Searches for neutral Higgs bosons decaying to tau pairs and measurement of the $Z+b$-jet cross section with the CMS detector

Andrew James Gilbert
Imperial College London
Department of Physics

A thesis submitted to Imperial College London
for the degree of Doctor of Philosophy
The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Abstract

The Compact Muon Solenoid (CMS) is a general-purpose particle detector at the CERN Large Hadron Collider. It is designed to search for the Higgs boson and evidence of new physics and to test the predictions of the standard model (SM) at the TeV scale. This thesis describes analyses of proton-proton collision data recorded by CMS during 2011 and 2012.

A study of Z boson production in association with b jets, using 2.1 fb\(^{-1}\) of data recorded at a centre-of-mass energy of 7 TeV, is presented. The cross sections for production with exactly one, or at least two, b jets are measured, and the event kinematics are compared to the predictions of the MADGRAPH event generator interfaced with PYTHIA for hadronisation and parton showering.

Searches for neutral Higgs bosons decaying to tau pairs are also presented. One search is in the context of the SM Higgs boson, for mass hypotheses in the range 90–150 GeV, and the other in the minimal supersymmetric standard model (MSSM), in which three neutral Higgs bosons are predicted and the search range is from 90 GeV to 1 TeV. Both searches use 4.9 fb\(^{-1}\) of data collected at 7 TeV and 19.7 fb\(^{-1}\) collected at 8 TeV. In the SM search an excess of events above the background expectation is observed and found to be compatible with the SM expectation for the 125 GeV Higgs boson. The observed (expected) local significance of this excess is 3.0 (3.1) standard deviations at 125 GeV. No significant excess is observed in the MSSM search. Upper limits at the 95% confidence level are determined, both in the \(m_A\)-\(\tan\beta\) parameter space of the \(m_h^{\max}\) scenario and on the production cross sections in a model-independent interpretation.
Declaration

I declare that the work contained in this thesis is my own, and all results and figures taken from other sources are indicated in the text and referenced appropriately. The analyses presented in this thesis were developed in close collaboration with other members of the CMS experiment.

For the Z+b-jet analysis I contributed to the inclusive Z+b cross section measurement [1] and the preliminary result [2] containing the separate Z+1b and Z+2b measurements. The latter analysis is presented in this thesis. For this I was responsible for implementing the analysis selection, developing a system of matrix equations to determine the cross sections at the hadron level, evaluating the effects of systematic uncertainties and deriving the final results.

As part of the H → ττ working group I contributed to the preliminary standard model analysis results presented at the ICHEP 2012 [3], HCP 2012 [4] and Moriond 2013 [5] conferences, as well as to the legacy results publication [6]. In addition I contributed to the minimal supersymmetric standard model analysis presented at HCP 2012 [7] and to a follow-up result in November 2013 [8]. For these results I contributed to the development and optimisation of the analysis strategy, the event selection and categorisation, studies of systematic uncertainties, and the statistical interpretation of the results. The analyses presented in this thesis are based on the most recent public results in each search.

Andrew Gilbert
Acknowledgements

I am grateful to the Imperial College HEP Group and the STFC for giving me the opportunity to pursue this research and for the two years of valuable experience I gained working at CERN. Thanks to my supervisor, David Colling, for introducing me to the world of high energy physics and for his encouragement and many useful discussions throughout my studies. Thanks also to Mónica for her invaluable guidance and advice and Sasha for his insight and enthusiasm. I owe a great deal to Anne-Marie for all her help in my first year at CERN and for teaching me the right way to do an analysis. I am grateful to everyone who read through the drafts of this thesis and provided many helpful ideas and suggestions. Thanks to everyone else in the CMS Z+b and H → ττ groups, with whom it was a pleasure to work. To Darren, Greg, Indrek, Matthew and Patrick, who I shared this adventure with - and more good times than I could possibly list here - I wish you all the very best. Finally, and most importantly, I thank my parents and my sister for their constant love and support.

Andrew Gilbert
## Contents

List of Figures 8

List of Tables 12

1 **Introduction** 13
   1.1 The standard model ................................................. 13
   1.1.1 Particles and forces ............................................. 13
   1.1.2 The Higgs mechanism ............................................ 17
   1.2 Higgs boson searches .............................................. 20
   1.3 Beyond the standard model ....................................... 23

2 **The CMS detector** 28
   2.1 The Large Hadron Collider ........................................... 28
   2.2 CMS design and geometry .......................................... 32
   2.3 Tracker .............................................................. 33
   2.4 Electromagnetic calorimeter ....................................... 35
   2.5 Hadron calorimeter ................................................. 36
   2.6 Muon system ......................................................... 38
   2.7 Trigger and computing ............................................ 39

3 **Event reconstruction and simulation** 42
   3.1 Tracks and vertices ................................................. 42
   3.2 Electrons ............................................................ 44
   3.3 Muons ................................................................. 45
   3.4 Particle flow ........................................................ 47
   3.5 Jets ................................................................. 49
      3.5.1 Clustering ...................................................... 50
      3.5.2 Jet energy corrections ....................................... 50
      3.5.3 b-Jet tagging .................................................. 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6</td>
<td>Missing transverse energy</td>
<td>53</td>
</tr>
<tr>
<td>3.7</td>
<td>Hadronic taus</td>
<td>55</td>
</tr>
<tr>
<td>3.8</td>
<td>Monte Carlo simulation</td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td>Measurement of the Z+b-jet cross section</td>
<td>59</td>
</tr>
<tr>
<td>4.1</td>
<td>Motivation</td>
<td>60</td>
</tr>
<tr>
<td>4.2</td>
<td>Data samples and simulation</td>
<td>61</td>
</tr>
<tr>
<td>4.3</td>
<td>Event selection</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>Kinematic observables</td>
<td>66</td>
</tr>
<tr>
<td>4.5</td>
<td>Background estimation</td>
<td>69</td>
</tr>
<tr>
<td>4.6</td>
<td>Efficiency and acceptance corrections</td>
<td>72</td>
</tr>
<tr>
<td>4.6.1</td>
<td>b-Tagging efficiency</td>
<td>73</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Lepton selection efficiency</td>
<td>74</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Reconstruction efficiency</td>
<td>75</td>
</tr>
<tr>
<td>4.6.4</td>
<td>Lepton acceptance</td>
<td>79</td>
</tr>
<tr>
<td>4.7</td>
<td>Results</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>Searches for neutral Higgs bosons decaying to tau pairs</td>
<td>85</td>
</tr>
<tr>
<td>5.1</td>
<td>Data samples and simulation</td>
<td>86</td>
</tr>
<tr>
<td>5.2</td>
<td>Event selection</td>
<td>88</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Lepton selection and vetoes</td>
<td>89</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Jets and b-tagging</td>
<td>96</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Missing transverse energy</td>
<td>99</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Topological selection</td>
<td>100</td>
</tr>
<tr>
<td>5.3</td>
<td>Di-tau invariant mass reconstruction</td>
<td>102</td>
</tr>
<tr>
<td>5.4</td>
<td>Event categorisation</td>
<td>105</td>
</tr>
<tr>
<td>5.5</td>
<td>Background estimation</td>
<td>108</td>
</tr>
<tr>
<td>5.6</td>
<td>Systematic uncertainties</td>
<td>113</td>
</tr>
<tr>
<td>5.7</td>
<td>Statistical interpretation</td>
<td>121</td>
</tr>
<tr>
<td>5.8</td>
<td>SM search results</td>
<td>123</td>
</tr>
<tr>
<td>5.9</td>
<td>MSSM search results</td>
<td>137</td>
</tr>
<tr>
<td>6</td>
<td>Conclusions</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>Acronyms</td>
<td>159</td>
</tr>
</tbody>
</table>
## List of Figures

In the following figures, uniformly-binned distributions are typically given with the y-axis label “Events”, which is taken to imply event yields per the bin width of the variable in question. The normalisation of distributions with variable-width bins are specified explicitly.

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Tree-level Feynman diagrams for the main Higgs boson production modes at the LHC</td>
</tr>
<tr>
<td>1.2</td>
<td>Cross sections as a function of mass for the main Higgs boson production modes in proton-proton collisions at $\sqrt{s} = 8$ TeV.</td>
</tr>
<tr>
<td>1.3</td>
<td>Branching fractions as a function of mass for the main decays of the Higgs boson.</td>
</tr>
<tr>
<td>1.4</td>
<td>Distributions of the diphoton invariant mass $m_{\gamma\gamma}$ from the CMS $H \rightarrow \gamma \gamma$ search and the four lepton invariant mass $m_{4l}$ from the ATLAS $H \rightarrow ZZ$ search.</td>
</tr>
<tr>
<td>1.5</td>
<td>Feynman diagrams for the radiative corrections to the Higgs boson mass.</td>
</tr>
<tr>
<td>1.6</td>
<td>Masses of the $h$, $H$ and $H^\pm$ bosons in the $m_h^{\text{max}}$ scenario of the MSSM as a function of $m_A$.</td>
</tr>
<tr>
<td>1.7</td>
<td>Tree-level Feynman diagrams for the production of neutral MSSM Higgs bosons via gluon-gluon fusion and in association with b quarks.</td>
</tr>
<tr>
<td>1.8</td>
<td>Cross sections for neutral Higgs boson production in the $m_h^{\text{max}}$ scenario of the MSSM.</td>
</tr>
<tr>
<td>1.9</td>
<td>Branching fractions for the neutral CP-odd Higgs boson in the $m_h^{\text{max}}$ scenario of the MSSM.</td>
</tr>
<tr>
<td>2.1</td>
<td>A schematic of the LHC accelerator complex at the CERN laboratory.</td>
</tr>
<tr>
<td>2.2</td>
<td>Cross sections for several processes in proton-proton and proton-antiproton collisions as a function of centre-of-mass energy.</td>
</tr>
<tr>
<td>2.3</td>
<td>The cumulative integrated luminosity of proton-proton collisions delivered by the LHC to the CMS experiment in the 2010–2013 run period.</td>
</tr>
<tr>
<td>2.4</td>
<td>Distributions of the average number of inelastic pp interactions per bunching crossing in the 2011 and 2012 datasets.</td>
</tr>
<tr>
<td>2.5</td>
<td>Cutaway view of the CMS detector highlighting the layout and position of the major sub-detector components.</td>
</tr>
</tbody>
</table>
2.6  The layout of the tracker in the $r$-$z$ plane. ........................................ 34
2.7  Cutaway view of the electromagnetic calorimeter. .................................. 36
2.8  Layout of one quadrant of the hadron calorimeter in the $r$-$z$ plane. ....... 37
2.9  Layout of one quadrant of the muon system in the $r$-$z$ plane. ............... 39
2.10 Diagram illustrating the components of the L1 trigger, and the sequence of
     reconstruction between each stage. ..................................................... 40
3.1  The resolution of the primary vertex $z$ co-ordinate and vertex reconstruction
     efficiency as a function of the number of constituent tracks. ................... 44
3.2  Distributions of the variables $\Delta \eta_{\text{inv}}, \Delta \phi_{\text{inv}}, \sigma_{\text{inv}}$ and $H/E$
     for genuine electrons and misidentified jets. ...................................... 46
3.3  Measurements of the muon reconstruction and identification efficiency in the
     barrel and endcap regions in $\sqrt{s} = 7$ TeV data and simulation as a function
     of $p_T$. .......................................................................................... 48
3.4  The relative jet energy scale correction as a function of jet $\eta$ and the jet energy
     scale uncertainty as a function of jet $p_T$. ......................................... 52
3.5  Ratio of b-tagging efficiencies for the medium working point of the CSV dis-
     criminator as a function of jet $p_T$. .................................................. 53
3.6  Distributions of the particle flow $E_T^{\text{miss}}$ in $Z \rightarrow \mu \mu$ and $\gamma$+jets events
     in $\sqrt{s} = 8$ TeV data and simulation. ............................................. 54
3.7  Illustration of the components of a proton-proton collision as implemented in
     MC event generators. ......................................................................... 58
4.1  Tree-level Feynman diagrams for $Z+b$ production in the four-flavour scheme.
............................................................................................................. 60
4.2  Tree-level Feynman diagrams for $Z+b$ production in the five-flavour scheme.
............................................................................................................. 60
4.3  Total production cross sections for the b-associated modes $b\bar{b}H$ and $b\bar{b}A$
     in the MSSM for $\sqrt{s} = 7$ TeV. ....................................................... 61
4.4  Expected and observed distributions of the number of jets and the number of
     secondary vertices associated to the leading jet after the dilepton selections.
     .............................................................. ............................... .......................... 64
4.5  Expected and observed distributions of $D_{SSV}$ for jets having at least one sec-
     ondary vertex and the number of b-tagged jets after the dilepton selections. .. 65
4.6  Expected and observed distributions of the $E_T^{\text{miss}}$ and $m_{\ell\ell}$ in events passing
     the $Z+2b$ selection. ............................................................................. 66
4.7  Expected and observed distributions of the $p_T$ of the leading and sub-leading
     b-tagged jets for events passing the $Z+2b$ selection. ............................. 67
4.8  Expected and observed distributions of the difference in azimuthal angle be-
     tween the Z boson candidate and $b\bar{b}$ system for events passing the $Z+2b$
     selection. ......................................................................................... 68
4.9  Expected and observed distributions of the $p_T$ of the dilepton and bb systems
     for events passing the $Z+2b$ selection. .............................................. 69
4.10 Expected and observed distributions of $m_{ee}$ under the $Z+2b$ selection for the $ee$ and $\mu\mu$ selections. ............................................. 70
4.11 Expected and observed distributions of $m_{SV}$ under the $Z+2b$ selection for the leading and sub-leading b-tagged jets. ............................................. 71
4.12 Comparisons of the CMS measured cross sections for $Z+1b$ and $Z+2b$ production with predictions from MCFM, AM@NLO and MADGRAPH. ................. 84
5.1 The number of reconstructed vertices in $\mu\tau_h$ events in the 7 TeV and 8 TeV datasets. ............................................. 89
5.2 The $p_T$ and $\eta$ distributions of the electron candidate in $e\tau_h$ events. ................. 91
5.3 The $p_T$ and $\eta$ distributions of the muon candidate in $e\mu$ events. ................. 92
5.4 The $p_T$ and $\eta$ distributions of the $\tau_h$ candidate in $\mu\tau_h$ events. ................. 92
5.5 The invariant mass of $\tau_h$ candidates in $\mu\tau_h$ events in the 8 TeV dataset. ................. 94
5.6 The efficiency of the isolation requirement for genuine $\tau_h$ compared to jets from the QCD background. ............................................. 95
5.7 Distribution of the $\beta$ variable for real and pileup jets and curves of the real jet efficiency compared to pileup jet rejection for the multivariate pileup discriminator. 97
5.8 Distributions of the number of b-tagged jets and the leading b-tagged jet $p_T$ in $e\mu$ events in the 8 TeV dataset. ............................................. 98
5.9 Distribution of the transverse mass between the selected muon and $\vec{E}_{T}^{miss}$ in the $\mu\tau_h$ channel. ............................................. 101
5.10 A diagram illustrating the definition of the $p_{T}^{vis}$ and $p_{T}^{miss}$ variables. ................. 101
5.11 Distributions of the variable $D_{\zeta}$ and the $t\bar{t}$ rejection BDT output in $e\mu$ channel events. ............................................. 102
5.12 Comparison of the visible and SVFit mass distributions in the $\mu\tau_h$ channel after the baseline selection. ............................................. 104
5.13 Diagram summarising the categorisation of events in the SM analysis. ................. 105
5.14 Distributions of the dijet invariant mass $m_{jj}$ and pseudorapidity separation $|\Delta\eta_{jj}|$ in $\mu\tau_h$ events which contain at least two jets. ................. 106
5.15 Distributions of the di-tau transverse momentum $p_{T}^{\tau\tau}$ in $\mu\tau_h$ channel events and the missing transverse energy $E_{T}^{miss}$ in $e\tau_h$ channel events. ................. 108
5.16 Distributions of $m_{\tau\tau}$ in the same-charge region for the 1-jet low-$p_{T}\tau$ and loose VBF tag categories of the $\mu\tau_h$ channel. ............................................. 111
5.17 Distribution of the number of reconstructed jets, having $p_T > 30$ GeV and $|\eta| < 4.7$, in $e\mu$ channel events. ............................................. 112
5.18 Example of the effect of the $\tau_h$ energy scale uncertainty on the shape of the $Z \rightarrow \tau\tau$ prediction in the 1-jet high-$p_{T}\tau$ category of the $\mu\tau_h$ channel. ................. 116
5.19 Distributions of $m_{\tau\tau}$ for the $Z \to ee$ and $Z \to \mu\mu$ processes in the 0-jet low-$p_T^{\ell_h}$ categories of the $e\tau_h$ and $\mu\tau_h$ channels respectively. .......................... 117
5.20 The result of a fit to the upper tail of the $m_{\tau\tau}$ distribution for the $W^+\text{jets}$ prediction in the B-Tag category of the $\mu\tau_h$ channel. ......................... 120
5.21 Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $\mu\tau_h$ channel. ................................................................. 125
5.22 Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $\mu\tau_h$ channel. ................................................................. 126
5.23 Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $e\tau_h$ channel. 127
5.24 Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $e\tau_h$ channel. 128
5.25 Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $e\mu$ channel. 129
5.26 Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $e\mu$ channel. 130
5.27 The combination of all $m_{\tau\tau}$ distributions from each category and channel. . . . 132
5.28 Expected 95% CL upper limits on the signal strength parameter $\mu$. .............. 133
5.29 Expected and observed 95% CL upper limits on the signal strength parameter $\mu$ as a function of $m_{h}.$. ......................................................... 134
5.30 Expected and observed local p-values under the background-only hypothesis, as a function of $m_{h}.$. ......................................................... 134
5.31 Scans of the negative log-likelihood, $-2\Delta \ln L$ as a function of signal strength $\mu$. 135
5.32 Best-fit values of the signal strength parameter $\mu$ in each channel-category combination and distribution of a test statistic $q_{\mu}$, which asses the compatibility of the individual measurements with the hypothesis of a single value of $\mu$. . . . 136
5.33 Likelihood scans in the two-dimensional parameter space $(\kappa_V, \kappa_I)$ for the coupling of the Higgs boson to vector bosons and fermions respectively. . . . 137
5.34 Observed and predicted $m_{\tau\tau}$ distributions for the MSSM 7 TeV analysis in the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels. ......................................................... 139
5.35 Observed and predicted $m_{\tau\tau}$ distributions for the MSSM 8 TeV analysis in the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels. ......................................................... 140
5.36 Expected and observed 95% CL upper limits in the $m_A$-tan $\beta$ parameter space of the $m_{h}^{\text{max}}$ scenario. ......................................................... 141
5.37 The excluded and allowed regions of the $m_A$-tan $\beta$ parameter space in the $m_{h}^{\text{max}}$ and $m_{h}^{\text{mod+}}$ scenarios. ......................................................... 142
5.38 Observed and expected 95% CL upper limits on $\sigma \cdot \mathcal{B}(\Phi \to \tau\tau)$ for a neutral Higgs boson $\Phi$ produced via gluon-gluon fusion or in association with b-quarks. 142
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Summary of the hadronic tau decay modes.</td>
<td>55</td>
</tr>
<tr>
<td>4.1</td>
<td>Electron identification requirements for the barrel and endcap regions.</td>
<td>63</td>
</tr>
<tr>
<td>4.2</td>
<td>The expected and observed event yields in the Z+1b and Z+2b selections.</td>
<td>67</td>
</tr>
<tr>
<td>4.3</td>
<td>Values, expressed as percentages, of the b-tagging efficiency factors given with the total estimated uncertainties.</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>Values, expressed as percentages, of the lepton selection efficiency factors given with the total estimated uncertainties.</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Values, expressed as percentages, of the reconstruction efficiency factors given with the total estimated uncertainties.</td>
<td>78</td>
</tr>
<tr>
<td>4.6</td>
<td>Lepton acceptance factors that correct the yields in both channels to a common acceptance of $p_T &gt; 20$ GeV and $</td>
<td>\eta</td>
</tr>
<tr>
<td>4.7</td>
<td>Lepton acceptance factors that are used to entirely remove the lepton $p_T$ and $\eta$ requirements.</td>
<td>80</td>
</tr>
<tr>
<td>4.8</td>
<td>Fractional uncertainties on the hadron-level event yields in each channel and b-jet multiplicity bin.</td>
<td>82</td>
</tr>
<tr>
<td>4.9</td>
<td>Cross-section results for each b-jet multiplicity bin and channel.</td>
<td>83</td>
</tr>
<tr>
<td>5.1</td>
<td>Simulated samples used in the SM and MSSM analyses and the corresponding theoretical cross sections.</td>
<td>88</td>
</tr>
<tr>
<td>5.2</td>
<td>Summary of the lepton selections in each channel.</td>
<td>96</td>
</tr>
<tr>
<td>5.3</td>
<td>Summary of sources of systematic uncertainty in the SM analysis.</td>
<td>114</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary of sources of systematic uncertainty in the MSSM analysis.</td>
<td>115</td>
</tr>
<tr>
<td>5.5</td>
<td>The observed and expected yields in the event categories of the SM analysis.</td>
<td>131</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 The standard model

The standard model (SM) is a renormalisable quantum field theory which describes the electromagnetic (EM), weak nuclear and strong nuclear forces and their interaction with matter. The forces are represented by spin-1 bosonic fields and the matter particles by spin-$\frac{1}{2}$ fermionic fields. The dynamics of these fields and their interactions are described by the Lagrangian density $\mathcal{L}_{SM}$, which may be divided into four parts,

$$\mathcal{L}_{SM} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}},$$

(1.1)

each of which will be described in this section. At the heart of the SM is the principle of local gauge invariance and the correspondence between symmetries of the Lagrangian and conserved currents in particle interactions.

1.1.1 Particles and forces

There are twelve fundamental spin-$\frac{1}{2}$ fermions, six leptons and six quarks, arranged in three generations. There is one negatively charged lepton in each generation, the electron ($e$), muon ($\mu$) and tau ($\tau$), each with a corresponding uncharged neutrino ($\nu_e, \nu_\mu, \nu_\tau$). The quarks are the up ($u$), charm ($c$), and top ($t$), each carrying a charge $+\frac{2}{3}e$, paired with the down ($d$), strange ($s$) and bottom ($b$), each carrying a charge $-\frac{1}{3}e$. For every particle there is also a corresponding antiparticle with opposite quantum numbers, including electric charge.

The gauge group of the SM is $SU(3)_C \times SU(2)_L \times U(1)_Y$. These symmetries imply the presence of spin-1 gauge bosons, which are the carrier particles of the fundamental forces. The $SU(3)_C$
symmetry governs the strong interaction between the massless gluon mediator bosons \((g)\) and the quarks, all of which carry colour charge. The leptons do not carry colour charge and so do not interact via the strong force. The corresponding quantum field theory is Quantum Chromodynamics (QCD), characterised by the properties of confinement, in which quarks are never observed in isolation but rather in bound hadron states, and asymptotic freedom [9,10], in which the coupling strength decreases at high energy. More detailed introductions to QCD may be found in [11,12].

The \(SU(2)_L \times U(1)_Y\) symmetry reflects the unification of the weak and EM forces, mediated by the \(W^\pm\) and \(Z\) bosons and the photon respectively. Introduced in the 1960s by Glashow [13], Weinberg [14] and Salam [15], it was a landmark achievement in the development of the SM. In particular, the electroweak theory implied the presence of neutral weak-current interactions. These were subsequently discovered with the Gargamelle bubble chamber experiment at CERN in 1973 [16]. This was followed by the discovery of the \(W^\pm\) and \(Z\) bosons by the UA1 and UA2 Collaborations at CERN in 1983 [17–20].

An important feature of the weak force is that it is chiral; it couples only to the left-handed components of the fermion fields, thus maximally violating parity conservation. The left- and right-handed components, \(\psi_L\) and \(\psi_R\), of a fermion field \(\psi\) may be extracted by the projection operators \(\frac{1}{2}(1 - \gamma^5)\) and \(\frac{1}{2}(1 + \gamma^5)\) respectively. In these operators \(\gamma^5\) is defined as the product of the four Dirac gamma matrices, \(\gamma^5 = i \gamma^0 \gamma^1 \gamma^2 \gamma^3\) [11]. As a consequence, there are no right-handed neutrinos in the SM because they would not interact via any of the EM, weak or strong forces. However, since the formulation of the SM, neutrino flavour oscillations have been observed in solar, atmospheric, reactor and beam sources. This implies that neutrinos have very small but non-zero masses and therefore the presence of right-handed neutrinos in the SM Lagrangian. This observation is omitted in the subsequent discussion; however, a more detailed description of the impact of neutrino mass is found in [21].

The weak \(SU(2)_L\) transformations, also called weak isospin, act on doublets \(L_m\) of the left-handed lepton pairs, where

\[
L_m = \begin{pmatrix} v_m \\ l_m \end{pmatrix}_L := \begin{pmatrix} v_e \\ e \end{pmatrix}_L, \begin{pmatrix} v_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} v_\tau \\ \tau \end{pmatrix}_L,
\]

and \(m\) runs over the three generations. Conversely the right-handed charged leptons transform as singlet states

\[
l_{Rm} := e_{Rm}, \mu_{Rm}, \tau_{Rm}.
\]
Correspondingly for the quarks there are three left-handed doublets,

\[ Q_m = \begin{pmatrix} u'_m \\ d'_m \end{pmatrix}_L \equiv \begin{pmatrix} u' \\ c' \\ t' \end{pmatrix}_L, \]

and six right-handed singlets,

\[ u'_{Rm} := u'_R, c'_R, t'_R \]

\[ d'_{Rm} := d'_R, s'_R, b'_R. \]

The prime notation is introduced to distinguish the weak eigenstates from the physical mass states, which mix via the Cabibbo-Kobayashi-Maskawa (CKM) matrix in charged-current interactions \[21\]. The mixing between the mass and flavour eigenstates of the neutrinos is specified by a lepton analogue of the CKM matrix, known as the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix \[22,23\].

The generators of the \( SU(3)_C \) group are the \( 3 \times 3 \) Gell-Mann matrices denoted as \( \frac{1}{2} \lambda^a \), where \( a = 1 \ldots 8 \). The generators of the \( SU(2)_L \) transformation are built from the \( 2 \times 2 \) Pauli spin matrices as \( T^i = \frac{1}{2} \sigma^i \) where \( i = 1, 2, 3 \). The \( U(1)_Y \) is not the group of the EM interaction, but is known as the weak hypercharge with generator \( Y \). It will be shown that the \( U(1)_{EM} \) generator emerges in the spontaneous symmetry breaking and can be defined as a mixture of the hypercharge and one of the weak isospin generators as \( Q_{EM} = T^3 + Y \). From this relationship the hypercharges \( y \) of the \( L_m, I_{Rm}, Q_m, u'_{Rm} \) and \( d'_{Rm} \) states are trivially found to be \( -\frac{1}{2}, -1, +\frac{1}{2}, +\frac{3}{2} \) and \( -\frac{3}{2} \) respectively. The part of the Lagrangian containing the fermion dynamics, \( \mathcal{L}_{fermion} \), is constructed by first considering the complete set of local gauge transformations on the doublet and singlet states, denoted generically as \( \chi_L \) and \( \psi_R \):

\[ \chi_L \rightarrow \exp \left( \frac{i}{2} \mathbf{\theta} \cdot \mathbf{\lambda} + \frac{i}{2} \mathbf{\zeta} \cdot \mathbf{\sigma} + i y \rho \right) \chi_L \]  

(1.7)

\[ \psi_R \rightarrow \exp \left( \frac{i}{2} \mathbf{\theta} \cdot \mathbf{\lambda} + i y \rho \right) \psi_R. \]  

(1.8)

The parameters \( \mathbf{\theta}, \mathbf{\zeta} \) and \( \rho \) are taken to be functions of the space-time co-ordinates. Covariant derivatives for each field require the introduction of the gauge fields for each symmetry group. These are the gluon fields \( G^a_\mu \) for \( SU(3)_C \), the weak boson fields \( W^i_\mu \) for \( SU(2)_L \) and the weak hypercharge field \( B_\mu \) for \( U(1)_Y \). Under infinitesimal gauge variations they are required to
transform as

\[
G^a_\mu \rightarrow G^a_\mu + \frac{1}{g_c} \partial_\mu \theta^a - f^{abc} \theta^b G^c_\mu
\]  
(1.9)

\[
W^i_\mu \rightarrow W^i_\mu + \frac{1}{g} \partial_\mu \xi^i - e^{ijk} \xi^j G^k_\mu
\]  
(1.10)

\[
B_\mu \rightarrow B_\mu + \frac{1}{g} \partial_\mu \rho,
\]  
(1.11)

where \(g_c\), \(g\) and \(g'\) are the coupling constants associated with each symmetry group and \(f^{abc}\) and \(e^{ijk}\) are the \(SU(3)_C\) and \(SU(2)_L\) structure constants respectively. The covariant derivatives are then defined as:

\[
D_\mu L_m = \left[ \partial_\mu - i \frac{g}{2} W^\mu \cdot \sigma + \frac{ig'}{2} B^\mu \right] L_m
\]  
(1.12)

\[
D_\mu l_{Rm} = \left[ \partial_\mu + ig' B^\mu \right] l_{Rm}
\]  
(1.13)

\[
D_\mu Q_m = \left[ \partial_\mu - i \frac{g_c}{2} G^\mu \cdot \lambda - i \frac{g}{2} g' W^\mu \cdot \sigma - i \frac{g'}{6} B^\mu \right] Q_m
\]  
(1.14)

\[
D_\mu u'_{Rm} = \left[ \partial_\mu - i \frac{g_c}{2} G^\mu \cdot \lambda - \frac{2ig'}{3} g' B^\mu \right] u'_{Rm}
\]  
(1.15)

\[
D_\mu d'_{Rm} = \left[ \partial_\mu - i \frac{g_c}{2} G^\mu \cdot \lambda + \frac{ig'}{3} B^\mu \right] d'_{Rm}
\]  
(1.16)

From these definitions the fermion part of the Lagrangian is given as

\[
L_{\text{fermion}} = i \bar{L}_m \slashed{D} L_m + i \bar{l}_{Rm} \slashed{D} l_{Rm} + i \bar{Q}_m \slashed{D} Q_m + i \bar{u}'_{Rm} \slashed{D} u'_{Rm} + i \bar{d}'_{Rm} \slashed{D} d'_{Rm},
\]  
(1.17)

where \(\slashed{D} = \gamma^\mu D_\mu\). Kinetic terms for the gauge fields are built from the field strength tensors, defined as

\[
G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^b_\mu G^c_\nu
\]  
(1.18)

\[
W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g e^{ijk} W^j_\mu W^k_\nu
\]  
(1.19)

\[
B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu,
\]  
(1.20)

and give the gauge part of the Lagrangian as

\[
L_{\text{gauge}} = -\frac{1}{4} G^a_{\mu\nu} \cdot G^{a\mu\nu} - \frac{1}{4} W^i_{\mu\nu} \cdot W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}.
\]  
(1.21)
The physical charged bosons $W_{\pm}^\mu$ are found as combinations of $W_{\mu}^1$ and $W_{\mu}^2$, and the neutral $Z_{\mu}$ boson and the photon $A_{\mu}$ are found as a mixture of $W_{\mu}^3$ and $B_{\mu}$:

$$W_{\mu}^\pm = \frac{1}{\sqrt{2}}(W_{\mu}^1 \mp iW_{\mu}^2)$$  
(1.22)

$$Z_{\mu} = W_{\mu}^3 \cos \theta_w - B_{\mu} \sin \theta_w$$  
(1.23)

$$A_{\mu} = W_{\mu}^3 \sin \theta_w + B_{\mu} \cos \theta_w,$$  
(1.24)

where $\theta_w$ is known as the weak mixing angle and is related to the coupling constants by

$$\sin \theta_w = \frac{g'}{\sqrt{g^2 + g'^2}} \quad \cos \theta_w = \frac{g}{\sqrt{g^2 + g'^2}}.$$  
(1.25)

The Lagrangians in equations 1.17 and 1.21 describe all interactions between the gauge fields and the fermions, but crucially do not contain mass terms for the weak force bosons or any of the fermions. Attempts to add gauge boson mass terms of the form $-M^2 W_{\mu} W^\mu$ break the gauge invariance of the Lagrangian. Fermion mass terms of the form

$$-m \bar{\psi} \psi = -m (\bar{\psi}_L \psi_L + \bar{\psi}_R \psi_R)$$  
(1.26)

contain field pairs that transform differently under the $SU(2)_L$ and $U(1)_Y$ groups and so are also not gauge invariant. Furthermore, the fact that the photon is massless and the $W^\pm$ and $Z$ bosons are known to be massive implies that the electroweak symmetry must be broken.

1.1.2 The Higgs mechanism

The masses of the weak vector bosons can be generated by a process known as the Higgs mechanism. It was first suggested in 1964 in three landmark papers by Englert and Brout [24]; Higgs [25]; and Guralnik, Hagen and Kibble [26]. The principle is that while the addition of mass terms explicitly breaks the symmetry of the Lagrangian, it would be possible to add a field, symmetric under the gauge transformations, that acquires a non-zero expectation value in the vacuum state and breaks the symmetry. A consequence of this spontaneous symmetry breaking is that by Goldstone’s theorem [27] each broken generator introduces a new massless scalar particle. In the Higgs mechanism these extra degrees of freedom are “eaten” by the gauge fields which become massive.
To break the electroweak symmetry, but leave the EM interaction intact, the simplest field that can be introduced is a complex $SU(2)_L$ doublet

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}.$$ \hfill (1.27)

The kinetic Lagrangian for the field is given by the general form

$$\mathcal{L}_{\text{kin}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi^\dagger \phi)$$ \hfill (1.28)

where under local $SU(2)_L \times U(1)_Y$ symmetry transformations it carries a weak hypercharge of $y = \frac{1}{2}$. This transformation is given as

$$\begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \longrightarrow \exp \left( i \zeta \cdot \frac{\sigma}{2} + i \rho \frac{1}{2} \right) \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$$ \hfill (1.29)

which requires the covariant derivative to be

$$D_\mu = \partial_\mu - i g W_\mu \cdot \sigma - i g' B_\mu.$$ \hfill (1.30)

The potential term $V(\phi^\dagger \phi)$ is chosen to take the form

$$V(\phi^\dagger \phi) = \lambda (\phi^\dagger \phi)^2 - \mu^2 (\phi^\dagger \phi)$$ \hfill (1.31)

where $\lambda$ and $\mu^2$ are real, $\lambda > 0$ is required for the vacuum to be stable and $\mu^2 > 0$ should be chosen to induce spontaneous symmetry breaking. In the vacuum state the field now acquires a non-zero expectation value, which may be arbitrarily chosen as

$$\langle 0 | \phi | 0 \rangle = \phi_0 = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}} \nu \end{pmatrix} \quad \nu \text{ real, } \nu^2 = \frac{\mu^2}{\lambda}.$$ \hfill (1.32)

By considering infinitesimal transformations about this vacuum state it can be shown that the only remaining symmetry is the linear combination $T_3 + Y$, which is the $U(1)_{\text{EM}}$ group as expected. The remaining generators $T_1$, $T_2$ and $T_3 - Y$, denoted collectively as $k$, do not leave
the vacuum state invariant. The field $\phi$ may then be re-parameterized as an expansion about $\phi_0$:

$$
\phi_0 = \exp \left( \frac{i}{\sqrt{2}v} \zeta \cdot k \right) \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + H) \end{pmatrix}.
$$

(1.33)

The Goldstone fields $\zeta$ are eliminated by an appropriate choice of gauge and become the longitudinal degrees of freedom for the $W^\pm$ and $Z$ bosons. This leaves the vacuum state containing a single scalar field $H$:

$$
\phi_0 = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + H) \end{pmatrix}.
$$

(1.34)

Substituting this expression into the Lagrangian given by equations 1.28, 1.30 and 1.31, and keeping at most the quadratic terms in the fields, gives

$$
\mathcal{L}_{\text{Higgs}} = \frac{1}{2} \partial_\mu H \partial^\mu H - \mu^2 H^2 + \frac{1}{4} g^2 v^2 W^\pm_\mu W^{-\mu} + \frac{1}{8} v^2 (g^2 + g'^2) Z_\mu Z^\mu
$$

(1.35)

where the definitions of $W^\pm_\mu$ and $Z_\mu$ in equations 1.22 and 1.23 have been substituted. From this the $W^\pm$ fields are identified with masses $\frac{gv}{2}$ and are charged under $U(1)_{\text{EM}}$. The neutral $Z$ boson acquires a mass

$$
m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2}.
$$

(1.36)

The scalar Higgs boson field $H$ is also massive, with $m_H = \sqrt{2}\mu^2$. Notably no mass term appears for the photon field $A_\mu$.

In order to generate masses for the fermions it is noted that the field $\phi$ permits gauge-invariant terms with left-handed doublet and right-handed singlet states, such as $\overline{\psi}_L \psi_R \phi$, and these are referred to as Yukawa interactions. After spontaneous symmetry breaking the Yukawa sector of the Lagrangian is found to be

$$
\mathcal{L}_{\text{Yukawa}} = -\frac{1}{\sqrt{2}} (v + H) \left( f^l_m \overline{l_{Lm}} l_{Rm} + f^u_m \overline{u_{Lm}} u_{Rm} + f^d_m \overline{d_{Lm}} d_{Rm} \right),
$$

(1.37)

where the $f_m$ are diagonal matrices of coupling parameters. From this, mass terms of the form $-m_f \overline{\psi} \psi$ can be identified where $m_f = \frac{v}{\sqrt{2}}$, as well as interaction terms with the Higgs field of the form $\frac{m_f}{v} \overline{\psi} \psi H$ which are proportional to the mass of the fermion in question.
1.2 Higgs boson searches

In the SM the mass of the Higgs boson is a free parameter. Precision electroweak measurements performed with data from the Large Electron-Positron (LEP), Tevatron and SLAC Large Detector (SLD) experiments have been used to constrain the most likely value of $m_H$ [28], under the assumption that the SM is a complete theory. A best fit of $94^{+29}_{-24}$ GeV is determined where the given uncertainty is due to experimental effects only.

Direct searches were performed by the experiments at the CERN LEP collider, at $e^+e^-$ centre-of-mass energies ($\sqrt{s}$) between 90 and 209 GeV. In such collisions the Higgs boson is primarily produced in association with a Z boson. The most sensitive search channels exploited the decays to $b\bar{b}$ and $\tau^+\tau^-$ pairs. The combination of results from all four experiments [29] led to the exclusion of masses with $m_H < 114.4$ GeV at the 95% confidence level (CL). Searches were also performed by the CDF and D0 Collaborations at the Tevatron accelerator, which collided protons with antiprotons at $\sqrt{s} = 1.96$ TeV. These searches were performed in a mass range of 90–200 GeV and focussed on the decays to $b\bar{b}$, $W^+W^-$, $\gamma\gamma$ and $\tau^+\tau^-$ pairs, with the $b\bar{b}$ and $W^+W^-$ channels offering the most sensitivity. The combined results [30] yielded exclusions of $m_H$ in the ranges 90–109 GeV and 140–184 GeV, also at the 95% CL.

![Figure 1.1](image-url)  

**Figure 1.1:** Tree-level Feynman diagrams for the main Higgs boson production modes at the LHC in descending order of cross section: (a) gluon-gluon fusion, (b) vector boson fusion, (c) $W/Z$-associated production and (d) $t\bar{t}$-associated production.

The main Higgs boson production modes at the LHC, illustrated in figure 1.1, are gluon-gluon fusion, vector boson fusion (VBF), $W/Z$-associated and $t\bar{t}$-associated production. Cross sections
for each process are given in figure 1.2 as a function of $m_H$ for proton-proton collisions at $\sqrt{s} = 8$ TeV. The dominant mode is gluon-gluon fusion, which proceeds via a quark loop and has a cross section of $\mathcal{O}(10 \text{ pb})$ for $m_H < 200$ GeV. The remaining processes have cross sections one to two orders of magnitude smaller, however they have distinctive topologies which can be exploited. VBF is characterised by the presence of two high-momentum outgoing quarks which hadronise to form jets, typically in the forward detector regions. Production in association with a $W$ or $Z$ boson via the “Higgsstrahlung” process, or in association with a $t\bar{t}$ pair, leads to multi-lepton and multi-jet final states with reduced SM backgrounds.

![Cross sections as a function of mass for the main Higgs boson production modes in proton-proton collisions at $\sqrt{s} = 8$ TeV][1]. The precision of the QCD and electroweak (EW) corrections to each process is also given, which may be next-to-leading order (NLO) or next-to-next-to-leading order (NNLO), as well as the use of next-to-next-to-leading logarithmic (NNLL) resummation.

The branching fractions of the different decay channels are given in figure 1.3 as a function of $m_H$ in the range 80–200 GeV. At higher mass the decays to $W^+W^-$ and $ZZ$ pairs increasingly dominate as they become kinematically favourable, whereas for $m_H < 130$ GeV several processes are experimentally accessible. These include the decays to two photons via a fermion loop, to $W^+W^-$, $ZZ$, $b\bar{b}$ and $\tau^+\tau^-$ pairs. The decays to fermions are particularly important for unambiguously establishing the Yukawa couplings in the SM. As the Higgs coupling is proportional to mass, the $b\bar{b}$ decay dominates with a branching fraction of $\sim 80\%$ in this mass range. However, the hadronic final state is difficult to disentangle from the large QCD jet background at the LHC, meaning the $\tau^+\tau^-$ channel has higher sensitivity, but is not without its own experimental
challenges. Tau leptons have a lifetime of $2.9 \times 10^{-15}$ s [21] so cannot be detected directly, but rather via the weak decays to hadrons (64.8%), electrons (17.8%) and muons (17.4%) in which one or two neutrinos are also produced.

On 4 July 2012 the ATLAS and CMS Collaborations announced the discovery of a new boson with mass around 125 GeV, based on analysis of approximately 5 fb$^{-1}$ of data collected at $\sqrt{s} = 7$ TeV and 5–6 fb$^{-1}$ at $\sqrt{s} = 8$ TeV [32, 33]. This was prompted by excesses of events observed in the $H \rightarrow ZZ$ and $H \rightarrow \gamma\gamma$ channels at both experiments, with combined significances in excess of five standard deviations from the background-only expectation. Figure 1.4a gives the weighted distribution of the diphoton invariant mass from the CMS results and figure 1.4b the four-lepton invariant mass from the ATLAS results. Though these channels have smaller branching fractions, both have final states with clean signatures that benefit from the excellent photon and lepton energy resolution provided by the detectors.

With the analysis of subsequent data a picture of increasing consistency with the SM has been building. Results have been presented on the production rates and couplings in several channels [34–36], studies of the spin-parity quantum numbers [35–37], as well as limits on the decay width [38] and invisible branching fraction [39, 40]. Taken together, these results confirm that the new particle is indeed a Higgs boson, and future studies will further test the compatibility with the SM.

Figure 1.3: Branching fractions as a function of mass for the main decays of the Higgs boson. The solid bands reflect the uncertainties in the branching fraction calculations [31].
Introduction

Although extremely successful and tested to a high degree of precision [21], the SM fails to address a number of issues and observations. For example:

- It does not contain a candidate for the large fraction of non-radiating, non-baryonic dark matter in the universe. Recent results from a study of the cosmic microwave background with the Planck space telescope indicate this fraction is around 85% [41]. A large number of direct and indirect dark matter detection experiments are currently in operation [42].

- If the neutrino masses are generated in the same way as the other fermions then it is unclear why there is an orders-of-magnitude difference in the strength of the Yukawa couplings. Other neutrino mass-generating mechanisms have been proposed [21] which address this.

- The $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge group is a direct product of three simple groups and the associated electroweak and strong coupling constants do not intersect at a high energy scale, as would be hoped for in a unified theory.

- The force of gravity, best described by general relativity, does not appear in the SM.

Of particular relevance to the Higgs sector is the hierarchy problem. It is generally accepted that the SM is an effective theory up to some energy scale $\Lambda$ where the effect of new physics becomes important. At most this could be the Planck scale ($\sim 10^{19}$ GeV) where the quantum
effects of gravity are unavoidable [11]. It can be shown that the scale $\Lambda$ enters the calculation of the loop corrections to $m_H$ from both boson and fermion contributions [43], as illustrated in figure 1.5. These corrections to the tree-level mass are quadratically divergent, with $\Delta m_H^2 \sim \Lambda^2$. This means that observing a Higgs boson with mass $\mathcal{O}(100\text{ GeV})$, while $\Lambda$ is many orders of magnitude higher, requires a precise fine-tuning of the bare mass term $2\mu^2$ to achieve the necessary cancellation.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{feynman_diagrams.png}
\caption{Feynman diagrams for the radiative corrections to the Higgs boson mass from (a) fermion couplings and (b) the scalar self coupling.}
\end{figure}

Many beyond-the-standard-model (BSM) solutions to the hierarchy problem have been proposed, and one of the most popular is supersymmetry (SUSY), in which a new symmetry between fermions and bosons is introduced [44]. An important feature of the Higgs boson mass correction is that fermion and boson loops contribute to $\Delta m_H^2$ with opposite sign. In SUSY models every SM fermion has a boson superpartner and vice versa. The effect is that the divergent terms in $\Delta m_H^2$ from each particle-superpartner pair cancel. If SUSY is unbroken then the superpartners have exactly the same masses as their SM partners. The fact that none have been observed by experiment suggests that SUSY is in fact a broken symmetry in which the superpartner masses are larger than their SM counterparts. However the SUSY breaking scale cannot be much larger than $\mathcal{O}(1\text{ TeV})$, otherwise further fine-tuning would be required. As further motivation the lightest SUSY particles, if stable, would provide a candidate for dark matter [44]. The behaviour of the running coupling constants is also modified such that all three intersect at a common scale of $\mathcal{O}(10^{16}\text{ GeV})$ [45].

The simplest addition of SUSY to the SM results in the minimal supersymmetric standard model (MSSM). In the Higgs sector of the MSSM two complex doublet fields are required in order to generate the appropriate mass terms for the up-type and down-type fermions. These are chosen as

$$
\phi_u = \left( \begin{array}{c} \phi_u^+ \\ \phi_u^0 \end{array} \right), \quad \phi_d = \left( \begin{array}{c} \phi_d^0 \\ \phi_d^- \end{array} \right).
$$

(1.38)
Introduction

Analogously to the SM case, an appropriately chosen potential for these fields leads to spontaneous symmetry breaking, where the vacuum expectation values may be chosen as

\[
\phi_u^0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v_u \end{pmatrix}, \quad \phi_d^0 = \frac{1}{\sqrt{2}} \begin{pmatrix} v_d \\ 0 \end{pmatrix}
\]

(1.39)

where \(v_u\) and \(v_d\) are related to the SM value by \(v^2 = v_u^2 + v_d^2\). Of the eight initial degrees of freedom, three become the longitudinal states of the \(W^\pm\) and \(Z\) bosons which leaves five massive Higgs fields. These are the two neutral CP-even states \(h\) and \(H\), one neutral CP-odd state \(A\) and a charged pair \(H^\pm\). At tree level all properties of the MSSM Higgs sector are specified by two free parameters. These are typically chosen as the mass of the CP-odd Higgs boson, \(m_A\), and the ratio of the vacuum expectation values, \(\tan \beta = v_u/v_d\). The other masses are then determined as

\[
m_{H^\pm}^2 = m_A^2 + m_W^2 \quad (1.40)
\]

\[
m_{H,h}^2 = \frac{1}{2} \left( m_A^2 + m_Z^2 \pm \sqrt{(m_A^2 + m_Z^2)^2 - 4m_Z^2 m_A^2 \cos^2 2\beta} \right). \quad (1.41)
\]

The equation for \(m_{H,h}\) leads to an upper bound on the mass of the light CP-even state \(h\),

\[
m_h \leq m_Z \cos 2\beta \leq m_Z. \quad (1.42)
\]

However, the loop corrections to this mass term can be quite large, depending on several other SUSY parameters, and allow for values of \(m_h\) up to around 135 GeV [43]. As the number of MSSM parameters that may enter here is quite large, it is typical to study Higgs sector predictions in scenarios where the relevant parameters are fixed to benchmark values. The results can then be interpreted within the \(m_A\)-\(\tan \beta\) plane. The \(m_{h_{\text{max}}}^\text{max}\) scenario sets these parameters such that \(m_h\) obtains its maximum value through these radiative corrections [46]. Figure 1.6 shows the \(h, H \text{ and } H^\pm\) masses as a function of \(m_A\) for representative \(\tan \beta\) values of 3 and 30. In the decoupling limit, where \(m_A \gg m_Z\), it is found that \(m_A \approx m_{h_{\text{max}}} \approx m_{H^\pm}\) and that the light CP-even \(h\) has couplings close to those in the SM. The couplings of the \(H\) and \(A\) bosons to down-type fermions are enhanced by a factor \(\sim \tan \beta\) relative to the SM value.

In the MSSM the main neutral Higgs boson production modes are via gluon-gluon fusion and \(b\)-associated production, for which example tree-level Feynman diagrams are given in figure 1.7. The cross sections at \(\sqrt{s} = 8\) TeV for both modes are given in figures 1.8a and 1.8b as a function of mass for \(\tan \beta = 5\) and \(\tan \beta = 30\) respectively [31]. The gluon-gluon fusion production dominates at lower values of \(\tan \beta\), while at high \(\tan \beta\) the \(b\)-associated mode can be greatly enhanced. The decays of the \(A\) and \(H\) bosons are dominated by the \(b\bar{b}\) and \(\tau^+\tau^-\) channels for
Figure 1.6: Masses of the h, H and H⁺ bosons in the $m_h^{\max}$ scenario of the MSSM as a function of $m_A$. Values are given for tan $\beta = 3$ and tan $\beta = 30$ [43].

Figure 1.7: Tree-level Feynman diagrams for production of neutral MSSM Higgs bosons via (a) gluon-gluon fusion and (b) in association with b quarks.
larger values of $\tan \beta$. Figure 1.9 gives the branching fractions for the CP-odd $A$ boson at $\tan \beta$ values of 10 and 50. In the latter, the $\tau^+\tau^-$ decay is found to have a branching fraction of $\sim 11\%$ for $m_A$ up to 600 GeV, providing a strong experimental motivation for searches in this channel.

**Figure 1.8:** Cross sections for neutral Higgs boson production in the $m_h^{\text{max}}$ scenario of the MSSM for (a) $\tan \beta = 5$ and (b) $\tan \beta = 30$ [31].

**Figure 1.9:** Branching fractions for the neutral CP-odd Higgs boson in the $m_h^{\text{max}}$ scenario of the MSSM for (a) $\tan \beta = 10$ and (b) $\tan \beta = 50$ [31].
Chapter 2

The CMS detector

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [47] is a 26.7 km circumference synchrotron designed to collide beams of protons at centre-of-mass energies up to 14 TeV. It is operated by the European Organization for Nuclear Research (CERN), near Geneva, Switzerland, and housed in the circular tunnel formerly occupied by the Large Electron Positron (LEP) [48] experiment. The tunnel is located 80–150 m underground and straddles the Franco-Swiss border. As well as proton-proton (pp) collisions, the LHC is capable of accelerating beams of lead ions to an energy of 1.38 TeV per nucleon and producing both ion-ion (PbPb) and proton-ion (pPb) collisions.

A schematic of the LHC accelerator complex is given in figure 2.1. The production of the proton beams begins with a bottle of hydrogen gas. An electric field strips the electrons from the hydrogen atoms leaving protons which are accelerated to an energy of 50 MeV in the Linac 2 accelerator. Proton bunches then transition to the Proton Synchrotron Booster (PSB) which increases the energy to 1.4 GeV. The Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) then raise the energy to 25 GeV and 450 GeV respectively. From the SPS protons are injected as two counter-rotating beams into the LHC ring. At design operation each beam consists of up to 2808 bunches spaced 25 ns apart and made up of $O(10^{15})$ protons each. Eight radio frequency (RF) cavities are responsible for accelerating the beams to collision energy. Keeping the beams circulating requires 1232 niobium-titanium superconducting dipole magnets. Each magnet is 14.3 m long and cooled by superfluid helium to operate at a temperature of 1.9 K and generate magnetic fields up to 8.4 T. The beams cross at four points around the LHC where collisions are recorded by the ALICE [49], ATLAS [50], CMS [51] and LHCb [52] detectors.
The LHC is a discovery machine built to explore physics at the TeV scale. ATLAS and CMS are general-purpose detectors designed to search for the Higgs boson, and thus elucidate the mechanism of electroweak symmetry breaking, and to search for evidence of BSM signatures. The ALICE detector is designed to analyse the results of heavy-ion collisions and to probe the strong interaction at extreme energy densities. The LHCb detector is used to study the properties and decays of b-flavour hadrons. For example, anomalies in the rates of rare decays could provide indirect evidence for new physics and measurements of CP violation may lead to a better understanding of the matter-antimatter asymmetry in the universe.

Many processes of interest are expected to occur only rarely at the LHC. Figure 2.2 gives the cross sections for several processes in pp collisions as a function of centre-of-mass energy. The cross section for SM Higgs boson production, for example, is approximately nine orders of magnitude smaller than the total pp cross section. Therefore, to accumulate a sufficient number of interesting events, the LHC operates with a high instantaneous luminosity, $L$, defined as

$$L = \frac{N_b n \beta_{	ext{rev}} y}{4 \pi \epsilon_n \beta^*} F,$$  

(2.1)
where \( N_b \) is the number of protons in each bunch, \( n_b \) is the number of bunches, \( f_{\text{rev}} \) is the revolution frequency, \( \gamma \) is the Lorentz factor, \( \epsilon_n \) is the normalised emittance, \( \beta^* \) is the beta function at the collision point and \( F \) is a reduction factor due to the crossing angle. The LHC is designed to operate at instantaneous luminosities up to \( 10^{34} \text{ cm}^{-2} \text{ s}^{-1} \). The first beams were successfully circulated around the entire LHC ring in September 2008. However, before the first collisions could be achieved a faulty connection between two dipole magnets caused a sudden quench and significant mechanical damage which involved a leak of several tonnes of helium into the tunnel. This necessitated a year-long period of extensive repair and reinforcement. For safety reasons the decision was also taken to limit the maximum centre-of-mass energy,
initially to 7 TeV. The first proton-proton collisions were achieved in November 2009 at the injection energy of 450 GeV. Figure 2.3 summarises the integrated luminosity of collisions delivered to the CMS detector during the first LHC run, which concluded in early 2013. In March 2010 collisions at 7 TeV began and continued through to the end of 2011. In this time the LHC achieved a peak luminosity of $3.7 \times 10^{33}$ cm$^{-2}$ s$^{-1}$, delivering 6.1 fb$^{-1}$, of which 5.6 fb$^{-1}$ was recorded by CMS. In 2012 the collision energy was increased to 8 TeV and a luminosity of $7.7 \times 10^{33}$ cm$^{-2}$ s$^{-1}$ was reached by the end of the run, with 23.3 fb$^{-1}$ delivered. During this time the LHC operated with a bunch spacing of 50 ns. It is noted that of the luminosity delivered to CMS, only the data in which all sub-detectors were known to be operating normally is certified for use in physics analysis. This amounts to 5.1 fb$^{-1}$ in 2011 and 19.7 fb$^{-1}$ in 2012. Following a period of upgrades and maintenance, operations are expected to resume in 2015, with the LHC operating close to or at the design centre-of-mass energy of 14 TeV and instantaneous luminosity of $10^{34}$ cm$^{-2}$s$^{-1}$.

![Figure 2.3: The cumulative integrated luminosity of proton-proton collisions delivered by the LHC to the CMS experiment in the 2010–2013 run period. Adapted from [55].](image)

The high luminosity operation of the LHC makes multiple proton-proton collisions in each bunch crossing likely. Figure 2.4 shows distributions of the average number of inelastic interactions per bunch-crossing in the 2011 and 2012 CMS datasets. The distribution for 2011 is also given separately for the earlier part of the year where the instantaneous luminosity was lower, with an average of around 6 interactions observed, and the later part of the year where this average increased to 12. The increase in luminosity for the 2012 run caused a further increase to an average of 20 interactions per bunch crossing. In events where a proton-proton interaction leads to a process of interest the additional interactions, which typically produce softer jets, are referred to as pileup.
The CMS detector was designed for high performance in the search for the Higgs boson and for evidence of new physics at the TeV scale. It is composed of several sub-detector layers arranged in a barrel and endcap configuration and centred on the beam axis, as illustrated in figure 2.5. The detector is 22 m long, 15 m in diameter and weighs 12 500 metric tons [51]. An important feature is the 3.8 T superconducting solenoid, 13 m long and 6 m in diameter, which causes charged particles to follow curved trajectories and allows for a precise measurement of their momentum. Within the bore of the magnet coil is a silicon tracking detector and both electromagnetic and hadronic calorimeters. Outside the coil gaseous muon chambers are interspersed between the iron plates of the return yoke.

Measurements of physical quantities are given in a co-ordinate system which has its origin at the nominal interaction point in the centre of the detector. From here the x-axis points to the centre of the LHC ring, the y-axis points vertically upwards and the z-axis is collinear with the beam direction. In the x-y, or transverse plane the azimuthal angle $\phi$ is measured with respect to the x-axis. Measurements of momentum or energy in the transverse plane are of particular
Figure 2.5: Cutaway view of the CMS detector highlighting the layout and position of the major sub-detector components [51]. The detector is 22 m long and 15 m in diameter.

interest and are denoted as $p_T$ or $E_T$ respectively. The polar angle $\theta$ is measured from the $z$-axis and pseudorapidity is defined as $\eta = -\ln \left[ \tan \left( \frac{\theta}{2} \right) \right]$. Distance in the $\eta$-$\phi$ plane is given by $\Delta R = \sqrt{\Delta \phi^2 + \Delta \eta^2}$.

### 2.3 Tracker

The detector closest to the beam axis is the tracker [51], which records the trajectories of charged particles, including electrons, muons and charged hadrons. With the bending due to the magnetic field it is possible to make precise measurements of particle momentum, with a design resolution of about 1% at 100 GeV. It is also important for the reconstruction of collision vertices, as well as secondary vertices from the decay of longer-lived particles, such as B mesons.

The tracker is comprised of layers of silicon pixel and strip detectors which extend over a pseudorapidity range $|\eta| < 2.5$. A charged particle hit is recorded in a pixel or strip by the production of electron-hole pairs, which drift under an applied electric field and give rise to a current pulse. Silicon technology was chosen in order to provide high granularity in the face of a large particle flux while also being radiation hard. In nominal operations, $O(1000)$ particles
are produced and traverse the tracker volume in each bunch crossing. It also provides a fast response to ensure particle tracks can be associated with the correct bunch crossing.

The layout of the tracker is given in figure 2.6. The innermost component is the pixel detector, comprised of three layers in the barrel region and two in each endcap. The barrel layers are positioned at radii of 4.4 cm, 7.3 cm and 10.2 cm from the beam axis. At the first layer the charged particle flux is \( \approx 10^8 \text{ cm}^{-2} \text{ s}^{-1} \) at design luminosity. Each pixel is 100 \( \mu \text{m} \times 150 \mu \text{m} \) in area to provide the resolution and low occupancy necessary in such a high flux environment. A spatial resolution of 15–20 \( \mu \text{m} \) is achieved in both the \( r \)-\( \phi \) plane and the \( z \) direction, which allows for a three-dimensional vertex reconstruction. In total there are 66 million pixels covering an area of approximately 1 m\(^2\).

**Figure 2.6:** The layout of the tracker in the \( r \)-\( z \) plane. The pixel and microstrip layers are indicated as horizontal and vertical lines [51].

Surrounding the pixel detector are layers of silicon microstrip detectors. These contain a total of 9.6 million silicon strips, with a total active area of 198 m\(^2\). These are arranged into the tracker inner barrel (TIB), comprising four layers which extend to \( r = 55 \text{ cm} \), and the tracker inner disks (TID) comprised of three layers in each endcap. Each strip is 10–20 cm long and 80–180 \( \mu \text{m} \) wide. The strips are aligned parallel to the beam direction in the barrel and radially in the endcaps, and provide up to four \( r \)-\( \phi \) measurements of a particle trajectory. Surrounding the TIB/TID is the tracker outer barrel (TOB) which extends to \( r = 116 \text{ cm} \) and \( z \pm 118 \text{ cm} \) and contains six barrel layers. The tracker endcaps (TEC) are comprised of 9 disks extending out to \( z \pm 282 \text{ cm} \). In the barrel layers the strip pitch varies from 80–180 \( \mu \text{m} \) from the innermost to the outermost layer and the resolution varies correspondingly from 23–52 \( \mu \text{m} \). Several layers utilise
two strip modules mounted back-to-back, offset with a stereo angle of 100 mrad which allows for a measurement of the \( z \) co-ordinate in the barrel region with a resolution of 230–530 µm.

### 2.4 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [51] measures the energy of photons and electrons and covers the region \( |\eta| < 3.0 \). It is composed of around 75 000 lead tungstate (PbWO\(_4\)) crystals separated into a barrel (EB) and two endcap (EE) regions.

High energy electrons or photons entering a crystal initiate an electromagnetic shower, producing a cascade of lower energy particles in which electrons and photons undergo bremsstrahlung and pair production respectively. The shower continues until photon energies fall below the pair-production threshold and ionisation begins to dominate for electrons. The charged particles in the shower ionise the atoms in PbWO\(_4\) which then de-excite by emitting the scintillation light that is converted to a current by photodetectors. Avalanche photodiodes and vacuum phototriodes are used in the EB and EE respectively.

The layout of the ECAL is given in figure 2.7. The EB covers the pseudorapidity range \( |\eta| < 1.479 \) and contains 61 200 crystals: 360 in the \( \phi \) direction and 170 in the \( \eta \) direction. It is constructed from 36 identical supermodules each spanning half the barrel length. Each crystal has an area in \( \eta-\phi \) of approximately 0.0174 \( \times \) 0.0174 (22 \( \times \) 22 mm\(^2\)) and is 230 mm long, equivalent to 25.8 radiation lengths. They are also tilted by 3° with respect to the axis from the detector origin to avoid particles travelling along the gaps between crystals. In the endcaps crystals are arranged in an \( x-y \) grid each with an area of 28.6 \( \times \) 28.6 mm\(^2\). Preshower detectors made from two discs of lead and two silicon strip layers are mounted in front of each endcap. This corresponds to about 3 radiation lengths of absorber material. They are designed to initiate showering and provide sufficient resolution to distinguish single photons from pairs produced in neutral pion decay.

The ECAL design was driven by the need for accurate photon and electron reconstruction, in particular for Higgs boson decay to a pair of photons, one of the benchmark discovery channels.

Lead tungstate was chosen as the scintillation material for its radiation hardness, short radiation length (0.89 cm) and small Molière radius (2.2 cm). This means that almost the entire electron or photon energy is deposited within the crystals. The decay time of the scintillation light is also short, with about 80% emitted within the 25 ns between bunch crossings. These properties lead to a calorimeter with excellent energy resolution, granularity, and timing precision.
The CMS detector

Figure 2.7: Cutaway view of the electromagnetic calorimeter. The barrel and endcap crystal layouts are indicated [51]. The inner radius of the barrel is 1.29 m and the endcaps are positioned 3.1 m from the interaction point.

The energy resolution of the ECAL can be parameterized as:

$$\frac{\sigma}{E} = \frac{S}{\sqrt{E}} \oplus \frac{N}{E} \oplus C$$ (2.2)

where $E$ is the energy of the incident particle and $S$, $N$ and $C$ are known as the stochastic, noise and constant terms respectively. The stochastic term encapsulates fluctuations in the scintillation and lateral containment of the shower; the noise term originates from the electronics and digitisation; and the constant term from non-uniform longitudinal response and inter-calibration errors. These have been measured in an electron beam test as $S = 0.028 \text{ GeV}^{1/2}$, $N = 0.12 \text{ GeV}$ and $C = 0.003$, although without the presence of a magnetic field or material in front of the ECAL.

2.5 Hadron calorimeter

Surrounding the ECAL is the sampling hadron calorimeter (HCAL) [51], designed to detect and measure the energy of strongly interacting particles. This is achieved with alternating layers of absorber and scintillator material. Brass is used as the absorber in most of the HCAL as it has a fairly short nuclear interaction length (16.42 cm) and is non-magnetic [21]. Hadron
showers cause light pulses in the plastic scintillator tiles that are fed to hybrid photodiodes by wavelength-shifting fibres.

The HCAL consists of several sub-detectors, arranged as in figure 2.8. The hadron barrel (HB) covers the region $|\eta| < 1.3$ and is located inside the magnet coil. It is segmented and read-out in $\eta$-$\phi$ towers of area $0.087 \times 0.087$, thus equivalent to the area of a $5 \times 5$ array of ECAL crystals. The limited volume between the ECAL and magnet does not provide a sufficient containment of hadronic showers, so an additional calorimeter is located on the outside of the solenoid. This hadron outer (HO) detector uses the magnet coil as an absorber and is important for sampling highly-penetrating or late-starting showers. The HB alone provides between 5.8 and 10.6 interaction lengths of absorber, with the minimum at $\eta = 0$. However, in combination with the HO this increases to a minimum of 11.8 interaction lengths. The hadron endcaps (HE) cover the range $1.3 < |\eta| < 3$ and provide approximately 10 interaction lengths. The tower granularity in $\eta$-$\phi$ space varies between $0.087 \times 0.087$ and $0.17 \times 0.17$ depending on $\eta$. The hadron forward (HF) detectors extend the HCAL coverage to $|\eta| = 5.2$ and experience the highest particle fluxes in the detector. Radiation-hard quartz fibres are used as the active medium, embedded in a steel absorber. A signal is generated when charged showering particles emit Cherenkov radiation in the fibre, which is detected by photomultiplier tubes.

Figure 2.8: Layout of one quadrant of the hadron calorimeter in the $r$-$z$ plane. The HB, HE, HO and HF components of the detector are highlighted [51]. The tracker and ECAL sub-detectors are visible in the inner part of the detector and the muon chambers are visible in the outer part.
The HCAL is designed to provide good energy resolution for the measurement of hadronic jets, with single charged pion resolution measured in a test beam [58] and found to be approximately

$$\frac{\sigma}{E} = \frac{94.3\%}{\sqrt{E}} \oplus 8.4\%.$$

The hermetic design and shower containment of the HCAL is driven by the need for accurate measurement of the transverse energy balance in an event and to ensure unambiguous identification of muons by minimising hadronic punch-through into the muon chambers.

## 2.6 Muon system

The muon system [51] utilises gaseous particle detectors positioned outside of the solenoid and covering the range $|\eta| < 2.4$. It is designed to meet three criteria: the efficient identification of muons, their precise momentum measurement and the capability to act as a trigger. Muons, having considerably higher mass than electrons, lose little energy via bremsstrahlung and are minimally ionising. This means they leave hits in the tracker, then typically pass through the calorimeters and the solenoid mass with only a small loss of energy.

The layout of the muon system is given in figure 2.9. In the central region of the barrel ($|\eta| < 1.2$) drift tube (DT) chambers are arranged in four cylindrical layers of stations positioned between the plates of the magnet return yoke. These contain a total of about 172,000 wires, each about 2.4 m long and contained in a tube of cross section area $13 \times 42$ mm$^2$. The tubes are filled with a mixture of argon and carbon dioxide gas, which is ionised when traversed by a muon. Free electrons drift towards the anode wire giving rise to an electrical signal. Each chamber consists of between 8 and 12 layers of tubes, some oriented parallel to the beam axis to measure the muon $\phi$ direction and some perpendicular to measure the $z$ co-ordinate.

In the endcap region ($0.9 < |\eta| < 2.4$), where the expected muon and background rates are higher than in the barrel, cathode strip chambers (CSCs) are used. These have a fast response, fine segmentation and are radiation hard. Each is a wedge-shaped multi-wire proportional chamber containing six gas layers with cathode strips running radially outward to measure hits in the $r-\phi$ plane, and anode wires running perpendicular to measure $\eta$. Using all chamber layers the hit position resolution is approximately 80 \(\mu m\).

Both the DTs and CSCs are augmented by resistive plate chambers (RPCs) in the range $|\eta| < 1.6$. These are constructed from parallel anode and cathode plates with a gas gap in between, with the muon ionisation detected by arrays of metallic strips that run parallel to the beam axis. Although
these have a poorer position resolution than the DT and CSC detectors the time response is very quick at about 1 ns. This means they can be used as a dedicated and independent muon trigger and to correctly identify the bunch crossing in which a muon originates.

Optimal muon momentum resolution is achieved by combining hits in the muon chambers with those from the tracker, as described in the next chapter. The muon system alone provides a resolution of 9–11% for muons with $p_T < 200$ GeV and $|\eta| < 2.4$. The performance is limited by muons scattering in the detector material before reaching the first muon chamber.

### 2.7 Trigger and computing

To date, proton bunches have crossed inside CMS at a rate of 20 MHz, which will likely increase to 40 MHz in future operation. It is not possible for the data acquisition (DAQ) system to read-out every event at this rate, nor is it feasible to write each $\sim 1$ MB event to tape. Therefore a trigger system is employed to select interesting events and bring the recorded rate down to $\mathcal{O}(100 \text{ Hz})$.

The trigger consists of two stages. The first is the Level-1 (L1) trigger, which runs on custom hardware and uses information from the calorimeters and muon system only. The architecture of
the L1 trigger is illustrated in figure 2.10. It starts with local triggers which compute calorimeter energy deposits and identify hit patterns in the muon chambers. A regional trigger combines local trigger information in a particular section of the detector to produce sorted lists of the relevant objects, for example, the highest $p_T$ electron or muon candidates. The global muon and calorimeter triggers collate this information from across the entire detector, and finally the global trigger issues a decision to reject or accept the event based on the complete L1 information. The time between a bunch crossing and this decision is limited to 3.2 $\mu$s, therefore the event processing is pipelined at each sub-detector in front-end electronics. If the event is accepted the full detector information is read-out. At this stage the event rate is reduced to around 100 kHz.

![Diagram illustrating the components of the L1 trigger, and the sequence of reconstruction between each stage [51].](image)

The second trigger stage is the high-level trigger (HLT) which is operated on a standard processor farm of several thousand CPU cores and makes use of the complete detector information, including hit patterns from the tracker. This allows for a more accurate determination of object momentum and identification. As such the algorithms used in the HLT are more sophisticated and generally closer to those used in the full offline reconstruction. The collection of trigger algorithms running at any one time is referred to as a trigger menu. In response to increasing instantaneous luminosity over the run period, the menu was updated several times with higher $p_T$ and energy thresholds in order to maintain a stable rate. During the 2012 run the HLT operated with a total output capacity of around 1 kHz. Of this, 300 Hz was promptly reconstructed and
the remainder stored on disk in a process known as data parking [60]. This parked data was reconstructed after the end of the run when more computing capacity was available.

Despite the rate reduction in the HLT, CMS produces several petabytes of data each year, and even larger sets of simulated events. To manage this, CMS and the other LHC experiments employ a global data storage and analysis network known as the Worldwide LHC Computing Grid (WLCG) [61]. The WLCG integrates the computing facilities of universities and research laboratories around the world. The system is built around a number of tiers. The Tier 0 centre at CERN performs the full event reconstruction and makes a backup to tape. All data are then distributed to at least one of the Tier 1 centres, which are spread across the globe, to keep a custodial copy. Tier 2 centres provide resources for specific analyses to thousands of researchers around the world.
Chapter 3

Event reconstruction and simulation

This chapter describes the reconstruction of collision events with the CMS detector. Emphasis is given to the algorithms which are relevant to both analyses in this thesis, as well as to common aspects of event selection. Section 3.1 describes the reconstruction and clustering of charged tracks. Sections 3.2 and 3.3 describe how tracks are combined with ECAL deposits and muon chamber hits to identify electron and muon candidates respectively. The particle flow algorithm, detailed in section 3.4, uses the complete detector output to provide a fully particle-based interpretation of an event. The reconstruction of jets, missing transverse energy and hadronic taus is based on the particle flow output, and is outlined in sections 3.5, 3.6 and 3.7 respectively. Both analyses also rely on the predictions of Monte Carlo (MC) event simulation and this is described in section 3.8.

3.1 Tracks and vertices

The trajectories of charged particles, and thus position and momentum measurements, are determined from the patterns of hits in the inner tracker layers. This track reconstruction is performed with the combinatorial track finder (CTF) algorithm [62], which can be separated into four steps:

- The identification of track seeds consisting of two or three hits in the innermost layers. This defines an initial estimate of the track path, which is helical in the approximately uniform magnetic field within the tracker.

- A Kalman filter [63] is used to extrapolate from the seed along the expected trajectory, incorporating any matched hits in each tracker layer, with the track parameters updated each time a new hit is found. This typically repeats until the final tracker layer is reached.
- The Kalman filter is again used with the complete set of hits to determine the best estimate of the trajectory.
- Fake tracks, not originating from a charged particle, are rejected by the application of quality criteria.

This process is repeated up to six times and at the end of each iteration the hits associated with identified tracks are removed. The tracking efficiency for pions and muons was estimated in early low-pileup LHC data and found to be $>98\%$ for tracks with $p_T > 500\, \text{MeV}$ and $>99\%$ for tracks with $p_T > 2\, \text{GeV}$ [64].

Given the complete set of reconstructed tracks, the positions, or vertices, of each proton-proton interaction can be determined. This makes use of the beamspot which is the centre of a three-dimensional profile of the luminous region where the collisions occur. Vertices from different interactions are resolved by a clustering algorithm, which uses as input the $z$ co-ordinates of tracks at the point of closest approach to this beamspot. Only tracks compatible with an origin in the luminous region are considered. The clustering uses the deterministic annealing (DA) algorithm [65] which identifies the most probable vertex positions and assigns each track to the most likely originating vertex. For each output vertex which contains at least two tracks, the adaptive vertex fitter [66] is used to determine the best fit of the three-dimensional vertex position as well as the fit quality. In this fit each track is assigned a probability $p_i$ for originating at this point. The $p_i$ are used to define the number of degrees of freedom for the fit as $n_{\text{dof}} = -3 + 2 \sum_i p_i$. This variable is useful in assessing the mutual compatibility of the vertex tracks and can aid in the selection of genuine interaction vertices. The vertex position resolution and reconstruction efficiencies have been measured in early LHC data, as well as in simulation [67], and are given in figure 3.1 as a function of the number of constituent tracks. The resolution is found to improve significantly with higher track multiplicities and higher average track $p_T$. Similarly, the reconstruction efficiency quickly approaches 100\% as the number of tracks increases.

Analyses typically require vertices to pass quality criteria in order to select genuine pp interactions and reject beam-induced backgrounds with an efficiency greater than 99\% [67]. These requirements are that the distance $\Delta z$ between the vertex and nominal interaction point be less than 24 cm, the corresponding distance in the transverse plane be less than 2 cm, and $n_{\text{dof}} > 4$ in the vertex fit. In each event the vertex having the highest scalar sum of track $p_T$ is assumed to be the vertex of the hard-scattering interaction and is referred to as the primary vertex.
Event reconstruction and simulation

Figure 3.1: (a) The resolution of the primary vertex \( z \) co-ordinate and (b) vertex reconstruction efficiency as a function of the number of constituent tracks. The resolution performance is given for three ranges of average track \( p_T \) [67].

3.2 Electrons

Electrons are reconstructed by matching ECAL energy clusters with charged tracks. However, this is complicated by the emission of bremsstrahlung photons as electrons traverse the material of the inner tracker which can be up to 2 radiation lengths depending on the \( \eta \) direction. These photons may also convert to \( e^+e^- \) pairs before reaching the calorimeter surface. This causes the energy deposited in the ECAL to be spread out in the \( \phi \) direction. Approximately 35\% of electrons will radiate at least 70\% of their energy in this way [68]. Therefore, dedicated “supercluster” algorithms [69] are used to combine the ECAL energy clusters from both the initial electron and the bremsstrahlung photons.

In the barrel region the “hybrid” clustering algorithm is used. This starts with the identification of a single seed crystal with \( E_T > 1 \) GeV, around which a “domino” of \( 3 \times 1 \) or \( 5 \times 1 \) crystals in \( \eta \)-\( \phi \) is formed and centred on the seed. Additional dominoes are then formed by stepping in both \( \phi \) directions about the seed up to \( \Delta \phi \approx 0.3 \) rad. Dominoes with \( E_T < 100 \) MeV are discarded, and the remainder grouped into sub-clusters that together form the supercluster. In the endcaps the “Multi5\times5” algorithm collects \( 5 \times 5 \) arrays of crystals that lie within \( \Delta \eta < 0.07 \) and \( \Delta \phi < 0.3 \) rad.

The energy loss in the tracker material makes the CTF algorithm sub-optimal for the reconstruction of electron tracks, so a tailored track-finding algorithm is employed. For high \( p_T \) electrons the optimal seeding is driven by the energy-weighted mean impact point of the ECAL.
supercluster. Together with the measured $E_T$ this determines a loose $\phi$-$z$ search region in the
pixel tracker for each charge hypothesis. If a track seed of two compatible hit layers is found,
the electron trajectory is updated. Track building then proceeds as described in the previous
section, but with the Kalman filter replaced by the dedicated Gaussian sum filter (GSF) [70].
This better incorporates the non-Gaussian energy loss caused by the bremsstrahlung emission.

Backgrounds to genuine electrons are expected to originate in hadronic jets, for example where
a $\pi^0$ and $\pi^\pm$ overlap, or where a $\pi^\pm$ showers early in the ECAL [68]. Several variables are useful
in distinguishing electrons from this background. These include:

- $\Delta \eta_{in}$ and $\Delta \phi_{in}$, the separation in the $\eta$ and $\phi$ directions between the supercluster and track
direction evaluated at the primary vertex position and extrapolated to the ECAL.
- $\sigma_{\eta_{in}}$, the energy-weighted $\eta$ width of the cluster which is typically small for prompt
electrons as the shower width is not significantly affected by the spreading due to the
magnetic field.
- $H/E$, the ratio of hadronic to electromagnetic energy in the region of the seed cluster.

Distributions of each variable for simulated electrons and jets are given in figure 3.2. A back-
ground to prompt electron production comes from photons that convert to $e^+e^-$ pairs within
the tracker volume. Such electrons may be distinguished by the absence of hits in the innermost
tracker layers and the matching to the track of the conversion partner. Specific identification
criteria are given in subsequent chapters.

### 3.3 Muons

Initially, muon tracks are reconstructed independently in the tracker and the muon system. For
the latter, the “standalone muon” algorithm [51] starts by building track segments from layers of
hits in individual chambers. These segments are used as position and momentum seeds for a
combined track fit with a Kalman filter from all DT, cathode strip and resistive plate chambers.
The beamspot position is taken as an additional constraint.

The muon momentum resolution can be greatly improved by utilising hits in both the muon
system and the tracker. The “global muon” reconstruction starts with each standalone muon
track and searches for a compatible tracker track. The global track is then determined from a
fit to both tracker and muon station hits, again using a Kalman filter and taking into account
the expected energy loss within the magnet and support structure. For muons with $p_T$ below
Figure 3.2: Distributions of the variables (a) $\Delta \eta_{\text{int}}$, (b) $\Delta \phi_{\text{int}}$, (c) $\sigma_{\eta \eta}$ and (d) $H/E$ for genuine electrons and misidentified jets [71]. Golden electrons are those with minimal bremmstrahlung emission whereas showering electrons typically lose a large fraction of their initial energy in this way.
200 GeV the best momentum resolution is driven by the fit in the tracker [72]. For muons with $p_T > 200$ GeV the global track fit provides an improved resolution over the tracker-only fit.

For low momentum muons, with $p < 5$ GeV, a reconstruction starting with the inner tracker is more efficient as these may not penetrate to the outer layers of the muon system. The “tracker muon” reconstruction considers all tracks with $p_T > 0.5$ GeV and extrapolates each through the detector to the muon system. A search is then performed in any coincident chambers for a matching track segment, taking into account the expected position uncertainty.

Muons reconstructed by these algorithms may be classified as prompt or non-prompt. For example, prompt muons may be produced in vector boson or quarkonia state decays. Non-prompt muons may originate from the in-flight decays of light hadrons, or from tau or heavy flavour quark decays. For global muon reconstruction, non-muon backgrounds are generally small but may originate from the punch-through of hadronic showers which reach the muon system. For the identification of prompt muons, additional selections may be applied to reduce the non-prompt and punch-through backgrounds. These “tight muon” criteria require:

- The $\chi^2$ per degree of freedom of the global track fit, using tracker and muon chamber hits, must be less than 10.

- The global track fit must include hits from at least one segment in the muon detector. Combined with the $\chi^2$ requirement this provides rejection of muons from hadronic punch-through from the calorimeters.

- Track segments must be found in at least two stations of the muon detector.

- At least one hit in the pixel detector and hits in at least 5 tracking layers are required, in order to reduce the background of muons from decays-in-flight.

The efficiency for muons to be reconstructed by the global algorithm and pass the tight identification has been measured using $J/\psi \to \mu^+\mu^-$ and $Z \to \mu^+\mu^-$ decays in data and simulation. The results are shown in figure 3.3 as a function of $p_T$ for two $\eta$ regions. The efficiency is found to reach a plateau of 96–99% for $p_T > 10$ GeV.

### 3.4 Particle flow

The particle flow (PF) algorithm [73–75] is designed to optimally combine the information from all sub-detectors to reconstruct every particle in an event. Each particle is then classified as either an electron, muon, photon, charged or neutral hadron. It exploits the excellent momentum
Event reconstruction and simulation

Figure 3.3: Measurements of the muon reconstruction and identification efficiency in the (a) barrel and (b) endcap regions in $\sqrt{s} = 7$ TeV data and simulation as a function of $p_T$ [72].

resolution of the tracker and the energy resolution and high granularity of the ECAL to assign the most accurate momentum measurements to charged hadrons and photons, respectively. The tracking information also makes it possible to separate charged and neutral hadrons in close proximity, which cannot be achieved with the HCAL alone. The output of the PF algorithm is a list of particle candidates that can be used as input to higher-level objects, for example, in jet clustering or the missing transverse energy calculation.

The algorithm starts with a set of basic elements: charged particle tracks, calorimeter energy clusters and muon chamber hits. These elements are linked into blocks and, depending on the composition, the block is interpreted as a particle of a particular type. Calorimeter energy deposits are first clustered so that neutral hadrons may be identified and separated from charged hadrons and so that bremsstrahlung photons may be recombined with the parent electrons. The clustering is performed separately in each sub-detector component, apart from the HF where each cell gives rise to a single cluster. First, the calorimeter cells with local energy maxima are designated as “cluster seeds”. Neighbouring cells are then aggregated into the cluster, provided they register an energy at least two standard deviations above the expected noise level. This level is 80 MeV in the ECAL barrel, 300 MeV in the ECAL endcaps and 800 MeV in the HCAL. The energy in each cell may be shared amongst more than one cluster.

Any given particle is likely to result in several PF elements. Therefore these must be linked to form particle candidates, and this is achieved by defining a distance metric between every pair of elements. For example, in linking between tracks and calorimeter clusters the track trajectory
is extrapolated from the outermost tracker layer to points within the ECAL and HCAL volumes. If these fall within the boundaries of a cluster then the track and cluster are linked. The clusters of bremsstrahlung photons emitted by electrons are linked by extrapolating the tangent of the track direction at each tracker layer. Muon system tracks are linked with those from the tracker when the global muon fit meets a maximum $\chi^2$ requirement.

The identification of a particle, or set of particles, from each block proceeds as follows:

- If a track element has a position and momentum compatible with a global muon it is removed, and the global muon is designated as a PF muon. The expected energy deposited in the calorimeters is also subtracted from the relevant clusters.

- Candidate electron tracks are refitted using the GSF as in the standalone reconstruction, and a multivariate discriminator assesses the compatibility with the linked ECAL clusters. This provides discrimination against charged hadrons. If the candidate passes, a PF electron is formed from the GSF track and linked ECAL clusters, which are also removed from the block.

- Each remaining track results in a charged hadron. This is assigned the pion mass and the momentum from the track fit. If the energy of the linked calorimeter clusters is compatible with this momentum, within uncertainties, the candidate momentum is updated to a weighted average between the track and cluster measurements.

- However, if the cluster energy exceeds the track momentum by more than the expected resolution, this excess is interpreted as the presence of overlapping neutral particles. If the excess is larger than the total linked ECAL energy a PF photon is created, with any remainder becoming a neutral hadron.

- All remaining ECAL and HCAL clusters not linked to a track give rise to photon and neutral hadron candidates, respectively.

### 3.5 Jets

Quarks and gluons are produced copiously in the QCD-dominated collisions at the LHC. However, these are not observed directly as they fragment and hadronise almost immediately, resulting in collimated showers of particles. Thus, to determine the properties of the original parton, these jets of particles must be combined and measured.
3.5.1 Clustering

The combination of objects into jets is specified by a clustering algorithm [76]. This requires both a definition of how the objects should be grouped, usually involving some measure of distance between objects, and how the four-momentum of the jet should be assigned. CMS utilises the class of algorithms known as sequential recombination. These are known to be collinear- and infrared-safe, meaning the number of jets is not affected by soft collinear gluon emission or parton splitting. The algorithm starts by determining a distance between every pair of objects $d_{ij}$ and the distance between each object and the beamline $d_{iB}$:

$$d_{ij} = \min\left(\frac{\Delta R_{ij}^2}{R^2}, \frac{\Delta \eta_{ij}}{R} \right)$$

$$d_{iB} = p_T^{2p}$$

where $p = -1, 0, 1$ and defines the algorithm behaviour, $\Delta R$ is the separation distance in the $\eta$-$\phi$ plane and $R$ is a fixed parameter which sets the jet scale. The smallest of the $d_{ij}$ and $d_{iB}$ parameters is then identified. If this is a $d_{ij}$ parameter then objects $i$ and $j$ are combined into a new single object and the process repeats. If it is a $d_{iB}$ parameter the object with label $i$ is taken to be a final-state jet and removed from the list. This process repeats until no objects remain. At CMS $p$ is typically chosen as $-1$, known as the anti-$k_T$ algorithm [77]. This tends to cluster jets around the hardest particles, leading to cone-like jet areas in $\eta$-$\phi$. The analyses described in subsequent chapters make use of jets clustered from PF candidates with the anti-$k_T$ algorithm, as implemented in the fastjet package [78], using a distance parameter $R = 0.5$. In addition, simulated jets may be determined at the level of hard-scattering partons or final-state hadrons. An important property of the clustering algorithm, resulting from the requirements of collinear- and infrared-safety, is that the same number of jets should be observed regardless of the level of simulated input objects.

3.5.2 Jet energy corrections

Due to a number of experimental effects, the measured jet energy typically differs from the true hadron-level energy. This can be caused by a non-linear calorimeter response, detector noise, overlap with non-instrumented regions or the presence of additional energy from pileup interactions. A correction is determined and applied to the reconstructed jet four-momentum, both in simulation and data, such that the energy response is equal, on average, to that at the
Event reconstruction and simulation

hadron level. This can be factorised into a set of sequential corrections [79]:

\[ P_{\mu}^{\text{corr}} = C_{\text{offset}}(p_T^{\text{RAW}}, \eta) \cdot C_{\text{rel}}(p_T^{\text{"}}, \eta) \cdot C_{\text{abs}}(p_T^{\prime}, \eta) \cdot P_{\mu}^{\text{RAW}} . \quad (3.3) \]

The first term, \( C_{\text{offset}} \), subtracts the contribution from pileup and detector noise. This is estimated on a per-jet basis using the \( p_T \)-density in the event, \( \rho \), and the jet area \( A \) [80]. The former is determined as the median \( p_T \)-density of all jets in the event and is unaffected by the presence of hard jets. It does, however, include the contribution from the underlying event (UE) of the hard-scattering interaction, which should not be subtracted. The subtraction is therefore of the form \((\rho - \langle \rho \rangle_{\text{UE}}) \cdot A\), where \( \langle \rho \rangle_{\text{UE}} \) is the average expected UE contribution.

The second term, \( C_{\text{rel}} \), is applied to the offset-corrected transverse momentum \( p_T^{\text{"}} \) and corrects the response to be flat as a function of \( \eta \). It is measured using a dijet \( p_T \)-balancing method [79]. A reference jet is required to be within the central region of the detector, where the response is uniform, and a probe jet at any value of \( \eta \). The average of the balance quantity, defined as \((p_T^{\text{probe}} - p_T^{\text{ref}})/p_T^{\text{average}}\), is used to determine the relative response in bins of the probe jet \( \eta \) and the dijet average \( p_T \). The relative response in data and simulation, as a function of jet \( \eta \), is given in figure 3.4a.

The absolute correction \( C_{\text{abs}} \) is applied to the relative-response corrected \( p_T^{\prime} \) and is designed to give a uniform response in \( p_T \). It is measured in samples of \( \gamma + \text{jets} \) and \( Z + \text{jets} \) events using the missing transverse energy projection fraction (MPF) method which was developed at the Tevatron experiments [81]. This exploits the excellent \( p_T \) resolution for leptons and photons, and the fact that these events should not contain any real missing energy. Any measured missing energy can be used to calibrate the \( p_T \) response of the jets in the event. The uncertainty in the jet energy scale, due to the uncertainties in each correction factor, is given in figure 3.4b. The main contributions are from uncertainty in the energy densities used for pileup subtraction and in the photon energy scale and extrapolation required in the MPF method. The total uncertainty ranges from around 4% at a \( p_T \) of 20 GeV to below 2% at 1 TeV.

### 3.5.3 b-Jet tagging

The identification of jets originating from the hadronisation of b quarks is important for studying many SM processes and for searches of BSM signatures. The b-flavour hadrons have relatively large masses (5–6 GeV) and long lifetimes (\( \tau \approx 1.5 \text{ ps} \)) [21]. These properties can be exploited to tag b jets from the dominant light quark (u, d, s) and gluon jet production. A number of algorithms have been developed for this purpose, each of which yields a single discriminator
Event reconstruction and simulation

Figure 3.4: (a) The relative jet energy scale correction as a function of jet $\eta$ and (b) the jet energy scale uncertainty as a function of jet $p_T$. Both results are for PF jets in data and simulation at $\sqrt{s} = 7$ TeV [79].

value per b jet. Three working points for each algorithm are designated as loose, medium and tight, corresponding to light-quark misidentification probabilities of 10%, 1% and 0.1%, respectively [82].

The b hadrons typically travel a distance set by the scale $c \tau \approx 450 \mu$m before decaying, meaning any charged decay products produce tracks with a measurable impact parameter with respect to the primary event vertex. The excellent position resolution of the pixel tracker means impact parameters are measured with a resolution of around 30 $\mu$m for 5 GeV tracks [51]. The impact parameter significance of each track associated to a jet is used in several tagging algorithms. This is defined as the ratio of the three-dimensional impact parameter to its estimated uncertainty [82]. Other algorithms attempt to reconstruct the secondary decay vertex using the tracks associated to the jet. The simple secondary vertex (SSV) algorithm uses the significance of the flight distance between the secondary and primary vertices as a discriminator. This is used in the $Z$+b-jet cross section measurement and is described in more detail in the next chapter. The performance of this algorithm is limited by the secondary vertex reconstruction efficiency which is around 65%. The combined secondary vertex (CSV) tagger is a more complex algorithm that uses track-based and secondary vertex variables, when one is present, to provide a higher b-jet efficiency. The CSV algorithm is used for b-jet identification in the $H \to \tau \tau$ searches and is described in chapter 5.

Not all variables and properties used in these algorithms are perfectly modelled in the simulation. Many analyses depend on a precise knowledge of the tagging and mistagging rates. Several
Event reconstruction and simulation

Techniques are employed to measure these in data, which can be compared to the efficiency in simulation. Ratios of the form $\frac{\varepsilon_{\text{data}}}{\varepsilon_{\text{sim}}}$ give scale factors that can be used to correct for efficiency differences in simulated events. High purity samples of b jets can be obtained by selecting jets with overlapping soft muons, exploiting the $\approx 11\%$ semi-leptonic branching fraction of b hadrons. Another source is from events with the kinematic signatures of $t\bar{t}$ production, which exploits the near $100\%$ branching fraction for $t \rightarrow Wb$. Efficiencies from the different methods are combined over a wide range of jet $p_T$ to give the scale factors. For example, figure 3.5 gives these factors for the medium working point of the CSV algorithm. The uncertainties in these factors include an uncertainty in the rate of gluon splitting to $b\bar{b}$ pairs, biases due to selection of semi-leptonic $b$-hadron decays and the modelling of pileup in simulation.

![Diagram](image)

**Figure 3.5:** Ratio of b-tagging efficiencies for the medium working point of the CSV discriminator as a function of jet $p_T$ [82]. The hatched areas correspond to the weighted average of several measurements in each $p_T$-bin. The solid line results from a fit to these measurements.

3.6 Missing transverse energy

Neutrinos, in addition to hypothetical BSM long-lived and weakly interacting particles, pass through the entire detector without interacting. Their presence can be inferred as a momentum imbalance in the transverse plane when all detected particles are evaluated. This is quantified as the missing transverse momentum $E_T^{\text{miss}}$, the negative vectorial sum of all visible particle $p_T$ in an event. The magnitude $E_T^{\text{miss}}$ is referred to as the missing transverse energy. Given the particle identification capabilities and improved resolution offered by the PF algorithm, the full list of PF candidates is a natural input choice for the $E_T^{\text{miss}}$. Figure 3.6 shows the $E_T^{\text{miss}}$ distribution.
Event reconstruction and simulation

in $Z \rightarrow \mu\mu$ and $\gamma$+jet selections for $\sqrt{s} = 8$ TeV data and simulation. As muons and photons are measured with good energy resolution, and these processes predominately do not involve genuine $E_T^{\text{miss}}$, the observed $E_T^{\text{miss}}$ reflects the resolution of the hadronic activity in these events.

Figure 3.6: Distributions of the particle flow $E_T^{\text{miss}}$ in (a) $Z \rightarrow \mu\mu$ and (b) $\gamma$+jets events in $\sqrt{s} = 8$ TeV data and simulation. The shaded band reflects the uncertainty in the muon, photon, jet and neutral hadron energy responses [83].

During data-taking a number of issues were identified that result in a small fraction of events containing anomalously high $E_T^{\text{miss}}$, up to values of several TeV [83]. These include energy deposited by beam-halo particles, noise in the HCAL readout electronics, heavily ionising charged particles depositing large energies in the ECAL avalanche photodiodes, and the reconstruction of fake tracks due to noise in the silicon strip tracker. A number of filters have been developed to identify and remove such events, which is important for obtaining the level of agreement with simulation seen in figure 3.6.

The $E_T^{\text{miss}}$ response and resolution can be degraded by a number of factors. These include minimum energy thresholds in the calorimeter read-out, non-instrumented detector regions and reconstruction inefficiencies. Furthermore, the presence of multiple pileup interactions has the effect of degrading the $E_T^{\text{miss}}$ resolution. A number of techniques for correcting both the response and resolution of the PF missing energy have been developed [84]. For example, a correction which accounts for the jet energy scale of the jets in the event can reduce bias in the response. A number of techniques have also been developed to reduce the impact of pileup on the resolution and are described in [83].
3.7 Hadronic taus

Hadronic tau decays produce characteristically narrow jets containing one or three charged particles and some number of neutrals, as well as a tau neutrino ($\nu_\tau$). The $\nu_\tau$ passes through the detector without interacting and contributes to the missing transverse energy of the event. Hadronic tau reconstruction uses the output of the PF algorithm to identify specific decay modes in the visible particles. This is referred to as the “hadron-plus-strips” algorithm [85], which combines PF charged hadrons with $\pi^0$ candidates formed from strips of clustered photons. The final states and branching ratios for the most common decay modes are summarised in table 3.1.

<table>
<thead>
<tr>
<th>Decay Channel</th>
<th>Resonance</th>
<th>Mass [MeV]</th>
<th>Branching Fraction [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^\pm \rightarrow \pi^\pm \nu_\tau$</td>
<td></td>
<td></td>
<td>11.6</td>
</tr>
<tr>
<td>$\tau^\pm \rightarrow \pi^\pm \pi^0 \nu_\tau$</td>
<td>$\rho$</td>
<td>770</td>
<td>26.0</td>
</tr>
<tr>
<td>$\tau^\pm \rightarrow \pi^\pm \pi^0 \pi^0 \nu_\tau$</td>
<td>$a_1$</td>
<td>1260</td>
<td>10.8</td>
</tr>
<tr>
<td>$\tau^\pm \rightarrow \pi^\pm \pi^\mp \pi^\pm \nu_\tau$</td>
<td>$a_1$</td>
<td>1260</td>
<td>9.8</td>
</tr>
<tr>
<td>$\tau^\pm \rightarrow \pi^\pm \pi^\mp \pi^\pm \pi^0 \nu_\tau$</td>
<td></td>
<td></td>
<td>4.8</td>
</tr>
<tr>
<td>Other hadronic modes</td>
<td></td>
<td></td>
<td>1.7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>64.7</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of the hadronic tau decay modes, with the branching fractions and intermediate resonances listed where relevant [21].

These decays are searched for within PF jets clustered by the anti-$k_T$ algorithm with distance parameter 0.5. Over 60% of the hadronic tau decays contain at least one $\pi^0$ meson. The photons produced in the $\pi^0 \rightarrow \gamma \gamma$ decay have a high probability of converting to electron-positron pairs within the tracker detector volume. The magnetic field in this region causes bending which results in an EM shower dispersed in the $\phi$ direction. The original $\pi^0$ candidate is recovered by clustering PF electrons and photons in strips of size $0.05 \times 0.20$ in $\eta-\phi$ space. A strip must contain at least one photon and have total photon $p_T > 2.5$ GeV to be considered a $\pi^0$ candidate. All possible combinations of the $\pi^0$ and charged hadron candidates within the jet are evaluated and the following modes identified:

- One charged hadron and one strip: the combined system having mass $0.4 < m < 1.3$ GeV if $p_T < 200$ GeV. To account for changes in resolution the upper limit increases with $p_T$ for candidates with $p_T > 200$ GeV, up to 2.1 GeV for $p_T = 800$ GeV. This selects decays to the intermediate $\rho$ meson, $\tau^\pm \rightarrow \rho^\pm \nu_\tau \rightarrow \pi^\pm \pi^0 \nu_\tau$. 
• Single charged hadron: corresponding to the decay $\tau^\pm \rightarrow \pi^\pm \nu_\tau$.

• Three charged hadrons: the vectorial sum having $p_T > 0.5$ GeV and $0.8 < m < 1.5$ GeV, thus being compatible with the mass of the $a_1(1260)$ resonance in the decay $\tau^\pm \rightarrow a_1 \nu_\tau \rightarrow \pi^+ \pi^- \pi^\pm \nu_\tau$. The charged tracks must originate from the same vertex, have a maximum impact parameter of 2 mm in the $z$ direction and sum to unit charge.

• One charged hadron and two strips: the combined system having mass $0.4 < m < 1.2$ GeV if $p_T < 200$ GeV. The upper limit increases with $p_T$ for candidates with $p_T > 200$ GeV up to 2.0 GeV for $p_T = 800$ GeV. This selects the decay $\tau^\pm \rightarrow a_1 \nu_\tau \rightarrow \pi^+ \pi^- \pi^0 \pi^0$.

All charged hadrons and strips must be contained within a narrow cone around the jet axis defined as

$$\Delta R = \begin{cases} 
0.10 & \text{if } p_T < 28 \text{ GeV} \\
\frac{2.8}{p_T} & \text{if } 28 \text{ } p_T < 56 \text{ GeV} \\
0.05 & \text{if } p_T > 56 \text{ GeV}
\end{cases}$$

where $p_T$ is that of the identified mode. Even after this decay mode reconstruction is applied, a large background from hadronic jets can be expected, which can be further mitigated by requiring the tau candidate to be isolated from other PF candidates. This is described in more detail in chapter 5 in the context of the $H \rightarrow \tau\tau$ analyses.

### 3.8 Monte Carlo simulation

The generation of simulated proton-proton collision events is factorised into several stages, illustrated in figure 3.7. A recent review of generators and the principles involved may be found in [86]. The core process is the hard-scattering of two incoming partons leading to some fixed set of outgoing particles. The momenta of the incoming partons are sampled from parton distribution functions (PDFs). These define the probability for a given parton species to carry some fraction of the proton momentum at a chosen factorisation scale $\mu_F$. The hard-scattering involves a large transfer of momentum, with both incoming and outgoing partons asymptotically free, meaning the process can be computed within the framework of perturbative QCD. This requires a matrix element calculation, usually performed at leading order (LO) in the process of interest, with either a multipurpose or dedicated event generator. Multipurpose generators such as PYTHIA [87], HERWIG++ [88] and SHERPA [89] can model the kinematics of a wide range of $2 \rightarrow 1$ and $2 \rightarrow 2$ processes. Dedicated generators such as ALPGEN [90], MADGRAPH [91] and also
SHERPA can calculate matrix elements with higher numbers of final state partons. In addition, next-to-leading order (NLO) calculations have been implemented in some generators to provide more accurate kinematic predictions and reduced uncertainties, for example, POWHEG [92, 93] or amc@nlo [94].

After the matrix element calculation, the parton showering process is used to evolve the event from the parton to the hadron level. This is an iterative process which generates the emission of multiple soft or collinear gluons from the incoming and outgoing partons. Care must be taken at this stage to ensure that no phase space is over- or under-counted in matching the matrix element to the parton shower for which several schemes have been developed [95]. The process continues until partons reach a minimum energy threshold, normally around 1 GeV, where interactions become non-perturbative [86]. At this stage a hadronisation model is required to combine the colour-charged partons into colour-neutral hadrons. Another effect a generator must incorporate is the interaction between other proton constituents not involved in the hard-scattering, known as the UE, which gives rise to typically soft QCD interactions and must be modelled. Dedicated packages are then used to simulate the decay of unstable particles, for example, EVTGEN [96] for hadrons and TAUOLA [97] for tau leptons. Most generators also provide a large number of parameters that control the different aspects of the event simulation and are tuned to reproduce experimental data. For example, PYTHIA is configured to use the Z2 tune in CMS simulation [98], the inputs to which include underlying event and minimum bias data from both the Tevatron and LHC experiments. Finally, to simulate the detector response the set of final-state particles is passed through a GEANT 4 [99] based model of the CMS detector.
Figure 3.7: Illustration of the components of a proton-proton collision as implemented in MC event generators [100]. The hard-scattering, at the centre of the diagram, of two incoming partons with a fixed number of outgoing particles is calculated at fixed order with a matrix element. Parton showering, in red, is then initiated for incoming or outgoing quarks and gluons, which iterates to lower and lower energies at which point the partons hadronise, indicated in green. The hadrons may subsequently decay to other stable particles. The proton remnants which interact to produce the underlying event are indicated in purple.
Chapter 4

Measurement of the Z+b-jet cross section

This chapter presents a study of Z boson production in association with b jets in proton-proton collisions. The dataset was recorded at a centre-of-mass energy of 7 TeV during the 2011 LHC run and corresponds to an integrated luminosity of 2.1 fb⁻¹.

Cross sections are determined for production with exactly one or at least two b jets, denoted Z+1b and Z+2b respectively. The symbol Z is taken to imply $Z/\gamma^*$ production, which is required to decay to either an electron or muon pair, denoted as $\ell\ell$. Events with b jets are identified through the use of a dedicated tagging algorithm. The measured yields under these selections are corrected for acceptance and efficiency effects using MC simulation. The cross sections are calculated for a Z boson with a mass in the range $76 < m_Z < 106$ GeV; a lepton acceptance of $p_T > 20$ GeV and $|\eta| < 2.5$; and a b-jet acceptance of $p_T > 25$ GeV and $|\eta| < 2.1$. The resulting cross sections are defined at the hadron level of final-state particles clustered into jets. As for the reconstructed jets, these are determined with the anti-$k_T$ algorithm and distance parameter of 0.5. A b jet is defined as any jet in which a b-flavour hadron is found within $\Delta R < 0.5$ of the jet axis.

The dominant backgrounds in this measurement come from Z+jets production in which light-or c-flavour jets are misidentified as b jets, and from $t\bar{t}$ events which contain genuine b jets. The Z and bb components of the final state may result from separate parton scatterings within a single proton-proton interaction, known as a multi-parton interaction (MPI), and this process is included as part of the signal definition.

Section 4.1 gives the theoretical and experimental motivations for these measurements and summarises previous results. Details of the data samples and MC simulation are given in section 4.2. The Z+1b and Z+2b event selection is described in section 4.3 and comparisons between data and simulation for a number of kinematic observables are given in section 4.4.
For the cross section determination, the data-driven background estimation is described in section 4.5, the corrections to hadron level in section 4.6 and the results in section 4.7.

4.1 Motivation

The measurement of $Z+b$ production is of interest at the LHC for several reasons. It is a benchmark process for neutral MSSM Higgs boson production in association with $b$ quarks. Cross sections for both processes can be calculated in the framework of perturbative QCD in two different schemes. The four-flavour scheme (4FS) allows only gluons and the four lightest quarks to participate in the hard scattering interaction. The $b$ quarks are then produced in pairs via gluon splitting. Calculations at NLO have been performed in this scheme [101,102]. Examples of LO Feynman diagrams in the 4FS are given in figure 4.1.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig4_1.png}
\caption{Tree-level Feynman diagrams for $Z+b$ production in the four-flavour scheme.}
\end{figure}

The five-flavour scheme (5FS) does allow $b$ quarks in the initial state, with the gluon splitting integrated into the PDF. Example tree-level Feynman diagrams are given in figure 4.2. Calculations at NLO have also been made in this scheme [103–105].

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig4_2.png}
\caption{Tree-level Feynman diagrams for $Z+b$ production in the five-flavour scheme.}
\end{figure}

Figure 4.3 gives the cross sections calculated in each scheme for the $b$-associated production of heavy MSSM Higgs bosons as a function of mass. Differences between the two of up to 30% are observed at large masses [106], although the comparison is limited by uncertainty in the QCD scale in this region. In the future, sufficiently precise measurements of the $Z+b$ cross section should help in resolving these differences.
Measurement of the Z+b-jet cross section

**Figure 4.3:** Total production cross sections for the b-associated modes (a) $b \bar{b}H$ and (b) $b \bar{b}A$ in the MSSM for $\sqrt{s} = 7$ TeV [106]. The upper (blue) uncertainty band is for the 5FS and the lower (red) band is for the 4FS with both reflecting the size of the QCD scale uncertainty.

Understanding the $Z+b$ process is also important as it is a major background in the search for the SM Higgs boson decaying to a $b \bar{b}$ pair and produced in association with a $Z$ boson. It also contributes as a background in other BSM processes which predict leptons and heavy-flavour jets in the final state. Finally, kinematic quantities related to the reconstructed leptons and $b$ jets may be used to validate the predictions of MC simulation. In this chapter the observed data are compared to the 5FS predictions of the MADGRAPH [91] matrix-element generator interfaced with PYTHIA for parton showering and hadronisation.

The $Z+b$ process has been studied previously at CMS and at other experiments. Measurements of the inclusive cross section, with any number of $b$ jets in the final state, have been made by the CMS [1,107], ATLAS [108], CDF [109] and D0 [110] Collaborations. The results presented in this chapter correspond to the preliminary results given in [2]. An update of this measurement using a larger dataset has recently been published [111]. A complementary measurement of angular correlations between $bb$ pairs in $Z+2b$ events has also been reported [112] by CMS.

### 4.2 Data samples and simulation

The data sample used in this analysis was recorded between March and August in the 2011 run, corresponding to an integrated luminosity of $2.13 \pm 0.09$ fb$^{-1}$, in which an average of 6.4
collisions per bunch crossing were observed. Events were recorded by dielectron and dimuon triggers. The dielectron trigger has $p_T$ thresholds of 17 GeV and 8 GeV on the leading and sub-leading electrons respectively. It also requires both electrons to pass a loose version of the offline identification and isolation criteria which are discussed in the next section. The dimuon trigger thresholds increased during the run period from 7 GeV for both muons to 13 GeV and 8 GeV, in order to maintain a stable trigger rate in the presence of increasing luminosity.

Simulated Z boson and t\bar{t} events are produced by the Madgraph matrix element generator [91] and include up to four additional jets. For the former process only the leptonic decays are simulated and the requirement $m_Z > 50$ GeV is applied at the generator level. The Z boson sample is normalised to the integrated luminosity of the data with the next-to-next-to-leading order (NNLO) cross section of $3048 \pm 130$ pb [113]. In subsequent figures and results the events in this sample are split into three subsets based on the quark content, denoted as

- Z+b for events containing one or more b quarks;
- Z+c for events containing one or more c quarks but no b quarks;
- Z+l for events containing no c or b quarks, thus those containing light-flavour jets only.

In this definition no $p_T$ requirement is placed on the b and c quarks, and they may originate either in the matrix element or the parton shower. While this avoids ambiguity in defining the flavour process of an event, it does mean, for example, the selected reconstructed jet in a Z+b event may not necessarily be a b-flavour jet. The Z+b process is also simulated by the aMC@NLO generator [102] and is used for the estimation of theory uncertainties as describe in section 4.6.

The t\bar{t}+jets sample is normalised using the NLO cross section of $155^{+23}_{-21}$ pb [114]. A small contribution is also expected from ZZ production, and this process is simulated with PYTHIA. Parton showering and hadronisation is performed by PYTHIA in all samples, which is configured with the tune Z2 [98] and also includes a simulation of MPI. A randomly sampled number of pileup interactions are added in each event. These events are then weighted so that the distribution of the number of pileup interactions matches that estimated in data, as given in section 2.1.

### 4.3 Event selection

The reconstruction of vertices, electrons, muons and jets follows the descriptions given in chapter 3. All selected events in data and simulation must contain a well-reconstructed primary vertex passing the standard quality criteria. Events are then required to pass either a $Z \rightarrow \mu^+\mu^-$ or $Z \rightarrow e^+e^-$ selection. The $Z \rightarrow \mu^+\mu^-$ selection requires two opposite-charge muons each with
Measurement of the $Z+b$-jet cross section

$p_T > 20 \text{ GeV}$ and $|\eta| < 2.1$. The muon candidates must be reconstructed by both the global and tracker-only algorithms and pass the standard tight identification criteria.

The $Z \rightarrow e^+e^-$ selection requires the presence of two opposite-charge electrons, each with $p_T > 25 \text{ GeV}$ and $|\eta| < 2.5$. Electrons reconstructed within the non-instrumented pseudorapidity gap ($1.4442 < |\eta| < 1.566$) between the ECAL barrel and endcap detectors are vetoed. The electrons must pass identification criteria utilising both track and supercluster information to minimize the background from hadrons and converted photons. The selections on these variables are given in table 4.1. In addition, there must be no more than one missing pixel layer hit in front of the innermost detected hit, and no opposite charge track in the vicinity of the candidate GSF track that matches geometrically as a conversion partner.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Barrel</td>
</tr>
<tr>
<td>$\Delta \eta_{in}$</td>
<td>$&lt; 0.006$</td>
</tr>
<tr>
<td>$\Delta \phi_{in}$</td>
<td>$&lt; 0.06$</td>
</tr>
<tr>
<td>$\sigma_{\eta\eta}$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>$H/E$</td>
<td>$&lt; 0.04$</td>
</tr>
</tbody>
</table>

**Table 4.1:** Electron identification requirements for the barrel and endcap regions.

Both electron and muon tracks must have transverse impact parameters of less than 200 $\mu\text{m}$ with respect to the beamspot position in order to reject backgrounds from non-prompt leptons, including cosmic muons. The leptons must also be isolated to minimise the background from misidentified jets. The isolation is defined by a sum of transverse energy (or $p_T$) within a cone of $\Delta R < 0.3$ centred on the lepton trajectory. Three isolation quantities are defined: one in each of the tracker ($I_{\text{trk}}$), ECAL ($I_{\text{ECAL}}$) and HCAL ($I_{\text{HCAL}}$) sub-detectors. For electrons, separate requirements are applied on the relative isolation, $I/p_T$, in each sub-detector. The maximum allowed value of $I/p_T$ varies from 2.5–10% depending on the sub-detector in question and whether the electron is in the barrel or endcap region. For muons, the combined relative isolation, defined as $(I_{\text{trk}} + I_{\text{ECAL}} + I_{\text{HCAL}})/p_T$, is required to be less than 15%.

Lepton pairs that pass these selections must also have invariant mass $m_{ee}$ between 76 GeV and 106 GeV. In the case of multiple candidate pairs within an event, the pair with mass closest to $m_Z$ is selected. This occurs rarely, affecting fewer than 0.01% of events selected in data. The efficiency for electrons and muons to pass the trigger, identification and isolation criteria is measured in data using the tag-and-probe method, as described in [115]. The ratio of efficiencies...
Measurement of the Z+b-jet cross section

![Graphs showing measured and expected distributions of (a) the number of jets and (b) the number of secondary vertices associated to the leading jet after the dilepton selections.]

Figure 4.4: Expected and observed distributions of (a) the number of jets and (b) the number of secondary vertices associated to the leading jet after the dilepton selections. Only jets with $p_T > 25$ GeV, $|\eta| < 2.1$ and which are separated from each selected lepton by at least $\Delta R = 0.5$ are counted. The shaded region reflects the statistical and jet energy scale uncertainties in the expected distribution.

between data and simulation gives rise to weight factors which are applied to simulated events so that the efficiency matches that in data.

Reconstructed jets, clustered from PF candidates, are required to have corrected $p_T > 25$ GeV, $|\eta| < 2.1$ and to be separated from each selected lepton with $\Delta R > 0.5$. The $\eta$ requirement is driven by the maximum extent of the tracker, which covers a region up to $|\eta| < 2.5$. By keeping the jet requirement at $|\eta| < 2.1$ this ensures a high fraction of tracks within the jet cone will pass through the tracker, as these are required for the identification of b jets. Loose jet identification criteria are also applied [116], primarily to reject those originating from calorimeter noise and from single isolated photons. The distribution of the number of reconstructed jets passing these selections is given in figure 4.4a, and the expectation is found to be in good agreement with the data up to high jet multiplicities.

Jets which originate from b quarks are identified through an algorithm which exploits the long b-hadron lifetime. This SSV algorithm [82], introduced in the previous chapter, proceeds by first constructing any additional vertices from the tracks associated to a given jet. In the high-efficiency variant, which is used in this analysis, the secondary vertices may be built from as few as two tracks. The distribution of the number of two-track secondary vertices associated to the highest $p_T$ jet in each event is given in figure 4.4b. For jets with at least one secondary vertex
Measurement of the Z+b-jet cross section

Figure 4.5: Expected and observed distributions of (a) $D_{SSV}$ for jets having at least one secondary vertex and (b) the number of b-tagged jets after the dilepton selections. The shaded region reflects the statistical, jet energy scale and b-tagging efficiency uncertainties in the expected distribution. The large majority of light-flavour jets do not contain a reconstructed secondary vertex, so cannot be assigned a discriminator value and are not accounted for in (a). The expected distributions in (b) incorporate the b-tagging efficiency weighting for the yields in each bin.

discriminator $D_{SSV}$ is defined as a monotonic function of the three-dimensional flight distance significance between a primary and secondary vertex,

$$D_{SSV} = \ln \left(1 + \frac{\sigma_{3D}}{d_{3D}}\right). \quad (4.1)$$

The parameters $d_{3D}$ and $\sigma_{3D}$ are the flight distance and uncertainty respectively. The distribution of this discriminator is given in figure 4.5a for the leading jet in each event which passes the above selection and contains at least one secondary vertex. Jets are considered b-tagged if $D_{SSV} > 1.74$. This requirement is chosen to give a mistagging rate for light-flavour jets of close to 1% and an efficiency for b jets of around 65%. The b-tagging and mistagging efficiencies have been measured in both data and simulation [82] and found to differ by 8–10% depending on jet $p_T$ and $\eta$. These data-to-simulation scale factors are used to weight events in simulation. The weights take into account the true flavour of the jets by matching to generator-level hadron jets. The distribution of the number of b-tagged jets is given in figure 4.5b, where the prediction from simulation incorporates these event weights and is seen to yield good agreement with the data.

With this b-tagging definition two event populations are defined: those with exactly one (Z+1b) and those with two or more (Z+2b) b-tagged jets. In the Z+2b selection the background due to $t\bar{t}$
Figure 4.6: Expected and observed distributions of (a) the $E_T^{\text{miss}}$ and (b) $m_{\ell\ell}$ in events passing the Z+2b selection. The shaded region reflects the statistical, jet energy scale and b-tagging efficiency uncertainties in the expected distribution.

production is enhanced. As these events typically contain larger real $E_T^{\text{miss}}$ than in Z production, an additional requirement of $E_T^{\text{miss}} < 50$ GeV is applied in this selection. The distribution of the $E_T^{\text{miss}}$ is given in figure 4.6a. The observed and predicted event yields after these selections are summarised in table 4.2, where the predicted yields are given with the statistical uncertainty only. Yields are given for the Z+2b selection both with and without the $E_T^{\text{miss}}$ requirement, from which it is seen that this reduces the $t\bar{t}$ rate by around 65% and the signal rate by only 2%. The $m_{\ell\ell}$ distribution after this selection is given in figure 4.6b.

4.4 Kinematic observables

This section gives comparisons of event kinematics between data and simulation in the Z+2b final state. It is important that this process is well modelled as it is a major background in the ZH(bb) [117] and several BSM searches. Though these searches typically use data-driven methods to evaluate backgrounds, simulated samples are still important for extrapolation between control and signal regions.

In each figure in this section the shaded region indicates the systematic uncertainties associated with the jet energy scale (JES), the b-tagging efficiency and the limited number of simulated
Measurement of the Z+b-jet cross section

<table>
<thead>
<tr>
<th>Process</th>
<th>Z+1b</th>
<th>Z+2b w/o $E_T^{\text{miss}}$</th>
<th>Z+2b $E_T^{\text{miss}} &gt; 50$ GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z+l</td>
<td>1283.6 ± 15.2</td>
<td>3.0 ± 0.7</td>
<td>3.0 ± 0.7</td>
</tr>
<tr>
<td>Z+c</td>
<td>2631.5 ± 21.7</td>
<td>22.1 ± 2.0</td>
<td>21.7 ± 2.0</td>
</tr>
<tr>
<td>Z+b</td>
<td>5165.1 ± 30.4</td>
<td>244.7 ± 6.6</td>
<td>239.9 ± 6.6</td>
</tr>
<tr>
<td>t\bar{t}</td>
<td>586.3 ± 7.3</td>
<td>177.6 ± 4.0</td>
<td>61.8 ± 2.4</td>
</tr>
<tr>
<td>ZZ</td>
<td>61.5 ± 0.4</td>
<td>8.6 ± 0.2</td>
<td>8.4 ± 0.2</td>
</tr>
</tbody>
</table>

Table 4.2: The expected and observed event yields in the Z+1b and Z+2b selections, where the latter is given both with and without the $E_T^{\text{miss}} < 50$ GeV requirement which is used to suppress the t\bar{t} background. The errors on the expected yields account for the statistical uncertainty only.

A mild excess of data compared to expectation is observed at low $p_T$ in both distributions.

Figure 4.7: Expected and observed distributions of the $p_T$ of the (a) leading and (b) sub-leading b-tagged jets for events passing the Z+2b selection. The shaded region reflects statistical and systematic uncertainties in the expected distribution.

The difference in azimuthal angle between the dilepton and bb systems, $\Delta \phi_{Z,bb}$, is given in figure 4.8. This shows agreement both for the majority of events at high $\Delta \phi_{Z,bb}$, where the two
systems are mostly back-to-back, and for the events at lower values. This distribution is useful in testing the simulation of MPI. In these events the b-jet kinematics result from a separate parton scattering and so are less correlated with the Z boson kinematics. These events are therefore expected to have a flat distribution in this variable. This suggests a good modelling of MPI by PYTHIA, which has also been studied by CMS in W+dijet and four-jet events [118].

![Graph](image)

**Figure 4.8**: Expected and observed distributions of the difference in azimuthal angle between the Z boson candidate and b\(\bar{b}\) system for events passing the Z+2b selection. The shaded region reflects the statistical and systematic uncertainties in the expected distribution.

Figures 4.9a and 4.9b give the \(p_T\) distributions of the dilepton and bb systems respectively. The \(p_T^{le}\) distribution is found to be harder in data than simulation, with the largest discrepancy in the region 80–120 GeV. A possible explanation for this trend is that the \(p_T^{le}\) distribution in the 5FS, which is used in the MADGRAPH simulation, has been shown to be softer than in the 4FS at NLO in which the b quarks are massive [102].
Measurement of the Z+b-jet cross section

Figure 4.9: Expected and observed distributions of the $p_T$ of the (a) dilepton and (b) $bb$ systems for events passing the $Z+2b$ selection. The shaded region reflects the statistical and systematic uncertainties in the expected distribution.

4.5 Background estimation

In both the $Z+1b$ and $Z+2b$ selections, background events are expected from $Z+\text{jets}$ production, in which $c$- or light-flavour jets are mistagged as $b$ jets, and from $t\bar{t}$ production in which genuine $b$ jets are tagged. Partially data-driven methods are employed to estimate these backgrounds rather than relying solely on simulation. The methods discussed in this section are a summary of those presented in [2]. An additional small contribution from $ZZ$ diboson production is also expected. Other potential backgrounds require at least one misidentified lepton or $b$ jet, such as from QCD multijet, $W+\text{jets}$, single top, $WW$ and $WZ$ production. However, with the chosen lepton selection and $\sim 1\%$ mistagging rate of the SSV discriminator these are found to make a negligible contribution.

The fraction of $t\bar{t}$ events is estimated by a fit to the binned $m_{\ell\ell}$ distribution in each channel and selection. This exploits the fact that the $m_{\ell\ell}$ distribution is relatively flat for $t\bar{t}$ events and may be constrained by the data away from the $Z$ boson mass peak. The fit is therefore performed in an extended range of $60 < m_{\ell\ell} < 120$ GeV. The high $t\bar{t}$ purity further from the $Z$ boson peak gives greater constraint on the fraction of $t\bar{t}$ events compared to the nominal mass window of $76 < m_{\ell\ell} < 106$ GeV. The results of the fits in the $Z+2b$ selection are given in figure 4.10. The $Z+\text{jets}$ and $t\bar{t}$ fit templates are taken from simulation. An alternative parameterization of the $t\bar{t}$ component with an exponential curve is found to yield a compatible $t\bar{t}$ fraction. The difference
between the two $t\bar{t}$ models is taken as a systematic uncertainty. The $t\bar{t}$ fractions in the mass range $76 < m_{\ell\ell} < 106$ GeV are found to be: $f^{ee+1b}_{t\bar{t}} = 5.8 \pm 0.9\%$ and $f^{\mu\mu+1b}_{t\bar{t}} = 5.1 \pm 0.7\%$ in the $Z+1b$ selection, and $f^{ee+2b}_{t\bar{t}} = 17 \pm 7\%$ and $f^{\mu\mu+2b}_{t\bar{t}} = 19 \pm 6\%$ in the $Z+2b$ selection.

![Figure 4.10](image-url) Expected and observed distributions of $m_{\ell\ell}$ under the $Z+2b$ selection for the (a) $ee$ and (b) $\mu\mu$ selections. The normalisation of the $Z+\text{jets}$ and $t\bar{t}$ processes corresponds to the result of a maximum likelihood fit [2].

The fraction of events selected due to the mistagging of light and $c$-flavour jets is estimated by fits to the distributions of the secondary vertex mass, $m_{SV}$, of the $b$-tagged jets. Templates for each jet flavour are taken from simulation and both dielectron and dimuon events are combined. In the $Z+1b$ selection the signal purity is defined as $f_b = 1 - f_c - f_l$ where $f_c$ and $f_l$ are the light- and $c$-jet fractions determined in the fit respectively. The signal purity in the $Z+2b$ selection can be written as

$$f_{bb} = 1 - f_{cc} - f_{cb} - f_{bc} - f_{cl} - f_{lc} - f_{bl} - f_{lb} - f_{ll}$$

(4.2)

where the subscript ordering indicates the flavour of the leading and sub-leading jets. Of these fractions, only $f_{cc}$, $f_{bl}$ and $f_{lb}$ are expected to have an appreciable contribution. The expected contribution from the other fractions is therefore neglected in the fit, with the estimate from simulation taken as a systematic uncertainty. It is found to good approximation that the independent fractions $f_l$ and $f_c$ in the leading and sub-leading jet samples are identical. Therefore the fraction $f_{cc}$ may be associated to the fraction $f_c$, and $f_{bl}$ and $f_{lb}$ to $f_l$. The $Z+2b$
purity may then be simplified to

\[ f_{bb} = 1 - f_c - 2 \cdot f_l \]  \hspace{1cm} (4.3)

where the fractions \( f_c \) and \( f_l \) are determined in a simultaneous fit to the \( m_{SV} \) distributions of the leading and sub-leading tagged jets, subject to the constraint that the same fractions are found in both distributions. Figure 4.11 gives the \( m_{SV} \) distributions for the leading and sub-leading jets where the normalisation of each component corresponds to the result of the maximum likelihood fit. The fraction \( f_l \) in the \( Z+1b \) selection is found to be \( 54.9 \pm 1.7\% \) and the fraction \( f_{bb} \) in the \( Z+2b \) selection is \( 83.0 \pm 4.5\% \).

![Figure 4.11: Expected and observed distributions of \( m_{SV} \) under the \( Z+2b \) selection for the (a) leading and (b) sub-leading b-tagged jets, where the normalisation of the \( Z+ \) jets and \( tt \) processes correspond to the result of a maximum likelihood fit [2].](image)

The contribution of the \( ZZ \) background is estimated from simulation using the cross section and uncertainty from the CMS measurement in \cite{119}. The numbers of b-tagged signal events, \( N_{Z+1b}^{tag} \) and \( N_{Z+2b}^{tag} \), are then defined as:

\[ N_{Z+1b}^{tag} = (f_b - f_{tt})N_{Z+1b}^{obs} - N_{ZZ} \]  \hspace{1cm} (4.4)
\[ N_{Z+2b}^{tag} = \frac{1}{\epsilon(E_{T}^{\text{miss}})}((f_{bb} - f_{tt})N_{Z+2b}^{obs} - N_{ZZ}) \]  \hspace{1cm} (4.5)

where \( N_{Z+1b}^{obs} \) and \( N_{Z+2b}^{obs} \) are the observed yields in data, \( N_{ZZ} \) is the \( ZZ \) yield estimated from simulation and all quantities are evaluated under the \( Z+1b \) and \( Z+2b \) selections respectively. The
term $\varepsilon(E_T^{\text{miss}})$ is the efficiency of the $E_T^{\text{miss}} < 50$ GeV selection and is calculated using simulated signal events. An uncertainty of 2% is assigned to this, based on studies of $E_T^{\text{miss}}$ response uncertainty in [120].

### 4.6 Efficiency and acceptance corrections

In order to extract cross sections at the hadron level, the background-subtracted yields $N_{Z+b}^{\text{tag}}$ and $N_{Z+2b}^{\text{tag}}$ must be corrected for four effects: the efficiency of the b-tagging, the efficiency of the lepton selection, detector resolution effects and the acceptance of the leptons. One complication is that these corrections must account for migrations of events between the $Z+1b$ and $Z+2b$ bins, which may occur due to imperfect jet $p_T$ resolution or b-tagging inefficiency. The result is the hadron-level and acceptance-corrected yields, denoted as $N_{Z+1b}^{\text{had-acc}}$ and $N_{Z+2b}^{\text{had-acc}}$. The full set of corrections may be written as a matrix equation,

$$
\begin{pmatrix}
N_{Z+1b}^{\text{had-acc}} \\
N_{Z+2b}^{\text{had-acc}}
\end{pmatrix}
= 
A_\ell^{-1} \varepsilon_r^{-1} \varepsilon_l^{-1} \varepsilon_b^{-1}
\begin{pmatrix}
N_{Z+1b}^{\text{tag}} \\
N_{Z+2b}^{\text{tag}}
\end{pmatrix},
$$

(4.6)
in terms of four $2 \times 2$ matrices: $A_\ell$ for the lepton acceptance, $\varepsilon_r$ for the detector resolution, $\varepsilon_l$ for the lepton efficiency and $\varepsilon_b$ for the b-tagging efficiency. Only the $\varepsilon_r$ and $\varepsilon_b$ matrices contain off-diagonal elements which account for the migration between b-jet bins. The $\varepsilon_l$ correction concerns only the lepton triggering and identification efficiency and so cannot give rise to bin migrations. The lepton acceptance correction $A_\ell$ is needed so that the cross sections determined in the ee and $\mu\mu$ channels may be combined under a common lepton selection of $p_T > 25$ GeV and $|\eta| < 2.1$. All components of these corrections are estimated from the $Z+b$ MADGRAPH sample. However, the matrix equation is designed such that the $Z+1b$ to $Z+2b$ ratio, which is extracted from the observed yields, does not depend on the ratio being correctly modelled in the simulation.

As these corrections rely on the true particle content in the simulated events, a matching definition between reconstructed and generator-level objects is chosen as:

- Selected reconstructed jets must be matched to b-flavour hadron jets within $\Delta R = 0.5$. At the hadron level, a jet is taken to have b-flavour if a b hadron is found within $\Delta R = 0.5$ of the jet axis.
Measurement of the Z+b-jet cross section

- Selected reconstructed leptons must be matched to the leptons from the Z boson decay within $\Delta R = 0.3$. At the generator level these leptons are recombined with any photons from final state radiation.

### 4.6.1 b-Tagging efficiency

The matrix $\epsilon_b$ is used to correct for the b-tagging efficiency of b-flavour reconstructed jets. It is applied to the yields $N_{Z+b}^{\text{tag}}$ and $N_{Z+2b}^{\text{tag}}$, which result from the background-subtracted observed yields as given in equations 4.4 and 4.5. It is derived by first considering how these yields relate to the number of reconstructed Z+1b and Z+2b events, $N_{Z+1b}^{\text{reco-sel}}$ and $N_{Z+2b}^{\text{reco-sel}}$, via efficiency factors of the form $\epsilon_b^{XY}$. These represent the efficiency for tagging $Y$ jets, given $X$ b jets in the acceptance, where these reconstructed b jets are defined by the matching procedure outlined above. The b-tagged yields are then determined as

\[
N_{Z+2b}^{\text{tag}} = \epsilon_b^{22} N_{Z+2b}^{\text{reco-sel}} \tag{4.7}
\]

\[
N_{Z+1b}^{\text{tag}} = \epsilon_b^{11} N_{Z+1b}^{\text{reco-sel}} + \epsilon_b^{21} N_{Z+2b}^{\text{reco-sel}} \tag{4.8}
\]

where it is noted that the number of tagged Z+1b events depends on both the number of reconstructed Z+1b and Z+2b events. The latter implies events where only one of the two b jets passes the tagging requirement. By construction, the sums of the $\epsilon_b$ factors for a given b-jet multiplicity must equal unity:

\[
\epsilon_b^{20} + \epsilon_b^{21} + \epsilon_b^{22} = 1 \tag{4.9}
\]

\[
\epsilon_b^{10} + \epsilon_b^{11} = 1. \tag{4.10}
\]

However, no constraint is made on the ratio of events between the two multiplicity bins. Given these constraints, and equations 4.7 and 4.8, it is possible to solve for the corrected yields which may be written as the matrix equation

\[
\begin{pmatrix}
N_{Z+1b}^{\text{reco-sel}} \\
N_{Z+2b}^{\text{reco-sel}}
\end{pmatrix}
= \epsilon_b^{-1}
\begin{pmatrix}
N_{Z+1b}^{\text{tag}} \\
N_{Z+2b}^{\text{tag}}
\end{pmatrix}
= \begin{pmatrix}
\frac{1}{\epsilon_b^{20}} & -\frac{\epsilon_b^{21}}{\epsilon_b^{22}} \\
0 & \frac{1}{\epsilon_b^{21}}
\end{pmatrix}
\begin{pmatrix}
N_{Z+1b}^{\text{tag}} \\
N_{Z+2b}^{\text{tag}}
\end{pmatrix}. \tag{4.11}
\]

The sole off-diagonal element of this matrix is responsible for subtracting from $N_{Z+1b}^{\text{tag}}$ the number of events in which there are two b jets in the acceptance but only one is tagged.
The $\varepsilon_b$ factors are determined from simulated $Z+b$ events in which the b-tagging and mistagging efficiencies have been corrected to match those in data via event weights, as described previously. These event weights are subject to the uncertainty in the data-to-simulation scale factors measured in [82]. These uncertainties range from 3–12% for the tagging of b jets and 8–15% for the mistagging of light jets, dependent on the jet $p_T$ and $\eta$. These uncertainties are assumed to be fully correlated across the kinematic range of the jets selected in the analysis. Therefore, the uncertainty in the $\varepsilon_b$ factors is evaluated by coherently raising or lowering the scale factors by their ±1σ extent. Table 4.3 gives the values of the $\varepsilon_b$ factors in the ee and $\mu\mu$ channels, along with the relative uncertainties due to the limited number of simulated events and the b-tagging and mistagging scale factors. The b-tagging efficiency is found to be the dominant source of uncertainty.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>b-tag Syst. [%]</th>
<th>Mistag Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_b^{11}$</td>
<td>49.7 ± 2.3</td>
<td>0.7</td>
<td>4.5</td>
<td>0.02</td>
</tr>
<tr>
<td>$\varepsilon_b^{21}$</td>
<td>49.5 ± 2.4</td>
<td>2.5</td>
<td>4.1</td>
<td>-</td>
</tr>
<tr>
<td>$\varepsilon_b^{22}$</td>
<td>26.8 ± 2.2</td>
<td>4.0</td>
<td>7.3</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>b-tag Syst. [%]</th>
<th>Mistag Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_b^{11}$</td>
<td>49.7 ± 2.3</td>
<td>0.6</td>
<td>4.5</td>
<td>0.02</td>
</tr>
<tr>
<td>$\varepsilon_b^{21}$</td>
<td>50.4 ± 2.0</td>
<td>1.9</td>
<td>3.4</td>
<td>-</td>
</tr>
<tr>
<td>$\varepsilon_b^{22}$</td>
<td>26.2 ± 2.2</td>
<td>3.4</td>
<td>7.5</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 4.3: Values, expressed as percentages, of the b-tagging efficiency factors given with the total estimated uncertainties. The relative uncertainties from the limited number of simulated events (Stat.) and the b-tagging and mistagging efficiencies are also given. The uncertainties on the $\varepsilon_b^{21}$ factors due to the mistagging variation were found to be negligible and are omitted.

### 4.6.2 Lepton selection efficiency

The matrix $\varepsilon_l$ corrects for the dilepton selection inefficiency and includes contributions from the triggering, identification and isolation criteria. Separate factors $\varepsilon_l^1$ and $\varepsilon_l^2$ are calculated for each b-jet multiplicity bin. Although jets are required to be separated from each lepton by at least $\Delta R = 0.5$, it is still possible that the number of jets in the event could bias the lepton selection efficiency.
Each factor is defined as

\[ \epsilon_l^{1(2)} = \frac{N_{\text{reco-sel}}^{Z+1(2)b}}{N_{\text{reco}}^{Z+1(2)b}}. \]

These are the fractions of Z+1b or Z+2b events in which the reconstructed leptons, matched to their generator-level counterparts, pass the trigger and selection requirements. This leads to a trivial matrix representation of

\[
\begin{pmatrix}
N_{\text{reco}}^{Z+1b} \\
N_{\text{reco}}^{Z+2b}
\end{pmatrix} = \epsilon_l^{-1}
\begin{pmatrix}
N_{\text{reco-sel}}^{Z+1b} \\
N_{\text{reco-sel}}^{Z+2b}
\end{pmatrix} = \begin{pmatrix}
\frac{1}{\epsilon_l} & 0 \\
0 & \frac{1}{\epsilon_l}
\end{pmatrix}
\begin{pmatrix}
N_{\text{reco-sel}}^{Z+1b} \\
N_{\text{reco-sel}}^{Z+2b}
\end{pmatrix}.
\]

One source of systematic uncertainty in the \( \epsilon_l \) factors comes from the tag-and-probe method used to determine the trigger, identification and isolation scale factors which correct these efficiencies in simulation. These result from statistical uncertainty in the tag-and-probe fits and from the choice of signal and background parameterization. These are typically 1% for each muon and 2% for each electron. Another contribution is from the uncertainty in the frequency of pileup interactions in data to which the simulation is corrected. This is estimated by shifting the mean of the distribution in the simulated sample by \( \pm 0.6 \) interactions, which reflects the \( \pm 1 \sigma \) uncertainty for the shape of the distribution in data. These results are summarised in table 4.4. The uncertainty due to pileup is found to be larger for the ee channel. This is mostly due to a greater sensitivity to pileup in the ECAL and HCAL isolation requirements compared to those for muons. Values of \( \epsilon_l \) between the one and two b-jet bins are found to be compatible within the statistical uncertainty of the sample, confirming that the bias due to the jet selection is indeed small.

### 4.6.3 Reconstruction efficiency

The matrix \( \epsilon_l \) accounts for detector reconstruction and resolution effects for both the selected leptons and b jets. It relates the numbers of reconstructed events, \( N_{\text{reco}}^{Z+1b} \) and \( N_{\text{reco}}^{Z+2b} \), to the numbers at the hadron level, denoted \( N_{\text{had}}^{Z+1b} \) and \( N_{\text{had}}^{Z+2b} \). This requires efficiency terms, \( \epsilon_l^{XY} \), which correspond to the efficiency for obtaining a \( Z + Yb \) event at the reconstructed level given a \( Z + Xb \) event at the hadron level. However, if the event does not pass the lepton selections at either the reconstructed or generator level, then \( X \) or \( Y \) is taken as zero, regardless of the b-jet multiplicity. Reconstructed leptons and jets must be matched to the corresponding generator-level objects as described previously. It is also necessary to consider the contribution of events
ee Selection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>Tag-and-probe [%]</th>
<th>Pileup [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon^1_{\ell}$</td>
<td>62.9 ± 2.7</td>
<td>0.4</td>
<td>4.0</td>
<td>1.7</td>
</tr>
<tr>
<td>$\epsilon^2_{\ell}$</td>
<td>63.7 ± 2.9</td>
<td>1.4</td>
<td>4.0</td>
<td>1.8</td>
</tr>
</tbody>
</table>

$\mu\mu$ Selection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>Tag-and-probe [%]</th>
<th>Pileup [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon^1_{\ell}$</td>
<td>84.7 ± 1.6</td>
<td>0.2</td>
<td>1.9</td>
<td>0.3</td>
</tr>
<tr>
<td>$\epsilon^2_{\ell}$</td>
<td>84.0 ± 1.7</td>
<td>0.8</td>
<td>1.9</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 4.4: Values, expressed as percentages, of the lepton selection efficiency factors given with the total estimated uncertainties. The relative uncertainties from the limited number of simulated events (Stat.) and the tag-and-probe and pileup re-weighting are also given.

which do not have any b jet in the acceptance at the hadron level, but nonetheless have a reconstructed b jet in the acceptance. This may occur, for example, when an event has a single hadron-level b jet with $p_T$ just below the 25 GeV threshold which, due to the detector energy resolution, is measured as being just above threshold. The number of such events with no b jet in the acceptance is denoted $N_{\text{had}}^{Z+b}$. The reconstructed-level yields may then be written as

$$N_{Z+b}^{\text{reco}} = \epsilon_{r}^{01} N_{Z+b}^{\text{had}} + \epsilon_{r}^{11} N_{Z+b}^{\text{had}} + \epsilon_{r}^{21} N_{Z+b}^{\text{had}}$$

(4.14)

$$N_{Z+b}^{\text{reco}} = \epsilon_{r}^{02} N_{Z+b}^{\text{had}} + \epsilon_{r}^{12} N_{Z+b}^{\text{had}} + \epsilon_{r}^{22} N_{Z+b}^{\text{had}}.$$  

(4.15)

Similar to the $\epsilon_b$ factors, the $\epsilon_r$ factors have the property of summing to unity in a given hadron-level jet bin:

$$\epsilon_r^{00} + \epsilon_r^{01} + \epsilon_r^{02} = 1$$

(4.16)

$$\epsilon_r^{10} + \epsilon_r^{11} + \epsilon_r^{12} = 1$$

(4.17)

$$\epsilon_r^{20} + \epsilon_r^{21} + \epsilon_r^{22} = 1,$$ 

(4.18)

and thus are not dependent on the ratios between b-jet bins being correctly modelled in the simulation. It is noted that the two hadron-level yields of interest are underdetermined in equations 4.14 and 4.15 due to the presence of the $N_{\text{had}}^{Z+b}$ term. It is not possible to write a third equation for a reconstructed-level yield without a b jet in the acceptance, as there is no means by which this can be found from the data. A solution is to apply a further constraint, chosen as the ratio of hadron-level events outside the acceptance, to those within the acceptance and having any
number of $b$ jets. This is defined as

$$R = \frac{N_{Z+1b}^{\text{had}}}{N_{Z+2b}^{\text{had}}} + N_{Z+2b}^{\text{had}},$$

and assumed to be fixed. This factor is determined from the $Z+b$ sample so is assumed to be correctly modelled by simulation. However, the impact on the final yields is limited, as it only enters the corrections as a product of the small off-diagonal $\epsilon_r$ efficiencies. Importantly, it does not depend on the $Z+1b$ to $Z+2b$ ratio in the sample being correct. With this constraint, equations 4.14 and 4.15 are equivalent to

$$N_{Z+1b}^{\text{reco}} = (\epsilon_{1r} + Re^{01}_{r}) N_{Z+1b}^{\text{had}} + (\epsilon_{2r} + Re^{01}_{r}) N_{Z+2b}^{\text{had}},$$

$$N_{Z+2b}^{\text{reco}} = (\epsilon_{1r} + Re^{02}_{r}) N_{Z+1b}^{\text{had}} + (\epsilon_{2r} + Re^{02}_{r}) N_{Z+2b}^{\text{had}}.$$

This may then be written in the form of a matrix equation for $N_{Z+1b}^{\text{had}}$ and $N_{Z+2b}^{\text{had}}$:

$$\begin{pmatrix} N_{Z+1b}^{\text{had}} \\ N_{Z+2b}^{\text{had}} \end{pmatrix} = \epsilon_r^{-1} \begin{pmatrix} N_{Z+1b}^{\text{reco}} \\ N_{Z+2b}^{\text{reco}} \end{pmatrix} = \begin{pmatrix} \epsilon_{1r} + Re^{01}_{r} & \epsilon_{1r} + Re^{01}_{r} \\
\epsilon_{2r} + Re^{02}_{r} & \epsilon_{2r} + Re^{02}_{r} \end{pmatrix}^{-1} \begin{pmatrix} N_{Z+1b}^{\text{reco}} \\ N_{Z+2b}^{\text{reco}} \end{pmatrix}.$$

The values and uncertainties in each $\epsilon_r$ parameter and the ratio $R$ are given in table 4.5. Some differences in parameter values between the $ee$ and $\mu\mu$ channels are observed, due to the difference in lepton acceptance biasing the kinematics of the jets. The dominant sources of systematic uncertainty in these factors come from the limited number of events and uncertainties in the JES and jet energy resolution (JER) in simulation. The effect of the JES uncertainty is found by varying the $p_T$ of all jets up and down coherently by the $1\sigma$ estimates as determined in [79]. The per-jet scale uncertainty ranges from around 5% at 25 GeV to 2% at 100 GeV, and incorporates a dedicated uncertainty to account for differences in scale for heavy and light flavour jets. For the JER, the default resolution in simulation is found to be better than that measured in data. An additional 10–20% smearing of the jet $p_T$, as a function of $p_T$ and $\eta$, is applied for the nominal efficiency calculation such that the resolution matches that expected in data. The uncertainty is then taken as the maximal variation between the factors measured with the nominal resolution and for the case of the corrected resolution with a further 10% smearing. Such a variation in resolution is expected to be conservative. Even given this, the contribution to the total uncertainty is minimal as this is dominated by the JES, which can be as large as 25%.
### Measurement of the Z+b-jet cross section

#### ee Selection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>JES Syst. [%]</th>
<th>JER Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{r}^{11}$</td>
<td>71.6 ± 1.3</td>
<td>0.3</td>
<td>±1.8</td>
<td>±0.05</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{21}$</td>
<td>10.2 ± 2.4</td>
<td>5.4</td>
<td>±23.0</td>
<td>±0.49</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{12}$</td>
<td>1.1 ± 0.3</td>
<td>5.1</td>
<td>±25.3</td>
<td>±0.97</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{22}$</td>
<td>69.9 ± 2.6</td>
<td>1.1</td>
<td>±3.6</td>
<td>±0.04</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{01}$</td>
<td>0.78 ± 0.17</td>
<td>1.7</td>
<td>±21.1</td>
<td>±0.8</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{02}$</td>
<td>0.017 ± 0.004</td>
<td>11.8</td>
<td>±23.5</td>
<td>-</td>
</tr>
<tr>
<td>$R$</td>
<td>12.93 ± 0.05</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

#### $\mu\mu$ Selection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>JES Syst. [%]</th>
<th>JER Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{r}^{11}$</td>
<td>81.8 ± 1.5</td>
<td>0.2</td>
<td>±1.8</td>
<td>±0.04</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{21}$</td>
<td>13.1 ± 2.1</td>
<td>4.5</td>
<td>±15.4</td>
<td>±0.24</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{12}$</td>
<td>1.1 ± 0.3</td>
<td>5.2</td>
<td>±27.7</td>
<td>±1.1</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{22}$</td>
<td>77.2 ± 2.3</td>
<td>0.9</td>
<td>±2.9</td>
<td>±0.10</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{01}$</td>
<td>0.78 ± 0.18</td>
<td>1.6</td>
<td>±22.8</td>
<td>±0.25</td>
</tr>
<tr>
<td>$\varepsilon_{r}^{02}$</td>
<td>0.008 ± 0.003</td>
<td>12.5</td>
<td>±31.3</td>
<td>-</td>
</tr>
<tr>
<td>$R$</td>
<td>12.17 ± 0.05</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 4.5:** Values, expressed as percentages, of the reconstruction efficiency factors given with the total estimated uncertainties. The relative uncertainties from the limited number of simulated events (Stat.) and the jet energy scale (JES) and jet energy resolution (JER) uncertainties are also given. A dash indicates the corresponding uncertainty has a negligible effect on the given efficiency factor. The effect of uncertainties on parameters denoted with a ± symbol are anti-correlated with those having a ± symbol.
4.6.4 Lepton acceptance

A final, optional, correction can be applied to the yields \(N_{\text{had}}^{Z+1b}\) and \(N_{\text{had}}^{Z+2b}\), to account for a different choice of kinematic acceptance on the generator-level leptons. A correction is also included for lepton final-state radiation such that the resulting acceptance is defined at the Born level. As for the lepton efficiency correction, separate factors are calculated for the \(Z+1b\) and \(Z+2b\) bins, denoted \(A_1^\ell\) and \(A_2^\ell\). As these factors solely involve the leptons, this correction does not allow migration between b-jet multiplicity bins.

The acceptance-corrected yields, \(N_{\text{had-acc}}^{Z+1b}\) and \(N_{\text{had-acc}}^{Z+2b}\), are given by the matrix equation

\[
\begin{pmatrix}
N_{\text{had-acc}}^{Z+1b} \\
N_{\text{had-acc}}^{Z+2b}
\end{pmatrix}
= \begin{pmatrix}
A_1^\ell & 0 \\
0 & A_2^\ell
\end{pmatrix}
\begin{pmatrix}
N_{\text{had}}^{Z+1b} \\
N_{\text{had}}^{Z+2b}
\end{pmatrix}.
\]

(4.23)

Two sets of values are calculated and are given in tables 4.6 and 4.7. Those in table 4.6 consider a minimal extrapolation of the lepton \(p_T\) and \(\eta\) distributions, such that the electron and muon results can be compared directly. This requires extrapolating the electron \(p_T\) threshold from 25 GeV down to 20 GeV in the ee channel and extrapolating the muon \(\eta\) out to 2.5 from 2.1 in the \(\mu\mu\) channel. In table 4.7 the factors completely remove the kinematic selections, and these are used to facilitate a comparison with the previous CMS measurement of the inclusive \(Z+b\) cross section in [1]. Both tables give the uncertainties in each factor due to the limited number of events, the choice of PDF, the renormalisation and factorisation scales, \(\mu_R\) and \(\mu_F\), and the choice of MC generator. The scale and PDF uncertainties are determined with the mcfm [105] NLO parton-level calculator. The former is calculated by evaluating \(A_1\) for the grid of values \((\mu_R = \frac{1}{2}, 1, 2 \cdot m_Z) \times (\mu_F = \frac{1}{2}, 1, 2 \cdot m_Z)\). The uncertainty is then the maximal variation with respect to the nominal \(\mu_R = \mu_F = m_Z\). The latter is evaluated by finding the maximal variation on \(A_1\) in the cteq6mE PDF set [121]. The model uncertainty is taken as the maximum difference in the values of the extrapolation factors between the MadGraph, amc@NLO and mcfm predictions. A larger total uncertainty in the factors is observed in the muon channel for the minimal extrapolation in table 4.6. This may be due to the fact that the extrapolation is in \(\eta\), compared to the electron channel which is in \(p_T\).
### Measurement of the Z+b-jet cross section

#### Table 4.6: Lepton acceptance factors that correct the yields in both channels to a common acceptance of $p_T > 20$ GeV and $|\eta| < 2.5$ for the leptons. The total uncertainty includes contributions from the limited number of simulated events, from PDF and scale variations and from a comparison with the MCFM and amC@NLO predictions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>MC Stat. [%]</th>
<th>PDF Syst. [%]</th>
<th>Scale Syst. [%]</th>
<th>Model Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1^e$</td>
<td>$84.4 \pm 1.5$</td>
<td>0.2</td>
<td>1.0</td>
<td>$^{+0.6}_{-0.8}$</td>
<td>1.2</td>
</tr>
<tr>
<td>$A_2^e$</td>
<td>$83.8 \pm 5.1$</td>
<td>0.7</td>
<td>5.6</td>
<td>$^{+0.6}_{-2.1}$</td>
<td>1.2</td>
</tr>
</tbody>
</table>

#### Table 4.7: Lepton acceptance factors that are used to entirely remove the lepton $p_T$ and $\eta$ requirements. The total uncertainty includes contributions from the limited number of simulated events, from PDF and scale variations and from a comparison with the MCFM and amC@NLO predictions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>PDF Syst. [%]</th>
<th>Scale Syst. [%]</th>
<th>Model Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1^\mu$</td>
<td>$88.2 \pm 2.7$</td>
<td>0.2</td>
<td>1.6</td>
<td>$^{+0.1}_{-2.6}$</td>
<td>3.3</td>
</tr>
<tr>
<td>$A_2^\mu$</td>
<td>$88.9 \pm 6.0$</td>
<td>0.6</td>
<td>5.9</td>
<td>$^{+1.9}_{-2.5}$</td>
<td>3.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>PDF Syst. [%]</th>
<th>Scale Syst. [%]</th>
<th>Model Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1^e$</td>
<td>$55.7 \pm 4.0$</td>
<td>0.4</td>
<td>1.0</td>
<td>$^{+0.4}_{-0.8}$</td>
<td>+6.5</td>
</tr>
<tr>
<td>$A_2^e$</td>
<td>$55.6 \pm 5.1$</td>
<td>1.3</td>
<td>5.1</td>
<td>$^{+3.8}_{-1.7}$</td>
<td>+6.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value [%]</th>
<th>Stat. [%]</th>
<th>PDF Syst. [%]</th>
<th>Scale Syst. [%]</th>
<th>Model Syst. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1^\mu$</td>
<td>$58.2 \pm 4.0$</td>
<td>0.3</td>
<td>1.6</td>
<td>$^{+0.1}_{-2.6}$</td>
<td>+6.5</td>
</tr>
<tr>
<td>$A_2^\mu$</td>
<td>$60.1 \pm 5.2$</td>
<td>1.1</td>
<td>5.7</td>
<td>$^{+1.6}_{-3.2}$</td>
<td>+6.5</td>
</tr>
</tbody>
</table>
4.7 Results

Based on the results of the previous section, the full matrix equation for the corrections to the observed event yields is

$$
\begin{pmatrix}
N_{\text{had-acc}}^{Z+b-jet + 1} \\
N_{\text{had-acc}}^{Z+b-jet + 2}
\end{pmatrix}
= \begin{pmatrix}
\frac{1}{\mathcal{A}_l} & 0 \\
0 & \frac{1}{\mathcal{A}_l}
\end{pmatrix}
\begin{pmatrix}
e^{11}_r & e^{01}_r + R_{e^{01}} \\
e^{12}_r & e^{02}_r + R_{e^{02}}
\end{pmatrix}^{-1}
\begin{pmatrix}
\frac{1}{\mathcal{A}_l} & 0 \\
0 & \frac{1}{\mathcal{A}_l}
\end{pmatrix}
\begin{pmatrix}
e^{11}_r & e^{01}_r + R_{e^{01}} \\
e^{12}_r & e^{02}_r + R_{e^{02}}
\end{pmatrix}
\begin{pmatrix}
N_{\text{tag}}^{Z+1b} \\
N_{\text{tag}}^{Z+2b}
\end{pmatrix}.
\tag{4.24}
$$

Table 4.8 summarises the uncertainties in these yields, which are divided into those correlated between channels and b-jet bins and those which are uncorrelated. The effect of each source of uncertainty is propagated through the background subtraction and efficiency corrections, taking into account correlated effects between different efficiency parameters. Experimental systematic uncertainties are found to dominate both the $Z+b$ and $Z+2b$ measurements, with the largest effects coming from uncertainties in the b-jet purity, $t\bar{t}$ estimation and b-tagging efficiency.

The hadron level yields determined in equation 4.24 are then used directly to determine the cross sections as

$$
\sigma_{Z+1(2)b} = \frac{N_{\text{had-acc}}^{Z+1(2)b}}{\mathcal{L}}
\tag{4.25}
$$

where $\mathcal{L}$ is the integrated luminosity of the data. The cross-section results in each channel are given in table 4.9 for a lepton acceptance of $p_T > 20$ GeV and $|\eta| < 2.5$. The results in the $ee$ and $\mu\mu$ channels are found to be compatible, so they are combined into single measurements by taking the weighted mean, and taking into account all correlated and uncorrelated uncertainties. For the exclusive $Z+1b$ selection this gives a cross section of $3.41 \pm 0.05_{\text{(stat.)}} \pm 0.27_{\text{(syst.)}} \pm 0.09_{\text{(theory)}}$ pb; and for the inclusive $Z+2b$ selection a cross section of $0.37 \pm 0.02_{\text{(stat.)}} \pm 0.07_{\text{(syst.)}} \pm 0.02_{\text{(theory)}}$ pb. Though not measured explicitly, the contribution to the latter cross section from the MPI process is expected to be less than 5% [111]. The $Z+1b$ and $Z+2b$ measurements may also be combined into a single inclusive $Z+b$ cross section of $3.78 \pm 0.05_{\text{(stat.)}} \pm 0.31_{\text{(syst.)}} \pm 0.11_{\text{(theory)}}$ pb, again taking into account which uncertainties are correlated or uncorrelated between jet bins.

To facilitate a comparison with the previous inclusive measurement by CMS, in which no lepton acceptance selections were applied, the cross section obtained with the $\mathcal{A}_l$ factors in table 4.7
### Table 4.8: Fractional uncertainties on the hadron-level event yields in each channel and b-jet multiplicity bin. The sources of uncertainty which are fully correlated between each measurement are listed first, followed by uncorrelated sources. These are the limited event yields in simulation and the electron and muon selection efficiencies, which are only correlated between b-jet bins within the respective channels. The total statistical, experimental and theoretical uncertainties are also given.

<table>
<thead>
<tr>
<th>Source</th>
<th>ee (%)</th>
<th>μμ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correlated sources</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b-Jet purity</td>
<td>3.5</td>
<td>10.3</td>
</tr>
<tr>
<td>t̅t̅ contribution</td>
<td>0.9</td>
<td>8.9</td>
</tr>
<tr>
<td>b-Tagging efficiency</td>
<td>4.0</td>
<td>7.4</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>3.9</td>
<td>6.9</td>
</tr>
<tr>
<td>Luminosity</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>E_{miss}^{T} selection</td>
<td>0.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Pileup</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>ZZ contribution</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Jet energy resolution</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Mistagging rate</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Theory (via \mathcal{A}_t)</strong></td>
<td>1.8</td>
<td>5.9</td>
</tr>
<tr>
<td><strong>Uncorrelated sources</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC sample stat.</td>
<td>1.2</td>
<td>5.1</td>
</tr>
<tr>
<td>Dilepton selection</td>
<td>4.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

**Statistical**

<table>
<thead>
<tr>
<th>Source</th>
<th>ee (%)</th>
<th>μμ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>2.4</td>
<td>10.0</td>
</tr>
<tr>
<td>Experimental systematic</td>
<td>9.1</td>
<td>18.9</td>
</tr>
<tr>
<td>Theoretical systematic</td>
<td>1.8</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Table 4.9: Cross-section results for each b-jet multiplicity bin and channel. The three errors correspond to the statistical, systematic and theory uncertainties respectively.

<table>
<thead>
<tr>
<th>Cross Section [pb]</th>
<th>ee</th>
<th>$\mu\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{Z+1b}$</td>
<td>$3.25 \pm 0.08 \pm 0.29 \pm 0.06$</td>
<td>$3.47 \pm 0.06 \pm 0.27 \pm 0.11$</td>
</tr>
<tr>
<td>$\sigma_{Z+2b}$</td>
<td>$0.39 \pm 0.04 \pm 0.07 \pm 0.02$</td>
<td>$0.36 \pm 0.03 \pm 0.07 \pm 0.03$</td>
</tr>
<tr>
<td>$\sigma_{Z+b}$</td>
<td>$3.64 \pm 0.09 \pm 0.35 \pm 0.08$</td>
<td>$3.83 \pm 0.07 \pm 0.31 \pm 0.14$</td>
</tr>
</tbody>
</table>

is found to be $5.72 \pm 0.09\text{ (stat.)} \pm 0.47\text{ (syst.)} \pm 0.39\text{ (theory)}$ pb. The previous measurement was determined in a mass range of $60 < m_Z < 120\text{ GeV}$, and must be corrected down by approximately 3% for a range of $76 < m_Z < 106\text{ GeV}$. After this correction, the previous result is $5.73 \pm 0.08 \pm 0.75^{+0.24}_{-0.55}$ pb, so in good agreement with the result presented here.

The measurement of the inclusive $Z+b$ cross section by the ATLAS experiment $\cite{108}$, for leptons with $p_T > 20\text{ GeV}$ and $|\eta| < 2.5$ and b jets with $p_T > 25\text{ GeV}$ and $|\eta| < 2.1$, gives a result of $3.55 \pm 0.82\text{ (stat.)} \pm 0.73\text{ (syst.)} \pm 0.12\text{ (lumi.)}$ pb. Though not directly comparable to the result presented here, this is found to be in agreement with the expectation from McfM of $3.88 \pm 0.58$ pb after taking into account corrections from the parton to hadron level.

In the updated results from CMS $\cite{111}$, using the full 2011 dataset of 5 fb$^{-1}$, the measured $Z+1b$ and $Z+2b$ cross sections are compared to the predictions of several MC generators. These are given in figure 4.12 for McfM, amc@NLO and madgraph. The amc@NLO and madgraph predictions are given for calculations in both the 4FS and 5FS. The measured cross sections are found to be compatible, within uncertainties, with the 4FS predictions for the $Z+2b$ final state. However, these predict lower cross sections for the $Z+1b$ final state than observed. This discrepancy is largest for the amc@NLO prediction. The measured cross sections are found to be compatible with the amc@NLO and madgraph 5FS predictions.
Measurement of the $Z+b$-jet cross section

**Figure 4.12:** Comparisons of the CMS measured cross sections for (a) $Z+1b$ and (b) $Z+2b$ production with predictions from MC@NLO, aMC@NLO and MADGRAPH [111, 122]. For the aMC@NLO and MADGRAPH predictions, separate values are given for the four- and five-flavour schemes. Uncertainties in the predicted cross sections account for the uncertainties in renormalisation and factorisation scale, the choice of PDF and the jet matching scale.
Chapter 5

Searches for neutral Higgs bosons decaying to tau pairs

This chapter describes two searches for Higgs boson production and subsequent decay to tau pairs. One search is in the context of the SM, which is important given the discovery by the ATLAS and CMS Collaborations of a Higgs boson with mass around 125 GeV [32, 33, 123]. Confirmation of the $\tau\tau$ decay would constitute evidence for the Yukawa couplings which are predicted to be responsible for fermion masses. The search is performed for Higgs boson mass hypotheses in the range 90–150 GeV, where the SM branching ratio to $\tau\tau$ is favourable at 2–8%. The other search is in the context of the MSSM, in which three neutral Higgs bosons are predicted: the CP-even states $h$ and $H$ and the CP-odd state $A$, which may all decay to tau pairs. The search is performed in the $m_A$-$\tan\beta$ parameter space of the $m_h^{\text{max}}$ scenario [46] for $m_A$ in the range 90 GeV to 1 TeV. The analysis is sensitive to production via gluon-gluon fusion and production in association with b-quarks. The cross section of the latter increases for larger values of $\tan\beta$ due to the enhanced down-type fermion Yukawa couplings. A model-independent search for a single Higgs boson, denoted $\Phi$, is performed in the same mass range.

Throughout this chapter the symbol $\tau_h$ refers to a hadronic tau decay, $\ell$ to an electron or muon and $L$ to either an electron, muon or $\tau_h$.

Searches for the SM $H \to \tau\tau$ decay have previously been performed by experiments at the LEP and Tevatron colliders [124–129]. The collaborations at LEP were able to search using the $Z$ boson associated production mode and found no significant indication of Higgs boson decays. Upper limits at the 95% CL were determined by the CDF and D0 Collaborations at 16 and 14 times the SM expectation for a Higgs boson with mass 125 GeV. Recent results from the ATLAS and CMS Collaborations [6,130] have found evidence for $H \to \tau\tau$ decays each with significances larger than three standard deviations, and compatible with the expectation for a 125 GeV Higgs
Searches for neutral Higgs bosons decaying to tau pairs

boson. The CMS search utilises all $\tau\tau$ final states and combines an inclusive analysis targeting gluon-gluon fusion and VBF production with a dedicated analysis searching for $W/Z$-associated production. The analysis described in this chapter follows the inclusive analysis but is restricted to the final states $\mu\tau_h$, $e\tau_h$ and $e\mu$ to which the author contributed.

Searches for MSSM neutral Higgs bosons have also been performed by the collaborations at LEP [131] and the Tevatron [132], and the ATLAS and CMS Collaborations [8,133], with no excess observed above the background expectation. The results presented here follow those published by CMS but again are restricted to the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels.

Both searches described in this chapter use the entire proton-proton collision datasets recorded by CMS during 2011 and 2012. This corresponds to an integrated luminosity of 4.9 fb$^{-1}$ at a centre-of-mass energy of 7 TeV and 19.7 fb$^{-1}$ at 8 TeV. Both analyses share a common triggering strategy, object and event selection. Event categorisation is used to enhance sensitivity to particular production modes and results are extracted from distributions of the di-tau invariant mass, $m_{\tau\tau}$.

The main irreducible background in all three channels is from the Drell-Yan production of a $Z$ boson which decays to tau pairs. Events with the production of a $t\bar{t}$ pair can also lead to the same final states, and this is a large background in the $e\mu$ channel. The main reducible backgrounds are from QCD multijet production and, in the $\mu\tau_h$ channels, from $W$ boson production in which an additional jet is misidentified as the $\tau_h$ candidate.

Section 5.1 outlines the triggering of events in data and the use of MC simulation. The event selection is described in section 5.2 and the estimation of the di-tau invariant mass in section 5.3. The split of events into exclusive categories is described in section 5.4. The estimation of each background process, using data-driven methods where possible, is detailed in section 5.5 and followed by a summary of the experimental and theoretical uncertainties affecting the signal and background estimations in section 5.6. The statistical procedures used to quantify the presence of signal in the data is given in section 5.7 and is followed by the results in the SM and MSSM searches in sections 5.8 and 5.9 respectively.

5.1 Data samples and simulation

Events for each channel are selected by dedicated trigger algorithms which require the appropriate pair of electron, muon or tau objects. For the $e\tau_h$ and $\mu\tau_h$ channels this starts with the requirement of a single electron or muon candidate in the L1 trigger. The $e\mu$ channel requires both an electron and muon at L1. In the HLT the electron and muon candidates are required to be loosely identified and isolated which exploits the tracking information available at this
stage. Additionally, a non-overlapping $\tau_h$ object must be identified in the $\ell\tau_h$ channels. For this a simplified version of the PF algorithm is run for both $\tau_h$ reconstruction and isolation. The isolation requirement is designed to be loose with respect to the analysis selection. As this is computationally expensive to run, this reconstruction only proceeds in the case that a good electron or muon candidate is identified first. Some trigger object $p_T$ thresholds were increased throughout the data-taking period to maintain stable rates in the presence of increasing instantaneous luminosity. In the $e\tau_h$ trigger the $p_T$ thresholds ranged from 15–22 GeV for the electron and 15–20 GeV for the $\tau_h$. In the $\mu\tau_h$ trigger the thresholds were 12–18 GeV and 10–20 GeV respectively. For the $e\mu$ channel events are accepted by either of two complementary triggers, with fixed thresholds throughout the entire run. One trigger has $p_T$ thresholds of 8 GeV for the electron and 17 GeV for the muon and the other vice versa. The object properties determined in the trigger reconstruction, such as $p_T$ and isolation, are only approximate to those in the full offline reconstruction. Consequently, the triggering of events that would pass the offline event selection is not fully efficient. The trigger efficiencies for offline electron, muon and $\tau_h$ candidates typically reach a plateau of between 80% and 95% when above the trigger $p_T$ thresholds [71,72].

Several MC generators are employed to produce simulated samples of signal and background events. The MadGraph [91] matrix element generator is used for $Z+$jets, $W+$jets, $t\bar{t}+$jets and diboson production. To increase the number of simulated $Z+$jets and $W+$jets events passing the most signal-sensitive category selections, additional samples are generated with fixed jet multiplicity in the matrix element for up to four jets. These are combined with the inclusive samples by weighting events to maintain the LO cross-section ratios between jet multiplicity bins. The Powheg [92,93,134] generator is used for single top-quark production. The SM gluon-gluon fusion and VBF production modes of the Higgs boson are also simulated with Powheg at NLO precision. Production in association with a vector boson and both MSSM modes are provided by Pythia [87] at LO only. All samples utilise Pythia for parton showering and hadronisation and Tauola [97] for tau decays. The Pythia tunes Z2 and Z2* [98] are used for 7 TeV and 8 TeV simulation respectively. Additional proton-proton interactions are also simulated with Pythia and added to these events. The events are then weighted according to the number of these simulated pileup interactions to match the distribution expected in data, as given in section 2.1.

Table 5.1 summarises the simulated processes used in the analysis and the cross sections used for normalisation. Signal processes are normalised using the cross sections and branching ratios determined by the LHC Higgs Cross Section Working Group (LHCHXSWG) [31,106,135]. An additional correction is applied in the SM gluon-gluon fusion signal samples to exploit recent improved predictions. Events are weighted to match the Higgs boson $p_T$ distribution
calculated at NNLO using the HRes [136] program. This also includes the resummation at next-to-next-to-leading logarithmic (NNLL) accuracy of terms of the form $\ln(m_{H^0}^2/p_T^2)$ which are particularly important for low Higgs boson $p_T$. An event weight for the difference between the finite and infinite top mass approximations is also applied [137].

<table>
<thead>
<tr>
<th>Process</th>
<th>Generator</th>
<th>Cross Section [pb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM ggH($\tau\tau$)</td>
<td>POWHEG</td>
<td>0.96</td>
</tr>
<tr>
<td>SM qqH($\tau\tau$)</td>
<td>POWHEG</td>
<td>0.077</td>
</tr>
<tr>
<td>SM ZH($\tau\tau$)+WH($\tau\tau$)+tH($\tau\tau$)</td>
<td>PYTHIA</td>
<td>0.063</td>
</tr>
<tr>
<td>SM ggH(WW)</td>
<td>POWHEG</td>
<td>0.34</td>
</tr>
<tr>
<td>SM qqH(WW)</td>
<td>POWHEG</td>
<td>0.028</td>
</tr>
<tr>
<td>MSSM ggΦ($\tau\tau$)</td>
<td>PYTHIA</td>
<td>-</td>
</tr>
<tr>
<td>MSSM b$b$Φ($\tau\tau$)</td>
<td>PYTHIA</td>
<td>-</td>
</tr>
<tr>
<td>$Z(LL)+$jets</td>
<td>MADGRAPH</td>
<td>3048</td>
</tr>
<tr>
<td>$W(L\nu)+$jets</td>
<td>MADGRAPH</td>
<td>31314</td>
</tr>
<tr>
<td>$t\bar{t}+jets$</td>
<td>MADGRAPH</td>
<td>164.4</td>
</tr>
<tr>
<td>$W(qq')Z(LL)+$jets</td>
<td>MADGRAPH</td>
<td>1.8</td>
</tr>
<tr>
<td>$W(L\nu)Z(LL)+$jets</td>
<td>MADGRAPH</td>
<td>0.9</td>
</tr>
<tr>
<td>$Z(LL)Z(LL)+$jets</td>
<td>MADGRAPH</td>
<td>0.06</td>
</tr>
<tr>
<td>$Z(LL)Z(qq)+$jets</td>
<td>MADGRAPH</td>
<td>0.8</td>
</tr>
<tr>
<td>$Z(LL)Z(vv)+$jets</td>
<td>MADGRAPH</td>
<td>0.3</td>
</tr>
<tr>
<td>Single-top (tW channel)</td>
<td>POWHEG</td>
<td>15.7</td>
</tr>
</tbody>
</table>

Table 5.1: Simulated samples used in the SM and MSSM analyses and the corresponding theoretical cross sections. SM Higgs boson samples were generated for $m_{H}$ hypotheses in the range 90–150 GeV in 5 GeV steps. The cross sections for these processes are given for $m_{H} = 125$ GeV. MSSM signal samples were generated for $m_{\Phi}$ in the range 90 GeV to 1 TeV.

### 5.2 Event selection

This section describes the baseline event selection in each channel, starting with the identification of electron, muon and $\tau_h$ candidates. This is followed by a description of lepton isolation and vetoes, which are important for reducing the contamination from backgrounds. The jet and missing transverse energy selection is given and the definition of topological variables which are used to reject events from specific background processes.
All events are required to contain at least one well-reconstructed vertex passing the standard quality criteria listed in chapter 3. The distributions of the number of reconstructed vertices passing these selections are given in figure 5.1, where the re-weighting of simulated events has also been applied. The higher rate of pileup interactions in the 2012 dataset has the potential to degrade the performance of the object reconstruction and selection. This motivates the use of more sophisticated pileup mitigation techniques, which are also described in this section.

Figure 5.1: The number of reconstructed vertices in $\mu \tau_h$ events in the (a) 7 TeV and (b) 8 TeV datasets. The predicted distribution and “bkg. uncertainty” band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given.

5.2.1 Lepton selection and vetoes

Electrons

The electron identification uses the four variables described in chapter 3, and additional variables [68] that, when combined, provide improved discrimination. Selections are applied on the output of a multivariate boosted decision tree (BDT) [138] discriminator which takes the following variables as input:

- Variables that use the tracking information exclusively: $f_{\text{brems}}$, the fraction of electron energy emitted as bremsstrahlung radiation due to interactions in the tracker volume; the
\(\chi^2\)-per-degree-of-freedom for both the CTF and GSF track fits; and the number of tracker layers in which hits are found.

- Variables that quantify the spatial matching between the supercluster and track: \(\Delta \eta_{\text{in}}\) and \(\Delta \phi_{\text{in}}\), the separations in \(\eta\) and \(\phi\) between the supercluster and track direction; and \(\Delta \eta_{\text{calo}}\), the separation in \(\eta\) between the supercluster seed crystal and the track evaluated at the calorimeter surface.

- Variables that quantify the cluster shape and energy distribution: \(\sigma_{\eta}\) and \(\sigma_{\phi}\), the energy-weighted \(\eta\) and \(\phi\) widths of the cluster; the supercluster \(\eta\) and \(\phi\) widths; the variable \(f_e = 1 - e_{1x5}/e_{5x5}\), where \(e_{1x5}\) (\(e_{5x5}\)) denotes the energy deposited in an array of \(1 \times 5\) (\(5 \times 5\)) cells in the vicinity of the supercluster seed; and \(r_9\), the fraction of the energy measured in the \(3 \times 3\) crystal array centred on the seed crystal compared to the total supercluster energy.

- The compatibility of the hadronic and EM energy and the track momentum: \(H/E\), the ratio of hadronic to EM energy; \(E/p\), the ratio of the supercluster energy to the track momentum as measured at the primary vertex; the energy-momentum compatibility \(1/E - 1/p\); \(E_{\text{cluster}}/P_{\text{out}}\), the ratio of the electron cluster energy to the momentum of the associated track evaluated at the outermost position; and the ratio of energy reconstructed in the pre-shower detector over the raw energy in the supercluster for electrons in the endcap region.

The BDT is trained to separate genuine electrons in \(Z \to e^+e^-\) events from fake electron candidates in \(Z+\text{jet}\) events in which the leptons from the \(Z\) decay are excluded. An electron is considered identified if the BDT score exceeds a \(p_T\) and \(\eta\) dependent threshold. Non-prompt electrons are produced by the conversion of photons which interact with the layers of the tracker. These are rejected by requiring there are no missing inner tracker hits and that the electron is not matched to a reconstructed conversion. The electron track must also be compatible with originating at the chosen primary vertex, by having small impact parameters in the transverse and beam directions, denoted \(d_{xy}\) and \(d_z\) respectively. Figure 5.2 gives the distributions of electron \(p_T\) and \(\eta\) in the \(e\tau_h\) channel selection.

**Muons**

Selected muons are required to be reconstructed by the PF algorithm; have valid tracker-only and global track fits; and pass the standard tight identification criteria. Requirements are also placed on the muon track impact parameters \(d_{xy}\) and \(d_z\) which, as for electrons, are calculated...
Searches for neutral Higgs bosons decaying to tau pairs

Events $2 \times 10^3$ at 8 TeV

$19.7 \text{ fb}^{-1}$ $h \tau_e$

Observed Bkg. uncertainty
$\tau^\tau \rightarrow Z \rightarrow ee$
Electroweak
$t\bar{t}$
QCD

[GeV]

Electron $p_T$

(a)

$\eta$

(b)

Figure 5.2: The (a) $p_T$ and (b) $\eta$ distributions of the electron candidate in $e\tau_h$ events in the 8 TeV dataset. The predicted distribution and “bkg. uncertainty” band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given.

with respect to the primary vertex. This is effective at reducing the background from cosmic muons and pileup. The distributions of muon $p_T$ and $\eta$ in the $e\mu$ final state are given in figure 5.3.

Hadronic taus

In addition to passing the decay-mode identification, $\tau_h$ candidates are required to pass specific criteria that reduce the misidentification of electrons or muons. Discrimination against muons is based on identifying hits in the muon systems co-incident with the tau direction. For $\tau_h$ that contain only one charged hadron, the fraction of energy deposited in the calorimeters must be at least 20% of the total. This is effective as muons are minimally-ionising and typically deposit small amounts of energy in the calorimeters. The rate of electrons passing the decay mode reconstruction is considerably higher, as these have similar properties to a one-charged-hadron tau decay. Additionally, electrons that emit photons via bremsstrahlung can mimic the one-charged-hadron-plus-strips modes. A multivariate BDT discriminator is trained to reject misidentified electrons, exploiting many of the same variables listed above for the electron selection. A tight working point is used for this discriminator in the $e\tau_h$ channel, which is found to have an efficiency of around 90% for real $\tau_h$ and 3-4% for misidentified electrons. The distributions of $p_T$ and $\eta$ for $\tau_h$ candidates in the $\mu \tau_h$ channel are given in figure 5.4.
Figure 5.3: The (a) $p_T$ and (b) $\eta$ distributions of the muon candidate in $e\mu$ events in the 8 TeV dataset. The predicted distribution and “bkg. uncertainty” band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given.

Figure 5.4: The (a) $p_T$ and (b) $\eta$ distributions of the $\tau_h$ candidate in $\mu\tau_h$ events in the 8 TeV dataset. The predicted distribution and “bkg. uncertainty” band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given.
The energy scale of simulated $\tau_h$ is also corrected to match that found in data. This is determined via fits to the $\tau_h$ invariant mass distribution. The tau energy scale is allowed to vary freely in these fits, which yield $p_T$ and decay-mode dependent corrections of up to 1.5%. The distribution in figure 5.5 gives the $\tau_h$ candidate mass for events passing the $\mu \tau_h$ selection. The expected contribution from the $Z \rightarrow \tau\tau$ process is split into three components which correspond to the different accepted decay modes. Candidates in the one-charged-hadron mode are predominantly assigned the pion mass by the PF algorithm, with a smaller number assigned the electron and muon masses. Distinct distributions due to the $\rho$ and $a_1$ intermediate resonances are also visible and show good agreement with observation.

**Lepton isolation**

Leptons are also required to be isolated to reduce an otherwise dominant background from misidentified jets in QCD production. The isolation is defined as the scalar sum of the transverse momenta of PF candidates within a cone in $\eta$-$\phi$ space centred on the lepton direction. Particles originating from both the primary and pileup vertices may be found in this cone, but the isolation quantity of interest is with respect to the hard-scattering process only. Therefore, charged PF candidates that have tracks with impact parameters greater than 0.1 cm with the primary vertex are excluded from the sum. However, it is not possible to distinguish photons and neutral hadrons that originate from pileup. Instead, an estimate of the neutral pileup contribution is made, based on the charged pileup contamination. The total isolation is then defined as:

$$I = \sum_{\text{charged}} p_T + \max \left( 0, \sum_{\text{neutral}} p_T + \sum_{\text{photon}} p_T - \frac{1}{2} \sum_{\text{charged}} p_T \right).$$ (5.1)

In this formula each sum considers the respective PF candidates within a cone defined by $\Delta R \leq 0.4$ for electrons and muons, and $\Delta R \leq 0.5$ for $\tau_h$. For $\tau_h$ candidates all the constituent particles are excluded from these sums. The charged pileup sum is multiplied by a factor $\frac{1}{2}$ to give the neutral pileup estimate. This factor is approximately the neutral-to-charged ratio of the hadronisation process as determined in simulation. Requirements on the relative isolation, defined as $R = I/p_T$ are used for electron and muon candidates.

The isolation requirement is particularly important for selecting genuine $\tau_h$ candidates from the large background of QCD jets. Figure 5.6 shows curves of the jet misidentification rate as a function of $\tau_h$ efficiency. These are determined by varying the maximum isolation threshold, $I^*$, in both simulated $Z \rightarrow \tau\tau$ and QCD-enriched data events. Separate curves are given for
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.5: The invariant mass of $\tau_h$ candidates in $\mu\tau_h$ events in the 8 TeV dataset. The predicted distribution and "bkg. uncertainty" band correspond to the result of the maximum likelihood fit described in section 5.8. The $Z \rightarrow \tau\tau$ component is subsequently split into the separate expected contributions from each decay mode. Those with one charged hadron only are typically assigned the pion mass by the PF algorithm and are therefore concentrated in a single bin. However, a small fraction of these are assigned electron and muon masses and are found in the first two bins. Candidates with one charged hadron and photons give a distribution which peaks near the mass of the intermediate $\rho(770)$ meson, and those with three charged hadrons a distribution which peaks near the mass of the $a_1(1260)$ meson.
candidates in the barrel and endcap regions, and the analysis selection of $I_{\tau_h} < 1.5$ GeV is highlighted. This working point gives an efficiency of around 65% and a background rejection of 95–97%.

![Graph](image)

**Figure 5.6:** The efficiency of the isolation requirement for genuine $\tau_h$ compared to jets from the QCD background.

**Lepton vetoes**

Events in each channel are vetoed in the presence of additional electrons or muons. This serves both to reduce specific backgrounds and to ensure there is no event overlap between channels, especially those targeting associated-production in which additional leptons are required. In the $e\tau_h$ and $\mu\tau_h$ channels events are vetoed if they contain an $e^+e^-$ or $\mu^+\mu^-$ pair respectively. This is primarily to reject events containing $Z \rightarrow \ell\ell$ decays in which the $\tau_h$ candidate is from an additional misidentified jet. The leptons considered for these vetoes must have $p_T > 15$ GeV and pass loose versions of the identification and isolation criteria. A further veto on any additional electron or muon candidate is applied in all three channels, again defined with relaxed selections.

The lepton selections in each channel are summarised in table 5.2. The efficiency for the identification, isolation and triggering of each object is known to be imperfectly modelled in the simulation. The tag-and-probe method [115] is used to measure these efficiencies, typically as a function of the object $p_T$ and $\eta$, in both data and simulation. The ratios of these give scale factors that are used to weight simulated events, thus correcting the efficiency. The tag-and-probe
Searches for neutral Higgs bosons decaying to tau pairs

method utilises unbiased samples of electron and muon candidates in $Z \rightarrow \ell \ell$ events. This is achieved by collecting events with a single lepton trigger, in which an additional lepton forms an $\ell^+ \ell^-$ pair with the triggering lepton which is compatible with the $Z$ boson mass. Corrections for the $\tau_h$ trigger efficiency are determined from $\mu \tau_h$ events recorded by a single muon trigger.

| & $\mu \tau_h$ & e$\tau_h$ & e$\mu$ |
|---|---|---|---|
| **HLT $p_T$ threshold** | $\mu(12–18) \&$ | e$(15–22) \&$ | $\mu(8) \& e(17)$ or |
| **ranges [GeV]** | $\tau_h(10–20)$ | $\tau_h(15–20)$ | $\mu(17) \& e(8)$ |
| **Offline selection (7 TeV)** | $p_T^\mu > 17$, $|\eta^\mu| < 2.1$ | $p_T^e > 20$, $|\eta^e| < 2.1$ | $p_T^\mu > 10(20)$, $|\eta^\mu| < 2.1$ |
| **Offline selection (8 TeV)** | $p_T^\mu > 20$, $|\eta^\mu| < 2.3$ | $p_T^e > 20$, $|\eta^e| < 2.3$ | $p_T^\mu > 20(10)$, $|\eta^\mu| < 2.3$ |
| **Isolation** | $R^\mu < 0.1$ | $R^e < 0.1$ | $R^\mu < 0.1(0.15)$ |
| | $I^\mu < 1.5$ GeV | $I^e < 1.5$ GeV | $R^e < 0.1(0.15)$ |
| **Track-vertex [cm]** | $d_{xy}^\mu < 0.045$ | $d_{xy}^e < 0.045$ | $d_{xy}^{\mu/e} < 0.02$ |
| | $d_z^\mu < 0.2$ | $d_z^e < 0.2$ | $d_z^{\mu/e} < 0.1$ |
| | $d_{xy}^{\tau_h} < 0.1$ | $d_{xy}^{\tau_h} < 0.1$ | - |
| **Lepton vetoes** | No loose $\mu^+\mu^-$ pair | No loose $e^+e^-$ pair | - |
| | with $p_T^e > 15$ GeV | with $p_T^\mu > 15$ GeV | - |
| | No additional loose e with $p_T > 10$ GeV and $|\eta| < 2.5$ | - | - |
| | No additional loose $\mu$ with $p_T > 10$ GeV and $|\eta| < 2.4$ | - | - |

Table 5.2: Summary of the lepton selections in each channel. The HLT $p_T$ requirements are given as ranges where thresholds were raised during running. In the $e\mu$ channel two triggers with complementary $p_T$ thresholds are used, with the corresponding offline $p_T$ selection at 10 GeV and 20 GeV. Electron and muon isolation requirements in the endcap region are also relaxed from 0.1 to 0.15 in the $e\mu$ channel. For the $\tau_h$ track impact parameter $d_z$ the charged constituent with the highest $p_T$ is used.

5.2.2 Jets and b-tagging

Jets are clustered from PF candidates as described in chapter 3. Those which overlap with either of the selected leptons are discarded. In the SM analysis the number of jets in an event is used for classification into categories, described in detail in section 5.4. For this, jets are required to
have corrected $p_T > 30$ GeV and $|\eta| < 4.7$, and must pass identification criteria to reject fakes that originate from detector noise [116]. Additionally, a multivariate BDT discriminator is used to reject jets originating from pileup interactions [139]. This takes as input the compatibility of any constituent tracks with the primary vertex as well as jet shape variables and the number of charged and neutral constituents. The jet shape variables are useful as the majority of pileup jets are found to be composed of multiple overlapping lower-$p_T$ jets. The track-based input variables are particularly powerful discriminators against pileup. One such variable is the jet $\beta$, defined as the scalar $p_T$ sum of tracks associated to both the jet and the primary vertex divided by the sum for all tracks associated to the jet. The predicted distributions of the jet $\beta$ are given in figure 5.7a for genuine and pileup jets in simulated $Z+jets$ events. While the $\beta$ fraction tends to be high for jets originating at the primary vertex, pileup jets have values concentrated around zero. This leads to improved performance within the tracker volume of $|\eta| < 2.5$. Figure 5.7b gives curves of the efficiency for genuine jet selection against pileup jet rejection that is achieved by placing a requirement on the BDT output discriminator. The curves are given for four regions of jet $\eta$, with the chosen working points in the analysis shown. These give an efficiency of around 99% for genuine jets within the tracker volume and an average of 95% outside this.

![Figure 5.7: (a) Distribution of the $\beta$ variable for real and pileup jets and (b) curves of the real jet efficiency compared to pileup jet rejection for the multivariate pileup discriminator described in the text. The shaded regions in (b) reflect the statistical uncertainty in the efficiency calculations. Both figures are derived from simulated $Z+jets$ events.](image-url)

The selection or veto of events containing b-tagged jets also forms part of the category definitions in the SM and MSSM analyses. This is particularly important in the latter, where a b-tag category
Searches for neutral Higgs bosons decaying to tau pairs
targets the production of a Higgs boson in association with b jets. In this analysis the CSV algorithm [82], which utilises track impact-parameter and secondary vertex information, is used exclusively. A jet is considered b-tagged if it has $p_T > 20 \text{ GeV}$, $|\eta| < 2.4$ and discriminator output greater than 0.679. This is a medium working point of the discriminator, chosen to give a misidentification rate for light-flavour jets of around 1% with an efficiency for b jets of around 75%. Figure 5.8 gives the distribution of the number of b-tagged jets and the $p_T$ distribution of the highest-$p_T$ tagged jet in $e\mu$ events.

![Graphs showing distributions of b-tagged jets and leading b-tagged jet $p_T$.](image)

**Figure 5.8**: Distributions of (a) the number of b-tagged jets and (b) the leading b-tagged jet $p_T$ in $e\mu$ events in the 8 TeV dataset. The predicted distribution and “bkg. uncertainty” band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given.

The efficiency for the tagging of b jets and the mistagging rate for light-flavour jets has been measured in both data and simulation. Differences are corrected for in the simulation through the application of efficiency and mistagging scale factors. The values of these factors and a description of the methods used to determine them can be found in [140] and [141] for 2011 and 2012 data respectively. The simulation is corrected by randomly reclassifying a fraction of tagged jets as non-tagged, or vice versa, as necessary to result in the correct average efficiency. The promotion or demotion probabilities for each jet are defined as

$$P(\text{demote}) = 1 - SF$$

$$P(\text{promote}) = \frac{(SF - 1)}{SF}$$

when $SF < 1$

when $SF > 1$. 
The scale factors, $SF$, are $p_T$, $\eta$ and jet-flavour dependent ratios of data and simulation efficiencies, and $\epsilon$ is the tagging efficiency in simulation.

### 5.2.3 Missing transverse energy

The standard CMS $E_T^{\text{miss}}$ is calculated using all the PF candidates in an event. This includes the candidates resulting from pileup interactions and has the effect of degrading the $E_T^{\text{miss}}$ resolution. Although not significantly affecting the response, each pileup interaction is equivalent to an additional 3–4 GeV smearing of each component of $E_T^{\text{miss}}$. This effect is mitigated by applying corrections to the hadronic recoil vector $u_T = -(\vec{q}_T + \vec{E}_T^{\text{miss}})$, where $\vec{q}_T$ is the boson momentum in the transverse plane. The correction is applied to the recoil instead of $\vec{E}_T^{\text{miss}}$ directly as it is less process dependent. It is applied through a series of multivariate BDT regressions [83]. The first corrects the azimuthal angle of $\vec{u}_T$ and the second corrects the magnitude. Both BDTs are trained with simulated $Z \rightarrow \mu^+\mu^-$ events where the true recoil vector is assumed to be the $Z$ boson $\vec{q}_T$. The BDT input variables are constructed from five $E_T^{\text{miss}}$-like objects, each sensitive to different components of the hadronic recoil, calculated from subsets of the PF candidates:

- all PF candidates;
- all charged PF candidates with tracks associated to the primary vertex;
- all charged PF candidates associated to the primary vertex and all neutral PF candidates in jets which have passed the pileup discrimination detailed above;
- all charged PF candidates not associated to the primary vertex and all neutral candidates in jets that fail the pileup discrimination;
- all charged PF candidates associated to the primary vertex and all neutral candidates, including those which are not clustered in jets, but subtracting the vectorial sum of all neutral candidates within jets that fail the pileup discrimination.

In each of these inputs the recoil vector $u_T$ is calculated and both the magnitude and azimuthal angle are used as inputs to the BDT. The scalar $p_T$ sum of the particles selected in each $E_T^{\text{miss}}$ object, the momentum vectors of the two highest $p_T$ jets and the number of reconstructed vertices are also used as inputs. Remaining differences in the $E_T^{\text{miss}}$ response and resolution between data and simulation are addressed by a correction to the recoil in simulation. This correction is derived from real and simulated $Z \rightarrow \mu^+\mu^-$ events and parameterized as a function of boson $\vec{q}_T$ and jet multiplicity.
5.2.4 Topological selection

After the selection of an opposite-charge lepton pair a large background from W+jets remains in the \( \ell \tau_h \) channels. A useful variable for distinguishing these events is the transverse mass \( m_T \) between the electron or muon and the \( E_T^{\text{miss}} \):

\[
m_T \equiv \sqrt{2 p_T^e E_T^{\text{miss}}(1 - \cos(\Delta\phi))}, \tag{5.2}
\]

where \( p_T^e \) is the lepton transverse momentum and \( \Delta\phi \) is the difference in azimuthal angle between the \( p_T^e \) and \( E_T^{\text{miss}} \). Figure 5.9 gives the distribution of \( m_T \) in the \( \mu \tau_h \) channel. In \( Z \to \tau\tau \) or \( H \to \tau\tau \) events the neutrinos from the \( \tau \) decay tend to be produced collinear to the visible products, thus leading to smaller values of \( m_T \). Conversely, in W+jets events the much larger mass of the W boson results in a neutrino travelling in the opposite direction to the lepton in the transverse plane, and therefore to larger values of \( m_T \). Events are required to have \( m_T < 30 \text{ GeV} \), which is found to give an optimal trade-off between signal efficiency and background rejection.

In the \( e\mu \) channel the dominant background is from \( t\bar{t} \) events in the fully-leptonic decay mode. Different background rejection strategies are chosen in the SM and MSSM analyses, although both rely on the topological variables \( p_T^{\text{vis}} \) and \( p_T^{\text{miss}} \) [142], illustrated in figure 5.10, defined as

\[
p_T^{\text{vis}} = p_T^e \cdot \hat{\zeta} + p_T^\mu \cdot \hat{\zeta}, \tag{5.3}
\]

\[
p_T^{\text{miss}} = E_T^{\text{miss}} \cdot \hat{\zeta}. \tag{5.4}
\]

The unit vector \( \hat{\zeta} \) is collinear with the line which bisects the electron and muon directions in the transverse plane. The distribution of the variable \( D_\zeta \equiv p_T^{\text{miss}} - 0.85 \cdot p_T^{\text{vis}} \), given in figure 5.11a, is found to give good separation between \( t\bar{t} \) and \( Z \to \tau\tau \) events. This variable exploits the fact that the tau decay neutrinos typically travel in the same direction as the visible part of the decay, meaning that the total neutrino system is more closely aligned with the \( \hat{\zeta} \) direction than in \( t\bar{t} \) events. The MSSM analysis selects events which have \( D_\zeta > -20 \text{ GeV} \). In the SM analysis \( p_T^{\text{vis}}, p_T^{\text{miss}} \) and several other discriminating variables are used as input to a multivariate BDT discriminator. The additional variables are: the azimuthal separation between electron and muon; the \( E_T^{\text{miss}} \), the transverse mass of the visible di-tau system and the \( E_T^{\text{miss}} \); the CSV output of the leading jet; and the impact parameter \( d_{xy} \). The BDT is trained to separate signal events from a combination of \( t\bar{t} \) and misidentified lepton backgrounds. The BDT output variable, given in figure 5.11b, provides improved signal and background discrimination over the \( D_\zeta \) variable alone. Events in the SM analysis are required to have a BDT score larger than \( -0.5 \).
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.9: Distribution of the transverse mass between the selected muon and \( p_T^{\text{miss}} \) in the \( \mu \tau_h \) channel. The dotted vertical line indicates the baseline selection applied in the analysis. The dashed vertical line indicates the control region used to estimate the W+ jets contribution as described in section 5.5.

Figure 5.10: A diagram illustrating the definition of the \( p_T^{\text{vis}} \) and \( p_T^{\text{miss}} \) variables.
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.11: Distributions of (a) the variable $D_\tau$ and (b) the $t\bar{t}$ rejection BDT output in $e\mu$ channel events. The former is used in the MSSM analysis and the latter in the SM analysis. Dotted vertical lines indicate the baseline selections applied in each analysis.

5.3 Di-tau invariant mass reconstruction

This section describes the di-tau mass estimation method used in the analysis and is a summary of the description given in [6]. The invariant mass of the visible di-tau system $m_{\tau\tau}^{vis}$ is a reasonable choice of discriminating variable from which a $H \rightarrow \tau\tau$ signal could be distinguished from the dominant $Z \rightarrow \tau\tau$ background. However, the neutrinos produced in tau decay can carry away a large fraction of the energy, reducing the separation power of the visible mass. An improved variable is found through an estimate of the tau pair invariant mass $m_{\tau\tau}$ which uses as input the four-momenta of the visible decay products and the missing transverse energy vector $E_T^{miss}$. One relatively simple method is the collinear approximation [143], in which $m_{\tau\tau}$ is calculated using the assumption that the neutrinos travel in the same direction as the visible part of the tau decay. However, this yields an unphysical solution in about 20% of events. The SVFIT algorithm, used in this analysis, is a likelihood-based method for estimating $m_{\tau\tau}$ using the inputs listed above. It provides an improved mass resolution compared to the collinear approximation and gives valid solutions in over 99.9% of events.

A hadronic tau decay can be specified by six parameters if all visible particles are treated as a single composite object. These can be chosen as the polar and azimuthal angles of the visible system in the tau rest frame; the three components of the boost vector that translate from the tau to the detector frame; and the invariant mass of the invisible system. Leptonic tau decays result
Searches for neutral Higgs bosons decaying to tau pairs

in two neutrinos; therefore, an additional variable is required and is chosen to be the invariant mass of the two-neutrino system. For a given di-tau final state, these six or seven parameters per tau decay are constrained by the measured four-momenta of the visible decays and the two components \( E_{\text{miss}}^x \) and \( E_{\text{miss}}^y \) of the \( \bar{E}_T \). The remaining degrees-of-freedom are chosen to be the parameters:

- \( x \), the fraction of the tau energy carried by the visible decay products in the laboratory frame;
- \( \phi \), the azimuthal angle of the tau direction in the laboratory frame;
- \( m_{\nu\nu} \), the invariant mass of the two-neutrino system, which is set to zero for hadronic decays where only one neutrino is present.

The most probable di-tau mass is then found by considering a likelihood function \( f(\tilde{z}, \tilde{\eta}, \tilde{a}_1, \tilde{a}_2) \), where the \( \tilde{a}_i = (x_i, \phi_i, m_{\nu\nu,i}) \) are the unknown decay parameters, \( \tilde{z} = (E_{\text{miss}}^x, E_{\text{miss}}^y) \) are the missing transverse energy components and \( \tilde{\eta} = (p_{1\text{vis}}^x, p_{2\text{vis}}^x) \) are the measured visible four-momenta. The probability of a given mass hypothesis \( P(m_{\tau\tau}^i) \) being the true \( m_{\tau\tau} \) is then defined as

\[
P(m_{\tau\tau}^i) = \int \delta(m_{\tau\tau}^i - m_{\tau\tau}(\tilde{\eta}, \tilde{a}_1, \tilde{a}_2)) f(\tilde{z}, \tilde{\eta}, \tilde{a}_1, \tilde{a}_2) \, d\tilde{a}_1 \, d\tilde{a}_2. \tag{5.5}
\]

The likelihood \( f(\tilde{z}, \tilde{\eta}, \tilde{a}_1, \tilde{a}_2) \) can be written as the product of three functions: one for each of the two tau decays \( \tilde{a}_1 \) and \( \tilde{a}_2 \) and one for the likelihood of the measured \( \bar{E}_T \), given the neutrino kinematics in the di-tau hypothesis and the expected experimental energy resolution.

The hadronic decay function is derived from the two-body phase space [21]:

\[
L_{\tau \tau}^{\text{had}} = \frac{\frac{d\Gamma}{dx \, d\phi}}{x} \propto \frac{1}{1 - m_{\text{vis}}^2/m_{\tau}^2}, \tag{5.6}
\]

and is defined within the physically allowed region \( m_{\text{vis}}^2/m_{\tau}^2 \leq x \leq 1 \). The leptonic decay function is derived from matrix elements in [144] with the likelihood given as

\[
L_{\tau \tau}^{\text{lep}} = \frac{\frac{d\Gamma}{dx \, dm_{\nu\nu} \, d\phi}}{x} \propto \frac{m_{\nu\nu}}{4m_{\tau}^2} \left[ (m_{\tau}^2 + 2m_{\nu\nu}^2)(m_{\tau}^2 - m_{\nu\nu}^2) \right], \tag{5.7}
\]

and defined within the physically allowed regions \( 0 \leq x \leq 1 \) and \( 0 \leq m_{\nu\nu} \leq m_{\tau} \sqrt{1-x} \). Neither \( L_{\tau \tau}^{\text{had}} \) or \( L_{\tau \tau}^{\text{lep}} \) depends directly on the \( x \) or \( \phi \) variables. However, the former contributes to the probability in equation 5.5 via the integration limits and the latter enters the third likelihood
function which concerns the measured $E_T^{\text{miss}}$. This is defined as

$$L_{\nu}(E_{x}^{\text{miss}}, E_{y}^{\text{miss}}) = \frac{1}{2\pi \sqrt{|V|}} \exp \left[ -\frac{1}{2} \left( \frac{(E_{x}^{\text{miss}} - \sum p_{x}^{\nu})^T}{E_{x}^{\text{miss}} - \sum p_{x}^{\nu}} \right) V^{-1} \right]$$

where the covariance matrix $V$ encapsulates the expected $E_T^{\text{miss}}$ resolution and the momentum sums are over all neutrinos in the final state hypothesis. The missing energy resolution is dependent on pileup but typically of the order 10-15 GeV for each of the two components. Figure 5.12 shows the $m_{\tau \tau}$ distributions obtained with the svfit algorithm for simulated $Z \to \tau \tau$ and $H \to \tau \tau$ events for $m_{H} = 125$ GeV. In comparison with the distributions obtained for $m_{\tau \tau}^{\text{vis}}$ this illustrates the improvement in separation between signal and background. This yields an increase of $\sim 40\%$ in the expected sensitivity of the analysis. The resolution of the svfit $m_{\tau \tau}$ estimate varies with channel and event category, but is typically 15% in the $e \tau_{h}$ and $\mu \tau_{h}$ channels and 20% in the $e \mu$ channel.

![Figure 5.12](image-url)

**Figure 5.12:** Comparison of the (a) visible and (b) svfit mass distributions in the $\mu \tau_{h}$ channel after the baseline selection defined in section 5.2. Each figure gives the expected distributions for $Z \to \tau \tau$ and $H \to \tau \tau$ decay where $m_{H} = 125$ GeV [6].
5.4 Event categorisation

To improve the sensitivity to signal, events passing the baseline selection are divided into mutually exclusive categories. These are also chosen to enhance the signal purity of particular Higgs boson production modes, so different schemes are used for the SM and MSSM analyses.

The SM scheme for each channel is outlined in figure 5.13. The primary division is into categories with zero, one or two jets. For this, only jets with $p_T > 30$ GeV, $|\eta| < 4.7$ and which are separated from both selected leptons by at least $\Delta R = 0.5$ are counted. Events which contain a b-tagged jet, as defined in section 5.2, with $p_T > 20$ GeV and $|\eta| < 2.4$, are rejected in order to reduce the background from $t\bar{t}$ production.

![Diagram summarising the categorisation of events in the SM analysis for each channel.](image)

**Figure 5.13:** Diagram summarising the categorisation of events in the SM analysis for each channel. Solid lines indicate selections common to multiple categories, with dashed lines delimiting categories which require zero, one or two jets to be present. The baseline selection, described in section 5.2, is common to all categories.

The two-jet (or VBF-tag) categories have additional requirements which exploit the characteristic jet topology in the VBF production mode. The two highest-$p_T$ jets in the event must have a large invariant mass, $m_{jj}$, and large pseudorapidity gap, $|\Delta \eta_{jj}|$. Events with additional jets in
this gap are vetoed, as central hadronic activity is suppressed by the colourless W or Z exchange in VBF production. These requirements lead to significantly reduced background expectations, especially from $Z \rightarrow \tau\tau$ in which VBF production is rare [145]. In the 8 TeV analysis two VBF-tag categories are defined, denoted “loose” and “tight”. The tight VBF-tag category requires $m_{jj} > 700 \, \text{GeV}$ and $|\Delta \eta_{jj}| > 4.0$. The $p_T$ of the di-tau system, defined as

$$p_T^{\tau\tau} = |p_T^L + p_T^{L'} + E_T^{\text{miss}}|,$$

is also required to be greater than 100 GeV. This further reduces backgrounds and selects events in which the Higgs boson is boosted in the transverse plane. Events which do not pass this tight selection but have $m_{jj} > 500 \, \text{GeV}$ and $|\Delta \eta_{jj}| > 3.5$ enter the loose VBF-tag category. In the 7 TeV analysis a single VBF-tag category is defined with the same selection as the loose VBF-tag category at 8 TeV. Figure 5.14 gives the distributions of the $m_{jj}$ and $|\Delta \eta_{jj}|$ variables from $\mu \tau_h$ events which contain at least two jets.

The 0-jet and 1-jet categories contain all the events which do not pass the VBF-tag criteria. The 1-jet categories give greater sensitivity to gluon-gluon fusion Higgs boson production, in which high $p_T$ jets are produced more frequently than in many of the backgrounds. The 0-jet categories
Searches for neutral Higgs bosons decaying to tau pairs

have large background expectations and low sensitivity to a SM Higgs boson signal. However, these are important for providing constraints on the predictions of the major backgrounds, which are propagated into the more signal-sensitive categories.

The 0-jet and 1-jet categories are further subdivided into low and high $p_T^{\ell}$ categories based on the $p_T$ of the $\tau_h$ candidate in the $\ell\tau_h$ channels or the muon in the $e\mu$ channel. The high $p_T^{\ell}$ categories improve the sensitivity by reducing the $Z \to \tau\tau$ background, as a Higgs boson with $m_H > m_Z$ decays to leptons with higher average $p_T$ than in the $Z \to \tau\tau$ decay. This also reduces backgrounds from $W+\text{jets}$ and QCD multijet production, where the jets misidentified as $\tau_h$ are typically softer.

In the $e\tau_h$ and $\mu\tau_h$ channels, an additional 1-jet high-$p_T^{\tau_h}$ boosted category is defined which requires the di-tau boost $p_T^{\tau\tau} > 100$ GeV. As with the tight VBF-tag category, this reduces the background contribution and selects events with improved $m_{\tau\tau}$ resolution, giving better separation between the $Z$ and Higgs boson mass distributions. The distribution of $p_T^{\tau\tau}$ under the baseline selection in the $\mu\tau_h$ channel is given in figure 5.15a and shows good compatibility between prediction and observation.

In the 1-jet categories of the $e\tau_h$ channel an additional requirement of $E_T^{\text{miss}} > 30$ GeV is applied. This improves the overall sensitivity of the channel by suppressing an otherwise large contribution from $Z \to e^+e^-$, which would lead to a peak in $m_{\tau\tau}$ within the signal search region. The distribution of the $E_T^{\text{miss}}$ in this channel is given in figure 5.15b. This $E_T^{\text{miss}}$ requirement also suppresses the fraction of signal and $Z \to \tau\tau$ events with lower $p_T^{\tau\tau}$. Consequently, the 1-jet high $p_T^{\tau_h}$ category, where $p_T^{\tau\tau} \leq 100$ GeV, offers little sensitivity and so is not included in the $e\tau_h$ channel. The $E_T^{\text{miss}}$ requirement is not applied in the 0-jet categories where the sensitivity to signal would be low regardless. By allowing a large $Z \to e^+e^-$ contribution here it is possible to constrain some of the systematic uncertainties related to this background that apply to all categories. This is discussed further in section 5.6.

In the MSSM analysis events passing the baseline selection may enter one of two categories, denoted “B-Tag” and “No B-Tag”. These are chosen to give sensitivity to $b$-associated and gluon-gluon fusion Higgs boson production respectively. The B-Tag category requires the presence of at least one CSV-medium $b$-tagged jet with $p_T > 20$ GeV and $|\eta| < 2.4$, and no more than one jet, of any flavour, with $p_T > 30$ GeV and $|\eta| < 4.7$. Events entering the No B-Tag category are required to have no $b$-tagged jets with $p_T > 20$ GeV. It is noted that further sub-division of categories, as in the SM analysis, using variables such as the number of jets and $p_T^{\tau\tau}$, would increase the sensitivity to an MSSM Higgs boson signal. However, the acceptance for signal events in such
Searches for neutral Higgs bosons decaying to tau pairs

![Figure 5.15: Distributions of (a) the di-tau transverse momentum $p_T^{\tau\tau}$ in $\mu\tau_h$ channel events and (b) the missing transverse energy $E_{miss}^T$ in $e\tau_h$ channel events. The predicted distribution and "bkg. uncertainty" band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given. The expected contribution from SM Higgs boson production is indicated by a blue dashed line.

Figure 5.15: Distributions of (a) the di-tau transverse momentum $p_T^{\tau\tau}$ in $\mu\tau_h$ channel events and (b) the missing transverse energy $E_{miss}^T$ in $e\tau_h$ channel events. The predicted distribution and "bkg. uncertainty" band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given. The expected contribution from SM Higgs boson production is indicated by a blue dashed line.

categories is dependent on the MSSM model of interest, so would limit the potential for results interpretation.

5.5 Background estimation

The expected contributions from major backgrounds are derived from data, wherever possible, to minimise systematic uncertainty. In particular, this reduces uncertainties due to mis-modelling in the simulation. The methods for estimating the shape and yield of each background in each channel and category are described in this section.

$Z \rightarrow \tau\tau$

The largest source of background events is the Drell-Yan production of $Z \rightarrow \tau\tau$. An “embedding” technique is used to create a sample of such events from $Z \rightarrow \mu\mu$ events in data, in which the muons are replaced by simulated taus. These events are required to contain two opposite-charge PF-identified muons that pass a loose isolation requirement. A $\tau^+\tau^-$ pair is then generated and assigned the four-momenta of the reconstructed muon pair. The tau decays are then simulated by
Searches for neutral Higgs bosons decaying to tau pairs

TAUOLA and the visible decay products processed through the full CMS detector simulation and PF reconstruction. The muons of the original event are removed and replaced with these visible tau decays, after which the $E_T^{\text{miss}}$, jets, $\tau_h$ candidates and lepton isolation sums are re-computed. This embedded sample is normalised to the expected rate under the baseline selection but without the topological $m_T$, $D_\tau$ or BDT selections applied. The rate is determined using the $Z$+jets simulation and the cross section given in table 5.1, with the latter scaled by the difference between data and simulation in a $Z \rightarrow \mu\mu$ control region. The predicted rate in a given category is then the product of this baseline normalisation and the efficiency for embedded events to pass the category selection. The embedded sample has the advantage that, apart from the tau decays, the entire event originates from data. Therefore, systematic uncertainties associated to the $E_T^{\text{miss}}$ and jet reconstruction are much smaller than in pure simulation. This is because the uncertainties which affect these, such as pileup, the underlying event, and detector noise and response, are not needed.

$Z \rightarrow \ell\ell$

A smaller, though still important, background in the $e\tau_h$ and $\mu\tau_h$ channels comes from $Z \rightarrow \ell\ell$ production in which one lepton is misidentified as a hadronic tau decay. This process is estimated with simulation, and the normalisation corrected using the same $Z \rightarrow \mu\mu$ control region as above. The background from $Z \rightarrow e^+e^-$ is particularly prominent in the $e\tau_h$ channel due to the 3–4\% probability for electrons to pass the tau anti-electron discrimination. The $Z \rightarrow \ell\ell$ process leads to another small background in which a jet in the event is misidentified as the $\tau_h$, and the other light lepton is not identified by the lepton vetoes.

$W$+jets

The background from $W$ boson production in which an additional jet is misidentified as the $\tau_h$ is large in both the $e\tau_h$ and $\mu\tau_h$ channels. The $m_T$, shape for this background is taken from simulation. The normalisation in each category is determined from data in a $W$+jets-rich control region with the transverse mass $m_T > 70$ GeV. This is illustrated in figure 5.9 for the baseline selection in the $\mu\tau_h$ channel. The expected number of events in each category is then given as

$$N_{W+\text{jets}}^{m_T<30} = \frac{N_{\text{sim}}^{m_T<30}}{N_{\text{sim}}^{m_T>70}} \left( N_{\text{data}}^{m_T>70} - \sum_i N_{i}^{m_T>70} \right).$$  

(5.10)
The simulated sample is normalised to the observed number of events in this region, $N_{mT>70}^{\text{data}}$, after first subtracting the sum of the small contributions from other backgrounds, $N_{i}^{mT>70}$. The yield in the low $m_T$ signal region, $N_{mT<30}^{W+jets}$, is then extrapolated using the fraction $N_{mT<30}^{\text{sim}}/N_{mT>70}^{\text{sim}}$ determined in simulation. As the number of simulated $W+jets$ events entering the VBF-tag categories is limited, the $m_{jj}$ and $|\Delta \eta_{jj}|$ requirements are relaxed in order to determine the low-$m_T$ to high-$m_T$ extrapolation and to construct smoother $m_{\tau\tau}$ shape templates.

**QCD multijet**

Another background in the $e\tau_h$ and $\mu\tau_h$ channels comes from QCD multijet production in which both reconstructed leptons are misidentified jets. The normalisation and shape of this background is extracted from data by exploiting events in which the $\ell \tau_h$ pair have the same charge and where the fraction of QCD multijet events is high. In the 0-jet and 1-jet SM categories, with the exception of 1-jet high-$p_T\tau_h$ boosted, and both MSSM categories, the normalisation is determined as

$$N_{\text{OS iso}} = N_{\text{SS iso}} \cdot \frac{N_{\text{OS anti-iso}}}{N_{\text{anti-iso}}} \cdot \frac{N_{\text{SS anti-iso}}}{N_{\text{anti-iso}}} \left( N_{\text{SS data}} - \sum_i N_{\text{SS i}} \right). \quad (5.11)$$

In this equation $N$ denotes an event yield; OS and SS are the opposite- and same-charge selections; “iso” implies the standard isolation requirement $R_{e/\mu} < 0.1$ for the electron or muon candidate; and “anti-iso” a selection in which the lepton is not well isolated, with $0.2 < R_{e/\mu} < 0.5$. The predicted category yield, $N_{\text{OS iso}}$, is determined as the number of observed events in the same-charge region, $N_{\text{SS iso}}$, subtracting the sum of the expected contributions from other backgrounds, $N_{\text{i}}$. This yield is then multiplied by the ratio $N_{\text{OS anti-iso}}/N_{\text{anti-iso}}$, which accounts for the difference in rate between opposite- and same-charge QCD multijet events. This ratio is measured in the independent anti-isolated control region and found to be $1.06 \pm 0.10$. This exploits the fact that the OS/SS ratio for QCD events does not depend strongly on the degree to which the lepton is isolated. There is an insufficient number of events in data to use the same method in the 1-jet high-$p_T\tau_h$ boosted and VBF tag categories. For these categories, the QCD yield under the baseline event selection is first determined and then multiplied by the efficiency of the category selection in the anti-isolated same-charge region. This selection gives a high purity in QCD multijet events.

In the SM 0-jet low-$p_T\tau$ and the MSSM No B-Tag categories the $m_{\tau\tau}$ shape is determined in a similar way to the normalisation, that is, from the same-charge distribution in data after subtracting the shapes of other backgrounds. For all other categories, where the number of
same-charge events is lower, the shape is taken directly from the same-charge and anti-isolated selection in data. The high QCD multijet purity makes it unnecessary to subtract the negligible contributions from other backgrounds. As an example, figure 5.16 compares the observed data and background expectation in the same-charge region of the 1-jet low-$p_T^\tau$ and loose VBF tag categories in the $\mu\tau_h$ channel. The shape of the QCD multijet expectation is taken from the anti-isolated selection and is found to give a good agreement with the data. In the 1-jet high-$p_T^\tau$ boosted and tight VBF-tag categories the isolation on the $\tau_h$ candidate is also relaxed to obtain a sufficiently populated template, though the contribution to the total background in these categories is very small.

![Figure 5.16](image.png)

**Figure 5.16**: Distributions of $m_{\tau\tau}$ in the same-charge region for the (a) 1-jet low-$p_T^\tau$ and (b) loose VBF tag categories of the $\mu\tau_h$ channel. The shape of the QCD prediction is taken from the same-charge anti-isolated region in data.

Backgrounds from $W$+jets and QCD multijet production are also expected in the $e\mu$ channel, where one or both leptons are misidentified jets, although to a lesser extent than in the $\ell\tau_h$ channels. Both contributions are evaluated together using a misidentified-$\ell$ control region in which one or both leptons are required to pass a loose selection but fail the nominal selection. Events in these samples are weighted by $p_T$ and $\eta$ dependent efficiencies for these loosely-selected leptons to pass the nominal selection. The total misidentified-$\ell$ background is determined as $N_e + N_\mu - N_{e\mu}$, where $N_e$ and $N_\mu$ are the efficiency-corrected yields from the loose electron and muon control regions, respectively. To avoid double-counting events where both leptons are misidentified, the yield $N_{e\mu}$ is subtracted. The misidentified-$\ell$ rate is determined under
the baseline selection, and a sample of same-charge $e\mu$ events is used to determine both the efficiency to pass each category selection and the shape of the $m_{\tau\tau}$ distributions.

$t\bar{t}+\text{jets, single-top and diboson}$

The background from $t\bar{t}$ production is present in all channels, but particularly the $e\mu$ channel. The shape is determined from simulation and the yield is corrected using an enriched control region requiring at least two b-tagged jets in the $e\mu$ channel. Figure 5.17 shows the distribution of the number of jets in the inclusive $e\mu$ channel selection. A good agreement is observed up to high jet multiplicities where the $t\bar{t}$ process dominates. Smaller background expectations from diboson and single-top production are determined from simulation and normalised using the cross sections given in table 5.1.

![Figure 5.17](image)

**Figure 5.17:** Distribution of the number of reconstructed jets, having $p_T > 30$ GeV and $|\eta| < 4.7$, in $e\mu$ channel events. The predicted distribution and "bkg. uncertainty" band correspond to the result of the maximum likelihood fit described in section 5.8. The ratio of the observed and predicted distributions is also given. The expected distributions from VBF and gluon-gluon fusion Higgs boson production are indicated by black and blue dashed lines respectively.
5.6 Systematic uncertainties

This section describes the sources of uncertainty that affect the signal and background predictions of the $m_{\tau\tau}$ distributions. The experimental uncertainties typically concern either object selection or the methods used to estimate the backgrounds described in the previous section. The former are more important for the signal prediction, whereas the latter have a larger effect on the background estimation. Theoretical uncertainties affect the predictions of both signal and background but are larger for the signal. Most uncertainties affect only the rate of signal and background in each category, but a smaller number may affect both the shape and rate. A summary of all sources of uncertainty and the processes they affect is given in table 5.3 for the SM analysis and table 5.4 for the MSSM analysis. Throughout this section the values of uncertainties given as percentages refer directly to variations in the predicted signal and background yields, unless otherwise specified.

$\tau_h$ trigger, identification and energy scale

In the $\ell\tau_h$ channels uncertainties related to the $\tau_h$ simulation are particularly important, as they affect both the signal and the $Z\rightarrow\tau\tau$ background. The uncertainty due to the tau identification and trigger selection is 8%, which originates from tag-and-probe studies performed on $Z\rightarrow\tau\tau$ events selected by a single-muon trigger [146]. An additional 3% is applied in the high $p_{T}^{\ell_h}$ categories due to limited sample sizes in the tag-and-probe measurement. Uncertainty in the $\tau_h$ energy scale affects both the rate and shape of the same backgrounds. The fits to the $\tau_h$ invariant mass, described in section 5.2, constrain this to within 3%. The difference in anti-lepton selection means these uncertainties are treated as independent in the two channels. The effect of the scale uncertainty on the shape of the $Z\rightarrow\tau\tau$ distribution is illustrated in figure 5.18. The magnitude of this variation on the binned yields is found to be comparable to the signal expectation on the upper tail of the distribution.

Electron, muon and jet selection and misidentification

Uncertainty in the misidentification rates of electrons and muons reconstructed as $\tau_h$ affect the $Z\rightarrow\ell\ell$ prediction. Tag-and-probe measurements of these rates are limited by small event yields when applying the anti-lepton discrimination, leading to uncertainties of 20% and 30% for electrons and muons respectively [6]. The $Z\rightarrow\ell\ell$ predictions in the SM boosted and VBF-tag
<table>
<thead>
<tr>
<th>Uncertainty</th>
<th>Affected processes</th>
<th>0-Jet</th>
<th>1-Jet</th>
<th>2-Jet (VBF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau energy scale</td>
<td>signal, $Z \to \tau\tau$, $t\bar{t}$, diboson</td>
<td>1–27%, shape</td>
<td>1–13%, shape</td>
<td>1–7%, shape</td>
</tr>
<tr>
<td>Tau ID (&amp; trigger)</td>
<td>signal, $Z \to \tau\tau$, $t\bar{t}$, diboson</td>
<td>8–11%</td>
<td>8–11%</td>
<td>8–11%</td>
</tr>
<tr>
<td>e misidentified as $\tau_h$</td>
<td>$Z \to ee$</td>
<td>20%</td>
<td>36–47%</td>
<td>71–75%</td>
</tr>
<tr>
<td>$\mu$ misidentified as $\tau_h$</td>
<td>$Z \to \mu\mu$</td>
<td>30%</td>
<td>30–40%</td>
<td>75–100%</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$</td>
<td>$Z \to \ell\ell$</td>
<td>20%</td>
<td>20%</td>
<td>20–80%</td>
</tr>
<tr>
<td>Electron ID &amp; trigger</td>
<td>signal, $Z \to LL$, $t\bar{t}$, diboson</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Muon ID &amp; trigger</td>
<td>signal, $Z \to LL$, $t\bar{t}$, diboson</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Electron energy scale</td>
<td>signal, $Z \to \tau\tau$</td>
<td>1%, shape</td>
<td>1%, shape</td>
<td>1%, shape</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>signal, $t\bar{t}$, $Z \to \ell\ell$, diboson</td>
<td>1–16%</td>
<td>1–6%</td>
<td>3–25%</td>
</tr>
<tr>
<td>$E_T^{miss}$ scale</td>
<td>signal, $t\bar{t}$, W+jets, diboson</td>
<td>1–7%</td>
<td>1–12%</td>
<td>1–10%</td>
</tr>
<tr>
<td>b-Tagging efficiency</td>
<td>$t\bar{t}$, diboson</td>
<td>2–10%</td>
<td>2–10%</td>
<td>3–15%</td>
</tr>
<tr>
<td>Norm. $Z$ production</td>
<td>$Z \to LL$</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>$Z \to \tau\tau$ category</td>
<td>$Z \to \tau\tau$</td>
<td>3–5%</td>
<td>3–5%</td>
<td>10–15%</td>
</tr>
<tr>
<td>$Z \to \ell\ell$ mass scale</td>
<td>$Z \to \ell\ell$</td>
<td>shape</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Norm. W+jets</td>
<td>W+jets</td>
<td>20%</td>
<td>10–17%</td>
<td>20–55%</td>
</tr>
<tr>
<td>Norm. $t\bar{t}$</td>
<td>$t\bar{t}$</td>
<td>8–10%</td>
<td>8–10%</td>
<td>9–34%</td>
</tr>
<tr>
<td>Norm. diboson</td>
<td>diboson</td>
<td>15%</td>
<td>15%</td>
<td>15–52%</td>
</tr>
<tr>
<td>Norm. QCD multijet</td>
<td>QCD multijet</td>
<td>6–20%</td>
<td>10–10%</td>
<td>22–100%</td>
</tr>
<tr>
<td>Shape QCD multijet</td>
<td>QCD multijet</td>
<td>shape</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Norm. misid. $e/\mu$</td>
<td>misidentified $e/\mu$</td>
<td>30%</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>Shape misid. $e/\mu$</td>
<td>misidentified $e/\mu$</td>
<td>shape</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Luminosity 7 TeV/8 TeV</td>
<td>signal &amp; diboson</td>
<td>2.2/2.6%</td>
<td>2.2/2.6%</td>
<td>2.2/2.6%</td>
</tr>
<tr>
<td>PDF ($qq$)</td>
<td>signal</td>
<td>1–10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>PDF (gg)</td>
<td>signal</td>
<td>1–4%</td>
<td>2–4%</td>
<td>2–4%</td>
</tr>
<tr>
<td>Scale variation</td>
<td>signal</td>
<td>1–8%</td>
<td>1–20%</td>
<td>2–31%</td>
</tr>
<tr>
<td>UE &amp; PS</td>
<td>signal</td>
<td>1–9%</td>
<td>1–5%</td>
<td>1–12%</td>
</tr>
<tr>
<td>Limited number of events</td>
<td>all</td>
<td>shape</td>
<td>shape</td>
<td>shape</td>
</tr>
</tbody>
</table>

**Table 5.3:** Summary of sources of systematic uncertainty in the SM analysis. The processes to which each source applies are given, in addition to the typical effect on the acceptance in each category. Adapted from [6].
<table>
<thead>
<tr>
<th>Uncertainty</th>
<th>Affected processes</th>
<th>No B-Tag</th>
<th>B-Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau energy scale</td>
<td>signal, $Z \rightarrow \tau\tau$</td>
<td>2–4%, shape</td>
<td>2–4%, shape</td>
</tr>
<tr>
<td>Tau ID (&amp; trigger)</td>
<td>signal, $Z \rightarrow \tau\tau, \tau\bar{\tau}$, diboson</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>$e$ misidentified as $\tau_h$</td>
<td>$Z \rightarrow ee$</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>$\mu$ misidentified as $\tau_h$</td>
<td>$Z \rightarrow \mu\mu$</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$</td>
<td>$Z \rightarrow \ell\ell, W+\text{jets}$</td>
<td>20%, shape</td>
<td>20%, shape</td>
</tr>
<tr>
<td>Electron ID &amp; trigger</td>
<td>signal, $Z \rightarrow LL, \tau\bar{\tau}$, diboson</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Muon ID &amp; trigger</td>
<td>signal, $Z \rightarrow LL, \tau\bar{\tau}$, diboson</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Electron energy scale</td>
<td>signal, $Z \rightarrow \tau\tau$</td>
<td>1%, shape</td>
<td>1%, shape</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>signal, $\tau\bar{\tau}, Z \rightarrow \ell\ell$, diboson</td>
<td>1%</td>
<td>1–11%</td>
</tr>
<tr>
<td>$E_T^{\text{miss}}$ scale</td>
<td>signal, $\tau\bar{\tau}, W+\text{jets}$, diboson</td>
<td>1–7%</td>
<td>1–7%</td>
</tr>
<tr>
<td>b-Tagging efficiency</td>
<td>signal, $\tau\bar{\tau}$, diboson, $Z \rightarrow \ell\ell$</td>
<td>1–5%</td>
<td>1–6%</td>
</tr>
<tr>
<td>Mistagging efficiency</td>
<td>signal, $\tau\bar{\tau}$, diboson, $Z \rightarrow \ell\ell$</td>
<td>1–3%</td>
<td>2–9%</td>
</tr>
<tr>
<td>Norm. $Z$ production</td>
<td>$Z \rightarrow LL$</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>$Z \rightarrow \tau\tau$ category</td>
<td>$Z \rightarrow \tau\tau$</td>
<td>-</td>
<td>1–5%</td>
</tr>
<tr>
<td>$Z \rightarrow \ell\ell$ mass scale</td>
<td>$Z \rightarrow \ell\ell$</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Norm. $W+\text{jets}$</td>
<td>$W+\text{jets}$</td>
<td>10%</td>
<td>30%</td>
</tr>
<tr>
<td>Norm. $\tau\bar{\tau}$</td>
<td>$\tau\bar{\tau}$</td>
<td>8–10%</td>
<td>8–10%</td>
</tr>
<tr>
<td>Embedding $\tau\bar{\tau}$ contamination</td>
<td>$\tau\bar{\tau}$</td>
<td>-</td>
<td>2–13%</td>
</tr>
<tr>
<td>Norm. diboson</td>
<td>diboson</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>Norm. QCD multijet</td>
<td>QCD multijet</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Shape QCD multijet</td>
<td>QCD multijet</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Norm. misid. $e/\mu$</td>
<td>misidentified $e/\mu$</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>Shape misid. $e/\mu$</td>
<td>misidentified $e/\mu$</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>Luminosity 7 TeV/8 TeV</td>
<td>signal &amp; diboson</td>
<td>2.2/2.6%</td>
<td>2.2/2.6%</td>
</tr>
<tr>
<td>PDF</td>
<td>signal</td>
<td>2–10%</td>
<td>2–10%</td>
</tr>
<tr>
<td>Scale variation (gluon fusion)</td>
<td>signal</td>
<td>5–25%</td>
<td>5–25%</td>
</tr>
<tr>
<td>Scale variation (b-associated)</td>
<td>signal</td>
<td>8–15%</td>
<td>8–15%</td>
</tr>
<tr>
<td>Limited number of events</td>
<td>all</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>High-$m_\tau$, template fit</td>
<td>QCD multijet, $\tau\bar{\tau}$, $W+\text{jets}$, diboson</td>
<td>shape</td>
<td>shape</td>
</tr>
</tbody>
</table>

Table 5.4: Summary of sources of systematic uncertainty in the MSSM analysis. The processes to which each source applies are given, in addition to the typical effect on the acceptance in both categories. Adapted from [6].
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.18: Example of the effect of the \( \tau_h \) energy scale uncertainty on the shape of the \( Z \rightarrow \tau \tau \) prediction in the 1-jet high \( p_T^{\tau_h} \) category of the \( \mu \tau_h \) channel. The nominal \( Z \rightarrow \tau \tau \) and the expected SM \( H \rightarrow \tau \tau \) shapes are shown as stacked distributions. The effect of the +3% and −3% variations in energy scale are indicated as dotted and dashed lines respectively.

categories also suffer from low event yields in simulation, giving statistical uncertainties up to 70%. Control of this background is particularly important in the 0-jet categories, where it is large compared to the SM signal yield expectation and peaks within the \( m_H \) search range. Measurements in control regions lead to a 2% uncertainty in the mass scale of \( Z \rightarrow \ell \ell \) events which is applied in all event categories. Figure 5.19 gives the expected \( Z \rightarrow \ell \ell \) and signal contributions in the 0-jet low-\( p_T^{\tau_h} \) categories of the \( e\tau_h \) and \( \mu \tau_h \) channels. It is seen that the effect of the mass scale uncertainty on the predicted \( Z \rightarrow \ell \ell \) yield in each \( m_{\tau\tau} \) bin is large compared to the expectation from signal, especially for the \( e\tau_h \) channel, and this is one of the reasons the sensitivity of the 0-jet categories is low.

The uncertainty in the misidentification rate of jets as \( \tau_h \) is typically 20% and primarily affects the prediction of \( Z \rightarrow \ell \ell \) events in which an additional jet is the \( \tau_h \) candidate. It does not affect the QCD and W+jets yield predictions, as these are extracted from data. However, the uncertainty does increase with \( p_T \), up to a maximum of 50% for misidentified \( \tau_h \) with \( p_T \geq 200 \text{ GeV} \). This is relevant for the W+jets shape prediction in the MSSM analysis, where events in the high \( m_{\tau\tau} \) tail typically have high jet \( p_T \). This is accounted for by a shape uncertainty which allows the rate in this region to vary independently of the low \( m_{\tau\tau} \) region.
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.19: Distributions of $m_{\tau\tau}$ for the (a) $Z \to ee$ and (b) $Z \to \mu\mu$ processes in the 0-jet low-$p_T^{\tau_h}$ categories of the $e\tau_h$ and $\mu\tau_h$ channels respectively. The effect of the $+2\%$ and $-2\%$ variations in mass are indicated as dotted and dashed lines respectively. The expectation from SM $H \to \tau\tau$ in these categories is also given and is scaled by a factor of 5 in (a), where the rate of $Z \to \ell\ell$ is considerably larger than in (b).

The identification, isolation and trigger efficiencies for simulated muons and electrons have uncertainties up to 2% and affect all channels. The effect of the energy scale uncertainty for electrons and muons is negligible in the $e\tau_h$ and $\mu\tau_h$ channels, where the $m_{\tau\tau}$ shape uncertainty is dominated by the tau energy scale. However, the electron energy scale is important in the $e\mu$ channel, where an uncertainty of 1% affects the shape for signal and the $Z \to \ell\ell$ background.

$E_T^{\text{miss}}$, jet energy scale and b-tagging

The uncertainty in the $E_T^{\text{miss}}$ scale can affect the predicted yields in several ways. In the $e\tau_h$ and $\mu\tau_h$ channels it affects the efficiency of the $m_{\tau\tau}$ selection, whereas in the $e\mu$ channel it affects several input variables to the $t\bar{t}$-rejection BDT. It affects the 1-jet categories in the $e\tau_h$ channel directly where the $E_T^{\text{miss}} > 30$ GeV selection is applied. The $E_T^{\text{miss}}$ scale uncertainty is constrained to within 5% by the recoil fits described in section 5.2 and translates to category yield uncertainties of between 1% and 12%. The calculation of the jet energy scale uncertainty follows a standard prescription [79] and leads to category acceptance uncertainties in the range 5–15%, with the largest found in the VBF-tag categories. The b-tagged jet vetoes in the SM analysis and the b-tag category selection in the MSSM analysis are another source of uncertainty. The effect on the category acceptance is found by varying the reclassification probabilities up
and down by their uncertainties. The largest effect is for \( \tau \bar{\tau} \) events which contain predominantly real b jets, giving uncertainties in the range 1-10% for this background.

**Background estimation methods**

Although the use of the embedded samples for the \( Z \to \tau \tau \) background removes many of these uncertainties, the process itself is subject to imperfections in the removal of the original \( Z \to \mu \mu \) event. A 5% uncertainty in the category selection efficiency is derived by comparing pure \( Z \to \tau \tau \) simulation with simulated \( Z \to \mu \mu \) events on which the embedding procedure is run. In the B-Tag category of the MSSM analysis the fraction of \( \tau \bar{\tau} \) events selected in data by the embedding procedure, and therefore contaminating the \( Z \to \tau \tau \) sample, is appreciable. This is accounted for by reducing the normalisation of the \( \tau \bar{\tau} \) prediction by the expected number of contaminating events, as determined by running the embedding procedure on the simulated \( \tau \bar{\tau} \) sample. This reduction in rate is channel dependent and is between 2–13%. An uncertainty equal to the magnitude of this correction is taken as an uncertainty in the rate of \( \tau \bar{\tau} \) events in this category.

For the \( W + \text{jets} \) background, the extrapolation from high to low \( m_\tau \) has an uncertainty of 10–25% depending on category. This is determined by comparing the \( m_\tau \) shape of data and simulation in a sample of \( Z \to \mu \mu \) events in which one muon is treated as invisible in the event reconstruction, thus mimicking a \( W \) boson decay. In the boosted and VBF-tag categories where the number of simulated \( W + \text{jets} \) events is low, the tau isolation or category selection is relaxed in order to generate better-populated templates. This was found to have a small bias on the \( m_\tau \) shape in a control region with a looser \( m_\tau \) selection [6].

For the QCD background in the \( e \tau_h \) and \( \mu \tau_h \) channels, the same-charge to opposite-charge extrapolation is assigned a 10% uncertainty, based on comparisons in an inverted lepton isolation region. This covers both small differences due to the dependence of the extrapolation factor on \( p_T^{\tau_h} \) and a larger statistical uncertainty. A difference in the \( m_\tau \) shape below 60 GeV is also observed in this region and used to derive a correction to the QCD template in the nominal selection. The QCD yield uncertainties in the 1-jet high-\( p_T^{\tau_h} \) boosted and VBF-tag categories are increased by up to 70% to account for the statistical uncertainty in the control-region extrapolation method.

**Luminosity and normalisation**

The uncertainty in the integrated luminosity recorded is 2.2% for the 7 TeV analysis and 2.6% for the 8 TeV analysis, as described in [56, 57], and applies to predictions of signal rates and
those backgrounds which do not use a data-driven method. An uncertainty of 3% is assigned to all $Z \rightarrow LL$ processes which results from the fit in the $Z \rightarrow \mu \mu$ control region that is used to determine the inclusive Drell-Yan normalisation. The diboson and single-top uncertainty is 15% based on recent measurements by CMS [147,148].

**Limited event yields**

The uncertainty in the background shape prediction due to limited event samples is also taken into account. In a given histogram bin the total statistical uncertainty is treated as a single uncorrelated source of uncertainty, in a similar method to that proposed in [149]. In the MSSM analysis the same issue can result in high $m_{\tau\tau}$ bins populated with low or zero background events, which would lead to instabilities when extracting a signal contribution in this region. To overcome this limitation, several of the background templates are replaced by binning analytic functions that are found to describe the mass tail distributions. The functional form used is:

$$f(m_{\tau\tau}) = \exp \left(-\frac{m_{\tau\tau}}{a + b \cdot m_{\tau\tau}}\right),$$

(5.12)

where $a$ and $b$ are free parameters, determined in the fit to each mass template, in a region typically above 200–500 GeV. This function was found to give minimal bias on the extracted signal strength compared to several other common parameterizations. The fit uncertainties in $a$ and $b$ are treated as uncertainties in each background shape. An example is given in figure 5.20 for the $W+$jets template in the B-Tag category of the $\mu\tau_h$ channel. It is shown that with the standard template taken from the simulation there are no events populating the final two mass bins. This is remedied by the fit, which is found to be in good agreement with the standard template within the fit uncertainties.

**Theory uncertainties**

In the SM analysis theoretical uncertainties in each of the gluon-gluon fusion, VBF and $W/Z$-associated production modes are evaluated. The effect of each source is then translated to the uncertainty in the acceptance in each category. The parton distribution function uncertainty is evaluated by calculating the acceptance with several different PDF sets and taking the maximal variation. This leads to typical values of 1% for VBF and 2% for gluon-gluon fusion production. These are in addition to the 2.8% and 7.5% uncertainties in the inclusive cross sections for VBF and gluon-gluon fusion production determined in [31,106,135]. The renormalisation and factorisation scale uncertainty varies depending on category, but is of the order 1–5%
Searches for neutral Higgs bosons decaying to tau pairs

\[ \frac{dN}{dm_{\tau\tau}} \left[ \text{[GeV]} \right] \]

W+jets Simulation

- Nominal Template
- Fit Result
- \( \pm 1\sigma \) Parameter 1
- \( \pm 1\sigma \) Parameter 2

\( m_{\tau\tau} \) [GeV]

Figure 5.20: The result of a fit to the upper tail of the \( m_{\tau\tau} \) distribution for the W+jets prediction in the B-Tag category of the \( \mu T_h \) channel. The template taken directly from simulation is indicated by markers with the corresponding statistical uncertainty shown by error bars. The result of the fit is indicated by the solid line and the ranges of the uncertainties by red and blue dotted lines.

for VBF and 5–20% for gluon-gluon fusion. The latter is largest in the boosted and VBF-tag categories. The gluon-gluon fusion events passing the VBF-tag are not expected to be as well modelled by the POWHEG simulation, as it includes at most one jet in the matrix element calculation. The acceptances from several other generators: MADGRAPH, POWHEG+MINLO [150] and aMC@NLO [94] are compared in order to derive an additional uncertainty of about 30%.

Finally, uncertainty in the underlying event simulation and parton showering amounts to 2–10% depending on the category jet requirements. In the MSSM analysis the signal uncertainties vary with the \( m_A \) and \( \tan \beta \) point under consideration. PDF uncertainties range from 2–10% and scale uncertainties range from 5–25% for gluon-gluon fusion and 8–15% for b-associated production [8].
5.7 Statistical interpretation

This section outlines the statistical procedure used to quantify or reject the presence of a signal in data. These methods were developed by the LHC Higgs Combination Group to provide a common strategy for both the CMS and ATLAS Collaborations and to facilitate the combination of individual search results [151]. The methods are first described in general terms and the results of their application to the SM and MSSM Higgs boson searches are given in subsequent sections.

The expected Higgs boson event yields in a given model can be denoted as $s$ and the expectations from background as $b$. This can refer equally to single event counts, or to predicted binned distributions for use in a shape-based analysis. An additional factor $\mu$ is introduced as a signal strength modifier, which allows for models with a uniform scaling of the signal rate, $\mu \cdot s$. The background-only hypothesis is then defined by $\mu = 0$, and any signal hypothesis by $\mu > 0$. The term “data” will refer to a corresponding observed event count or counts, which could originate from an actual experiment or from simulation. The yields $s$ and $b$ are, in general, functions of some parameters $\theta$ representing experimental and theoretical uncertainties: $s(\theta)$ and $b(\theta)$. The nominal values $\hat{\theta}$ of these nuisance parameters are usually determined by external measurements, with uncertainties described by probability density functions (pdfs) $p(\hat{\theta} | \theta)$.

From these components the likelihood for an observed dataset, $\mathcal{L}(\text{data} | \mu, \theta)$, is defined as

$$\mathcal{L}(\text{data} | \mu, \theta) = \text{Poisson}(\text{data} | \mu \cdot s(\theta) + b(\theta)) \cdot p(\hat{\theta} | \theta)$$  \hspace{1cm} (5.13)

where for a binned likelihood model the Poisson term is simply the product of Poisson probabilities over each bin $i$:

$$\text{Poisson}(\text{data} | \mu \cdot s(\theta) + b(\theta)) = \prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-\mu s_i - b_i}.$$  \hspace{1cm} (5.14)

A ratio of likelihoods can be used to define a test statistic, a single number which can distinguish between two hypotheses. Such a test statistic can be used to set upper limits on the rate of signal production. Historically, a number of definitions have been used in Higgs boson searches. The one chosen by the LHC experiments is known as the profile likelihood ratio

$$q_\mu = -2 \ln \frac{\mathcal{L}(\text{data} | \mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} | \hat{\mu}, \hat{\theta})}, \text{ with the constraint } 0 \leq \hat{\mu} \leq \mu,$$  \hspace{1cm} (5.15)

where $\mu$ is the signal hypothesis being tested; $\hat{\theta}_\mu$ are the values of the nuisance parameters that maximise the likelihood, given the fixed signal strength $\mu$; and $\hat{\mu}$ and $\hat{\theta}$ are the values which give the global maximum of the likelihood. The constraint $0 \leq \hat{\mu}$ is added to prevent an unphysical
negative signal strength. The constraint $\hat{\mu} \leq \mu$ is chosen to prevent the exclusion of any $\mu$ lower than the best fit $\hat{\mu}$, thus ensuring the construction of a one-sided confidence interval. Large values of $q_\mu$ indicate a value of $\mu$ the data disfavors, whereas values close to zero indicate good compatibility with the signal hypothesis in question. The probability of finding a value of $q_\mu$ at least as large as the observed value, $q_\mu^{\text{obs}}$, is defined as:

$$\text{CL}_{a+b} = \int_{q_\mu^{\text{obs}}}^{\infty} f(q_\mu \mid \mu, \hat{\theta}_\mu) dq_\mu,$$

where $f(q_\mu \mid \mu, \hat{\theta}_\mu)$ is the probability distribution function for $q_\mu$. The tested value of $\mu$ is then said to be excluded at a confidence level $\alpha$, where $\alpha = 1 - \text{CL}_{a+b}$. The 95% CL is typically chosen when setting upper limits. One issue with this definition is that in some cases it will lead to the exclusion of low signal strengths, where an analysis may not expect to have sensitivity. For example, this may happen with a downward fluctuation of the data where the signal expectation is very small compared to the background expectation. To protect against this an additional probability $\text{CL}_b$ can be introduced, defined similarly to equation 5.16, but under the assumption of the background-only hypothesis, $f(q_\mu \mid 0, \hat{\theta}_0)$. Instead, the ratio of these probabilities, denoted $\text{CL}_s$, where

$$\text{CL}_s = \frac{\text{CL}_{a+b}}{\text{CL}_b},$$

is used to set the 95% CL exclusion limit, and this is commonly referred to as the modified frequentist approach [152].

The distributions $f(q_\mu \mid \mu, \hat{\theta}_\mu)$ and $f(q_\mu \mid 0, \hat{\theta}_0)$ can be determined by generating toy MC datasets from their respective models, in which the nuisance parameters are fixed to the values found in the fits to the observed data. The value of $q_\mu$ is then determined for each toy dataset. The effect of systematic uncertainties is incorporated by sampling a set of pseudo-measurements $\tilde{\theta}$ in each toy using the chosen nuisance pdfs. It is often instructive to compare the observed exclusion limit to the expectation under the assumption of the background-only hypothesis. This can be determined by generating background-only toy datasets and determining the 95% CL limit in each. These values form a cumulative pdf from which the median exclusion and uncertainty bands can be extracted.

A profile likelihood ratio can also be used to calculate the p-value for an observed excess of events given the background-only hypothesis. For this a slightly modified definition of the test
Searches for neutral Higgs bosons decaying to tau pairs

Where the constraint $\hat{\mu} \geq 0$ is chosen to prevent a downward fluctuation being considered evidence against the background-only hypothesis. The $p$-value for the observed data is then given as

$$p_0 = \int_{q_{0}^{\text{obs}}}^{\infty} f(q_0 \mid 0, \hat{\theta}_0) dq_0$$

where the distribution $f(q_0 \mid 0, \hat{\theta}_0)$ can be determined by generating pseudo-data from the background-only hypothesis. The $p$-value is typically converted to a significance, $Z$, by determining the number of standard deviations of a one-sided normal distribution that would yield an equal tail probability.

A major advantage of the profile likelihood test statistic is that in the limit of a large data sample, the distribution $f(q_\mu)$ follows a known formula [153]. This so-called asymptotic limit approximation removes the need for the computationally intensive step of generating and fitting toy datasets, which can take an appreciable time for models with many bins and nuisance parameters. This method relies on the properties of the Asimov dataset, a single representative dataset in which the observed rates match exactly with the prediction of the model under the nominal set of nuisance parameters. Furthermore, it is possible to derive a formula for the median expected limit and uncertainty bands using only the properties of the Asimov dataset, thus completely removing the need for any toy MC [153].

5.8 SM search results

The signal extraction requires a simultaneous maximum-likelihood fit of the $m_{\tau\tau}$ distribution in every channel and category. In this fit each source of uncertainty detailed in section 5.6 is treated as one of the nuisance parameters $\theta$. Parameters that affect only the normalisation of templates are assigned log-normal likelihood pdfs, chosen to prevent negative yields being found in the fit. Uncertainties that affect the $m_{\tau\tau}$ shape are assigned a Gaussian pdf, with variations of the nuisance parameter resulting in a smooth morphing of the template. The signal strength parameter $\mu$ is chosen to uniformly scale the cross section of each signal production mode, with $\mu = 1$ being the expected SM rate.
It is noted that the $H \to WW$ decay may result in the same final states as the $H \to \tau\tau$ decay. From consideration of the relevant branching fractions this contribution is only significant, relative to the expected $H \to \tau\tau$ yield, for the final states that do not contain a $\tau_h$. Therefore, of the three channels considered in this thesis, only the $e\mu$ channel is affected. In all subsequent results, unless otherwise stated, the expected SM contribution from the $H(125\,\text{GeV}) \to WW$ process is treated as a background, regardless of the $m_H$ value being tested.

Many nuisance parameters affect processes in multiple categories. For example, the tau identification efficiency and energy scale uncertainties are correlated between all categories in a given channel. This means the high yield 0-jet categories are able to constrain the values of these parameters, reducing the uncertainty in situ on the $Z \to \tau\tau$ background in the more signal-sensitive categories. In the tight VBF-tag categories this is the dominant background and leads to an improvement in sensitivity compared to fitting this category in isolation. Conversely, the $W+\text{jets}$ background uncertainty is dominated by statistical uncertainty, which is independent in each category. Therefore, it is only constrained within a given category, mostly by events in the high $m_{\tau\tau}$ region, where $W+\text{jets}$ events dominate.

Figures 5.21 to 5.26 give the complete set of $m_{\tau\tau}$ distributions in the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels that enter the maximum likelihood fit. Figures 5.21 and 5.22 give the distributions for the 7 TeV and 8 TeV $\mu\tau_h$ analyses respectively, and similarly figures 5.23 and 5.24 for the $e\tau_h$ channel and figures 5.25 and 5.26 for the $e\mu$ channel. The shape and normalisations of each background process reflect the best-fit nuisance parameter values after the maximum likelihood fit of $\mathcal{L}(\text{data} \mid \mu, \hat{\theta}_\mu)$. The signal component shown is that expected in the SM for a 125 GeV Higgs boson. The uncertainty band reflects the total background uncertainty taking into account the nuisance parameter constraints and correlations found in the fit. Table 5.5 summarises the expected and observed event yields in each category and includes the expected contribution from each Higgs boson production mode.

To better visualise the agreement between the observed data and the background-only hypothesis, figure 5.27 shows the combination of all $m_{\tau\tau}$ distributions. Each distribution is weighted by the ratio $S/(S + B)$, where $S$ and $B$ are the expected signal and fitted background yields respectively, counted in a mass range containing the central 68% of signal events. The signal yield expectation is for a SM Higgs boson with mass 125 GeV. The inset figure shows the observed data with background subtracted, showing an excess of events broadly compatible with the SM Higgs boson expectation. The $S/(S + B)$ ratio and signal mass window found in each category is also given in table 5.5.
Figure 5.21: Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $\mu\tau$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
Figure 5.22: Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $\mu\tau$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.23: Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $e^+\tau^-$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
Figure 5.24: Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $e\tau_h$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
Searches for neutral Higgs bosons decaying to tau pairs

Figure 25: Observed and predicted $m_{\tau\tau}$ distributions for the 7 TeV analysis in the $e\mu$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
Figure 5.26: Observed and predicted $m_{\tau\tau}$ distributions for the 8 TeV analysis in the $\mu\tau$ channel. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for the SM prediction.
<table>
<thead>
<tr>
<th>Event category</th>
<th>Signal composition (%)</th>
<th>ggH</th>
<th>VBF</th>
<th>VH</th>
<th>Σ signal</th>
<th>Background</th>
<th>Data</th>
<th>$\frac{S}{\sqrt{S+B}}$</th>
<th>$\frac{S}{\sqrt{S+B}}$ (GeV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu T_{h}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.3 ± 3.2</td>
<td>11920 ± 110</td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85.7 ± 10.3</td>
<td>40240 ± 210</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.4 ± 2.5</td>
<td>1578 ± 38</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>66.0 ± 8.0</td>
<td>5748 ± 78</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.2 ± 1.5</td>
<td>2016 ± 40</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44.8 ± 5.6</td>
<td>8951 ± 94</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.3 ± 1.2</td>
<td>791 ± 23</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.4 ± 4.4</td>
<td>3122 ± 54</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ boosted 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.8 ± 0.6</td>
<td>281 ± 11</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ boosted 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.8 ± 2.5</td>
<td>1257 ± 31</td>
</tr>
<tr>
<td>VBF tag 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.5 ± 0.1</td>
<td>22 ± 2</td>
</tr>
<tr>
<td>Loose VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4 ± 0.4</td>
<td>81 ± 6</td>
</tr>
<tr>
<td>Tight VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3 ± 0.2</td>
<td>15 ± 2</td>
</tr>
<tr>
<td>$\sigma_{\mu}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.5 ± 1.7</td>
<td>6220 ± 80</td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.7 ± 4.3</td>
<td>17060 ± 140</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.2 ± 1.7</td>
<td>1200 ± 32</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33.0 ± 4.1</td>
<td>4479 ± 68</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0 ± 0.6</td>
<td>372 ± 15</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.3 ± 1.6</td>
<td>1218 ± 29</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ boosted 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8 ± 0.3</td>
<td>151 ± 7</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ boosted 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.8 ± 1.1</td>
<td>506 ± 17</td>
</tr>
<tr>
<td>VBF tag 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8 ± 0.1</td>
<td>14 ± 1</td>
</tr>
<tr>
<td>Loose VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3 ± 0.2</td>
<td>45 ± 3</td>
</tr>
<tr>
<td>Tight VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.6 ± 0.1</td>
<td>10 ± 2</td>
</tr>
<tr>
<td>$\sigma_{\eta}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21.9 ± 2.9</td>
<td>11287 ± 98</td>
</tr>
<tr>
<td>0-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>74.4 ± 9.1</td>
<td>40200 ± 190</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0 ± 1.1</td>
<td>1645 ± 31</td>
</tr>
<tr>
<td>0-jet high-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.4 ± 3.1</td>
<td>6008 ± 71</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.4 ± 1.3</td>
<td>2461 ± 40</td>
</tr>
<tr>
<td>1-jet low-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>49.7 ± 5.6</td>
<td>10910 ± 100</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.1 ± 0.7</td>
<td>923 ± 23</td>
</tr>
<tr>
<td>1-jet high-$p_T^h$ 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.5 ± 2.9</td>
<td>4050 ± 56</td>
</tr>
<tr>
<td>VBF tag 7 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.1 ± 0.1</td>
<td>19 ± 1</td>
</tr>
<tr>
<td>Loose VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.2 ± 0.3</td>
<td>98 ± 6</td>
</tr>
<tr>
<td>Tight VBF tag 8 TeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.7 ± 0.1</td>
<td>14 ± 1</td>
</tr>
</tbody>
</table>

Table 5.5: The observed and expected yields in the event categories of the SM analysis. The background expectation and uncertainty corresponds to the result of the maximum likelihood fit. The signal expectation is given for a SM Higgs boson of mass 125 GeV. The expected composition of gluon-gluon fusion (ggH), vector-boson fusion (VBF) and W/Z-associated (VH) production is also given. The ratios $S/(S+B)$ and $S/\sqrt{S+B}$ are calculated in an $m_{\tau\tau}$ window enclosing the central 68% of signal events, corresponding to a width $\sigma_{\eta}$. Adapted from [6].
Figure 5.27: The combination of all $m_{\tau\tau}$ distributions from each category and channel. The background distribution and uncertainty corresponds to the result of the maximum likelihood fit, and the signal expectation is for a SM 125 GeV Higgs boson. In the combination of $m_{\tau\tau}$ distributions, each is weighted by the ratio $S/(S+B)$ as defined in the text. The inset figure shows the observed distribution with the background subtracted.

Figure 5.28 shows expected 95% CL upper limits as a function of $m_{H}$ for the background-only hypothesis. These are calculated using the asymptotic approximation described previously. Figure 5.28a gives the expected limits for each channel, and figure 5.28b gives the expected limits for the 0-jet, 1-jet and VBF-tag combinations of categories. The VBF-tag and 1-jet categories are found to have comparable sensitivity for $m_{H} = 125$ GeV, and the 0-jet categories have considerably weaker sensitivity, as expected, due to the large backgrounds. The median expected limit for the combination of channels at $m_{H} = 125$ GeV is found for $\mu = 0.65$.

Figure 5.29 shows both the expected and observed 95% CL upper limits on $\mu$. Figure 5.29a shows the median expectation for the background-only hypothesis along with the 1\sigma and 2\sigma intervals. Due to an excess of events, the observed limit is less stringent than for the background-only expectation across a broad $m_{H}$ range. At 125 GeV the observed limit is at $\mu = 1.47$, compared
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.28: Expected 95% CL upper limits on the signal strength parameter $\mu$. Figure (a) gives the expected limit for the $e\tau_h$, $\mu\tau_h$ and $e\mu$ channels independently, and figure (b) for the 0-jet, 1-jet and VBF-tag combinations of categories.

with the expectation of 0.65. Figure 5.29b shows the same observed limit, but with the median expectation and uncertainty for the signal-plus-background hypothesis for $m_{H}\equiv 125$ GeV. The observed limit is seen to be compatible with this expectation.

This excess of events can be quantified by calculating the p-value $p_0$, the probability of the observed dataset given the background-only expectation. This is given in figure 5.30 as a function of $m_{H}$. The median p-value expectation for $\mu = 1$ is also given for each $m_{H}$ hypothesis. At $m_{H} = 125$ GeV the observed (expected) p-value is equivalent to a significance of 3.0 (3.1) standard deviations. A maximum significance of 3.2 standard deviations is found for $m_{H} = 115$ GeV.

Scans of the negative log-likelihood, $-2\Delta \ln \mathcal{L}$, as a function of $\mu$ are given in figure 5.31 for the $m_{H} = 125$ GeV hypothesis. Three scans are performed, in which different sets of nuisance parameters are profiled. In the first scan all parameters are profiled at each value of $\mu$. The best fit value and total uncertainty is found to be $\hat{\mu} = 0.88 \pm 0.33$. The uncertainty is determined from the points at which $-2\Delta \ln \mathcal{L} = 1$. The other two scans are used to calculate the contribution to the total uncertainty from statistical, experimental and theory systematic sources. In one of these scans the theory uncertainty parameters are fixed to their best-fit values while the other parameters are profiled. In the other scan all nuisance parameters are fixed to their best-fit values. By considering the differences in the uncertainty from each scan the best fit uncertainty may be decomposed as $\hat{\mu} = 0.88 \pm 0.25({\text{stat}}) \pm 0.20({\text{syst}}) \pm 0.07({\text{theory}})$. The statistical uncertainty is
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.29: Expected and observed 95% CL upper limits on the signal strength parameter $\mu$ as a function of $m_{H}$. The median expectation and $1\sigma$ and $2\sigma$ probability intervals are given for (a) the background-only hypothesis and (b) the signal-plus-background hypothesis under the assumption of a 125 GeV SM Higgs boson.

Figure 5.30: Expected (dashed line) and observed (solid line) local p-values under the background-only hypothesis, as a function of $m_{H}$. The expected values are determined for a signal strength $\mu = 1$ at each $m_{H}$ hypothesis. Horizontal lines are drawn to indicate corresponding significances as numbers of standard deviations.
found to dominate the measurement and the contribution from theory uncertainties is small compared to experimental effects.

Figure 5.31: Scans of the negative log-likelihood, $-2\Delta \ln L$ as a function of signal strength $\mu$. In the "stat+syst+theory" scan all nuisance parameters are profiled at each value of $\mu$. In the "stat+syst" scan only the nuisance parameters for the experimental uncertainties are profiled, and the theory uncertainties are fixed to their best-fit values. In the "stat only" scan all nuisance parameters are fixed to their best-fit values. The crossings of each curve with $-2\Delta \ln L = 1$ are used to define the uncertainty in the best-fit value.

It is also helpful to examine how separate channels and categories contribute to this signal strength measurement. Figure 5.32 gives the best-fit signal strengths in fits to individual combinations of channels and categories. Each signal strength is determined from a single fit to the observed data, but with a model in which a signal strength parameter is introduced for each channel and category combination, denoted $\mu_i$. These parameters are allowed to float freely and independently, while the correlation structure of all nuisance parameters $\theta$ is preserved. A test statistic $q_{\mu_i}$ can be defined which compares the likelihood of the nominal model with a single
signal strength parameter $\mu$, to this modified model:

$$q_\mu = -2 \ln \frac{\mathcal{L}(\text{data} | \hat{\mu}, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} | \hat{\mu}, \hat{\theta}_\mu_i)}.$$ (5.20)

The observed value of this test statistic is compared to the expected distribution under the one-signal-strength hypothesis in figure 5.32b. The expected distribution is determined from toy datasets, generated with an assumed signal strength from the combined fit of $\mu = 0.88$. The observed data is found to be compatible with this hypothesis.

![Figure 5.32: (a) Best-fit values of the signal strength parameter $\mu$ in each channel-category combination and (b) distribution of a test statistic $q_\mu$, which assess the compatibility of the individual measurements with the hypothesis of a single common value of $\mu$. The observed value of this test statistic is indicated by an arrow.](image)

The $H \to \tau\tau$ analysis is sensitive to Higgs boson couplings to both vector bosons and fermions. The former is via the VBF and W/Z-associated production modes, and the latter via the gluon-gluon fusion mode and the $\tau\tau$ decay itself. These coupling strengths can be denoted by the parameters $\kappa_V$ and $\kappa_f$ respectively, defined such that $\kappa_V = \kappa_f = 1.0$ in the SM. The signal-plus-background likelihood function can be redefined as a function of these two parameters instead of the single parameter $\mu$. Figure 5.33 shows scans of the likelihood in this two-dimensional parameter space. For these scans the expectation from the $H \to WW$ process is considered as part of the signal, to ensure a consistent measurement of the Higgs boson couplings. This provides increased sensitivity to the $\kappa_V$ coupling, mostly via VBF $H \to WW$ events entering the $e\mu$ VBF-tag categories, where the predicted number of events scales as $\kappa_V^4$. Figure 5.33a gives
the likelihood scan for the observed data and figure 5.33b for an Asimov dataset corresponding to the SM expectation. The observed value of \((\kappa_V, \kappa_f)\) is found to be compatible with the SM value of \((1, 1)\).

![Figure 5.33](image)

**Figure 5.33:** Likelihood scans in the two-dimensional parameter space \((\kappa_V, \kappa_f)\) for the coupling of the Higgs boson to vector bosons and fermions respectively. These scans are performed under the assumption \(m_H = 125\) GeV and with the \(H \rightarrow WW\) contribution treated as part of the signal. Scans for (a) the observed data and (b) an Asimov dataset for the SM expectation \(\kappa_V = \kappa_f = 1.0\) are given. Contours indicating the 68% and 95% CL regions are drawn.

### 5.9 MSSM search results

The statistical interpretation in the MSSM analysis follows the same methods as in the previous section. This includes the use of the profile likelihood ratio as a test statistic to compare background-only and signal-plus-background hypotheses. The maximum likelihood is found through a simultaneous fit to data of the \(m_{\tau\tau}\) distributions in all channels and categories. Upper limits in this search are determined in two contexts. The first is in the \(m_h^{\text{max}}\) scenario, where limits on the parameter \(\tan \beta\) are determined as a function of \(m_A\). The signal model includes the three neutral Higgs bosons \(h, H\) and \(A\) with masses and cross sections specified by the \(m_A\) and \(\tan \beta\) values in question. The second context is for model-independent limits on the cross section of a single neutral Higgs boson, denoted \(\Phi\), decaying to \(\tau^+\tau^-\) in either the gluon-gluon fusion or \(b\)-associated production mode. Figures 5.34 and 5.35 give the \(m_{\tau\tau}\) distributions for each category and channel for the 7 TeV and 8 TeV analyses respectively. As in the SM case,
the background expectation and uncertainties correspond to the result of the global maximum likelihood fit. The signal expectation is given for the $m_h^{\text{max}}$ scenario with $m_A = 160 \text{ GeV}$ and $\tan \beta = 8$.

A number of additional steps are needed to determine the $m_A$-$\tan \beta$ limits. Signal samples are generated only for the set of $m_A$ mass points to be tested, in the range 90 GeV to 1 TeV. The step size between points increases with $m_A$ to scale with the increasing $m_{\tau\tau}$ mass resolution. At each $m_A$-$\tan \beta$ hypothesis, the masses of the other two Higgs bosons are calculated using results from the LHC Higgs Working Group [106]. In each event category, templates for the $h$ and $H$ are generated by a horizontal morphing [154] between templates from the two samples closest in mass. The category acceptance is similarly interpolated from the neighbouring mass points. All three templates are scaled by the appropriate cross sections and branching ratios and combined into a single template. The 95% CL upper limit is determined for each point on the $m_A$-$\tan \beta$ grid, with the signal strength parameter $\mu$ uniformly scaling the entire signal model. The limit in $\tan \beta$ is then defined as the point on which this upper limit is found to occur at $\mu = 1.0$. Practically, this is determined by interpolation between the points either side of this threshold.

Observed and expected limits in this model are given in figure 5.36, which also shows the exclusion regions determined by the LEP experiments [131]. The observed limits in $\tan \beta$ are found to be compatible with the background-only expectation across the entire $m_A$ range. It is also important to consider how these limits may be modified, given the discovery of a Higgs boson with mass around 125 GeV and the evidence for decays to tau pairs. Therefore, an additional limit is given for a pseudo-dataset in which only a SM signal is present on top of the expected background. This is found to be less stringent than the background-only limit, especially in the low $m_A$ region where the SM and MSSM signal expectations overlap.

It should be noted that if the 125 GeV Higgs boson is assumed to be the light CP-even state $h$, then further regions of the $m_A$-$\tan \beta$ parameter space would be excluded in the $m_h^{\text{max}}$ scenario. Figure 5.37a shows a recent result [155] for the allowed region of the $m_h^{\text{max}}$ scenario under the assumption of a light Higgs boson with mass around 125 GeV. A number of updated benchmark scenarios are also proposed, for example, the $m_{H^{\text{mod}+}}$ scenario given in figure 5.37b. These scenarios have much larger allowed regions and will be studied in a future CMS publication.

Figure 5.38 gives model-independent upper limits on the production of a single neutral Higgs boson with mass $m_\Phi$. The limits on the cross section times branching fraction, $\sigma \cdot B(\Phi \rightarrow \tau\tau)$, are determined individually for gluon-gluon fusion and $b$-associated production using the 8 TeV dataset only. In the fit to extract gluon-gluon fusion limits the $b$-associated contribution
Figure 5.34: Observed and predicted $m_{\tau\tau}$ distributions for the MSSM 7 TeV analysis in the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for $m_A = 160$ GeV and $\tan\beta = 8$. 
Figure 5.35: Observed and predicted $m_{\tau\tau}$ distributions for the MSSM 8 TeV analysis in the $\mu\tau_h$, $e\tau_h$ and $e\mu\mu$ channels. The background distribution and uncertainty corresponds to the result of the global fit to data. The signal distribution is given for $m_A = 160$ GeV and $\tan\beta = 8$. 
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.36: Expected and observed 95% CL upper limits in the $m_A$-$\tan \beta$ parameter space of the $m_h^{\text{max}}$ scenario. The 1$\sigma$ and 2$\sigma$ probability intervals are given for the background-only expectation. The median expectation is also given for the background-only model with the addition of the expected SM Higgs boson contribution. Exclusion regions determined by the LEP Collaborations [131] are shown in green.

is allowed to float freely, and vice versa. This is required as neither the No B-Tag or B-Tag categories are completely pure in one production mode, and this avoids the need to impose any assumptions about the ratio of cross sections between the two processes. As for the $m_A$-$\tan \beta$ limits, additional limits are given for a pseudo-dataset in which only a SM signal is present on top of the expected background. The effect of this is found to be small compared to the background-only expectation. This is anticipated as, without a more advanced categorisation scheme, the sensitivity of the MSSM analysis to gluon-gluon fusion is lower than in the SM analysis.
Searches for neutral Higgs bosons decaying to tau pairs

Figure 5.37: The excluded and allowed regions of the $m_A$-$\tan \beta$ parameter space in the (a) $m_h^{\text{max}}$ and (b) $m_h^{\text{mod+}}$ scenarios [155]. The allowed regions are under the assumption that the light scalar $h$ is the Higgs boson observed with mass 125 GeV. These are indicated by light and dark green shaded areas which correspond to uncertainties in $m_h$ of 3 GeV and 2 GeV respectively.

Figure 5.38: Observed and expected 95% CL upper limits on $\sigma \cdot B(\Phi \to \tau \tau)$ for a neutral Higgs boson $\Phi$ produced via (a) gluon-gluon fusion or (b) in association with b-quarks. The median expectation is also given for the background-only hypothesis with the addition of the expected SM Higgs boson contribution, indicated by a blue line.
Chapter 6

Conclusions

This thesis has presented analyses of proton-proton collision data recorded by the CMS detector during the 2011 and 2012 runs. The cross sections for the production of a $Z$ boson in association with exactly one or at least two $b$ jets have been measured for $\sqrt{s} = 7$ TeV collisions using 2.1 fb$^{-1}$ of data. These are determined at the hadron level for a $Z$ boson with a mass in the range $76 < m_Z < 106$ GeV; a lepton acceptance of $p_T > 20$ GeV and $|\eta| < 2.5$; and a $b$-jet acceptance of $p_T > 25$ GeV and $|\eta| < 2.1$. The $b$ jets are required to be separated from each lepton by at least $\Delta R = 0.5$. The $Z+1b$ and $Z+2b$ cross sections are measured to be $3.41 \pm 0.05$ (stat.) $\pm 0.27$ (syst.) $\pm 0.09$ (theory) pb and $0.37 \pm 0.02$ (stat.) $\pm 0.07$ (syst.) $\pm 0.02$ (theory) pb respectively. Comparisons between data and simulation have also been made for several kinematic variables, with moderate discrepancies observed in some places. These provide useful tests of MC simulation, and are helpful in understanding the $Z+b$-jet background in searches for new physics.

Searches for a SM and for MSSM Higgs bosons decaying to tau pairs have also been presented. This channel provides a direct probe of the Yukawa couplings between fermions and the Higgs field that give rise to the fermion masses. The searches use the entire 2011 and 2012 collision datasets recorded by the CMS detector at centre-of-mass energies of 7 TeV and 8 TeV respectively. Results are determined from distributions of the di-tau invariant mass in the $\mu\tau_h$, $e\tau_h$ and $e\mu$ final states. Both the SM and MSSM analyses share a common baseline event selection. They also exploit event categorisation to improve sensitivity to signal and to specific Higgs boson production modes. In the SM search an excess of events is observed over the background-only expectation with a local significance greater than three standard deviations between mass hypotheses of 115 GeV and 125 GeV. At 125 GeV the observed (expected) significance is 3.0 (3.1) standard deviations and the best-fit signal strength is $0.88 \pm 0.33$ times the SM expectation. No significant excess of events above the background expectation is observed in the MSSM search. Upper limits at the 95% CL are determined in the $m_A\cdot\tan\beta$ parameter space for the
$m_h^{\text{max}}$ scenario. Additionally, model-independent limits on the cross section times branching fraction for a single Higgs boson produced via either gluon-gluon fusion or in association with b-quarks are determined for mass hypotheses in the range 90 GeV to 1 TeV.

The LHC has opened a new high-energy frontier in the study of the SM. The discovery of a Higgs boson with mass around 125 GeV by the CMS and ATLAS Collaborations is a triumph for the theory of electroweak symmetry breaking and has led to the 2013 Nobel Prize in Physics being awarded to Higgs and Englert. In 2015 the LHC will re-commence operation at a new record centre-of-mass energy of 13 TeV. New data will be used to test the compatibility of this Higgs boson with the SM to an even greater extent, as well as offering much improved sensitivity to signatures of new physics.
“The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to research, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Many discoveries are reserved for ages still to come, when memory of us will have been effaced. Our universe is a sorry little affair unless it has in it something for every age to investigate... Nature does not reveal her mysteries once and for all.”

— Seneca, Natural Questions Book 7, c. first century
Bibliography


[38] CMS Collaboration, “Constraints on the Higgs boson width from off-shell production and decay to ZZ to llll and llvv”, CMS Physics Analysis Summary CMS-PAS-HIG-14-002 (2014).


[110] D0 Collaboration, “A measurement of the ratio of inclusive cross sections \( \sigma(p\bar{p} \to Z + b \text{ jet})/\sigma(p\bar{p} \to Z + \text{jet}) \) at \( \sqrt{s} = 1.96 \text{ TeV} \)”, Phys. Rev. D83 (2011) 031105, doi:10.1103/PhysRevD.83.031105, arXiv:1010.6203.


[120] CMS Collaboration, “Search for the standard model Higgs Boson in the decay channel $H \to ZZ^{(*)} \to q\bar{q}\ell^+\ell^-$ at CMS”, CMS Physics Analysis Summary CMS-PAS-HIG-11-027 (2011).


[136] D. de Florian, G. Ferrera, M. Grazzini, and D. Tommasini, “Higgs boson production at the LHC: transverse momentum resummation effects in the H \( \rightarrow 2\gamma \), H \( \rightarrow WW \rightarrow \ell\ell\nu\nu \)
and $H \rightarrow ZZ \rightarrow 4\ell$ decay modes", \textit{JHEP} \textbf{06} (2012) 132, 


[141] CMS Collaboration, “Performance of b tagging at $\sqrt{s} = 8$ TeV in multijet, $t\bar{t}$ and boosted topology events”, \textit{CMS Physics Analysis Summary CMS-PAS-BTV-13-001} (2013).


[145] CMS Collaboration, “Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s} = 7$ TeV”, \textit{JHEP} \textbf{1310} (2013) 101, 

arXiv:1109.6034.


### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4FS</td>
<td>four-flavour scheme</td>
</tr>
<tr>
<td>5FS</td>
<td>five-flavour scheme</td>
</tr>
<tr>
<td>BDT</td>
<td>boosted decision tree</td>
</tr>
<tr>
<td>BSM</td>
<td>beyond-the-standard-model</td>
</tr>
<tr>
<td>CL</td>
<td>confidence level</td>
</tr>
<tr>
<td>CSC</td>
<td>cathode strip chamber</td>
</tr>
<tr>
<td>CSV</td>
<td>combined secondary vertex</td>
</tr>
<tr>
<td>CTF</td>
<td>combinatorial track finder</td>
</tr>
<tr>
<td>DA</td>
<td>deterministic annealing</td>
</tr>
<tr>
<td>DAQ</td>
<td>data acquisition</td>
</tr>
<tr>
<td>DT</td>
<td>drift tube</td>
</tr>
<tr>
<td>ECAL</td>
<td>electromagnetic calorimeter</td>
</tr>
<tr>
<td>GSF</td>
<td>Gaussian sum filter</td>
</tr>
<tr>
<td>HB</td>
<td>hadron barrel</td>
</tr>
<tr>
<td>HCAL</td>
<td>hadron calorimeter</td>
</tr>
<tr>
<td>HE</td>
<td>hadron endcaps</td>
</tr>
<tr>
<td>HF</td>
<td>hadron forward</td>
</tr>
<tr>
<td>HLT</td>
<td>high-level trigger</td>
</tr>
<tr>
<td>HO</td>
<td>hadron outer</td>
</tr>
<tr>
<td>JER</td>
<td>jet energy resolution</td>
</tr>
</tbody>
</table>
**ACRONYMS**

**JES** jet energy scale

**L1** Level-1

**LHCHXSWG** LHC Higgs Cross Section Working Group

**LO** leading order

**MC** Monte Carlo

**MPF** missing transverse energy projection fraction

**MPI** multi-parton interaction

**MSSM** minimal supersymmetric standard model

**NLO** next-to-leading order

**NNLO** next-to-next-to-leading order

**NNLL** next-to-next-to-leading logarithmic

**pdf** probability density function

**PDF** parton distribution function

**PF** particle flow

**PS** Proton Synchrotron

**PSB** Proton Synchrotron Booster

**RF** radio frequency

**RPC** resistive plate chamber

**SPS** Super Proton Synchrotron

**SSV** simple secondary vertex

**SM** standard model

**TEC** tracker endcaps

**TIB** tracker inner barrel

**TID** tracker inner disks

**TOB** tracker outer barrel

**UE** underlying event
**VBF**  vector boson fusion

**WLCG**  Worldwide LHC Computing Grid