two adjacent red pixels would mean two adjacent particles. Max pooling will simply pick the higher valued pixel during downsampling and information about the second particle will be lost. Thirdly, in the rectilinear image which is seen by the network (bottom right of Figure 6.14) there is a periodic boundary condition where the top pixels wrap around to the bottom ones. When convolution operations are performed on these images the padding must be periodic in the vertical direction (φ direction).

The dijet image distribution has a few notable features that originate from the

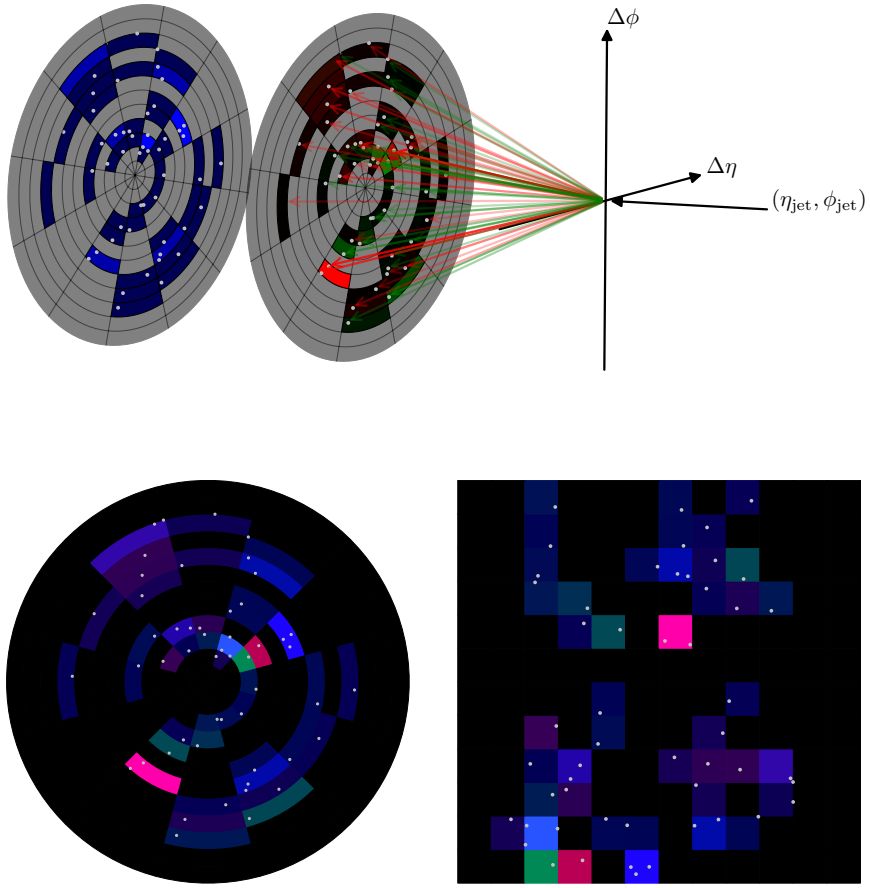

Figure 6.14: Construction of single 12×12-pixel three-channel jet image (top). Arrows correspond to individual jet constituents where red arrows are charged, green are neutral and the opacity of the arrows corresponds to candidate p_T. The multiplicity channel is drawn separately, and black pixels lightened so the charged and neutral channels can be seen clearly. The final image (bottom) shown in both $(\Delta\eta, \Delta\phi)$ coordinates (left) and the $(\Delta R, \varphi)$ coordinates seen by the network (right).

structure of CMS. Outside the tracker acceptance there is no charge measurement and therefore the charged p_T channel may be all zero. The coarse structure of the forward detector regions also gives a change in image properties: here the images become constrained to a grid of dots. This can be see in the mean images shown in Figure 6.15 where it manifests as a green grid in the mean of the image distributions (especially in VBF).

Images are preprocessed before being input to the DCNN for both training and inference such that each feature (image pixel channel value) distribution has mean

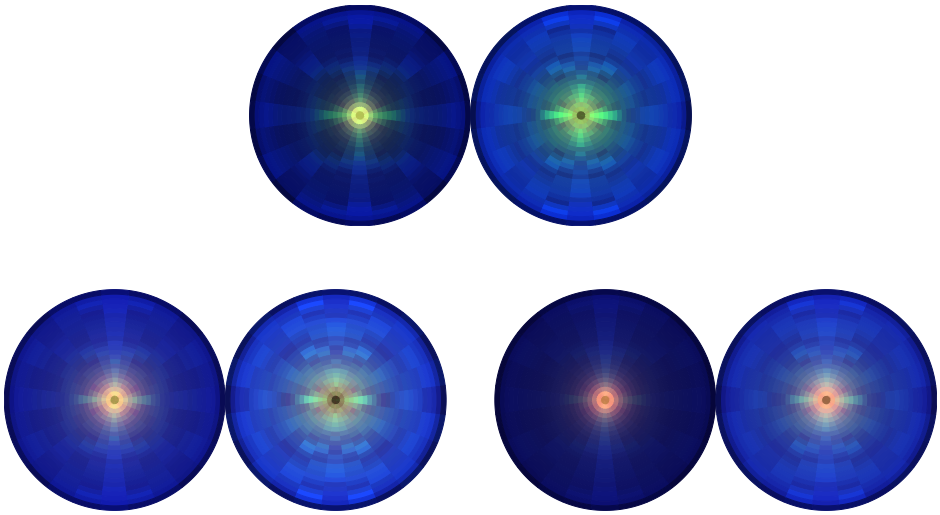

Figure 6.15: Mean dijet images for VBF events (top), gluon fusion events (bottom left) and Standard Model background processes (bottom right). In each dijet image the left hand image corresponds to the leading jet and the right corresponds to the subleading jet.

zero and standard deviation of one. This is achieved by subtracting the per-feature mean of the dataset and dividing by the per-feature standard deviation. In this case the class-balanced mean and standard deviation are used where each class has equal weight. This is important for when the training is carried out over class-balanced minibatches.

6.7.2 Dense CNN Model

Model Design

The overall structure of the model can be considered to built from three main parts:

- **Convolutional section** for learning dijet substructure features from dijet images.
- **Merge section** for processing and integrating engineered kinematic features with learned features from the convolutional section.
- **Main discriminant** fully-connected layers for integrating all information and producing the class logits.

The **convolutional section** consists of a 'spread layer' (SL) followed by three dense blocks (DB1, DB2, DB3) each of which are followed by transition units (TU1, TU2, TU3). All layer activation functions are leaky ReLUs to avoid the dying ReLU problem.

The spread layer is a depth-wise convolution layer that produces N-many feature maps for each channel where the filters do not mix the image channels. For each channel's associated feature maps half of them have their values evenly permuted in the vertical direction, this corresponds to a rotation by π in the polar image. The function of this layer is to spread-out the sparse image into a collection of feature maps that correspond to simple local spatial configurations of pixels such as radial or angular bands of deposition. The interleaved rotations of this layer's output feature maps allows for the comparison of pixels opposite each other around the jet axis much earlier in the network. This layer gives two hyperparameters to the model: the filter size and the number of features per input channel.

The dense blocks construct increasingly higher-level feature maps, and the transition units combine feature maps for feature reduction as well as downsampling with average pooling to avoid information loss associated with max pooling mentioned before. The structure of each of these parts is tuneable, and therefore gives another twelve hyperparameters to the model: three from each dense block and one from each transition unit.

The **merge section** consists of a set of fully-connected layers with the first one after the initial input a different size to the others. This is then concatenated with the output of the convolutional part. The function of this section is to embed the engineered features in a higher-dimensional space, form them into a vector the same size as the convolutional section output, and then combine them together with the jet structure features. This section has three hyperparameters: the size of the hidden layers, the number of layers, and the size of the first hidden layer relative to the others.

The **main discriminant** consists of a sequence of fully-connected layers that take the full vector of concatenated features as input and produce three class logits which correspond to the VBF, ggH, and SM background process classes. These logits are then mapped to class probabilities by a softmax function. The VBF class probability is then used to define tag categories.

In addition to the above, the formulation of the image can also be tuned like another hyperparameter to choose the most performant number of pixels.

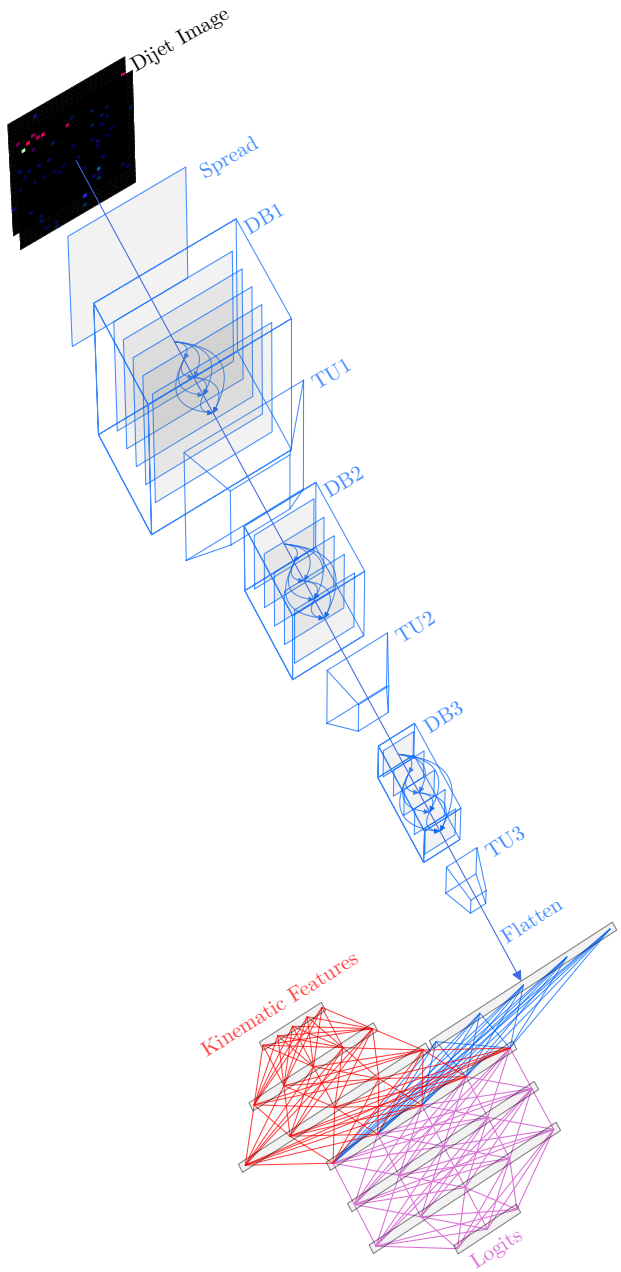

Figure 6.16: A schematic view of the dense CNN model architecture. The convolutional section is indicated by blue, the merge section by red, and the main discriminant by purple. Grey squares and rectangles show layers of neurons: a depthwise convolution layer in the spread layer, composite layers in the dense blocks, and fully-connected layers in the merge section and main discriminant.

Image Formulation and Network Architecture Search

Choosing the specific network architecture is a complex and time-consuming problem as there are many hyperparameter choices to explore and model training can take as long as 24 hours. To find an optimal model in the space of possible hyperparameter choices an optimisation scheme based on Bayesian optimisation is used.

As the training time is a severe bottleneck the hyperparameter space is sampled for trainings on the two-class subproblem of only VBF/ggH. The metric to be optimised for is the peak AUROC evaluated on the validation set during training, and each training can only take a maximum of 24 hours. This time requirement is to keep the model size and therefore training time to a practical level for use in a physics analysis. The optimisation is split into a series of steps targeting different parts of the hyperparameter space. All of these optimisations were carried out using the Imperial College London Research Computing Service [78].

First an approximate optimisation is carried out over all of the structure hyperparameters to ensure that the candidate model is not in a strongly suboptimal region to begin with. Next the choice of image channels and pixels is optimised to find the optimal image formulation. After this the spread unit, dense block and transition unit depth structure are optimised and then another optimisation is carried out for the filter sizes in the spread layer and dense blocks. Finally, the depth and size of the fully-connected layers of the merge unit and main discriminant are optimised.

The final optimised structure is as follows.

- Image: 24×24 pixels with charged p_T, neutral p_T, and multiplicity for each jet.
- SL: 16 features per channel with a filter size of 5×5.
- DB1: 3 layers with growth rate 20 and filter size 3×3.
- TU1: reduction factor of 0.59.
- DB2: 6 layers with growth rate 10 and filter size 3×3.
- TU2: reduction factor of 0.95.
- DB3: 16 layers with growth rate 4 and filter size 3×3.
- TU3: reduction factor of 0.5.
- Merge: first layer size factor 1.5, layer size 512, depth 2.
- Discriminant: layer size 512, depth 3.

The optimisation appears to have favoured more abstract and complex features later in the model where the field of view is larger. This is shown by the fact that much of the depth is introduced in DB3. There is also a strong feature reduction in TU3 that strongly curtails model complexity. High-dimensional output from the convolutional section leads to a large increase in parameters due to the fully-connected layer immediately after it.

Regularisation

The DCNN model is regularised with a single L_2 regulariser term for the convolutional section, dropout, and gradient clipping.

L_2 regularisation encourages the learning of smoother features in the convolutional section and was found in optimisation studies to give better performance than L_1 or $L_1 + L_2$. In the $L_1 + L_2$ optimisation the L_1 term became very small and effectively converged to the L_2 result. A three-dimensional Bayesian optimisation was carried out to tune the final L_2 hyperparameters for the convolution, merge and main discriminant sections. This optimisation showed that very small values are preferred in the non-convolutional sections, and they were therefore set to zero. The use of L_2 regularisation in the convolutional section will encourage smoother features. This may stop the model memorising particular configurations of pixels and prevent overfitting. It is desirable that the representation learned by the model is spread-out in the neurons as this leads to robustness and better generalisation [79]. This is another effect of L_2 regularisation.

Dropout is used in all sections, but the dropout in the convolutional section is different to standard dropout. In the convolutional section spatial dropout is used where entire feature maps are dropped rather than individual neurons. This stops the model from getting over-reliant over particular features and is another way of encouraging a more spread-out representation.

Gradient clipping is particularly crucial to keep training updates stable in this model. The scheme limits the maximum size of parameter updates during stochastic gradient descent by clipping the gradients to some maximum value. Specifically this is done by calculating the global L_2 norm of the gradient vector in the parameter space and scaling the gradient vector back if it is over the maximum value.

This is necessary when the loss surface as a function of the model parameters has cliff-like regions. If the gradient is evaluated at the cliff face the gradient may be very large which will lead to a very large parameter update kicking the parameters of the model far from the point it was evaluated at. Gradient clipping will stop this from happening by keeping the parameter update small and allowing the parameters of the model to settle at the bottom of the cliff. These cliff-like regions are common when dealing with sparse inputs where most values are zero, therefore this is most likely a consequence of the dijet image sparsity.

Loss

The loss function encodes an opinion of what constitutes good model performance and it is here that one can define the cost of particular sorts of misclassification over others. There are two components to this: inter-class costs and intra-class costs.

Inter-class cost defines the relative priority of misclassification between the classes. This was attempted in a less formal manner in the BDT-based model when ggH was left out of the combined BDT training. This would be equivalent to setting the misclassification cost for the ggH events to zero.

For the DCNN-based model a cost-sensitive version of cross entropy developed in [80] is used,

$$L_i = -\log\left(\frac{\xi_{pp}e^{o_p^i}}{\sum_{j=0}^{2}\xi_{pj}e^{o_j^i}}\right) \tag{6.8}$$

where i enumerates the events of the minibatch, o_j is the logit of class j, p is the true class index of the event, and ξ is the cost matrix which encodes misclassification costs. The cost matrix has the following definition,

$$\xi = \begin{pmatrix} c_{\mathrm{BG}} & c_{\mathrm{BG/ggH}} & c_{\mathrm{BG/VBF}} \\ c_{\mathrm{BG/ggH}} & c_{\mathrm{ggH}} & c_{\mathrm{ggH/VBF}} \\ c_{\mathrm{BG/VBF}} & c_{\mathrm{ggH/VBF}} & c_{\mathrm{VBF}} \end{pmatrix} \tag{6.9}$$

where each element c is a real number belonging to the interval $(0, 1]$ on the diagonal and $[0, 1]$ off-diagonal. This was selected to be symmetric to simplify the cost optimisation problem. If all the elements are set to one this is equivalent to ordinary cost-insensitive cross entropy.

When training the model $c_{\mathrm{BG/ggH}}$ is set to zero as this type of misclassification does not matter at all in the VBF tag. The training should not compromise on other types of misclassification to improve it. This leads to the largest performance increase in VBF discrimination and is crucial to being able to treat this as a three-class problem. The diagonal values are all set to one, tuning this was found to make little difference. These entries are equivalent to setting the general cost of misclassifying a class. The cost $c_{\mathrm{BG/VBF}}$ is kept at one as this type of discrimination is the primary objective of the model. Finally, $c_{\mathrm{ggH/VBF}}$ was tuned to maximise VBF/ggH discrimination while keeping VBF/BG discrimination at or above the level of the BDT-based model. This was found to be 0.5.

Intra-class costs are introduced by weighting events in each minibatch depending on their event weight and their class. Within each class in the minibatch each event is given a weight in proportion to its event weight scaled to belong to an interval $[1, a]$ where a is a positive real number greater than one. If this scaling is not used the difference in weight between events can be multiple orders of magnitude leading to instability and underperformance. The total weight for each class is then scaled to be equal to mitigate the effects of class imbalance.

The loss over the entire minibatch is expressed as a weighted sum of events,

$$L = \frac{1}{\sum_{j=0}^{N-1} w_j} \sum_{i=0}^{N-1} w_i L_i \tag{6.10}$$

where N is the number of events in the minibatch, and w_i is the weight of event i. This is the final loss that is minimised by the optimiser during training.

Training Process

The training process consists of two steps: first the convolutional part is trained on its own with two fully-connected layers, then the convolutional layer parameters are frozen and the full model is trained with the engineered features. This approach is used because it was found that when the convolutional section is trained simultaneously with the rest of the network, the training is less stable and the model under performs.

The objective of the convolutional section training is to develop the learned features that will be used in the full model. This is targeting VBF/ggH discrimination, therefore training is treated as a two-class problem with just the ggH and VBF samples.

This runs over 150000 batches of 900 events with event weights equal to one. Each batch is class-balanced and contains 450 events from each of ggH and VBF. During the training the data are augmented by randomly reflecting the images in the $\Delta\eta$ and $\Delta\phi$ directions to improve generalisation.

The optimisation algorithm used is Adam with Nesterov momentum (Nadam) with learning rate 0.001 and the loss used is cross entropy. Nadam was found to be the most performant in contrast to the CNN literature where SGD with Nesterov momentum is consistently better. This is likely due to adaptive learning rate algorithms such as Adam having parameter-specific learning rates that are increased if the parameter does not change much. This makes them better at handling sparse network inputs which can result some parameters of the network not being updated as often as others.

The trained convolutional model then has its parameters frozen, and the fully-connected layers after the convolutional section (after TU3) are removed. The output of TU3 will be the learned features produced in the first training step, they are then concatenated with the 'spread-out' engineered features of the merge section. Another training is then performed over the full training set to train the merge section and the final discriminant.

This training runs over 100000 batches of 900 events with scaled event weights as described in the loss subsection. Each batch is again class-balanced, but this time it has 300 events from the three classes: VBF, ggH, and SM background. The optimisation algorithm is Nadam with the same learning rate and the loss is the

cost-sensitive cross entropy described previously.

6.7.3 Model Performance

To determine how the learned features contribute to VBF/ggH discrimination performance the image-only network (stage one of the training) may be compared to the single-step BDT model. This performance is examined on the same test set with both the loose and full VBF preselections. On the loose preselection the image-only network achieves an AUROC of 0.800 which is already better than the 0.796 of the BDT (Table 6.4) However, upon moving to the full PS the image-only performance drops below the BDT. The score distribution and the associated ROC curve for the image-only model are shown in Figure 6.17. This drop is possibly due to the more stringent p_T cuts removing gluon jets which have a softer p_T spectrum.

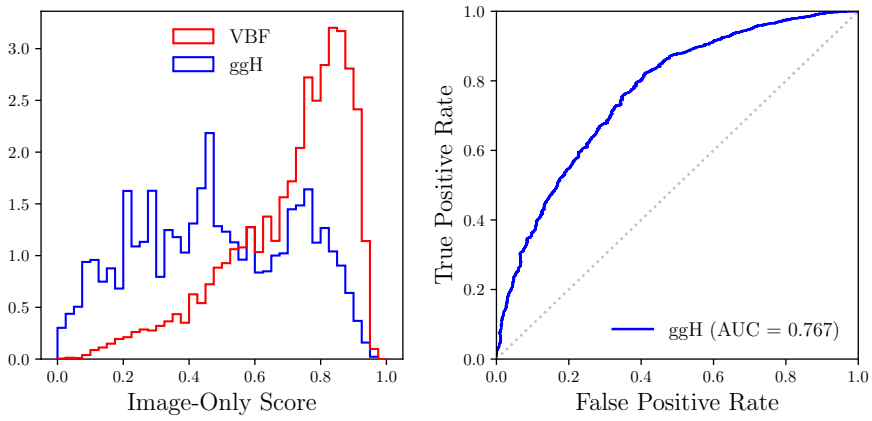

Figure 6.17: VBF/ggH discrimination performance for the image-only model with the full VBF preselection. The score distribution for VBF (red) and ggH (blue) is shown on the left. The associated ROC curve measuring VBF/ggH discrimination power is shown on the right.

The full model includes the same set of kinematic variables as the BDT and has been trained with the tuned costs over all of the backgrounds (stage two of the training). This model brings a substantial improvement in the VBF/ggH discrimination, especially with the full preselection. For the other background samples the performance is similar with the DCNN AUROCs being slightly higher. This information is summarised in Table 6.4 and the score distribution with the associated ROC curves are shown in Figure 6.18.

In conclusion, the DCNN-based model demonstrates a strong improvement over the BDT-based model in VBF/ggH discrimination and the learned features alone are strongly discriminating.

	Full PS		Loose PS	
Sample	BDT	DCNN	BDT	DCNN
ggH	0.778	0.837	0.796	0.845
QCD	0.901	0.907	0.892	0.885
γ-jet	0.868	0.870	0.869	0.882
$\gamma\gamma$	0.884	0.891	0.870	0.881
All	0.895	0.901	0.883	0.884

Table 6.4: Comparison table for BDT/DCNN AUROCs with full and loose preselections broken down by background sample.

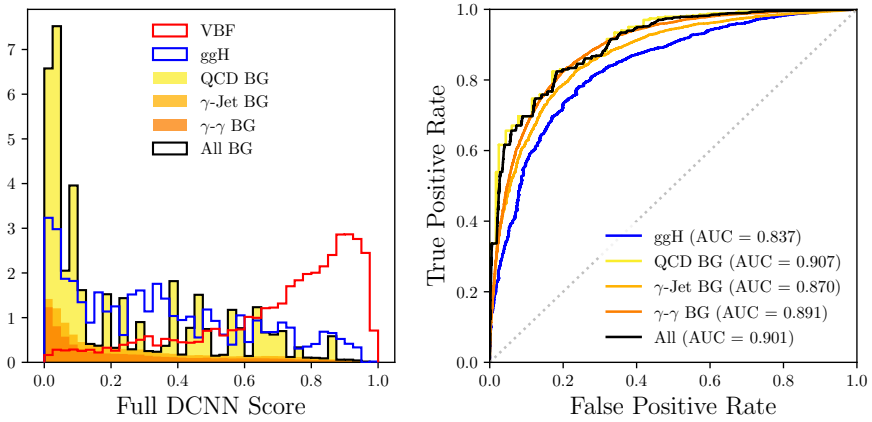

Figure 6.18: Discrimination performance for all of the background samples with the full model and the full VBF preselection. Score distributions (left) are shown as stacked histograms for the SM background, a blue line histogram for ggH, and a red line histogram for VBF. Associated ROC curves (left) are shown in the same colours for each sample, with all background together shown in black.

6.7.4 Categorisation and Tag Performance

To determine how the higher model performance of the DCNN translates into the VBF tagging and categorisation itself, the category boundaries are re-evaluated. This uses the same procedure as before for the boundary optimisation itself and the number of categories. The optimal number of categories for the DCNN-based VBF tag is found to be three. Approximate studies of the three VBF categories, their boundaries and their estimated performance is shown in Figure 6.19 and Table 6.5.

The proper measurement of these category performances requires the full machinery of the signal and background modelling, and the final fits of Chapter 7. These studies show a reduction in the ggH contamination of VBF 0, but the other categories are approximately unchanged. This demonstrates the challenge of improving

Category	Score Range	σ_{eff}	AMS	$B_{\text{ggH}}/(S+B_{\text{ggH}})$	$S/(S+B)$
VBF 0	$[1, 0.856)$	1.5	2.13	0.10	0.37
VBF 1	$[0.856, 0.704)$	2.0	1.15	0.27	0.11
VBF 2	$[0.704, 0.495)$	2.0	0.60	0.46	0.04

Table 6.5: Estimated category attributes for the DCNN-based VBF tag.

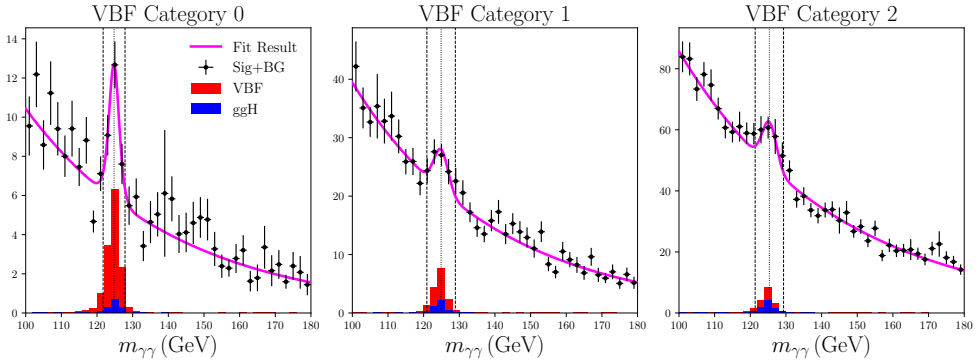

Figure 6.19: Mass distributions and fits for the three optimised DCNN-based tag categories.

ggH discrimination as it needs to be traded off against overall background rejection performance, and as a corollary; statistical significance.

6.7.5 Model Interpretation

A collection of techniques can be used to determine what sort of features the convolutional section is learning to form. This section will explore three: feature visualisation with synthetic images generated to maximally activate part of the network, real images that achieve the same thing, and direct inspection of the frontmost filters (filters of the spread layer).

Feature Visualisation

The technique of feature visualisation uses optimisation and differentiability of neural networks to interpret their inner workings, particularly CNNs [81]. A famous example of this process is inceptionism and the so-called 'deep dream' images [82]. The network is in effect forced to 'hallucinate' by creating a feedback loop between its visual input and neural activations that iteratively alters the visual input to maximise the activation.

The approach used in this thesis is based on SGD with momentum and no regularisation or other conditions. The process begins with an initial image tensor, x_{ijk},

with small random positive values, and a momentum tensor, v_{ijk}, of the same size initialised to zero. The loss in this case will be a single output neuron o_l. The image is then altered as follows:

$$\begin{aligned} v_{ijk} &= \mu v_{ijk} + \alpha \frac{\partial o_l}{\partial x_{ijk}} \\ x_{ijk} &= x_{ijk} + v_{ijk} \end{aligned} \tag{6.11}$$

where the partial derivative has been computed with backpropagation, and α is the learning rate. Negative pixel values are set to zero and the image is normalised as described in the image formulation section before the next iteration. This continues for 1000 iterations with the learning rate reduced by a factor of 0.75 every 100 iterations. The end results of this for the VBF and ggH logits of the image only network are shown in Figure 6.20.

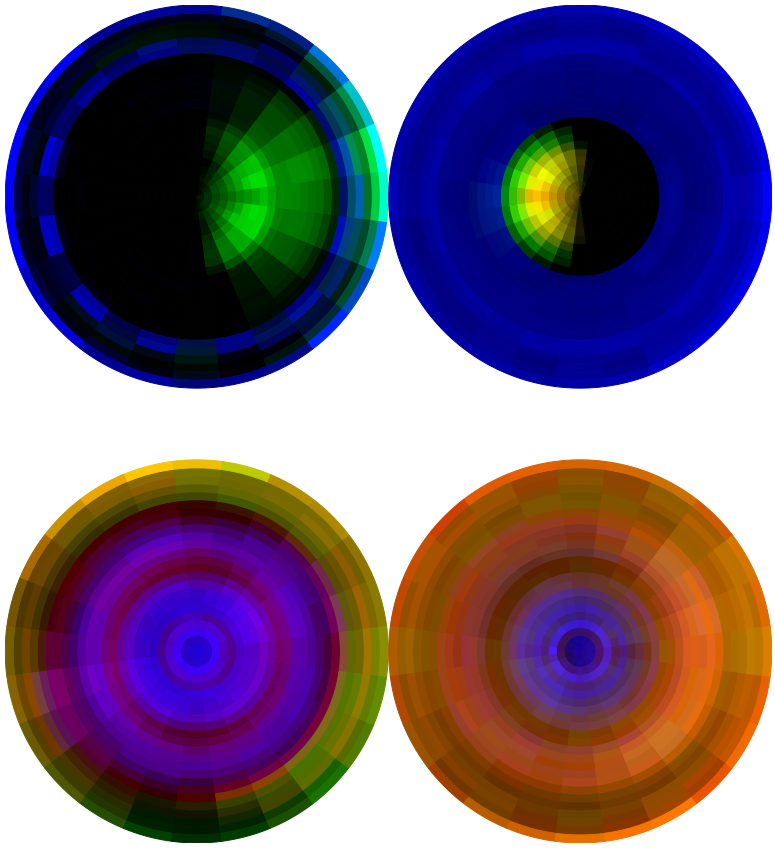

Figure 6.20: Generated images for feature visualisation which maximally activate the output neuron: VBF (top) and ggH (bottom).

These images show a clear difference in each class, and some interesting physical features. However, it should be noted that these will be somewhat unphysical as the optimisation will try to pack as much evidence for the target class in to the image as possible with no constraints on what is sensible. For example, if one takes a network such as inception and optimises for the dog class it will produce an image of a carpet of deformed dogs as the optimisation tries to include as many dog features as possible.

The VBF image shows clear signs of colour connection where the jet constituents are pulled in opposite $\Delta\eta$ directions. This causes asymmetry in the jet image where there are more non-zero pixels on one half and the leading and subleading images have this asymmetry in opposite directions. They are also more collimated with more p_T deposition in the jet core as expected for quark jets. This can be observed as more colour at the centre of the image and more empty pixels around the outside. The dominance of neural p_T deposition (green) and discrete rings of multiplicity (blue) indicate that the model is forming features that detect jets in the forward regions by detector coarseness and lack of charge.

In the ggH image a more circular structure can be seen with a larger proportion of charged deposition over a larger area. This is in line with the expectation that gluon jets are less collimated and higher multiplicity. The rings line up with the gaps in the coarse forward structure, this suggests that the model is learning to detect whether the jets are in the tracker acceptance.

Sets of images generated to show feature construction in different parts of the convolutional section can be found in Appendix C.

Pseudorapidity Inference

The structure of the feature visualisation images suggest that the model has learned to reconstruct a coarse estimate of the pseudorapidity properties of the jet: whether the jets are in the coarse forward regions or not. To test this hypothesis the second step of the training is run again with the leading and subleading jet pseudorapidities included in the engineered features. If there is no change in performance this would suggest that the model already has access to this information via the image input. This training shows no performance increase with the AUROCs of the two models being almost identical. This leads us to conclude that the network does indeed reconstruct this information or the information is not useful.

However, training a single BDT model with the pseudorapidities shows a small performance increase in discrimination power. This suggests that only a small portion of the DCNN performance gain comes from inferring the pseudorapidity.

Maximally-Activating Images

To get a better idea of what sort of physically sensible features are favoured by the network, and whether the conclusions about the feature visualisation are true, real event images are inspected that are maximally activating. A selection is applied on the leading jet so that it is in the tracker acceptance and the charged-p_T channel is present. The most activating image without this requirement shows two very forward jets with coarse structure. The two maximally activating images for the VBF and ggH classes are shown in Figure 6.21.

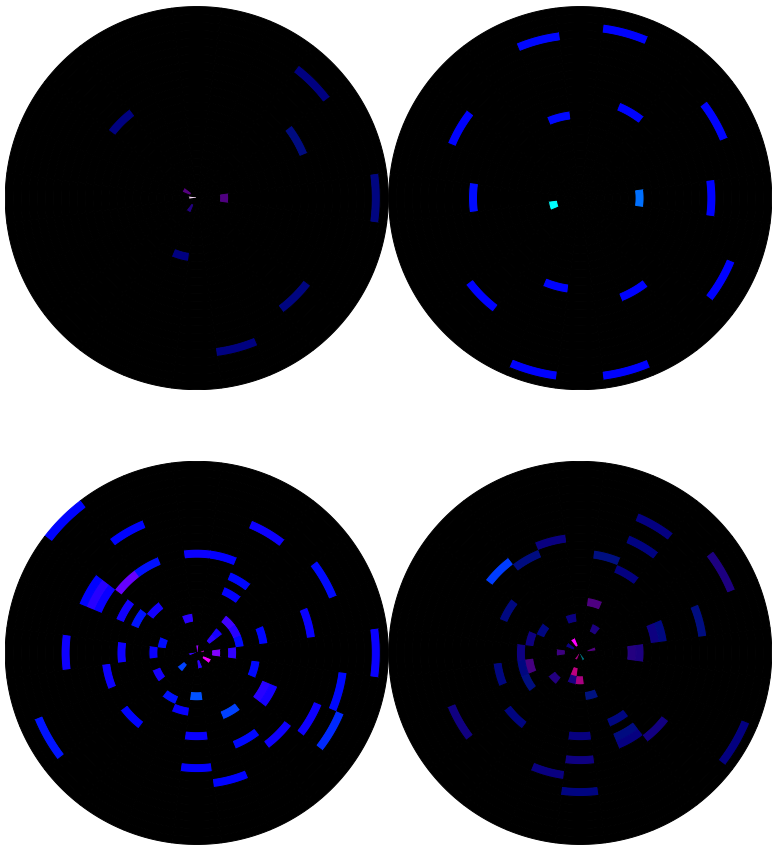

Figure 6.21: Real images that maximally activate the class logits. The top image is the maximally-activating VBF image, and the bottom is the maximally-activating ggH image.

The max-activating images for ggH and VBF resemble the feature visualisations with SGD. The VBF image is collimated, with a spread in one $\Delta\eta$ direction in the leading jet and the opposite in the subleading. This effect is easier to see in the leading jet, but is also present in the subleading image where there is one bright pixel to one

side. The effect of the coarse forward structure of CMS on the subleading jet image can also be observed. The ggH is far higher in multiplicity, much less collimated and more circular. This is in line with the image from feature visualisation.

Front Filters and Low-level Features

The lowest-level features that the network constructs are at the spread layer. This layer can be interpreted as forming local arrangements of pixels into more spread-out non-sparse feature maps. These convolution filters are shown in Figure 6.22.

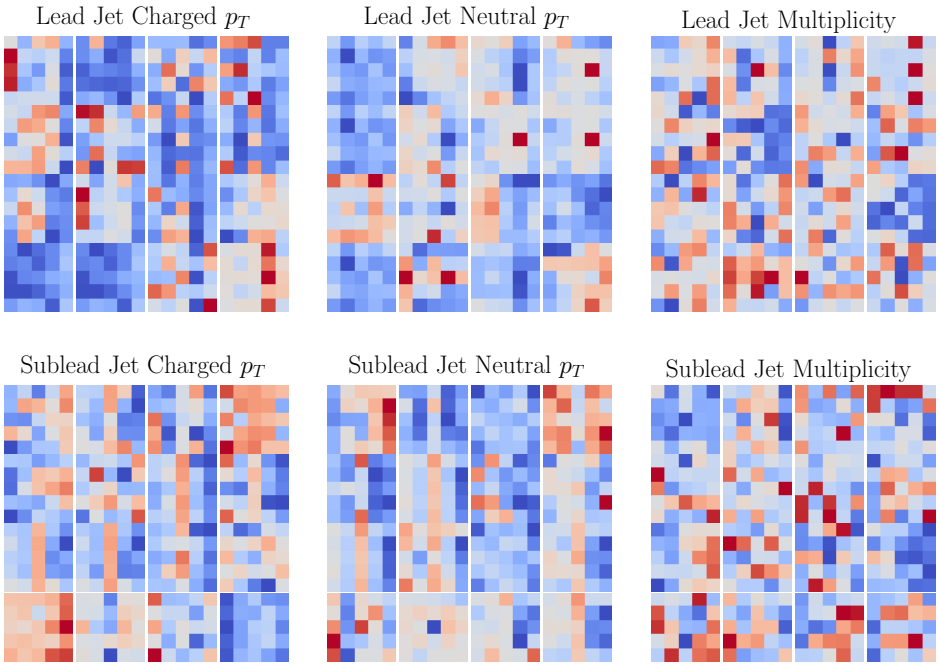

Figure 6.22: Front filters of the network grouped by the six image channels. Positive weight values are darker red and more negative values are darker blue, weights close to zero are shown as white. The vertical direction in the filters corresponds to φ and the horizontal direction corresponds to ΔR.

The front filters have a coarse and noisy appearance. This would be a sign of inappropriate training conditions in a CNN with normal images: either the learning rate is too high or the model requires regularisation. In this case the noisiness could be intrinsic to the dijet image problem because the images are sparse and are less smooth. Training without event weights, with large batches and a lower learning rate produced smoother filters and better generalisation performance.

The filters themselves detect radial and angular bands of pixels, this corresponds to horizontal and vertical stripes in the filter. There are also filters that detect separated

stripes of pixels these are two positive (red) stripes with a negative stripe (blue) in between. This behaves like an edge detection filter. Finally, there are filters where there is a single large-value weight offset from the centre of the filter. This appears to perform a small angular or radial shift in the jet image. Examples are shown in Figure 6.23.

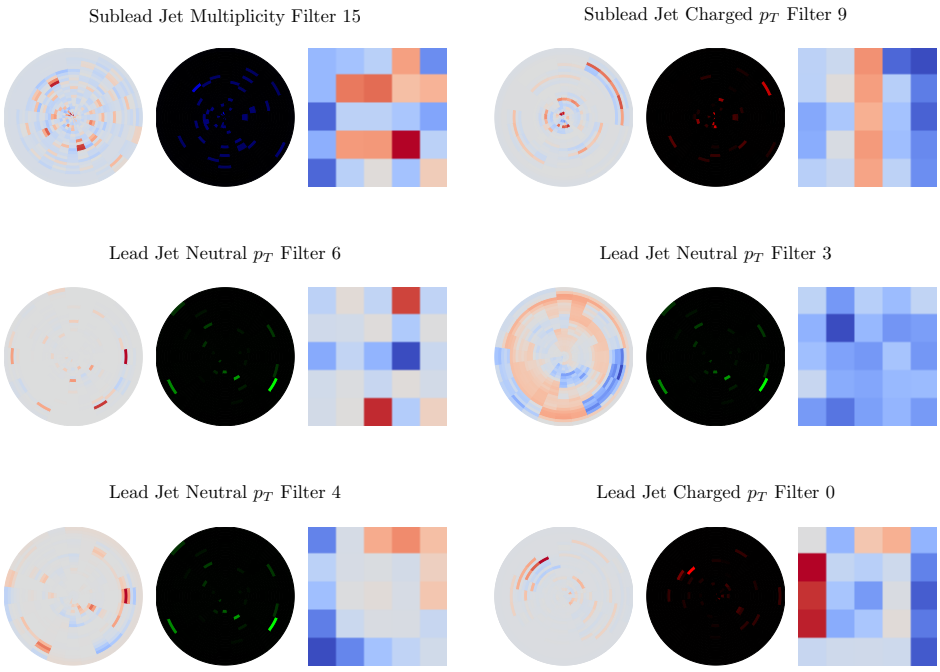

Figure 6.23: The effect of selected filters. Each subplot shows the effect of the filter and subsequent neural activation (left), the original preprocessed image (centre) and the filter (right). Clockwise starting from top left the filters are: radial gap detector, angular band smearer, general smearer, angular smear with gap in front, radial and angular smearer with shifts, double shifter.

Conclusion

It has been shown that the network is capable of constructing sophisticated and physically relevant features of dijet substructure from the input images. It is learning to infer information about jet pseudorapidity, but this accounts for only a small portion of the performance increase. Furthermore, these studies show that feature visualisation and max-activation images are powerful techniques. These can be applied to any part of the network to extract information about its behaviour.

6.7.6 Validation

The DCNN-based model may be particularly vulnerable to disagreement between simulation and data. The underlying QCD processes of the underlying event, parton showering, and hadronization are challenging to model and this may adversely affect the application of this model to real data. To determine whether this is the case the model is validated in the $Z \to e^+e^-$ control region and with QCD modelling variations described previously with special attention given to how the performance of the image-only model changes. The $Z \to e^+e^-$ data/simulation disagreement will be investigated with a specially trained network.

$Z \to e^+e^-$ Control Region

First the simulation/data agreement of the network score is evaluated for both the image-only and the full DCNNs (Figure 6.24).

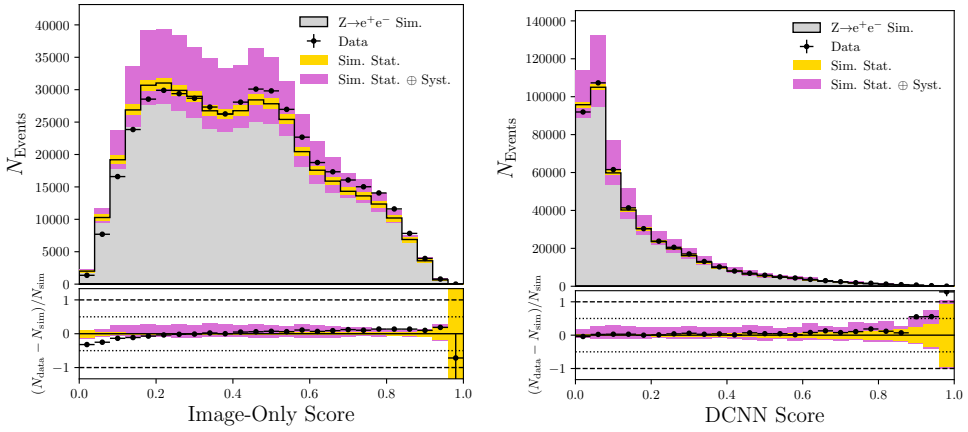

Figure 6.24: $Z \to e^+e^-$ validation plots for the output scores of the image-only network (left) and the full network (right).

Generally, the agreement is good with a very slight shift in the image-only model. This suggests that the real data is slightly more VBF-like. The systematics bands are slightly larger in the image-only model and at the lower end of the full DCNN score compared to the combined BDT. This difference may not be significant as it is located in a score region that is rejected by the DCNN VBF tag.

To determine how the distribution of images differs between simulation and data, an images-only DCNN is trained to discriminate between them. This network will have the same structure, regularisation, loss and training process as step one of the VBF DCNN model. Once trained the performance of this network is evaluated and the features it has built are analysed to learn about the image distribution disagreement.

The performance of the data/simulation discriminant is shown in Figure 6.25.

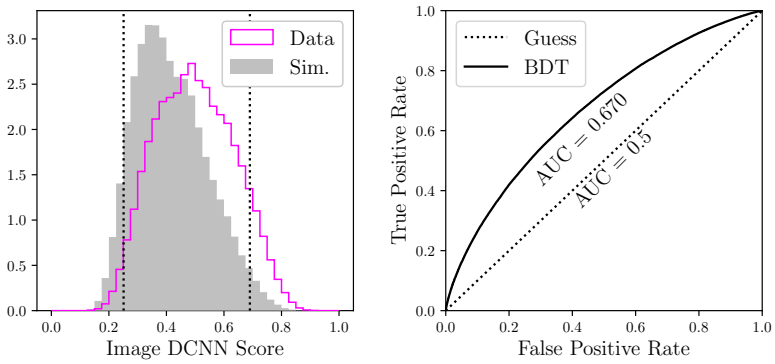

Figure 6.25: $Z \to e^+e^-$ Simulation/data discriminant performance. The score distribution for the simulation and data classes is shown on the left, and the associated ROC curve is shown on the right.

The AUROC suggests that there is indeed detectable disagreement in the image distributions. However, the network score validation's level of agreement suggests that although there is disagreement the network is fairly robust to it.

To extract the areas of disagreement between the image datasets from the network the same interpretation techniques described previously are used. First feature visualisation is used to produce images that contain a collection of discriminating features (Figure 6.26).

The difference is most pronounced in the leading jet where there is more charged p_T deposition in simulation and a different structure. The simulation is more rounded and ggH-like, the data has radial bands of neutral p_T along the $\Delta\eta$, $\Delta\phi$ axes and appears more VBF-like. The subleading jet appears to be similar in shape between the two classes but with more charged p_T in the simulation and a more collimated jet in data. Overall the data class visualisation has more VBF-like qualities and the simulation is more ggH-like.

To verify the feature visualisation, maximally activating images are inspected for the most simulation-like and most data-like events according to the model (Figure 6.27.

These selected samples have a class purity of approximately 85%, and they resemble the generated images. The image feature differences extracted with feature visualisation appear to be physically sensible. The difference in charged-p_T deposition is especially pronounced in these images, but a few other features are apparent. In the leading jet data-like sample there is a spread along the $\Delta\phi$ axis that is much less pronounced in the simulation-like sample. The smooth appearance of the neutral-p_T channel in the data-like subleading jet also suggests that there is more neutral

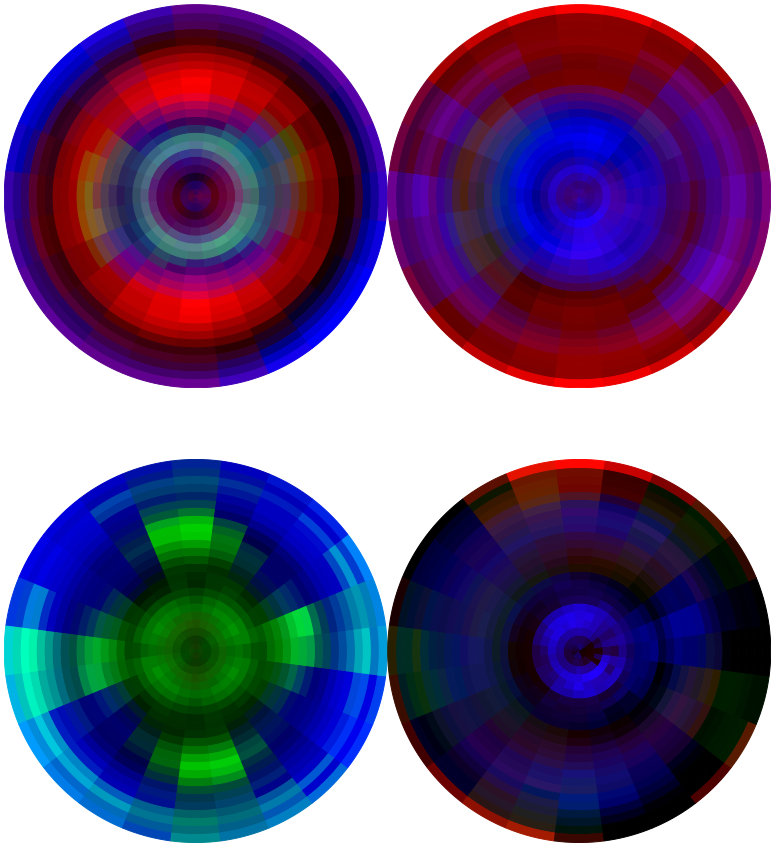

Figure 6.26: Sim/Data discriminant feature visualisation. The top dijet image is optimised for the simulation output neuron, and the bottom is optimised for the data neuron.

deposition within the tracker acceptance.

QCD Modelling Variations

The model is evaluated over samples that explicitly differ in their modelling of QCD with the same procedure as the BDT-based model. These consist of up and down variations that should encompass the behaviour of real data. The resulting score distribution variation and the variations in the ROC curves for both image-only and the full model are shown in Figure 6.28.

Remarkably, the image-only model is not only robust to these variations but actually has a higher AUROC and therefore better performance on the QCD variation samples. This robustness suggests that the learned features are physically well-

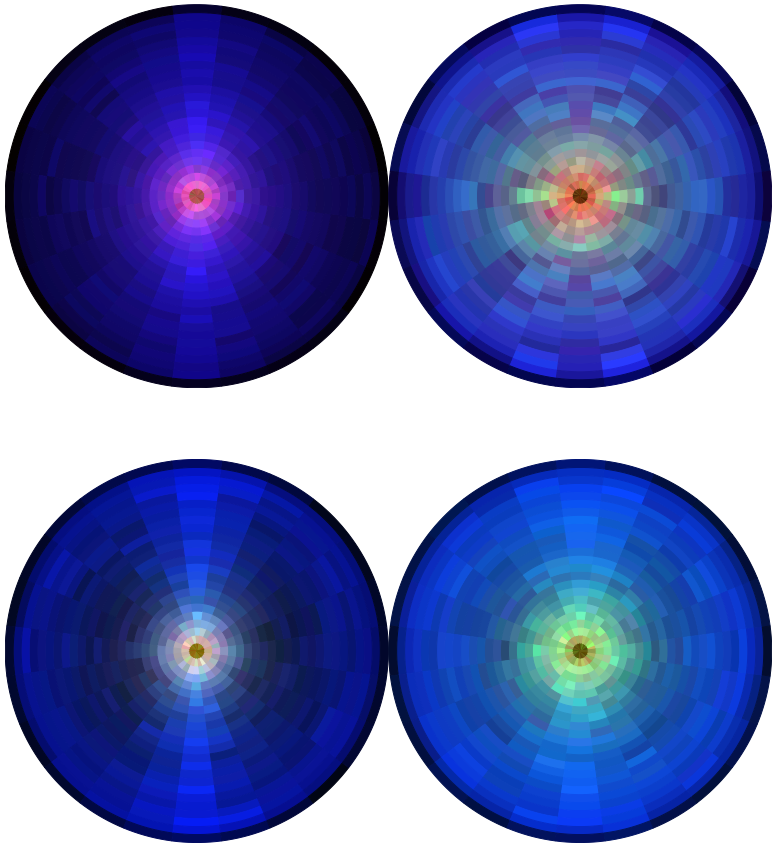

Figure 6.27: Mean images for top and bottom 5% maximally activating events. Score selection for simulation-like is shown at the top, selection for data-like is at the bottom.

motivated as the features found in data should be within these variations. The full model shows a small degradation in discrimination power, but at a level comparable to the combined BDT.

6.7.7 Conclusions

Dijet images have been demonstrated to encode useful discriminating features and that a dense CNN is capable of extracting them. These features are found to be robust to QCD modelling variations, and the overall response of the model is similar between simulation and data.

For model development Bayesian optimisation is a powerful technique for navigating possible hyperparameter choices. The final model is a powerful discriminant, but

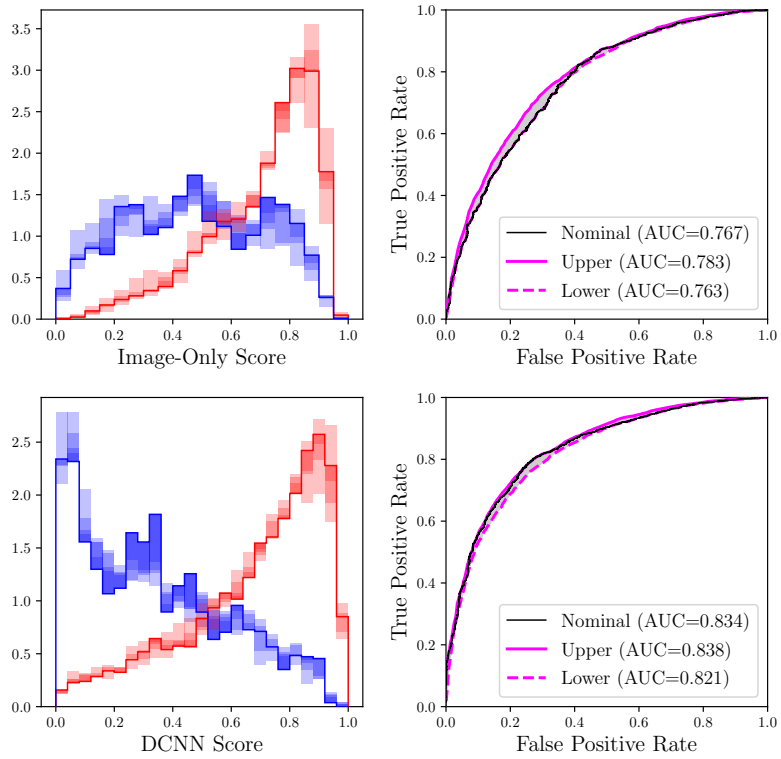

Figure 6.28: Variation in the score distributions and ROC curves with parton shower and underlying event variation. The top pair corresponds to the image-only model (step 1 of training) and the bottom plots correspond to the full model used in the tag.

compromises must be made due to the relationship between VBF/ggH discrimination and overall signal significance. To approach this tradeoff correctly and to tune for the right choice the notion of cost sensitivity needs to be introduced during training. This is the correct way to set the priority between these two aspects of VBF tag model performance.

The main impact on tag performance is reducing ggH contamination. This sort of technique will be more useful for targeting high-purity samples of particular processes rather than enhanced signal significance.

Finally, the common criticism that CNNs are black boxes is shown not to be the case. There are many techniques available to interpret them and only a few are used here. Feature visualisation and max-activation images can be used to look inside the network at single neurons, feature maps or even entire feature volumes to determine what they are looking for. Furthermore the above techniques can be used to analyse data/simulation compatibility and isolate features from regions of disagreement.

Chapter 7

Statistical Analysis and Results

7.1 Introduction

The set of categorised diphoton candidates are subjected to a statistical interpretation to measure the Higgs boson signal in data and to determine how its properties compare to SM expectations. This procedure consists of a statistical model that is fitted simultaneously to the $m_{\gamma\gamma}$ distribution in data for each tag category. Separate models are constructed and then combined for signal and background that take into account the contribution of different categories, production modes and systematic uncertainties.

This chapter begins by describing the construction of these models and the inclusion of the associated systematic uncertainties (following [8]). After this, the final fit of the model and the statistical interpretation of the result is described with emphasis on how the two different VBF tagging approaches affect the result.

7.2 Statistical Models

7.2.1 Signal Modelling

The signal modelling aims to construct a signal shape for each category using simulated data samples of the different Higgs production modes. This is achieved by constructing many parameterised signal shapes: one for each choice of production mode, category and whether the vertex has been correctly identified.

Furthermore, the Higgs boson mass m_H is not assumed to have a specific value. A parameterisation of the signal shape in terms of m_H is derived from a collection of

samples with different m_H. This is constructed by performing a simultaneous fit over all of the mass points where the parameters of the signal shape are all polynomials of m_H. The floating parameters of this fit are then the coefficients of the polynomial. This approach is used instead of interpolating the parameters as it leads to fewer parameters per fit and guarantees consistency across mass points.

This procedure is also used to determine relative normalisation of the RV and WV scenario shapes when they are combined. This is given by the vertex selection efficiency derived from simulated data. The value is then treated the same way as the other shape parameters in the simultaneous mass point fit.

Once constructed the different signal shapes from production modes are combined together to produce the signal model for each category. The production mode shapes are normalised to their expected signal yield and then summed.

7.2.2 Background Modelling

The background modelling aims to construct a smoothly falling background distribution that takes uncertainty about the choice of functional form into account. This is achieved with a data-driven approach based on the discrete profiling method [83]. This technique allows for the estimation of this systematic and its propagation to the final fits as a discrete nuisance parameter without assumptions about underlying processes or functional forms.

The candidate functions are expressed as a set of function families, each of these are expressed as a sequence where we pick a lowest and highest order form to consider. The maximum order is found via an F-test and the minimum order is found by requiring a minimal level of goodness-of-fit. The families considered are polynomials in the Bernstein basis, Laurent series, sums of exponentials, and sums of power laws.

The functions are fitted to the $m_{\gamma\gamma}$ sideband data using twice negative likelihood (2NLL) minimisation with an additional regularisation term n_p (number of parameters) to penalise function complexity.

7.3 Systematic Uncertainties

Systematic uncertainties are propagated to the final fits by including nuisance parameters primarily in the signal model. The desription here is taken from the work in [8] unless stated otherwise.

Systematic error in the signal model is handled in one of three main ways via nuisance parameters that have different effects on the $m_{\gamma\gamma}$ distribution:

- Shape uncertainties are propagated via nuisance parameters that alter the shape of the Gaussian signal peak position, width and normalisation.

- Yield uncertainties are handled by nuisance parameters that scale the $m_{\gamma\gamma}$ distribution and are subject to a log-normal constraint during the final fit.
- Categorisation uncertainties are implemented as category migration nuisance parameters which behave in a similar way to yield uncertainties but will reduce the yield in other categories simultaneously as the category of interest is increased.

There is an exception to this scheme with the vertex uncertainty. Here there is a nuisance parameter that controls the relative fraction of the RV/WV signal distributions rather than affecting the shape parameters directly.

An extra systematic uncertainty from choice of background functional form is propagated to the final fits via the background model as a discrete nuisance parameter.

7.3.1 Theoretical Uncertainties

Uncertainties from QCD theory calculations have their effects modelled in two ways: uncertainty on the overall yield for a process and uncertainty associated with analysis category migration. The migration uncertainties are extracted separately by scaling such that the overall yield is unchanged.

- **QCD scale uncertainty**: yield uncertainty is parameterised separately for renormalisation scale and factorisation scale. Category migrations associated with varying these parameters independently and together are also included.
- **PDF uncertainties**: There is an uncertainty associated with variation in the signal yield of each production process, and category migration uncertainties. The yield uncertainty is calculated using the procedure from PDF4LHC [84] and the migration uncertainties are computed from the NNPDF3.0 PDF set [85] and the MC2Hessian [86] procedure.
- **Strong force coupling (α_s) uncertainty**: evaluated with the same procedure as the PDF uncertainties.
- **Underlying event and parton shower uncertainty**: evaluated using simulated data where the modelling of the underlying event and the parton showers have been varied. This manifests primarily as variation in the jets of the analysis and is modelled as a category migration uncertainty. The probability that events move within the categories of the BDT-based VBF tag, or from this tag to Untagged are found to be 7 and 9%. This was re-evaluated for the DCNN-based VBF tag and found to be unchanged.
- **Gluon fusion contamination in $t\bar{t}H$ tag categories**: when the Higgs boson is produced by ggH it can be produced in association with a number of jets. As the number of jets becomes large the accuracy of theoretical predictions

becomes worse and introduces a systematic uncertainty in ggH contamination of the jet-based tag categories. This manifests in the $t\bar{t}H$ tags in multiple ways:

- **Uncertainty due to limited ggH sample size**: only a small quantity of simulated ggH events are accepted into the $t\bar{t}H$ tag. This introduces a significant statistical uncertainty on the ggH yield and contributes a 10% uncertainty.
- **Uncertainty due to modelling parton showers**: this is estimated by comparing simulation and data for events whose production is dominated by gluon-fusion-type diagrams ($t\bar{t}$+jets with semi-leptonic $t\bar{t}$ decays) binned by the number of jets. The largest discrepancy is in $N_{\text{jets}} \geq 5$ which corresponds to an uncertainty of 35%.
- **Uncertainty due to modelling gluon splitting**: estimated by calculating the difference in the ratio $\sigma(t\bar{t}b\bar{b})/\sigma(t\bar{t}jj)$ for simulation and data. The fraction of events in simulated ggH with b jets are then scaled by this difference. This gives a 50% variation in the ggH yield for the $t\bar{t}H$ tags.

- **Gluon fusion contamination in tag categories with jets and a high-p_T Higgs boson**: how ggH mismodelling manifests in the VBF categories, in particular:

  - **Uncertainty due to jet multiplicity mismodelling**: two nuisance parameters due to missing higher-order terms and two more nuisance parameters for category migration due to variation in jet multiplicity based on STWZ [87] and BLPTW [87–89] predictions.
  - **Uncertainty due to Higgs boson p_T mismodelling**: two nuisance parameters associated with migration between two bins in Higgs boson p_T; from 60 to 120 GeV and above 120 GeV. There is also a third nuisance due to uncertainty in top quark mass effects. This is negligible for p_T less than 150 GeV but increases to 35% at 500 GeV. These impact the highest-score VBF and Untagged categories where the ggH yield uncertainty is 6-8%.
  - **Uncertainty in the acceptance of ggH in VBF categories due to QCD effects**: effects from missing higher-order terms are estimated via variations in the renormalisation and factorisation scales in MCFM 5.8 [90]. Two nuisance parameters are introduced associated with the overall normalisation of Higgs bosons produced in association with two jets, and three or more jets. This allows for the impact of jet suppression arising from how the kinematic variables are used to form the VBF scores to be propagated to the analysis. The procedure is based on the Stewart-Tackmann method [91, 92] and the impact on the ggH yield in the VBF categories is 8-13%.

- **Uncertainty in the H $\rightarrow \gamma\gamma$ branching fraction**: the uncertainty on the theoretical prediction of the H $\rightarrow \gamma\gamma$ branching fraction is taken from [93] and is approximately 2%.

The theory uncertainties with the largest impact on measuring signal strengths and couplings are the H $\rightarrow \gamma\gamma$ branching fraction, and the renormalisation and factorisation scale uncertainties from the QCD scale.

7.3.2 Experimental Uncertainties

Uncertainty in the measurement and construction of physics objects at CMS give rises to associated experimental systematic uncertainties. These affect the shape of the signal distribution and are propagated through to the final fits via the signal model.

Photon Energy Measurement Uncertainties

Uncertainties in the photon energy measurement are a particularly important contribution and can affect the signal shape via both the photon energy scale and resolution.

- **Energy scale and resolution**: uncertainties associated with the photon energy scale and resolution corrections are estimated with events from the Z $\rightarrow$ e^+e^- control region reconstructed as photons. Data and simulation are compared in eight bins of R_9 and $|\eta|$ (high/low-R_9, and four $|\eta|$ bins). The uncertainties are quantified separately in four photon classes (EB/EE and high/low-R_9) and are propagated to the categorisation with four scale nuisance parameters (one per photon class) and eight shape nuisance parameters (one constant and one stochastic term per photon class) for each event category. This has a 0.15-0.5% effect on the photon energy, and an effect of 2.5% on the signal strength modifier.
- **Nonlinearity of photon energy**: There is potential residual data-simulation difference in the linearity of the ECAL response with photon energy scale. This is estimated by studying boosted Z $\rightarrow$ e^+e^- events and has the effect of shifting the peak position by a small amount per category, constituting a maximum uncertainty of 0.1% in each category except for the Untagged where it is 0.2%. The uncertainty is propagated to the final fits as a shape nuisance parameter that shifts the signal peak position.
- **Non-uniformity of light collection**: There is a systematic uncertainty associated with the modelling of the fraction of scintillation light reaching the ECAL crystal photodetector as a function of the longitudinal depth of the shower start. The size of this effect is 0.07% on the photon energy scale.

- **Electromagnetic shower modelling**: Mismodelling of electromagnetic showers in `GEANT4` introduces a small difference between electrons and photons and therefore another small uncertainty. It is estimated by comparing the latest version of shower simulation with a previous one and treating the small difference between them as representing the limit of accurate modelling. This gives a contribution of 0.05% uncertainty to the photon energy scale.
- **Modelling of the material budget**: The amount of material between the interaction point and the ECAL affects the behaviour of photons and electrons. The modelling of this material is a further source of systematic uncertainty. This uncertainty is estimated using simulated samples where the material has been uniformly varied by ±5% to cover the difference in the estimation between simulation and data. The uncertainty manifests as an effect on the photon energy scale of 0.24%.
- **Shower shape corrections**: Finally, there is an uncertainty due to mismodelling of shower shapes themselves. This is estimated by comparing between simulation samples with and without corrections on shower shape variables. This gives an uncertainty in the photon energy scale of 0.01-0.15% at maximum. This is propagated by separate signal shape nuisance parameters for each photon category.

Additional Experimental Uncertainties

Additional uncertainties that are not directly from the photon measurement arise from estimations of efficiencies, scale factors and selection variables. These are varied and their estimated effects propagated through the analysis chain. They are then applied as per-category yield and category migration nuisance parameters in the final fits.

- **Trigger efficiency**: uncertainty in the trigger efficiency estimation is evaluated with the $Z \to e^+e^-$ control region and the tag-and-probe technique. This leads to an impact on the event yields of 0.1% at maximum.
- **Photon preselection**: the uncertainty of the photon preselection efficiency is estimated as the ratio between the efficiency measured in simulation and data. This has an impact on event yields of 0.2-0.5% depending on category.
- **Photon ID BDT score**: Photon energy measurement uncertainties are propagated through the categorisation process to estimate their effect on category signal yields via the photon ID. The uncertainty is assigned to cover the observed difference between data and simulation in the $Z \to e^+e^-$ control region.
- **Photon energy resolution estimation**: This uncertainty is estimated by rescaling the energy resolution estimate around its nominal value by ±5% to

cover all disagreement between data and simulation. This variation is propagated through the categorisation and is implemented as a yield nuisance parameter.

- **Jet energy scale and smearing corrections**: Uncertainties in the correction of jet energy measurements are propagated through the event categorisation and are modelled as category migration nuisance parameters. These nuisance parameters correspond to migration within VBF categories, within VH categories, within $t\bar{t}H$ categories and from these tags to the Untagged categories. Jet energy scale corrections correspond to the following migrations:
  - 8-11% between VBF categories and 11% from VBF to Untagged;
  - 15% from VH to Untagged;
  - 5% from $t\bar{t}H$ to Untagged.

  The jet energy resolution has a migration effect of at most 3% across all tags except for VH where it can reach 20%.
- **Missing transverse energy**: uncertainty in the measurement of E_T^{miss} is estimated by varying the p_T of reconstructed physics objects entering the calculation of E_T^{miss} for the event. These variations correspond to the momentum scale and resolution uncertainties of each type of physics object. This corresponds to a 10-15% migration between Untagged and VH MET and is propagated to the final fits as a category migration nuisance parameter.
- **Pileup jet ID**: uncertainty associated with the PUJID in the VBF tags is analysed using $Z \to e^+e^-$ events with one jet whose momentum balances the dielectron in the transverse plane. Data and simulation are compared and the disagreement is used to estimate VBF category migrations. This effect is found to be at most 1% and is propagated to the final fits as a category migration nuisance parameter.
- **Lepton isolation and ID**: the associated uncertainty is estimated for both electrons and muons by measuring the difference in efficiency between simulation and data and varying an associated scale factor within this difference. This is measured using the tag-and-probe technique on both $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$. The associated variations manifest as yield uncertainties and are at most 1% for the $t\bar{t}H$ Leptonic category, 1.5% for the WH Leptonic category and 3% for the ZH Leptonic category. They are propagated to the final fit as category yield nuisance parameters.
- **Efficiency of b-tagging**: this uncertainty is evaluated by comparing data and simulation distributions of the b-tagging discriminant score. The uncertainty has a statistical component associated with the estimation of the relative amount of light and heavy quark initiated jets and confusion between them. This uncertainty is propagated in two different ways due to the difference in

approach of the $t\bar{t}H$ Hadronic and $t\bar{t}H$ Leptonic tags:

- the hadronic category uses a BDT that receives the b-tagger discriminant score as an input feature. The associated yield uncertainty is evaluated by altering the shape of the score in simulation and found to be at most 5%.
- the leptonic category uses a fixed selection on the b-tagger discriminant score. This uncertainty is evaluated by varying the efficiency in data and simulation within their uncertainties and is found to be 2%.

- **Vertex finding efficiency**: this uncertainty derives from mismodelling of the underlying event and disagreement between data and simulation from evaluating $Z \to \mu^+\mu^-$ events. The size of this uncertainty is around 2% and is propagated to the final fit as a nuisance parameter that changes the relative contribution of the RV and WV signal shapes.
- **Integrated luminosity**: this uncertainty is taken from [94] and modelled as a yield nuisance that affects all processes uniformly. The size of this effect is 2.5%.
- **Background modelling**: handled by the discrete profiling method and propagated to the final fits as a discrete nuisance that picks different functional forms.

The experimental systematic uncertainties with the largest impact on signal strength and couplings measurements are from the photon shower shape corrections which affects the photon ID and the photon energy resolution estimate, the photon energy scale and smearing, the jet energy scale and the integrated luminosity.

7.4 Results

Different measurements are extracted by performing a series of fits to the $m_{\gamma\gamma}$ distributions simultaneously across all event categories under different constraints. The fits are carried out in the range $100 < m_{\gamma\gamma} < 180\,\text{GeV}$ with a binning of 250 MeV, using a binned maximum-likelihood fit. The likelihood function to be used is the product of the individual category likelihoods and has the following form:

$$\mathcal{L}(\mu_c, m_H, \vec{n}|m_{\gamma\gamma}) = \prod_{c=0}^{N-1} \left(\mu_c S_c(m_H, \vec{n}_s|m_{\gamma\gamma}) + B_c(\vec{n}_s|m_{\gamma\gamma})\right), \tag{7.1}$$

where c enumerates the N-many event categories; S_c is the signal model of category c; B_c is the background model of category c; m_H is the Higgs signal peak position; $\vec{n}$ are the nuisance parameters of the model with $\vec{n}_s$ and $\vec{n}_b$ being the nuisance parameters of the signal and background models respectively; and μ_c is the category signal strength. The signal strengths may be constrained to be the same depending on the

measurement being performed: a global μ fit constrains them all to the same value that then floats in the fit, per-process μ fits will allow for different values between the production modes, but categories within them will use the same value.

This is maximised by finding the minimum twice negative log-likelihood (2NLL) of $\mathcal{L}$,

$$2\text{NLL} = -2\ln\mathcal{L}(\mu_c, m_H, \vec{n}|m_{\gamma\gamma}), \tag{7.2}$$

while taking into account constraints on the parameters.

7.4.1 Best Fit of Model to Data

A maximum-likelihood fit is performed with a single global μ and m_H to find their best fit values. These constitute the central values of a measurement of the global μ assuming the SM and the associated uncertainty is calculated via a likelihood scan of an associated test statistic. Example mass fit plots for both BDT and DCNN-based VBF tags are shown in Figure 7.1. A full set of mass fit plots for BDT-based and DCNN-based VBF tags can be found in Appendix D.

The expected category yields by production mode contribution are shown in Table 7.1 for fits with the BDT-based and DCNN-based VBF tags. A reduction in contamination from ggH is observed in the DCNN-based VBF tag categories, particularly in VBF 0 where it has been reduced by around a third from 15.5% to 9.5%. Downstream VH tags are only slightly affected, and the overall effect on the inclusive Untagged categories is small.

The effect of using the DCNN-based VBF tag over the BDT-based tag on category significances is shown in Table 7.2. An increase in overall expected significance of 13% is observed in the DCNN-based tag.

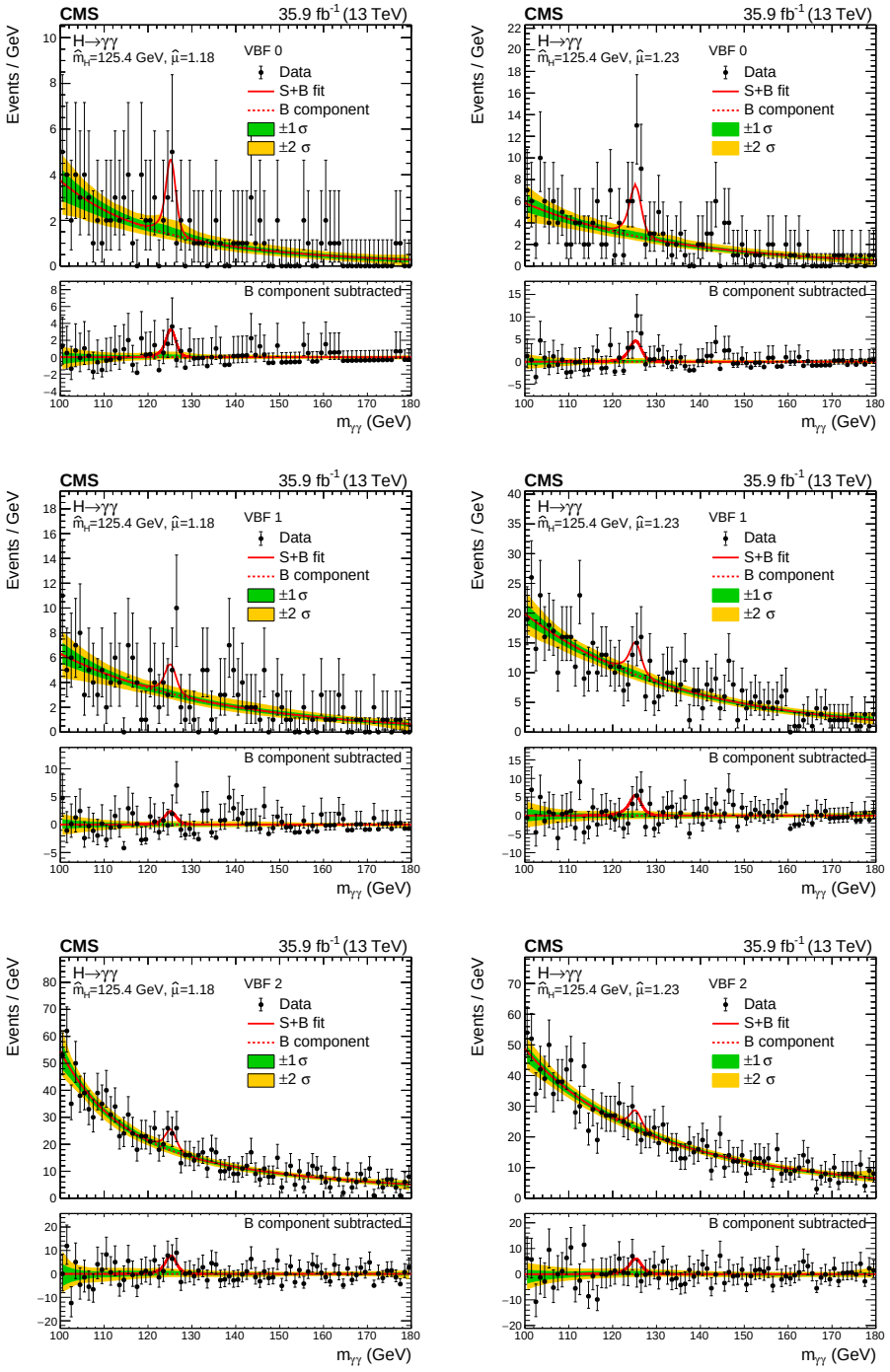

Figure 7.1: VBF category mass fits for the BDT-based VBF tag (left) and the DCNN-based VBF tag (right). Categories are shown in order from the most stringent to least: VBF 0 at the top to VBF 2 at the bottom.

Event categories	Expected SM 125 GeV Higgs boson signal (BDT-based VBF tag)													Bkg
	Total	ggH	VBF	$t\bar{t}H$	bbH	tHq	tHW	WH lep	ZH lep	WH had	ZH had	σ_{eff} (GeV)	σ_{HM} (GeV)	(GeV^{-1})
Untagged 0	32.5	72.0 %	16.6 %	2.6 %	0.6 %	0.7 %	0.3 %	0.6 %	0.3 %	4.2 %	2.2 %	1.32	1.26	21.8
Untagged 1	469.3	86.5 %	7.9 %	0.6 %	1.2 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.9 %	1.1 %	1.46	1.32	925.1
Untagged 2	678.3	89.9 %	5.4 %	0.4 %	1.2 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.4 %	0.8 %	1.93	1.67	2391.7
Untagged 3	624.3	91.3 %	4.4 %	0.5 %	1.0 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.2 %	0.7 %	2.61	2.27	4855.1
VBF 0	9.3	15.5 %	83.2 %	0.4 %	0.4 %	0.3 %	<0.05 %	<0.05 %	<0.05 %	0.2 %	<0.05 %	1.52	1.31	1.6
VBF 1	8.0	28.4 %	69.7 %	0.4 %	0.6 %	0.4 %	<0.05 %	0.1 %	<0.05 %	0.3 %	0.1 %	1.66	1.38	3.3
VBF 2	25.2	45.1 %	51.2 %	0.9 %	0.8 %	0.6 %	0.1 %	0.2 %	0.1 %	0.8 %	0.3 %	1.64	1.37	18.9
$t\bar{t}H$ Hadronic	5.6	7.0 %	0.7 %	81.1 %	2.1 %	4.3 %	2.1 %	0.1 %	0.1 %	0.7 %	1.9 %	1.48	1.30	2.4
$t\bar{t}H$ Leptonic	3.8	1.5 %	<0.05 %	87.8 %	0.1 %	4.7 %	3.1 %	1.5 %	1.2 %	<0.05 %	<0.05 %	1.60	1.35	1.5
ZH Leptonic	0.5	<0.05 %	<0.05 %	2.6 %	<0.05 %	<0.05 %	0.1 %	<0.05 %	97.3 %	<0.05 %	<0.05 %	1.65	1.43	0.1
WH Leptonic	3.6	1.3 %	0.6 %	5.2 %	0.2 %	3.0 %	0.7 %	84.5 %	4.3 %	0.1 %	0.1 %	1.64	1.43	2.1
VH LeptonicLoose	2.7	8.1 %	2.7 %	2.4 %	0.6 %	1.8 %	0.1 %	64.4 %	19.1 %	0.6 %	0.2 %	1.67	1.56	3.5
VH Hadronic	7.9	47.6 %	4.5 %	4.4 %	0.4 %	1.7 %	0.3 %	0.2 %	0.5 %	25.2 %	15.1 %	1.38	1.30	7.2
VH MET	4.0	18.7 %	2.6 %	15.4 %	0.4 %	2.1 %	1.2 %	26.8 %	30.4 %	1.4 %	0.9 %	1.56	1.39	3.5
Total	1875.0	86.9 %	7.1 %	1.0 %	1.1 %	0.2 %	<0.05 %	0.8 %	0.4 %	1.6 %	0.9 %	1.96	1.62	8237.8
Expected SM 125 GeV Higgs boson signal (DCNN-based VBF tag and Downstream Tags)														
Untagged 0	33.3	73.5 %	14.7 %	2.9 %	0.6 %	0.7 %	0.3 %	0.6 %	0.3 %	4.2 %	2.2 %	1.26	1.19	21.7
Untagged 1	466.5	87.2 %	7.3 %	0.6 %	1.2 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.8 %	1.1 %	1.46	1.31	910.0
Untagged 2	674.8	90.3 %	5.0 %	0.4 %	1.2 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.4 %	0.8 %	1.92	1.64	2415.6
Untagged 3	620.5	91.6 %	4.1 %	0.5 %	1.0 %	0.1 %	<0.05 %	0.5 %	0.3 %	1.2 %	0.7 %	2.62	2.29	4848.2
VBF 0	14.2	9.5 %	89.7 %	0.2 %	0.3 %	0.2 %	<0.05 %	0.1 %	<0.05 %	0.1 %	<0.05 %	1.70	1.41	3.4
VBF 1	17.2	25.0 %	73.2 %	0.3 %	0.5 %	0.4 %	<0.05 %	0.1 %	<0.05 %	0.3 %	0.1 %	1.78	1.43	10.6
VBF 2	19.6	44.2 %	51.7 %	0.7 %	0.9 %	0.6 %	<0.05 %	0.2 %	<0.05 %	1.2 %	0.5 %	1.78	1.40	23.0
VH Hadronic	7.9	47.3 %	4.5 %	4.8 %	0.4 %	1.7 %	0.3 %	0.3 %	0.5 %	25.2 %	14.9 %	1.46	1.38	7.2
VH MET	3.9	19.0 %	3.0 %	13.4 %	0.5 %	2.2 %	1.2 %	27.3 %	30.9 %	1.5 %	1.0 %	1.61	1.46	3.4
Total	1874.3	86.9 %	7.1 %	1.0 %	1.1 %	0.1 %	<0.05 %	0.8 %	0.4 %	1.6 %	0.9 %	1.96	1.61	8252.9

Table 7.1: Expected signal yields per category for the BDT-based VBF tag (top) and the DCNN-based VBF tag (bottom). Only the downstream tags are shown for the DCNN-based tag as the others are unaffected. The width values σ_{eff} and σ_{HM} correspond to the smallest interval containing 68.3% of the $m_{\gamma\gamma}$ distribution and the width at half maximum of the signal peak respectively.

	$S/\sqrt{S+B}$			
VBF tag	VBF 0	VBF 1	VBF 2	Total
BDT-based	2.02	1.25	1.35	2.73
DCNN-based	2.44	1.65	0.97	3.10

Table 7.2: VBF tag category expected signal significances comparing the BDT-based VBF tag to the DCNN-based VBF tag.

The final combined mass plots for both unweighted and weighted by sensitivity are shown in Figure 7.2. The change to a DCNN-based VBF tag does not have a significant effect on these plots.

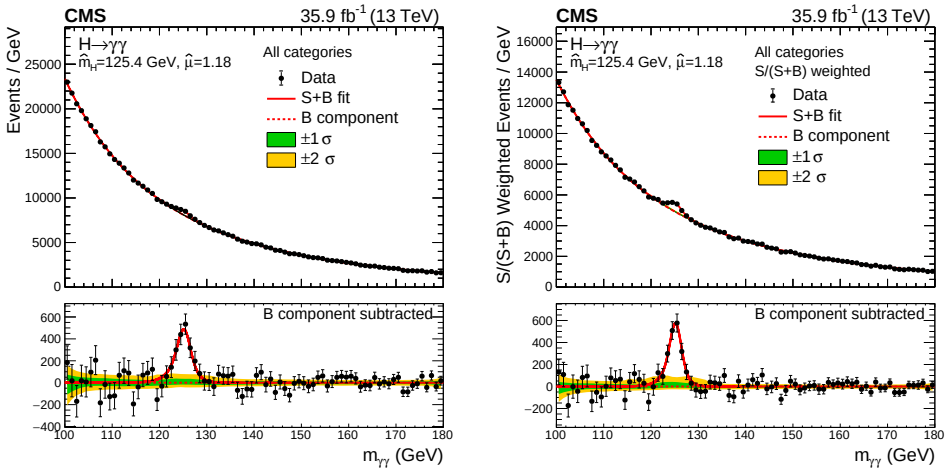

Figure 7.2: Diphoton mass distribution plots for all categories combined using the BDT-based VBF tag. The unweighted combined distribution is shown on the left, and the sensitivity weighted combination is shown on the right.

7.4.2 Signal Strength Likelihood Scans

The test statistic used is the twice-negative delta log-likelihood (2ΔNLL),

$$2\Delta\text{NLL} = -2\ln\mathcal{L}(\mu, \hat{m}_{H,\mu}, \vec{n}_\mu | m_{\gamma\gamma}) + 2\ln\mathcal{L}(\hat{\mu}, \hat{m}_H, \hat{n} | m_{\gamma\gamma}), \tag{7.3}$$

where $\hat{\mu}$, $\hat{m}_H$ and $\hat{n}$ are the best fit values for the signal strength modifier, Higgs mass and nuisance parameters respectively. The parameters μ, $\hat{m}_{H,\mu}$ and $\vec{n}_\mu$ are the global signal strength being profiled in the likelihood scan, the Higgs mass allowed to float for a given value of μ, and the nuisance parameter values also allowed to float.

This procedure is used to measure the global μ, the production mode μ values and the fermionic versus bosonic production μ values.

Global Signal Strength Likelihood Scan

To calculate the uncertainty associated with the measurement of the global signal strength a likelihood scan is performed with a test statistic and profiling in the value of μ. The contribution of statistical uncertainty is determined by performing the likelihood scan with the nuisance parameters associated with systematic uncertainties removed. The systematic contribution is then the difference in quadrature between these values and the total uncertainty from the full fit. The 2ΔNLL values of the global μ likelihood scans are shown in Figure 7.3.

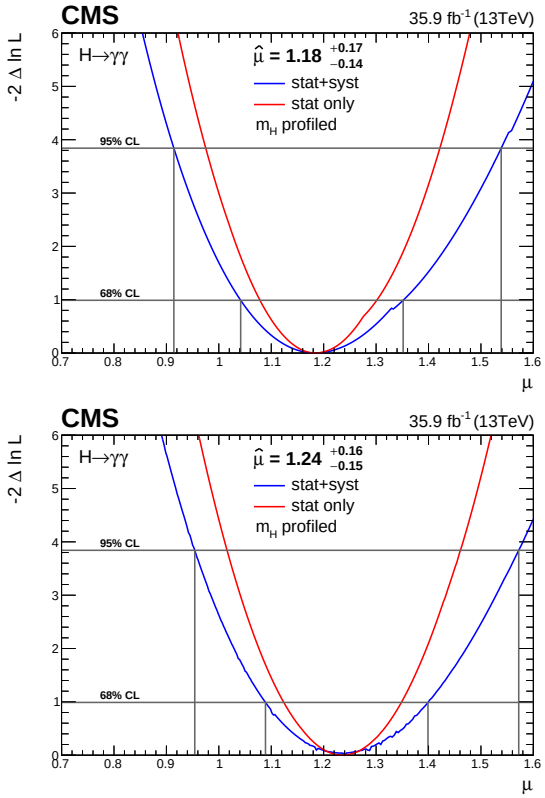

Figure 7.3: Likelihood scan of the global signal strength modifier μ with a 2ΔNLL test statistic for analysis with the BDT-based VBF tag (top) and the DCNN-based VBF tag (bottom).

The measured value for μ and its associated uncertainties in the BDT-based case are found to be $\hat{\mu} = 1.18^{+0.17}_{-0.14} = 1.18^{+0.12}_{-0.11}(\text{stat.})^{+0.09}_{-0.07}(\text{syst.})^{+0.07}_{-0.06}(\text{theo.})$. The best fit value for the Higgs boson mass is found to be $\hat{m}_H = 125.4 \pm 0.3 = 125.4 \pm 0.2(\text{stat.}) \pm 0.2(\text{syst.})$. A precise determination of the systematic effects on the mass value and therefore a precise determination of the mass itself are beyond the scope of this thesis.

The measured value for μ is found to be larger in the DCNN case with similar-

sized uncertainties: $\hat{\mu} = 1.24^{+0.16}_{-0.15} = 1.24^{+0.11}_{-0.11}(\text{stat.})^{+0.08}_{-0.08}(\text{syst.})^{+0.06}_{-0.07}(\text{theo.})$. The best fit value for the Higgs boson mass is unchanged.

Production Mode Signal Strength Modifiers

Likelihood scans specific to each production mode are carried out in a similar way to the global case, but with some differences. Instead of a single global μ and likelihood scan there are four, one for each production mode. For each case the corresponding μ is profiled and the others are allowed to float in the fit. The results of these likelihood scans are shown in Figure 7.4.

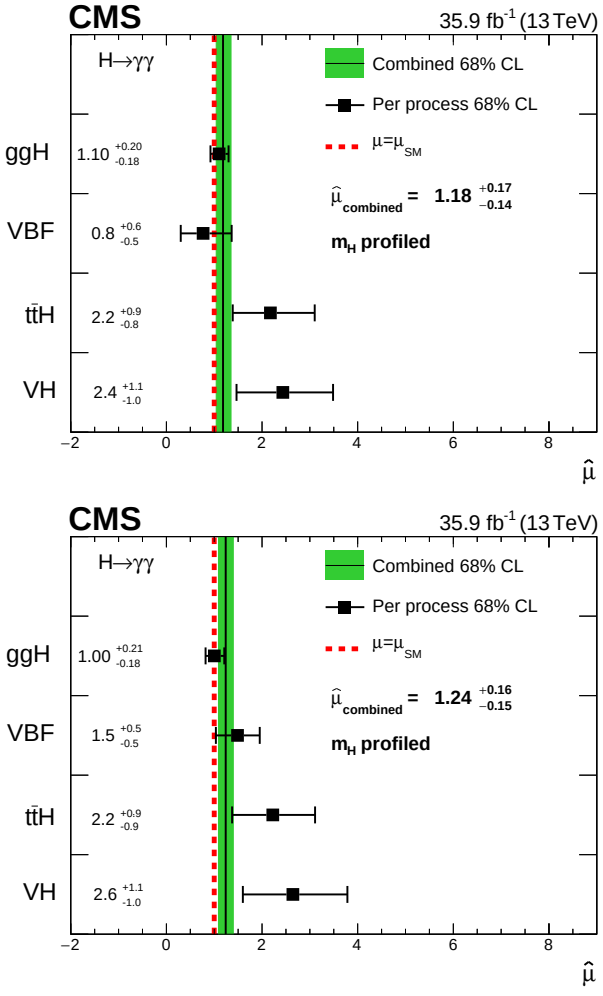

Figure 7.4: Likelihood scan results of the production mode signal strength modifiers μ with a 2ΔNLL test statistic. Analysis with the BDT-based VBF tag is shown at the top and the DCNN-based variant is at the bottom.

The DCNN-based VBF tag leads to a change in the VBF measurement from $\mu_{\text{VBF}} = 0.8^{+0.6}_{-0.5}$ to $\mu_{\text{VBF}} = 1.5^{+0.5}_{-0.5}$. The measured signal strength has increased, and there has been a small reduction in the uncertainty of its measurement. The other production modes are mostly unchanged except for the ggH and VH which correspond to tags downstream from VBF.

The same approach is used to extract the ratio of observed cross sections to the SM expectation as part of the simplified template cross section (STXS) framework Stage 0 [93]. This scheme is aimed at reducing the impact of theory uncertainties due to extrapolation to the full phase space from the fiducial region of the analysis. This imposes a criterion on the Higgs boson rapidity of $y < 2.5$ and splits the VH into separate WH, ZH and VH Hadronic categories. The results of measuring these ratios are shown in Figure 7.5.

The DCNN-based VBF tag has a similar effect in this scheme to the the production mode signal strengths.

Fermionic Versus Bosonic Production

A measurement of the fermionic versus bosonic signal strength is performed with a 2D likelihood scan. The procedure is similar to the above but with a signal strength for the bosonic production modes $\mu_{\text{VBF,VH}}$ and the fermionic modes $\mu_{\text{ggH,t}\bar{\text{t}}\text{H}}$. A best fit point is found and then a $2\Delta\text{NLL}$ test statistic is evaluated over a 2D space corresponding to different values of the two signal strength modifiers. The result with 68% and 95% confidence intervals is shown in Figure 7.6.

The best fit point for the BDT-based case was found to be $\mu_{\text{ggH,t}\bar{\text{t}}\text{H}} = 1.19^{+0.22}_{-0.18}$, $\mu_{\text{VBF,VH}} = 1.21^{+0.58}_{-0.51}$. The best fit point for the DCNN-based case was found to be $\mu_{\text{ggH,t}\bar{\text{t}}\text{H}} = 1.11^{+0.20}_{-0.18}$, $\mu_{\text{VBF,VH}} = 1.65^{+0.50}_{-0.42}$. A significant reduction in the uncertainty of the bosonic production mode μ is observed along with an increase in its value.

7.4.3 Couplings Measurements

Deviation in the Higgs couplings from the SM expectation are modelled within the κ framework as described in [95]. These differences are measured with two 2D likelihood scans: fermionic versus bosonic and photons versus gluons. The κ values not subject to the 2D likelihood scan are fixed at unity. The resulting plots for the BDT-based and DCNN-based VBF tags are shown in Figure 7.7.

With the BDT-based VBF tag the effective coupling to fermions is measured to be $\kappa_F = 1.04^{+0.51}_{-0.25}$ and the effective coupling to bosons to be $\kappa_V = 1.08^{+0.10}_{-0.08}$. The effective coupling to photons is measured to be $\kappa_\gamma = 1.25^{+0.16}_{-0.17}$ and the effective coupling to gluons to be $\kappa_g = 0.81^{+0.17}_{-0.13}$. With the DCNN-based VBF tag the effective coupling to fermions is measured to be $\kappa_F = 0.85^{+0.22}_{-0.17}$ and the effective coupling to

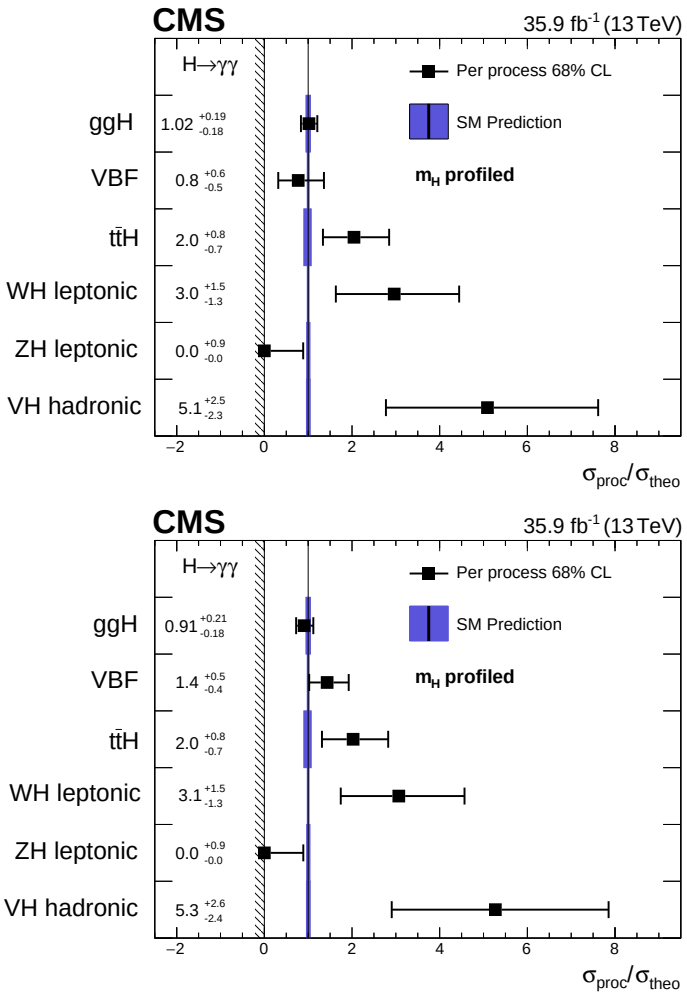

Figure 7.5: SM prediction to measured cross section ratios in the STXS Stage 0 framework. Analysis with the BDT-based VBF tag is shown on top and the DCNN-based variant is on the bottom.

bosons to be $\kappa_V = 1.06^{+0.07}_{-0.07}$. The effective coupling to photons is measured to be $\kappa_\gamma = 1.32^{+0.14}_{-0.13}$ and the effective coupling to gluons to be $\kappa_g = 0.75^{+0.13}_{-0.13}$.

The DCNN-based case demonstrates a reduction in uncertainty in these measurements, especially for the couplings to bosons.

7.4.4 Conclusions

A collection of measurements have been made comparing the BDT-based and DCNN-based VBF tags. The initial best fit to all categories shows an increase in expected signal purity and significance in the VBF production mode. In the likelihood scans

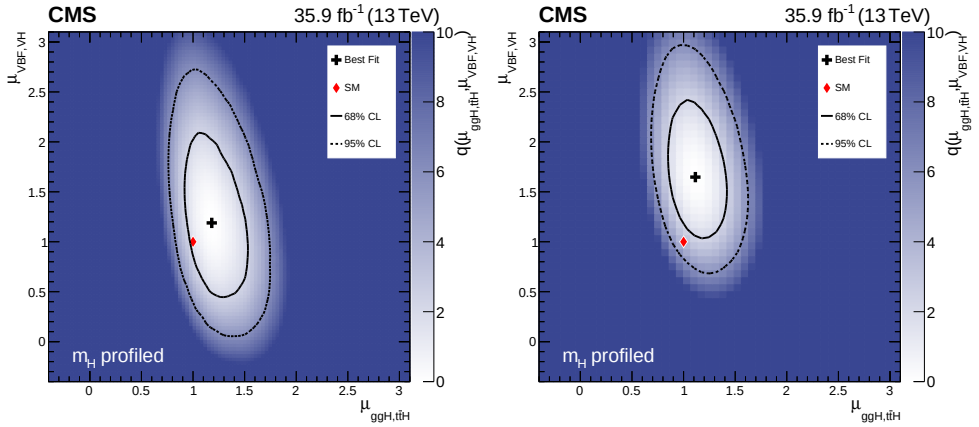

Figure 7.6: Two-dimensional likelihood scan of signal strength modifiers for bosonic (VBF, VH) and fermionic (ggH, $t\bar{t}H$) production modes. Analysis with the BDT-based VBF tag is shown on the left and the DCNN-based variant is on the right.

the DCNN is seen to bring improvement to some of the measurements. The greatest impact is seen in measurements of the VBF signal strength modifier itself, and on measurements of the bosonic signal strength and coupling modifiers. All measurements are compatible with the SM.

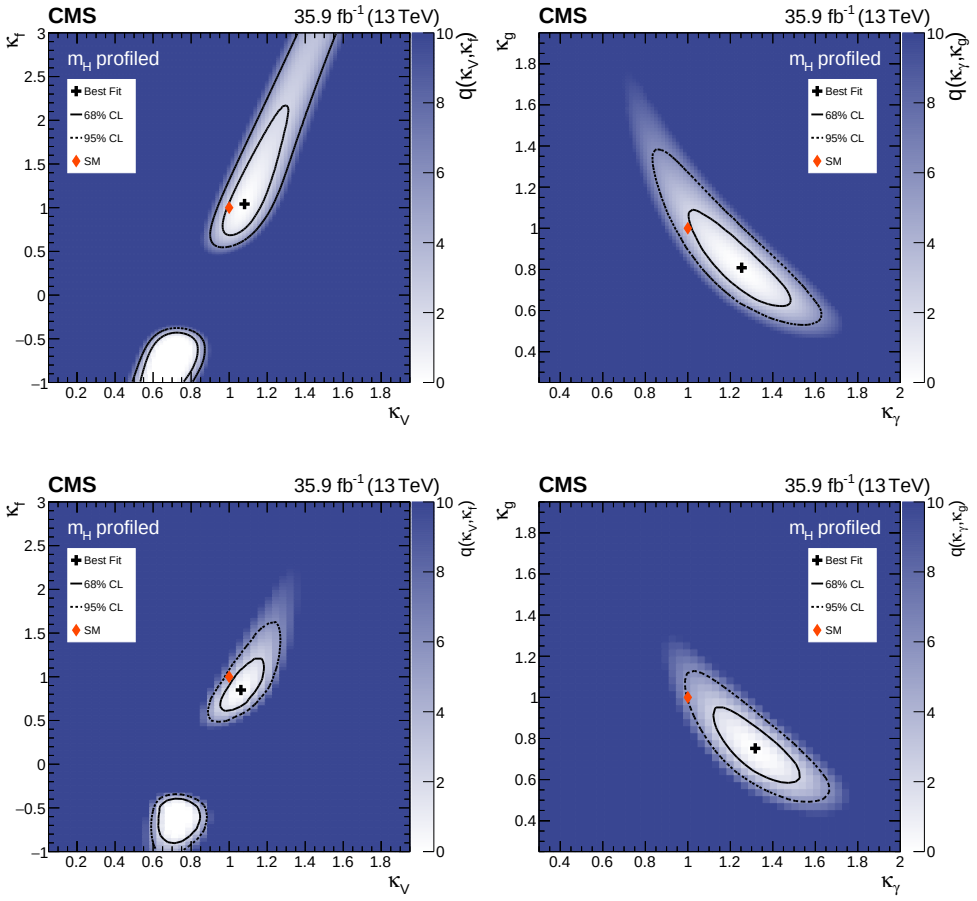

Figure 7.7: Two-dimensional likelihood scan of κ values for bosonic versus fermionic production modes (left) and effective gluon coupling versus effective photon coupling (right). The BDT-based VBF tag is shown on the top, and the DCNN-based tag is shown at the bottom.

Chapter 8

Conclusions

8.1 Summary of Results

Studies of the H $\to \gamma\gamma$ decay have been carried out using both a BDT-based VBF tag and a DCNN-based VBF tag with 35.9 fb^{-1} of $\sqrt{s} = 13$ TeV data. The Higgs boson is observed with high significance and measurements made of signal strength and coupling modifiers. These measurements are summarised in Table 8.1.

Measurement	BDT-based VBF		DCNN-based VBF	
μ (Global)	$1.18^{+0.17}_{-0.14}$		$1.24^{+0.16}_{-0.15}$	
μ_{VBF}	$0.8^{+0.6}_{-0.5}$		$1.5^{+0.5}_{-0.5}$	
$\sigma^{\text{VBF}}_{proc}/\sigma^{\text{VBF}}_{theo}$	$0.8^{+0.6}_{-0.5}$		$1.4^{+0.5}_{-0.4}$	
μ_F, μ_V	$1.19^{+0.22}_{-0.18}$	$1.21^{+0.58}_{-0.51}$	$1.11^{+0.20}_{-0.18}$	$1.65^{+0.50}_{-0.42}$
κ_F, κ_V	$1.04^{+0.51}_{-0.25}$	$1.08^{+0.10}_{-0.08}$	$0.85^{+0.22}_{-0.17}$	$1.06^{+0.07}_{-0.07}$
κ_γ, κ_g	$1.25^{+0.16}_{-0.17}$	$0.81^{+0.17}_{-0.13}$	$1.32^{+0.14}_{-0.13}$	$0.746^{+0.13}_{-0.13}$

Table 8.1: Measurement results.

The DCNN-based measurements have demonstrated significant improvement over the BDT-based approach [8] used since the Higgs boson was discovered [2]. Expected significance and purity in the VBF tag are improved leading to reduced uncertainty on the measurements, particularly those affected by the VBF tag. For this to be achieved the training had to be split into two steps: one for feature learning from the jet images, and one for the actual classification problem with the costs defined appropriately. Furthermore, the unique problems posed by the sparse nature of the jet images also required particular training conditions. Gradient clipping regularisation was needed to control large parameter updates early in the training, and an adaptive

optimiser was needed to handle how the sparseness affects the frequency of parameter updates.

The DCNN learns to construct physically sensible and robust features from correlations in structure between the jets. This was shown by data-simulation validation on $Z \to e^+e^-$, and by evaluating how the network performs over different modelling variations. The features themselves were extracted using feature visualisation to produce maximally-activating images for the class logits (and at other points in the network in Appendix C). These features were found to correspond to expected properties of VBF and ggH production, a particular example of this being the distortion of the jets by colour connection in VBF.

The DCNN-based measurements along with the validation and interpretation demonstrates that the DCNN-based VBF tag is superior and DCNNs have great potential in particle physics analyses. However, this is still a relatively new technology in the field of particle physics and will need to be studied closely for its precise systematic effects.

8.2 Future Development

The most immediate future improvement could be the application of a similar DCNN to VH hadronic signal extraction. This is currently achieved using a set of kinematic cuts and could be improved with a ML-based approach. When the associated vector boson decays hadronically there will be structure correlations from colour connection in the resulting dijet and characteristic charge deposition. These are features a DCNN could pick up on.

In 2026 the LHC will have been upgraded to the High-luminosity LHC (HL-LHC) [96] to provide much higher instantaneous luminosity resulting in a substantially larger dataset. This collision environment will pose particular challenges for the CMS detector's hardware and operation, as well as CMS analyses. Pileup will increase to 200 collisions per event and the radiation dose to the detector itself will be markedly increased. This has necessitated the future replacement of the CMS endcap calorimetry by a high-granularity silicon sampling calorimeter with many readout channels [97]. The difficulty of extracting physics objects in such an environment, and in accurately reconstructing them from such high-dimensional data, may be areas where deep learning approaches will prove to be especially important.

There are many avenues to investigate for the development of future ML algorithms and where to apply them. A few possible directions are outlined here: neural attention, generative adversarial networks (GANs), and ML algorithms more suited to the natural structure of jets.

Neural attention [98, 99] inserts a mechanism that allows the network to transform and focus on particular parts of the input depending on what features are present. This facilitates the construction of long-range dependencies between different input regions, for example different parts of an image. This could be useful in processing jet images by picking out dependencies between clusters of PF candidates in the jet more efficiently. However, given the non-standard behaviour of the polar jet images the way the transformations are applied will need to be handled carefully to be compatible with the periodic boundary condition.

GANs [100] are generative models that attempt to model the underlying data-generating process $P(\vec{x})$ itself and can achieve remarkable results (Figure 8.1).

Figure 8.1: Fake celebrity faces generated by a progressive GAN [101].

These consist of two competing networks: the generator and discriminator. The generator makes 'fake' examples given a vector in a space, and the discriminator tries to tell these fake examples from real ones. During training each learns from the other and once fully trained the generator ideally becomes a realistic generative model, and the discriminator develops many useful discriminating features.

Both the generator and discriminator may have application in particle physics analyses. The generator could be used to enhance under-populated simulation samples to improve cut optimisation or ML model trainings (an example may be found in [102]). The discriminator could be used to learn features from data that can then be used in another training, this was shown to work well in [103]. In a physics analysis this could be achieved by training a GAN on real data with a CNN as the discriminator, then using the frozen convolutional layers of the discriminator in another training over simulation. This would be like the two-step training of the VBF DCNN model

but with the discriminator features used instead of step one.

Finally, the components and structure of CNNs are geared towards images with locality and translational invariance in their features that can then be combined hierarchically. Recurrent neural networks (RNNs) make similar assumptions over a sequence. However, jets are not naturally images or lists. They have a tree structure where an initial parton gives rise to multiple daughter particles that then fragment or decay to more daughter particles.

The assumptions in an ML algorithm's construction determine how it makes predictions and is referred to as its 'inductive bias' [104]. The polar jet image formulation was a way of representing jets in a way that works with the inductive bias of CNNs, but it may be better to use an algorithm that is suited to trees or graphs. Examples of such algorithms are recursive neural networks (Chapter 10 of [44]) that extend the sequence processing of RNNs to tree structures, and graph CNNs [105] that have analogues for convolution and pooling given graph-based input.

8.3 Conclusions

All of the measurements presented in this thesis are compatible with predicted values for a Standard Model Higgs boson, but there is ample room for deviation within uncertainty. This uncertainty will be reduced as the available dataset increases in size over the Run II era and onwards. As the measurements become less dominated by statistical uncertainty, systematic uncertainties in precision measurements will become especially important, in particular contamination of categories such as VBF by ggH. Here the DCNN approach will be especially useful for constructing high-purity samples when overall significance is less of a priority.

The fields of Higgs physics and machine learning have now entered a new era of scale, precision and sophistication. Hopefully BSM signals are hidden within the uncertainties of our measurements and ML will grant us the power to extract them as our dataset expands. There is a bright future for both fields with great potential for collaboration between them and lots of work to be done.

Appendix A

VBF Tag Plots with Loose Preselection

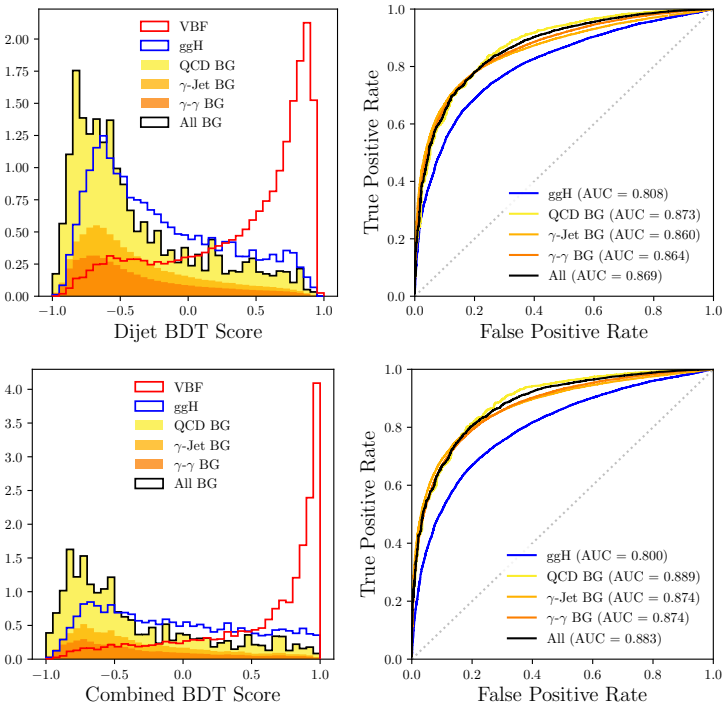

Figure A.1: Dijet BDT performance and combined BDT performance evaluated with the loose preselection.

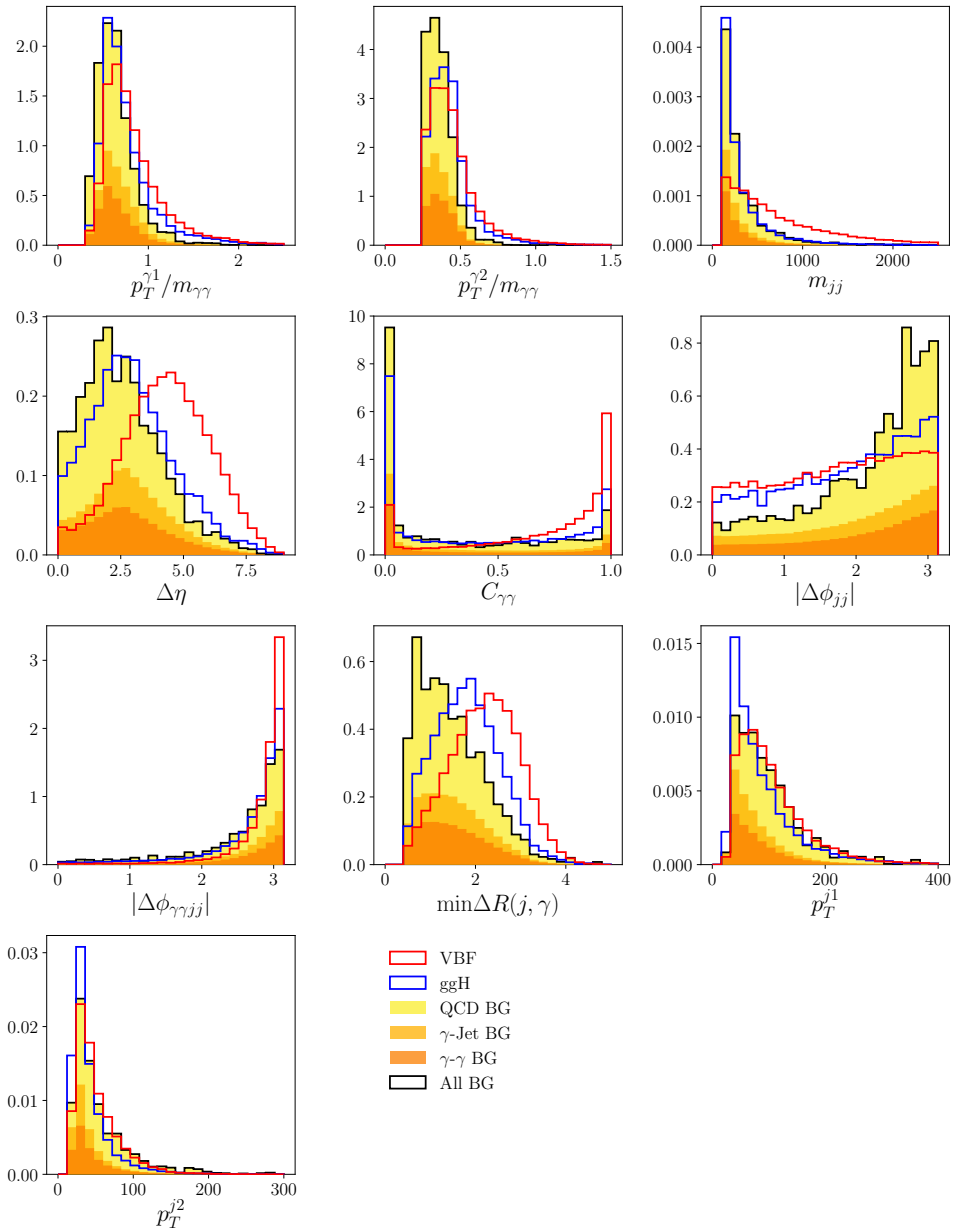

Figure A.2: Dijet BDT feature distributions with the loose VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM background.

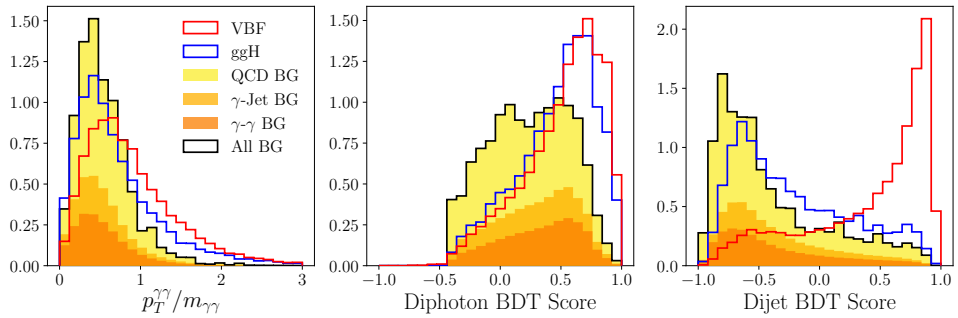

Figure A.3: Dijet BDT feature distributions with the loose VBF preselection. Distributions are all normalised to unity with solid red corresponding to VBF, blue line to ggH, and black line to SM background.

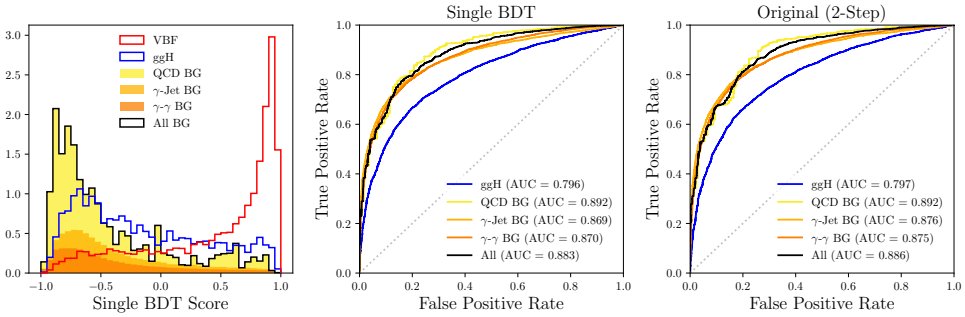

Figure A.4: Single BDT performance with the loose VBF preselection.

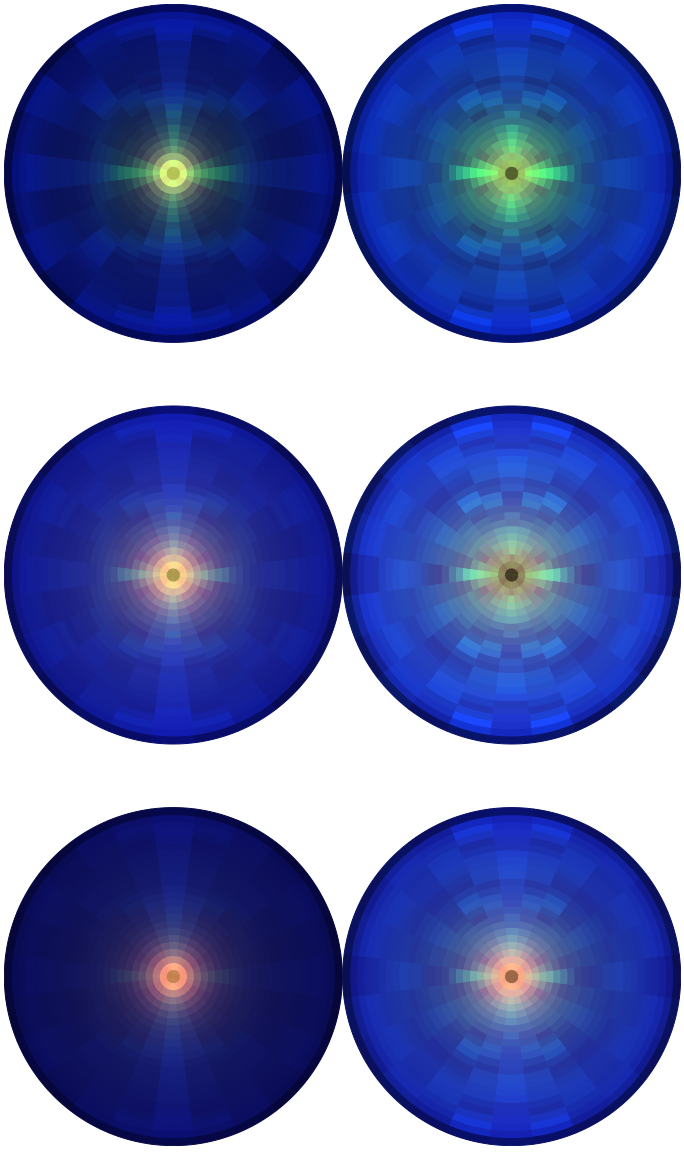

Figure A.5: Mean images in the loose VBF selection. From top to bottom: VBF, ggH and SM background processes.

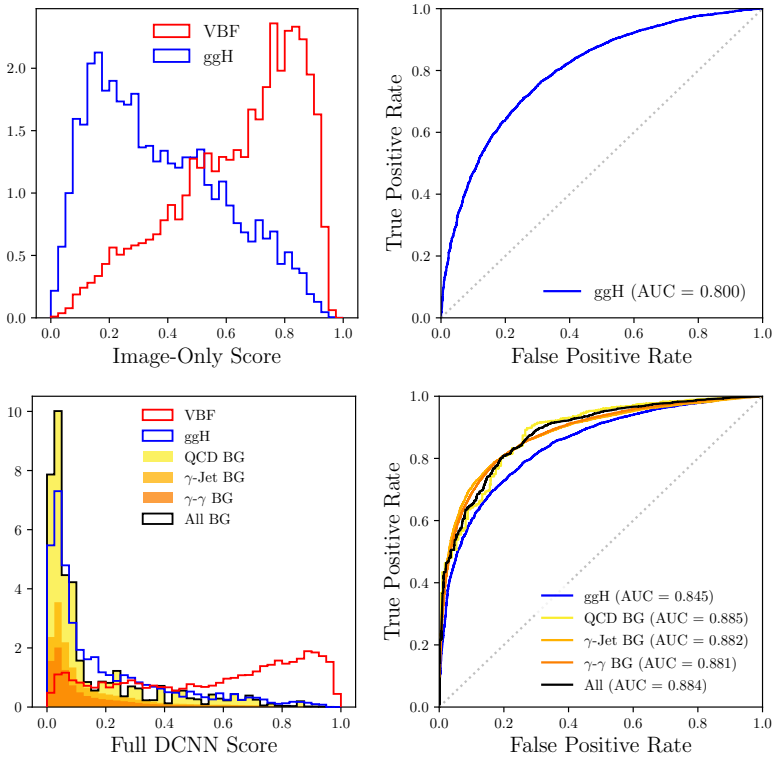

Figure A.6: Dense CNN model performance for images only (top) and the full model (bottom) in the loose VBF preselection.

Appendix B

VBF Tag $Z \to e^+e^-$ Validation Plots

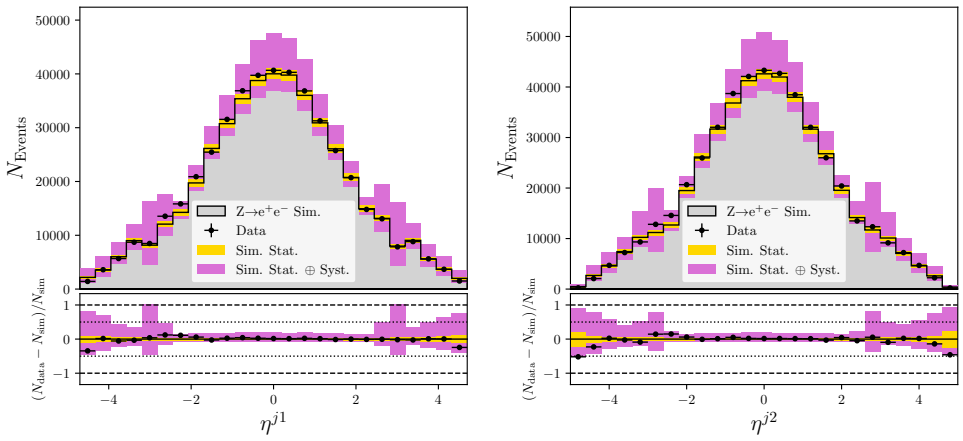

Figure B.1: $Z \to e^+e^-$ validation plots of pseudorapidity distributions for leading jet in p_T (left) and subleading jet (right).

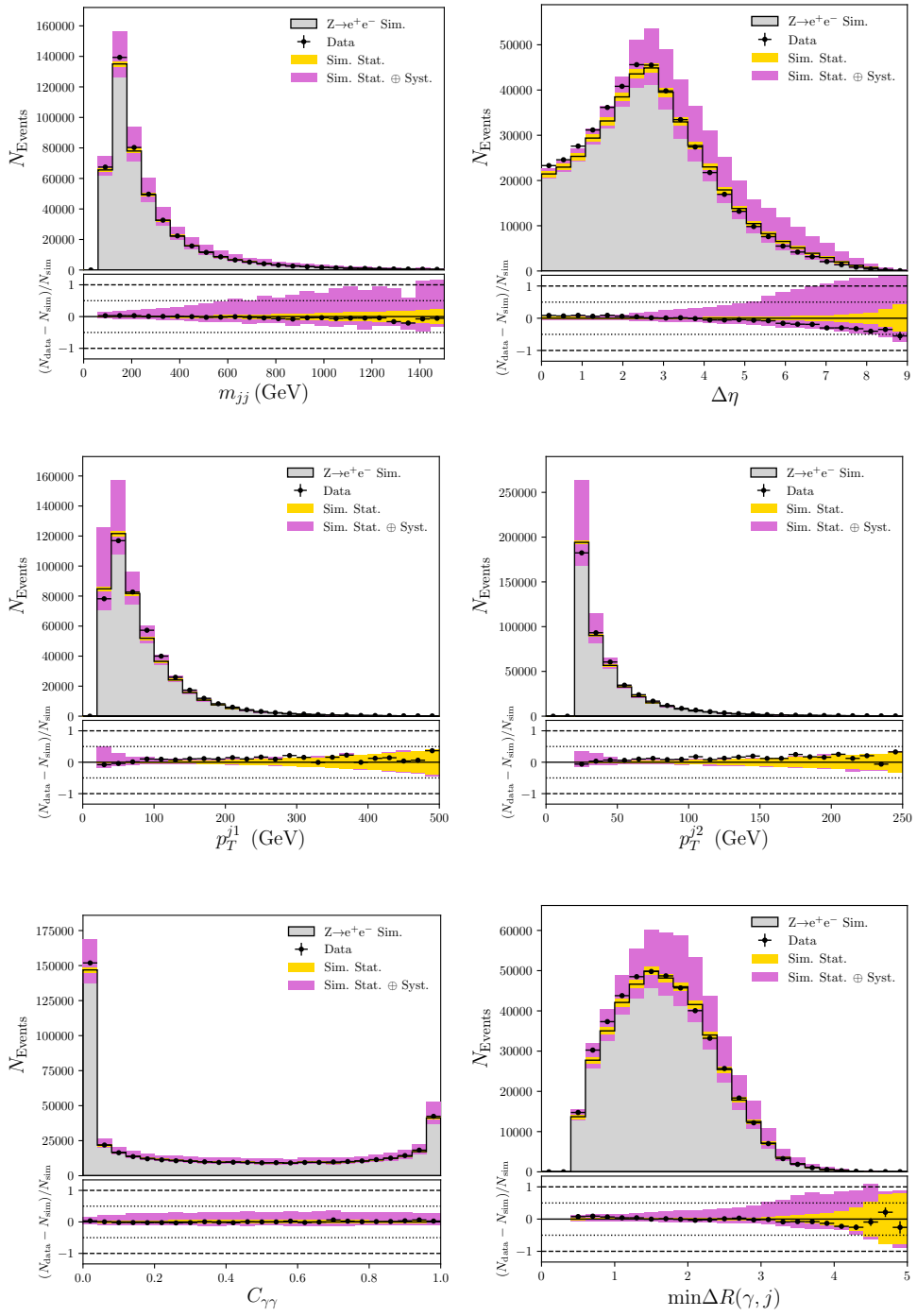

Figure B.2: $Z \to e^+e^-$ validation plots for kinematic features used by the VBF tag. Clockwise from top left: dijet mass, dijet pseudorapidity gap, subleading jet p_T, minimum ΔR between either photon and either jet, centrality, and leading jet p_T.

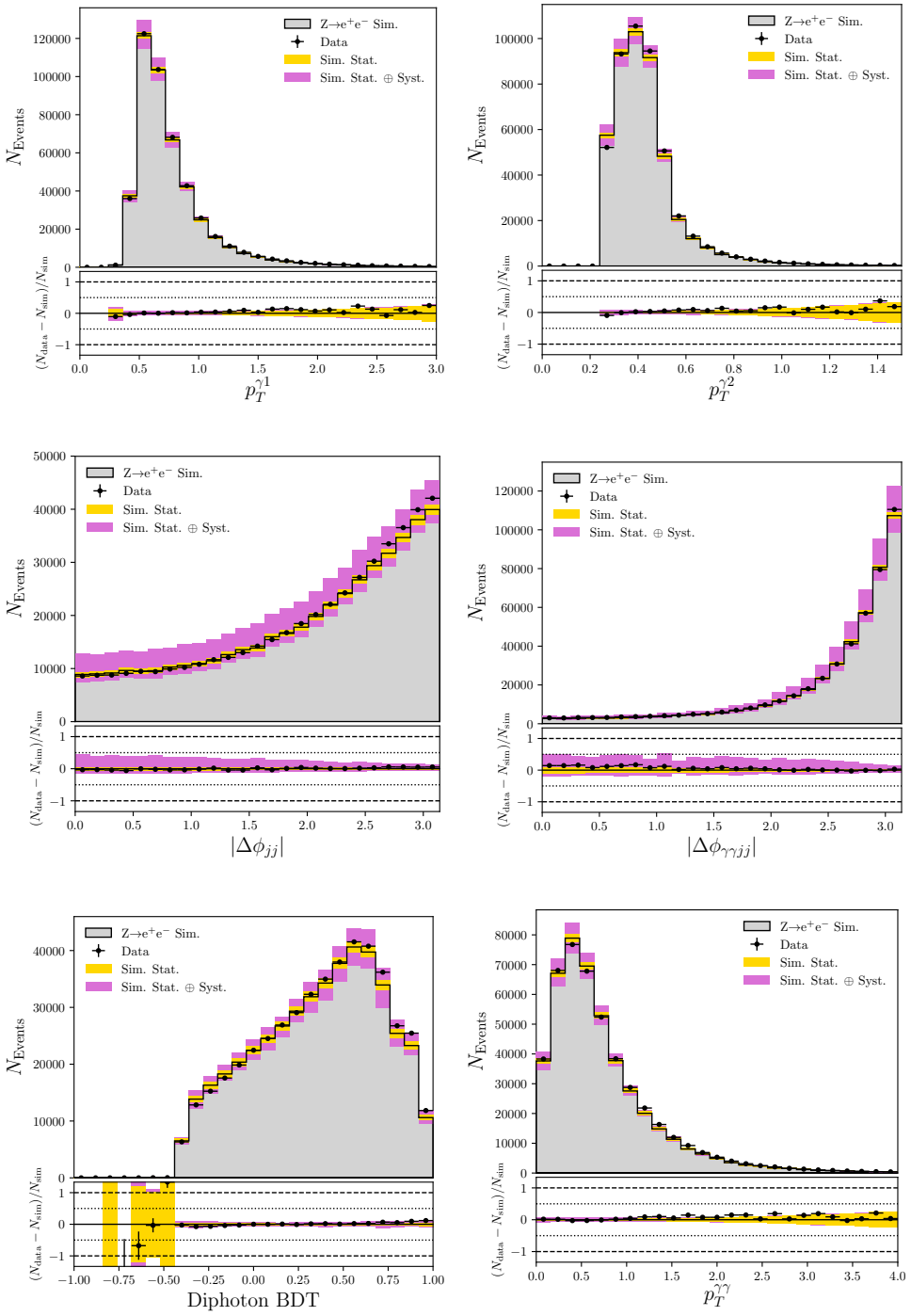

Figure B.3: $Z \to e^+e^-$ validation plots for kinematic features used by the VBF tag. Clockwise from top left: leading photon p_T scaled by the diphoton mass, subleading photon p_T scaled by the diphoton mass, azimuthal angular difference between dijet and diphoton, total diphoton p_T scaled by diphoton mass, diphoton BDT score, and azimuthal angular difference between the dijet jets.

Appendix C

Feature Visualisation of Different Network Layers

This appendix presents feature visualisation applied to different parts of the trained VBF DCNN model's convolutional section. This is to demonstrate how features are constructed and combined as one goes deeper into the network starting with the spread layer. Earlier layers are optimised for the mean over a feature map, later ones are for a single neuron as the receptive field becomes so large that trying to optimise them all does not show much structure. The reader is referred to Figure 6.16 for where the named places are located in the network.

These images are all normalised by dividing by the highest valued pixel in the dijet image. This will show the relative weighting of features from the leading vs subleading jet image channels.

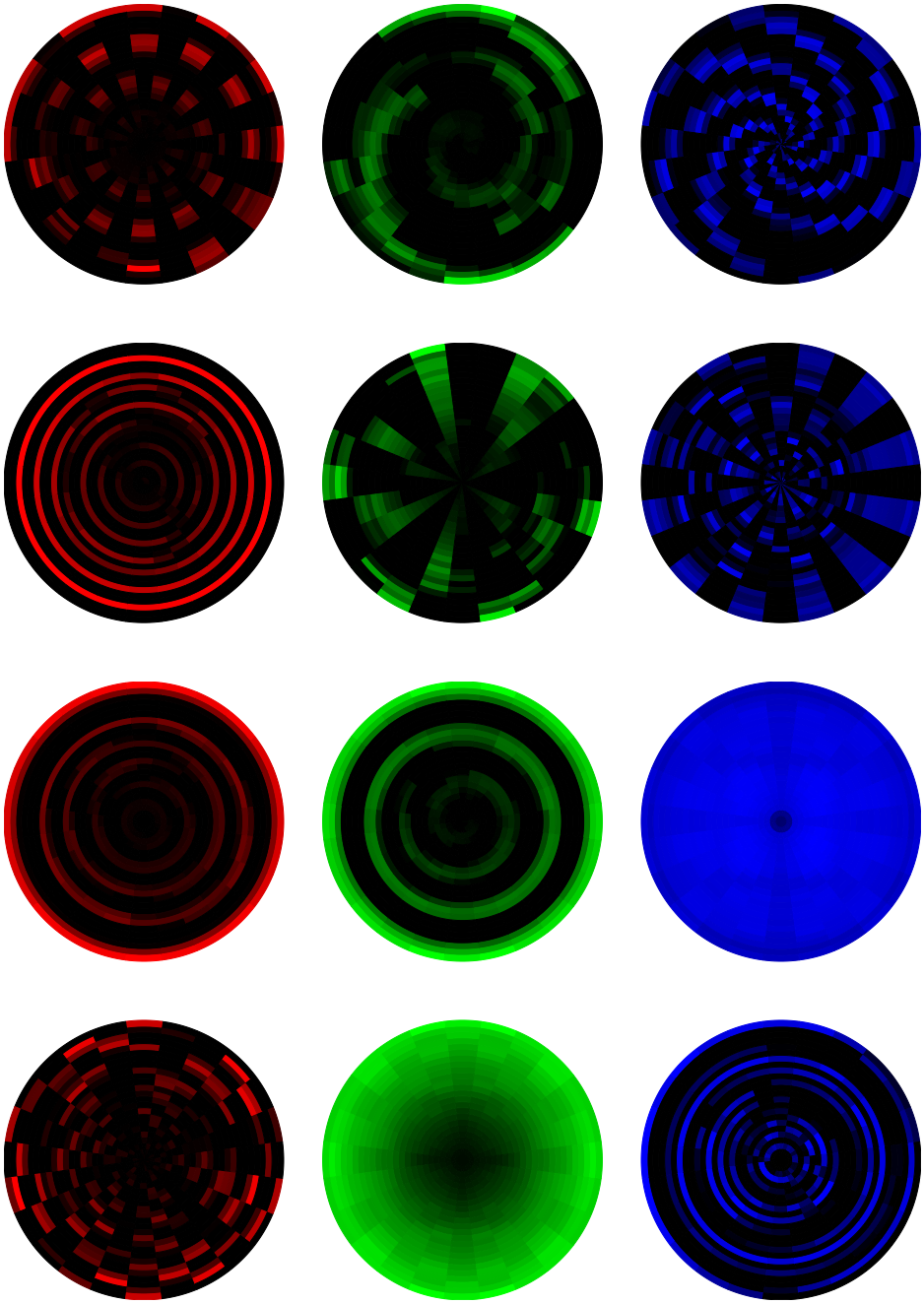

Figure C.1: Feature visualisation of the spread layer features. Red is the charged p_T channel, green is the neutral p_T channel and blue is the PF candidate multiplicity channel. This layer only constructs features in individual channels. Optimisation objective is the mean of the values over a whole feature map.

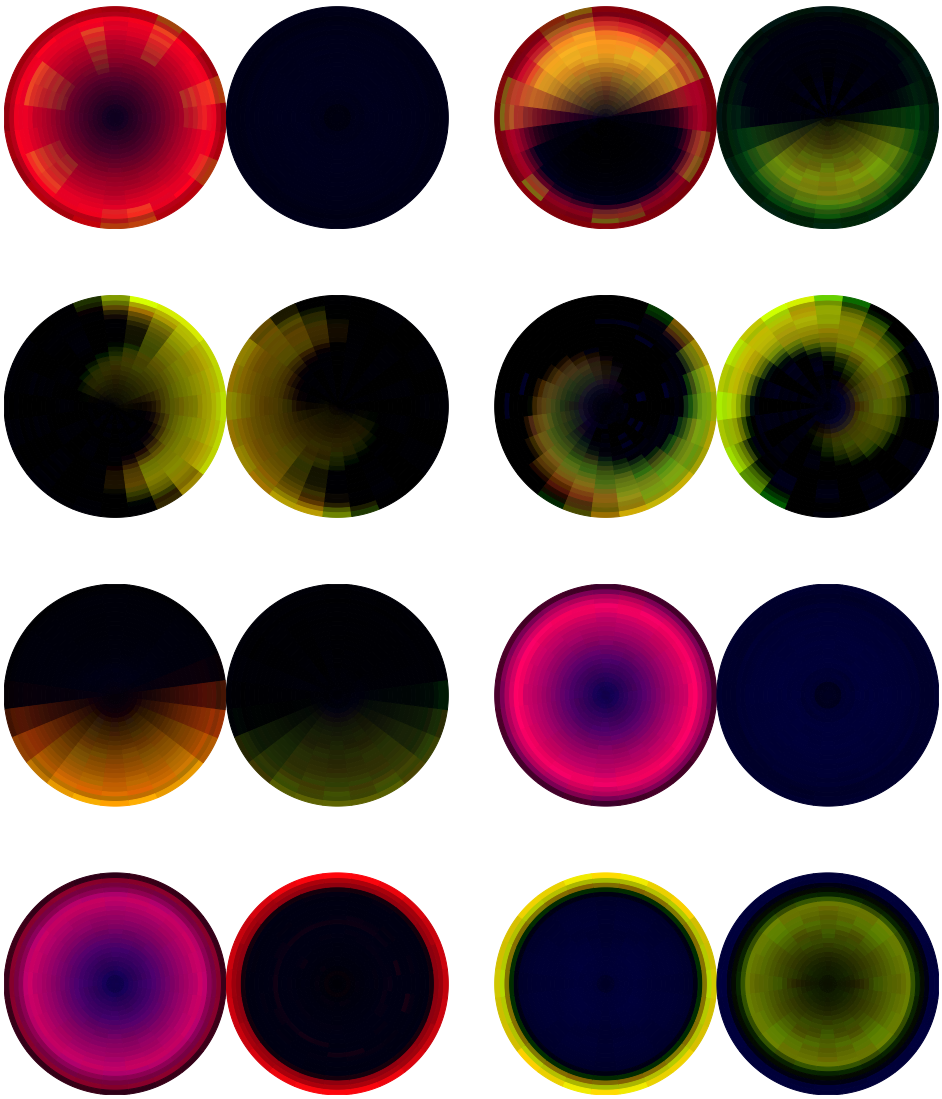

Figure C.2: Feature visualisation of the output of TU1. Here the low level features have been combined together to compare structure across channels, directly opposite around the jet axis and between the jets of the dijet. Optimisation objective is the mean of the values over a whole feature map.

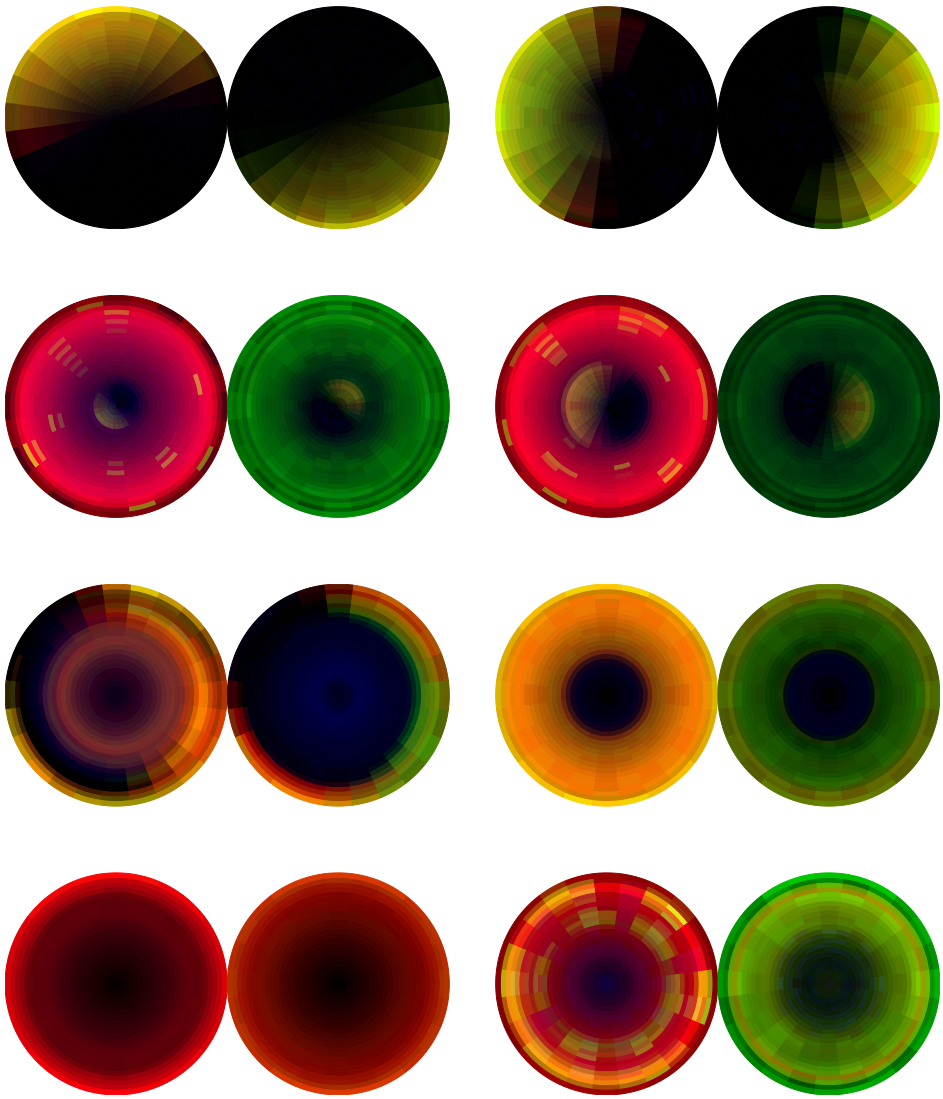

Figure C.3: Feature visualisation of the output of TU2. Here the features of TU1 are combined to make more complex features, but they are also reused (this is facilitated by the skip connections and is a capability of dense CNNs). Optimisation objective is the mean of the values over a whole feature map.

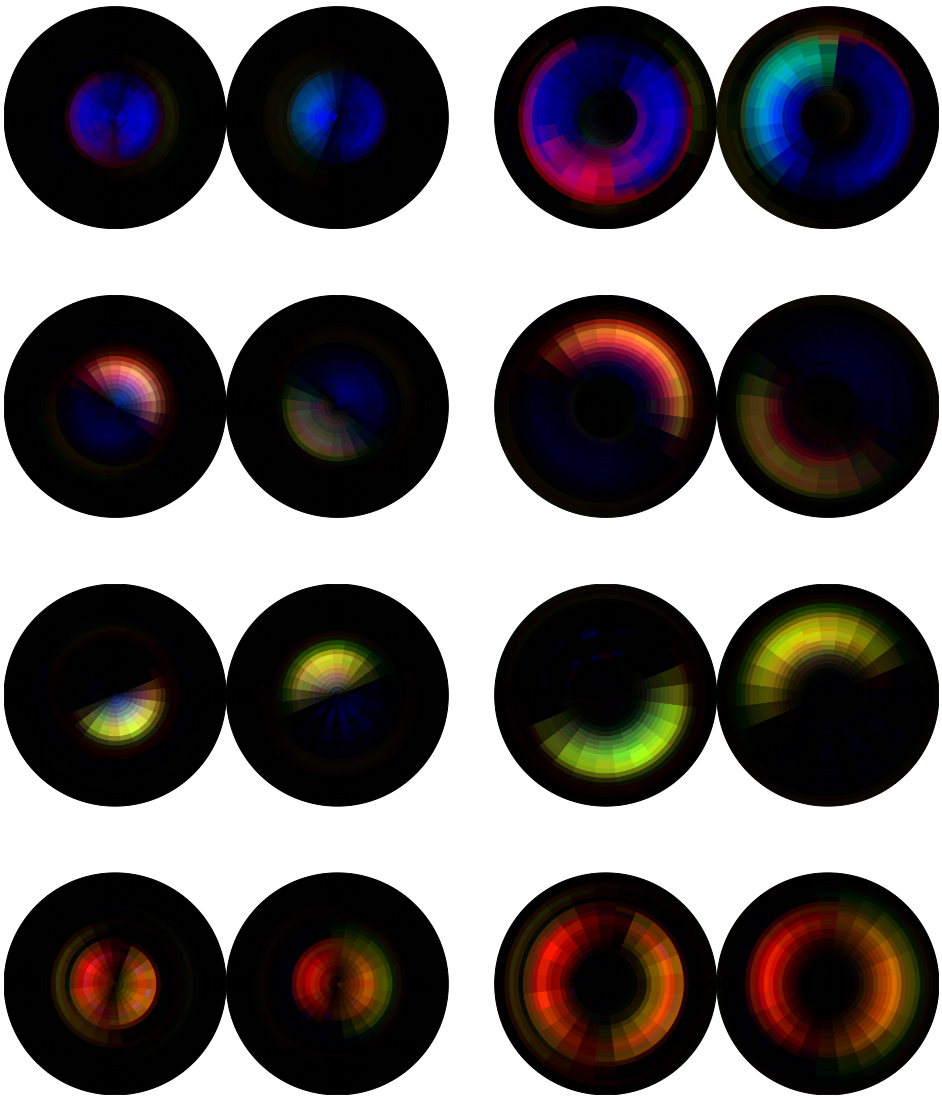

Figure C.4: Feature visualisations of individual neuron values after TU3. These constitute the learned features used in the main discriminant. These images are optimised to maximally activate a single neuron rather than the mean of the neurons of one feature map.

Appendix D

Per-Category Mass Plots

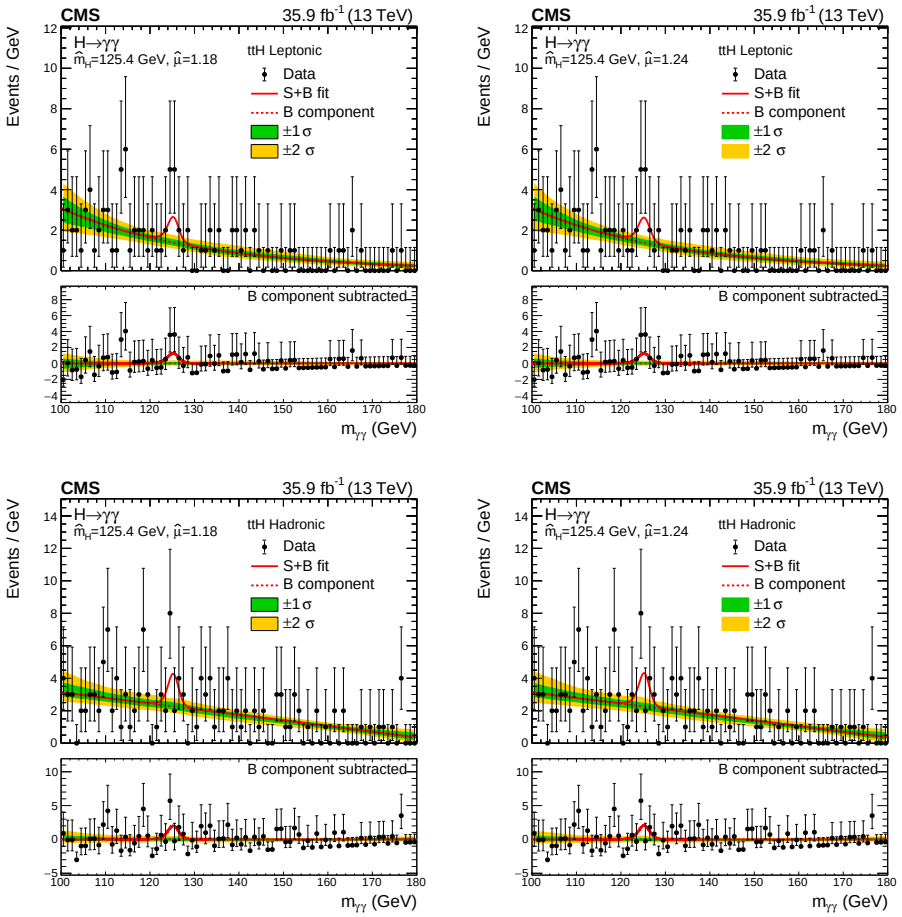

Figure D.1: Mass plots of the $t\bar{t}H$ tags. BDT-based VBF analysis is on the left and DCNN-based is on the right.

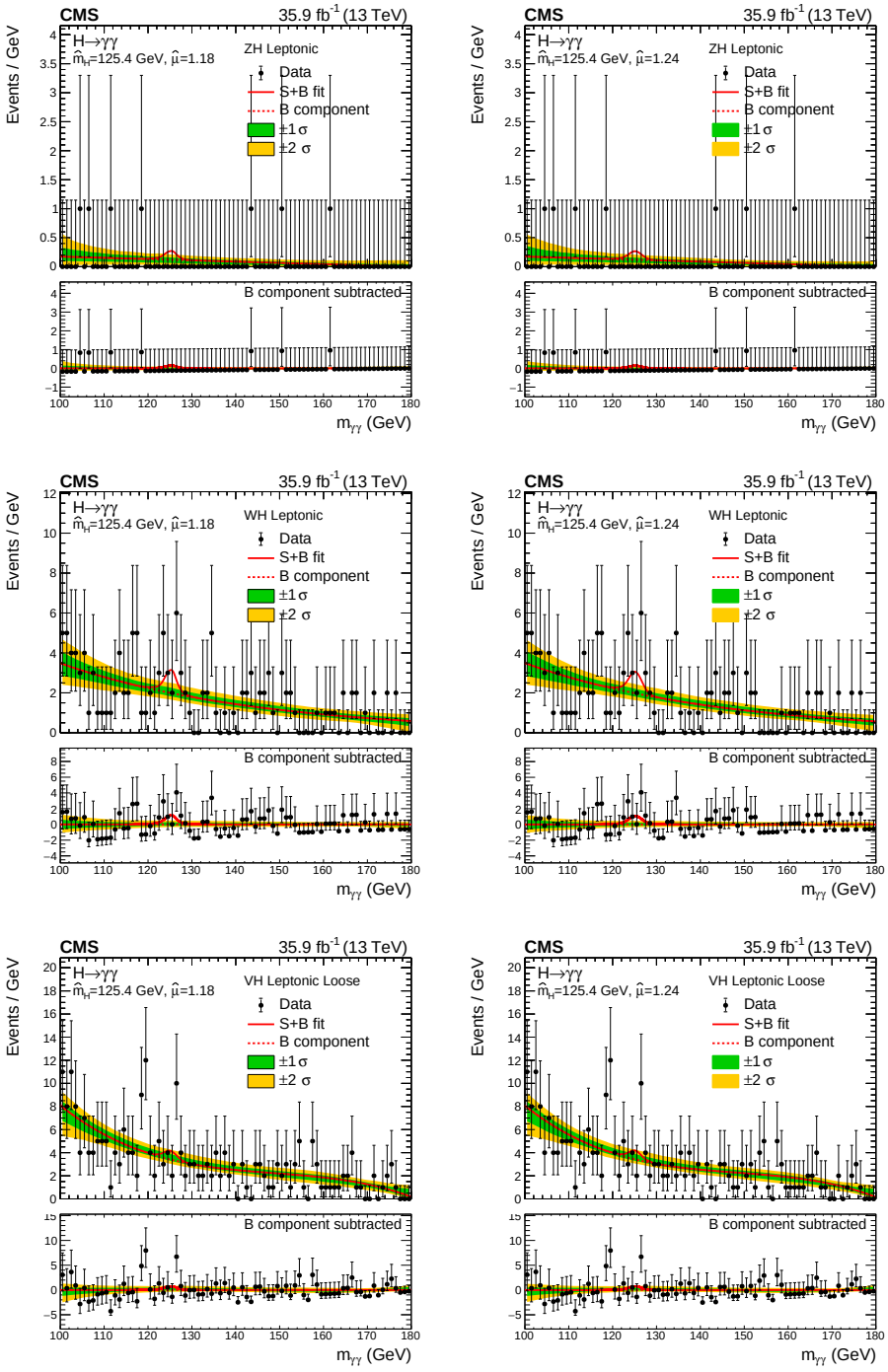

Figure D.2: VH leptonic tags. BDT-based VBF analysis is on the left and DCNN-based is on the right.

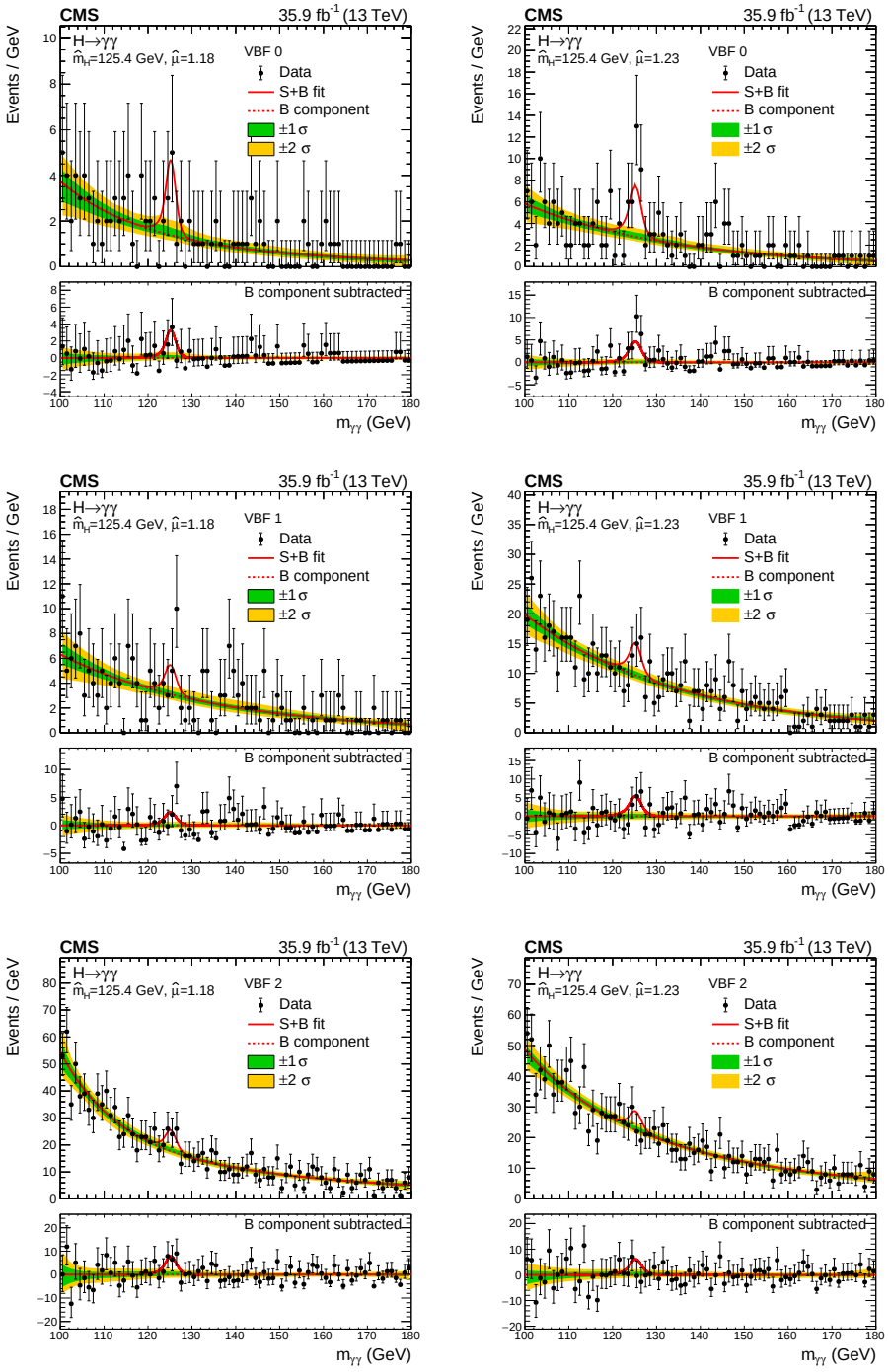

Figure D.3: VBF tag categories. BDT-based VBF analysis is on the left and DCNN-based is on the right.

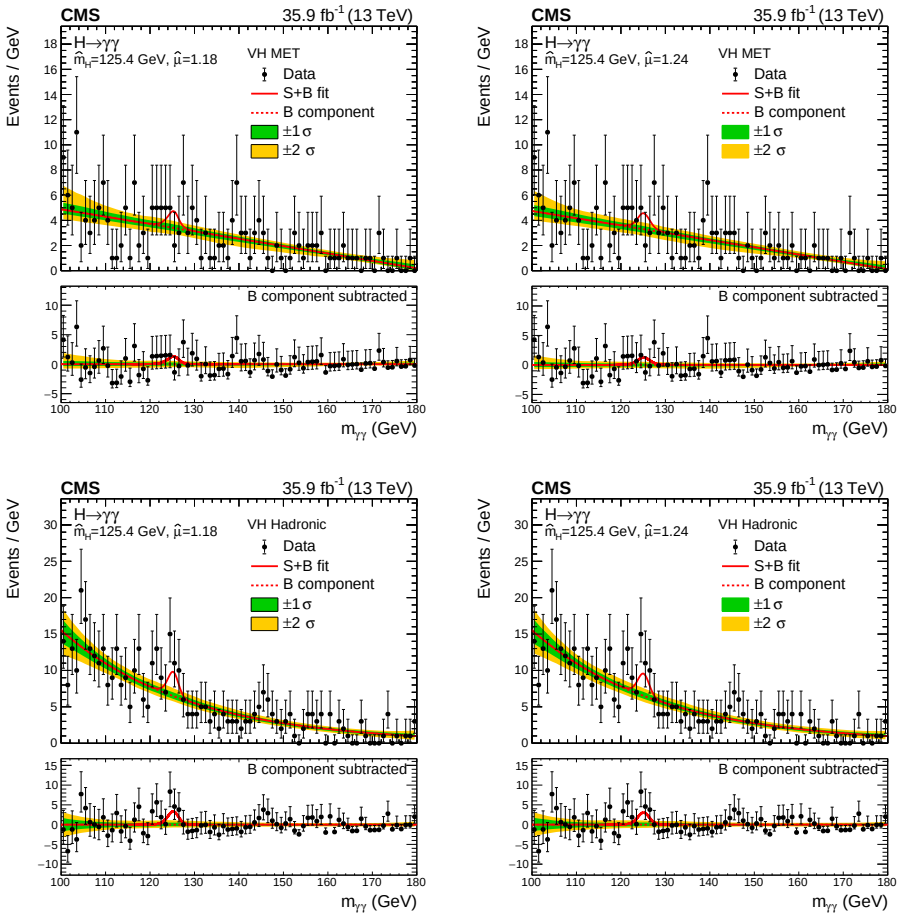

Figure D.4: VH MET and VH hadronic tags. BDT-based VBF analysis is on the left and DCNN-based is on the right.

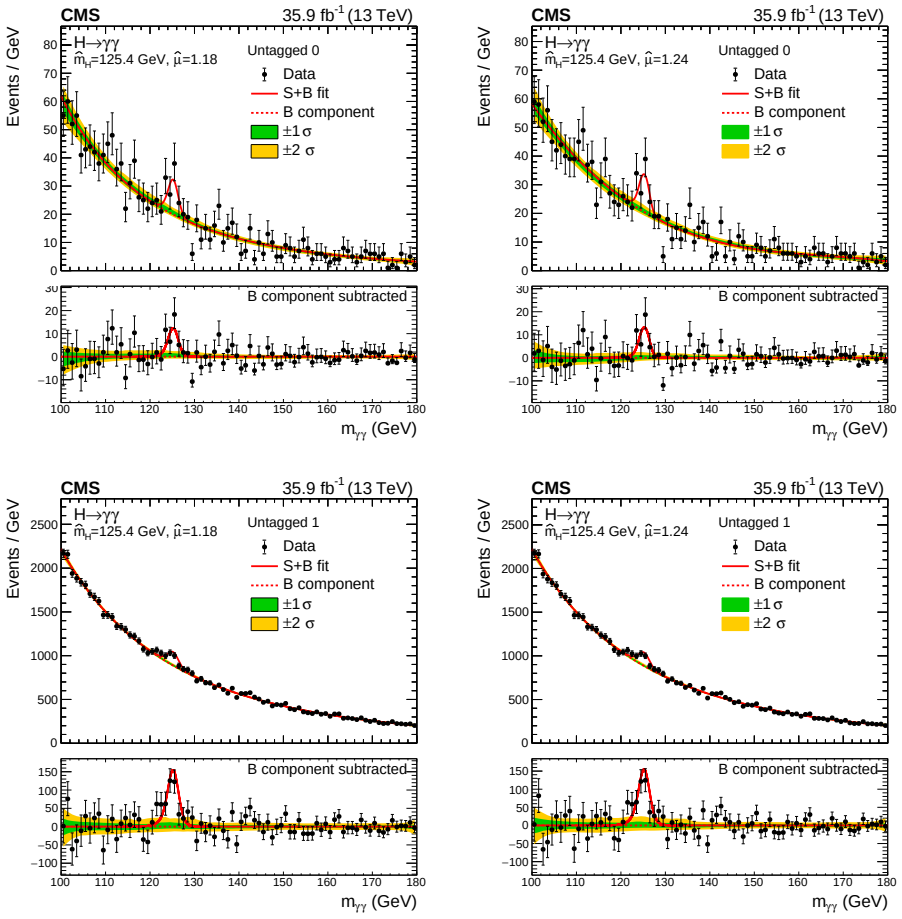

Figure D.5: Untagged categories 0 and 1. BDT-based VBF analysis is on the left and DCNN-based is on the right.

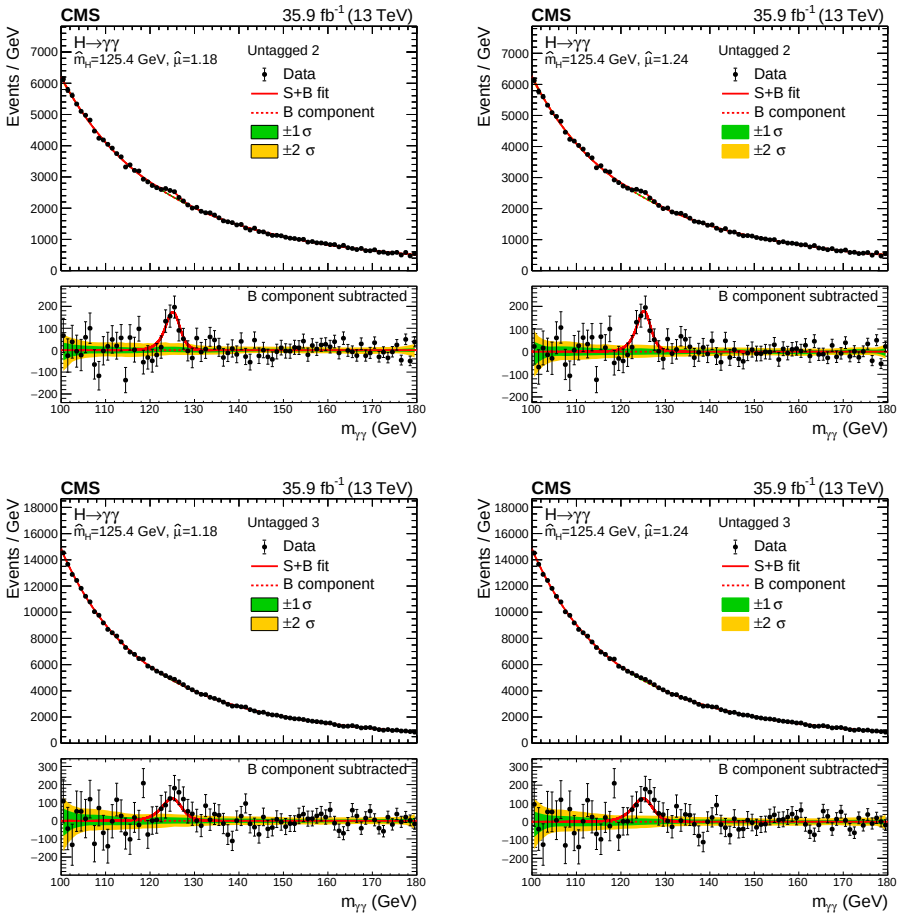

Figure D.6: Untagged categories 2 and 3. BDT-based VBF analysis is on the left and DCNN-based is on the right.

Bibliography

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

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

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

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

[5] Csaba Csáki and Philip Tanedo. "Beyond the Standard Model". In: *Proceedings, 2013 European School of High-Energy Physics (ESHEP 2013): Paradfurdo, Hungary, June 5-18, 2013*. 2015, pp. 169–268. DOI: 10.5170/CERN-2015-004.169. arXiv: 1602.04228 [hep-ph].

[6] H.-J. Yang, B. P. Roe, and J. Zhu. "Studies of boosted decision trees for MiniBooNE particle identification". In: *Nuclear Instruments and Methods in Physics Research A* 555 (Dec. 2005), pp. 370–385. DOI: 10.1016/j.nima.2005.09.022. eprint: physics/0508045.

[7] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[8] Albert M Sirunyan et al. "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: (2018). arXiv: 1804.02716 [hep-ex].

[9] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 9780201503975, 0201503972.

[10] Roger Penrose. *The road to reality : a complete guide to the laws of the universe*. London: Vintage, 2005. ISBN: 978-0-09-944068-0.

[11] Michele Maggiore. *A Modern introduction to quantum field theory*. 2005.

[12] Y. Nambu. "Quasi-Particles and Gauge Invariance in the Theory of Superconductivity". In: *Physical Review* 117 (Feb. 1960), pp. 648–663. DOI: 10.1103/PhysRev.117.648.

[13] J. Goldstone. "Field Theories with Superconductor Solutions". In: *Il Nuovo Cimento (1955-1965)* 19.1 (Jan. 1961), pp. 154–164. ISSN: 1827-6121. DOI: 10.1007/BF02812722. URL: https://doi.org/10.1007/BF02812722.

[14] P. W. Anderson. "Plasmons, Gauge Invariance, and Mass". In: *Phys. Rev.* 130 (1 Apr. 1963), pp. 439–442. DOI: 10.1103/PhysRev.130.439. URL: https://link.aps.org/doi/10.1103/PhysRev.130.439.

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

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

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

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

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

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

[21] M. Goldhaber, L. Grodzins, and A. W. Sunyar. "Helicity of Neutrinos". In: *Physical Review* 109 (Feb. 1958), pp. 1015–1017. DOI: 10.1103/PhysRev.109.1015.

[22] C. Patrignani et al. "Review of Particle Physics". In: *Chin. Phys.* C40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[23] N. Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Physical Review Letters* 10 (June 1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531.

[24] M. Kobayashi and T. Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49 (Feb. 1973), pp. 652–657. DOI: 10.1143/PTP.49.652.

[25] Sebastian Sapeta. "QCD and Jets at Hadron Colliders". In: *Prog. Part. Nucl. Phys.* 89 (2016), pp. 1–55. DOI: 10.1016/j.ppnp.2016.02.002. arXiv: 1511.09336 [hep-ph].

[26] S. Dittmaier et al. "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables". In: (2011). DOI: 10.5170/CERN-2011-002. arXiv: 1101.0593 [hep-ph].

[27] Vardan Khachatryan et al. "Limits on the Higgs boson lifetime and width from its decay to four charged leptons". In: *Phys. Rev.* D92.7 (2015), p. 072010. DOI: 10.1103/PhysRevD.92.072010. arXiv: 1507.06656 [hep-ex].

[28] A. Denner et al. "Standard Model Higgs-Boson Branching Ratios with Uncertainties". In: *Eur. Phys. J.* C71 (2011), p. 1753. DOI: 10.1140/epjc/s10052-011-1753-8. arXiv: 1107.5909 [hep-ph].

[29] Michael Benedikt et al. *LHC Design Report*. Tech. rep. CERN-2004-003-V-3. Geneva, 2004. URL: https://cds.cern.ch/record/823808.

[30] Christiane Lefèvre. *The CERN accelerator complex. Complexe des accélérateurs du CERN*. Tech. rep. CERN-DI-0812015. Geneva: CERN, Dec. 2008. URL: https://cds.cern.ch/record/1260465.

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

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

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

[34] *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/922757.

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

[36] CMS Collaboration. "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays". In: *Journal of Instrumentation* 5.03 (2010), T03021. URL: http://stacks.iop.org/1748-0221/5/i=03/a=T03021.

[37] *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/368412.

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

[39] M. Anderson et al. "A Review of clustering algoritms and energy corrections in the Electromagnetic Calorimeter". In: *CERN* CMS Internal Note (2010/008).

[40] Serguei Chatrchyan et al. "Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV". In: *JINST* 8 (2013). [JINST8,9009(2013)], P09009. DOI: 10.1088/1748-0221/8/09/P09009. arXiv: 1306.2016 [hep-ex].

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

[42] *The CMS muon project: Technical Design Report*. Technical Design Report CMS. Geneva: CERN, 1997.

[43] Vardan Khachatryan et al. "The CMS trigger system". In: *JINST* 12.01 (2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020. arXiv: 1609.02366 [physics.ins-det].

[44] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[45] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[46] Thomas M. Mitchell. *Machine Learning*. 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077, 9780070428072.

[47] Andrej Karpathy. "Stanford University CS231n: Convolutional Neural Networks for Visual Recognition". In: (). URL: http://cs231n.stanford.edu/syllabus.html.

[48] Ilya Sutskever et al. "On the Importance of Initialization and Momentum in Deep Learning". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. JMLR.org, 2013.

[49] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). URL: http://arxiv.org/abs/1412.6980.

[50] Timothy Dozat. "Incorporating Nesterov Momentum into Adam". In: 2016.

[51] J. R. Quinlan. “Induction of Decision Trees”. In: *Mach. Learn.* 1.1 (Mar. 1986), pp. 81–106. ISSN: 0885-6125. DOI: 10.1023/A:1022643204877. URL: http://dx.doi.org/10.1023/A:1022643204877.

[52] Leo Breiman. “Arcing classifier (with discussion and a rejoinder by the author)”. In: *Ann. Statist.* 26.3 (June 1998), pp. 801–849. DOI: 10.1214/aos/1024691079. URL: https://doi.org/10.1214/aos/1024691079.

[53] Jerome H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* 29 (2000), pp. 1189–1232.

[54] David H. Wolpert. “The Lack of a Priori Distinctions Between Learning Algorithms”. In: *Neural Comput.* 8.7 (Oct. 1996), pp. 1341–1390. ISSN: 0899-7667. DOI: 10.1162/neco.1996.8.7.1341. URL: http://dx.doi.org/10.1162/neco.1996.8.7.1341.

[55] Eric Brochu, Vlad M. Cora, and Nando de Freitas. “A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning”. In: *CoRR* abs/1012.2599 (2010). arXiv: 1012.2599. URL: http://arxiv.org/abs/1012.2599.

[56] Yann LeCun et al. “Efficient BackProp”. In: *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*. London, UK, UK: Springer-Verlag, 1998, pp. 9–50. ISBN: 3-540-65311-2. URL: http://dl.acm.org/citation.cfm?id=645754.668382.

[57] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), pp. 1929–1958. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2670313.

[58] Sepp Hochreiter et al. *Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies*. 2001.

[59] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.123. URL: http://dx.doi.org/10.1109/ICCV.2015.123.

[60] Andrew L. Maas. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: 2013.

[61] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *CoRR* abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[62] Bengt Nyman. *IMG˙7469 Peter Higgs Nobelpristagare i fysik December 2013.*

[63] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). arXiv: 1608.06993. URL: http://arxiv.org/abs/1608.06993.

[64] Serguei Chatrchyan et al. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

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

[66] Serguei Chatrchyan et al. "Measurement of the Inclusive W and Z Production Cross Sections in pp Collisions at $\sqrt{s} = 7$ TeV". In: *JHEP* 10 (2011), p. 132. DOI: 10.1007/JHEP10(2011)132. arXiv: 1107.4789 [hep-ex].

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

[68] Vardan Khachatryan et al. "Observation of the diphoton decay of the Higgs boson and measurement of its properties". In: *Eur. Phys. J.* C74.10 (2014), p. 3076. DOI: 10.1140/epjc/s10052-014-3076-z. arXiv: 1407.0558 [hep-ex].

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

[70] Serguei Chatrchyan et al. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV". In: *JINST* 7 (2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002. arXiv: 1206.4071 [physics.ins-det].

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

[72] Matthias Schröder and the CMS collaboration. "Performance of jets at CMS". In: *Journal of Physics: Conference Series* 587.1 (2015), p. 012004. URL: http://stacks.iop.org/1742-6596/587/i=1/a=012004.

[73] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[74] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[75] C Adam-Bourdarios et al. "The Higgs Machine Learning Challenge". In: *Journal of Physics: Conference Series* 664.7 (2015), p. 072015. URL: http://stacks.iop.org/1742-6596/664/i=7/a=072015.

[76] *Pileup Jet Identification*. Tech. rep. CMS-PAS-JME-13-005. Geneva: CERN, 2013. URL: https://cds.cern.ch/record/1581583.

[77] Patrick T. Komiske, Eric M. Metodiev, and Matthew D. Schwartz. "Deep learning in color: towards automated quark/gluon jet discrimination". In: *JHEP* 01 (2017), p. 110. DOI: 10.1007/JHEP01(2017)110. arXiv: 1612.01551 [hep-ph].

[78] *Imperial College Research Computing Service*. DOI: 10.14469/hpc/2232.

[79] A. S. Morcos et al. "On the importance of single directions for generalization". In: *ArXiv e-prints* (Mar. 2018). arXiv: 1803.06959 [stat.ML].

[80] Salman Hameed Khan et al. "Cost Sensitive Learning of Deep Feature Representations from Imbalanced Data". In: *CoRR* abs/1508.03422 (2015). arXiv: 1508.03422. URL: http://arxiv.org/abs/1508.03422.

[81] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization". In: *Distill* (2017). https://distill.pub/2017/feature-visualization. DOI: 10.23915/distill.00007.

[82] Mike Tyka Alexander Mordvintsev Christopher Olah. "Inceptionism: Going Deeper into Neural Networks". In: (2015). https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

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

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

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

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

[87] Iain W. Stewart et al. "Jet p_T resummation in Higgs production at $NNLL' + NNLO$". In: *Phys. Rev.* D89.5 (2014), p. 054001. DOI: 10.1103/PhysRevD.89.054001. arXiv: 1307.1808 [hep-ph].

[88] Xiaohui Liu and Frank Petriello. "Reducing theoretical uncertainties for exclusive Higgs-boson plus one-jet production at the LHC". In: *Phys. Rev.* D87.9 (2013), p. 094027. DOI: 10.1103/PhysRevD.87.094027. arXiv: 1303.4405 [hep-ph].

[89] Radja Boughezal et al. "Combining Resummed Higgs Predictions Across Jet Bins". In: *Phys. Rev.* D89.7 (2014), p. 074044. DOI: 10.1103/PhysRevD.89.074044. arXiv: 1312.4535 [hep-ph].

[90] John M. Campbell and R. K. Ellis. "MCFM for the Tevatron and the LHC". In: *Nucl. Phys. Proc. Suppl.* 205-206 (2010), pp. 10–15. DOI: 10.1016/j.nuclphysbps.2010.08.011. arXiv: 1007.3492 [hep-ph].

[91] Iain W. Stewart and Frank J. Tackmann. "Theory Uncertainties for Higgs and Other Searches Using Jet Bins". In: *Phys. Rev.* D85 (2012), p. 034011. DOI: 10.1103/PhysRevD.85.034011. arXiv: 1107.2117 [hep-ph].

[92] Shireen Gangal and Frank J. Tackmann. "Next-to-leading-order uncertainties in Higgs+2 jets from gluon fusion". In: *Phys. Rev.* D87.9 (2013), p. 093008. DOI: 10.1103/PhysRevD.87.093008. arXiv: 1302.5437 [hep-ph].

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

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

[95] J R Andersen et al. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties". In: (2013). Ed. by S Heinemeyer et al. DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347 [hep-ph].

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

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

[98] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[99] Max Jaderberg et al. “Spatial Transformer Networks”. In: *CoRR* abs/1506.02025 (2015). arXiv: 1506.02025. URL: http://arxiv.org/abs/1506.02025.

[100] I. J. Goodfellow et al. “Generative Adversarial Networks”. In: *ArXiv e-prints* (June 2014). arXiv: 1406.2661 [stat.ML].

[101] Tero Karras et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *CoRR* abs/1710.10196 (2017). arXiv: 1710.10196. URL: http://arxiv.org/abs/1710.10196.

[102] Hojjat Salehinejad et al. “Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks”. In: *CoRR* abs/1712.01636 (2017). arXiv: 1712.01636. URL: http://arxiv.org/abs/1712.01636.

[103] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *CoRR* abs/1511.06434 (2015). arXiv: 1511.06434. URL: http://arxiv.org/abs/1511.06434.

[104] P. W. Battaglia et al. “Relational inductive biases, deep learning, and graph networks”. In: *ArXiv e-prints* (June 2018). arXiv: 1806.01261.

[105] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. In: *CoRR* abs/1606.09375 (2016). arXiv: 1606.09375. URL: http://arxiv.org/abs/1606.09375.