Searching for Beyond the Standard Model Resonances in the $HH \to bbbb$ Final State Using the ATLAS Detector



Beojan Stanislaus
The Queen's College
Trinity Term 2020

Submitted for the degree of
Doctor of Philosophy
in
Particle Physics

This thesis is dedicated to my parents and my sister, Seophine.

An ATLAS analysis searching for Beyond the Standard Model resonances, X, that are produced and decay through the $pp \to X \to HH \to b\bar{b}b\bar{b}$ process is presented. These events are studied in the phase space region where the four b-quarks produce four well-separated jets.

A significantly stronger upper-limit is set on the $pp \to X \to HH$ cross-section for these resonances than was set by the Early Run 2 ATLAS analysis. Above 350 GeV these limits are also stronger than those of the combination of all ATLAS di-Higgs searches with 27.5 fb^{-1} to 36.1 fb^{-1} of data. No statistically significant excesses are found, with the largest excess having a global significance of $1.2^{+0.4}_{-0.3}\sigma$. In particular, with an improved pairing algorithm and more sophisticated background modelling, the 2.8σ excess found at 280 GeV by the Early Run 2 analysis is not confirmed. The results of this analysis suggest the upcoming version, which uses an evolution of the same technique on the same data to look for Standard Model di-Higgs production, will be able to set stringent limits on that process.

Contents

Analysis Overview 10

1. Theory 12

1.1. The Standard Model and the Higgs Mechanism 12

1.2. HH Production in the Standard Model 18

1.2.1. Variation with κ_λ 18

1.2.2. Resolving m_{HH} with Detector Effects 21

1.3. Beyond the Standard Model HH Production 23

1.4. The $HH \to b\bar{b}b\bar{b}$ Decay Channel 25

2. The ATLAS Detector 28

2.1. Inner Detector 30

2.1.1. Silicon Detectors 34

2.1.2. Transition Radiation Tracker 35

2.2. Calorimeter 36

2.2.1. Resolution and Clustering 38

2.3. Muon Spectrometer 39

3. Jets, Muons, and b-Tagging 41

3.1. Jets 41

3.1.1. Particle Flow 41

3.1.2. Anti-k_t Jet Clustering Algorithm 44

3.1.3. Jet Calibration 45

3.2. Muons, and the Muon-in-Jet Correction 47

3.3. b-Tagging . . . 48
3.3.1. Working Points . . . 51
3.3.2. b-jet Energy Correction . . . 53
3.4. b-Tagging on Particle Flow Jets . . . 54
3.4.1. Monte Carlo Sample and Preselection . . . 54
3.4.2. Performance Evaluation . . . 55

4. The $HH \to b\bar{b}b\bar{b}$ Analysis 59
4.1. Trigger . . . 59
4.1.1. Trigger Exclusivity (Bucketing) Strategy . . . 62
4.2. Dataset . . . 63
4.3. Event Selection . . . 64
4.3.1. Basic Requirements . . . 64
4.3.2. Tag Categories . . . 65
4.3.3. Higgs Reconstruction . . . 66
4.3.4. Background Rejection and Top Veto . . . 66
4.3.5. Kinematic Region . . . 70
4.4. Jet Pairing . . . 71
4.4.1. Background Peak with Early Run 2 Pairing Strategy . . . 73
4.4.2. Monte Carlo Signal . . . 77
4.4.3. Performance Evaluation of Old Pairing . . . 80
4.4.4. New Pairing Strategy . . . 81
4.4.5. Performance Comparison . . . 84
4.4.6. BDT Internals . . . 85
4.5. Background Estimation . . . 88
4.5.1. Validation . . . 92

4.6. Higgs Peak Sharpening 95

5. Statistical Analysis 97

5.1. Nuisance Parameters 101

5.2. Concrete Implementation 102

5.3. Expected (Pre-Unblinding) Limits 104

6. Sources of Uncertainty 106

6.1. Statistical Uncertainties 106

6.2. Systematic Uncertainties — Modelling 110

6.2.1. Jet and Flavour Tagging Uncertainties 110

6.2.2. Theory Uncertainties 123

6.3. Systematic Uncertainties — Background Estimation 123

6.3.1. Low / High H_T Decomposition 125

6.3.2. Orthogonal Polynomial Decomposition 130

6.3.3. Conclusion 140

7. Unblinded Results 143

8. Conclusions 150

A. Unblinded Cutflow 152

A.1. Signal Contamination in the Validation Region 152

Acknowledgements

I would first like to thank my supervisors, Çiğdem İşsever, Todd Huffman, and James Frost. Their guidance on all matters, both physics related and logistical / bureaucratical, has been invaluable throughout these past four years.

I would also like to thank Sue Geddes, the subdepartmental administrator, and Kim Proudfoot, the graduate studies assistant, who expertly handle all matters bureaucratic, and are in many ways the backbone of the subdepartment.

High Energy Physics is an exceptionally collaborative subject, and no work would be possible without the assistance of collaborators. The ATLAS $HH \to b\bar{b}b\bar{b}$ analysis described in this thesis was done in collaboration with around thirty others, from a number of institutions in the UK, Canada, Germany, Japan, and the United States. I would particularly like to thank Max Swiatlowski and Rafael Teixera de Lima, the analysis contacts who kept the process moving and the rest of us motivated. I would also like to thank Bill Balunas, who made many important contributions to the early parts of our analysis pipeline, Sean Gasiorowski who did much of the work on background modelling in this analysis, and Lucas Santiago Borgna who studied the triggers used and introduced the trigger exclusivity strategy. Many discussions were also had with Santiago Paredes, who has come through this process alongside me as one of the lead analysers on the boosted analysis. As a postdoc at Oxford, Bill also had many important discussions with me, helping understand what had happened when results were surprising.

The phenomenological study I (briefly) discuss was also a deeply collaborative work, though with a smaller group of people. Lydia Beresford expertly led this

work, while Santiago and Bill spent many hours on the machine learning strategy that forms the core of this paper. This work also would not have been possible with out Jesse Liu, who first suggested we consider using MadGraph and Delphes for this study, then joined himself and single-handedly handled our Monte Carlo needs. He was instrumental in pushing this work toward publication, even when this entailed joining our meetings at 6 A.M. Thanks are also owed to Cigdem and James for their guidance and organization, and Michael Spannowski for his theoretical contributions.

Not mentioned above are the other members of the Oxford ATLAS Exotics group. James Grundy and Iza Veliscek joined in the past two years, and have already made significant strides in the next stages of the ATLAS $HH \to b\bar{b}b\bar{b}$ analysis. I know I leave it in capable hands. Federico Celli and Migle Stankaityte work on the $H \to b\bar{b}$ + jet analysis, but their suggestions on the $HH \to b\bar{b}b\bar{b}$ analysis were nonetheless very helpful. Though not officially my supervisor, Tony Weidberg also kindly read my thesis, and helped ensure my description of the detector was accurate.

I would also like to thank the other students at Oxford (both ATLAS and otherwise) who accompanied me on this journey. Among others, these include Luca Ambroz, Cecilia Tosciri, Miha Zgubic, Mikkel Bjoern, Hannah Pullen, and Thomas Hancock. Thanks are also owed to Nurfikri Norjoharudeen who helped get me started on ATLAS, and Gabija Zemaityte with whom I shared an office through most of the past three years.

Thanks are due to CERN and the ATLAS experiment, who provide the framework necessary to enable such complex (and expensive!) experimental work. I would

particularly like to thank CERN for its hospitality during my LTA, and the UK Liaison Office, who helped make the LTA experience as smooth as could be.

Finances are of course an important aspect of a DPhil. I'd like to thank the Science and Technology Facilities Council, and the University of Oxford Subdepartment of Particle Physics for their generous funding. I would also like to thank my college, The Queen's College, who also provided a taberdarship, helping allay any financial concerns.

The Queen's College also deserves thanks for having hosted and educated me during my undergraduate studies. Particular thanks are owed to my undergraduate tutor, Richard Nickerson, whose highly enjoyable tutorials sparked my interest in high energy physics at colliders.

Finally, thanks are owed to my parents, who encouraged my ambitions and taught me the value of education, and without whose support it would never have been possible to complete this DPhil. I would also like to thank my sister, Seophine, who helped proof-read this thesis. I wish her good luck as she begins her own DPhil journey with the Oxford LHCb group.

Copyright

Figures 2.1, 2.3, 2.4, 2.5, 2.6, 2.8, 3.2, and 4.1 are copyright CERN for the benefit of the ATLAS Collaboration. Of these, Figures 2.1, 2.3. 2.6, 2.8, and 4.1 are used under the CERN audiovisual media terms of use, while Figures 2.4, 2.5, and 3.2 are taken from ATLAS publications covered by the Creative Commons Attribution license.

Analysis Overview

This thesis details a search for pairs of Higgs bosons, each decaying to a pair of b-quarks. The work in this thesis specifically focuses on relatively low mass ($< 1\,\text{TeV}$) Beyond the Standard Model resonant states, which decay to pairs of Higgs bosons which themselves decay into four well-separated b-quarks. This allows each b-quark to produce a distinct b-jet. The search was conducted using the ATLAS detector at the Large Hadron Collider (LHC).

The first chapter outlines the theory required to contextualize this search, that is, the fundamental principles of the Standard Model, the place of the Higgs field and boson in this theory, and the theoretical motivations for searches for pairs of Higgs bosons. It includes a description of phenomenological work I conducted looking at loss of m_{HH} resolution when reconstructing these events.

Likewise, the second chapter outlines the ATLAS detector, focusing primarily on the aspects of the detector relevant to the $HH \to b\bar{b}b\bar{b}$ search described.

The thesis then goes on to describe the objects used in the $HH \to b\bar{b}b\bar{b}$ analysis, including a description of my work commissioning flavour tagging on a new type of jet, the *particle flow* jet. The fourth chapter details the core of the $HH \to b\bar{b}b\bar{b}$ analysis itself, including the event selection which I implemented, and the novel jet pairing technique which I invented, refined, and implemented in this analysis.

The fifth chapter describes the statistical analysis procedure I implemented to obtain final results (in the form of a 95% confidence level upper limit on $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section) from this analysis, and is closely linked with the sixth chapter describing the statistical and systematic uncertainties in the analysis, and their

implementation in the statistical analysis procedure. It includes a description of my work evaluating the use of an orthogonal polynomial decomposition to model the uncertainty on the shape of the predicted background m_{HH} distribution.

The thesis then goes on to present the unblinded results of this analysis, and my conclusions.

1 Theory

1.1 The Standard Model and the Higgs Mechanism

The Standard Model is currently the most complete explanation of all of particle physics. It is one of the best tested models in history, and explains almost everything that has been observed to date. The only phenomena not explained by the Standard Model are gravity (explained by General Relativity, which has not yet been successfully unified with the Standard Model), Dark Matter, matter-antimatter asymmetry, and the masses of neutrinos.

Figure 1.1 shows the fundamental matter (*fermionic*, i.e. with half-integer spin) and force-carrying (*bosonic*, i.e. with integer spin) particles of the Standard Model. All nuclei are composed of protons and neutrons, which are themselves composed of quarks. These are surrounded by electrons. Muons and tau leptons are, in essence, heavier versions of the electron. These quarks and leptons are together called fermions, since they obey Fermi-Dirac statistics. This requires that they have half-integer spin. The fundamental fermions in the Standard Model all have spin $1/2$.

The remaining four (or five, counting W^+ and W^- separately) fundamental particles are bosons, so called because they obey Bose-Einstein statistics. This requires that they have integer spin. The bosons shown mediate three of the four forces: the gluon mediates the strong force, the photon (γ) the electromagnetic force, and the W and Z bosons the weak force. These force carrying bosons all have spin 1.

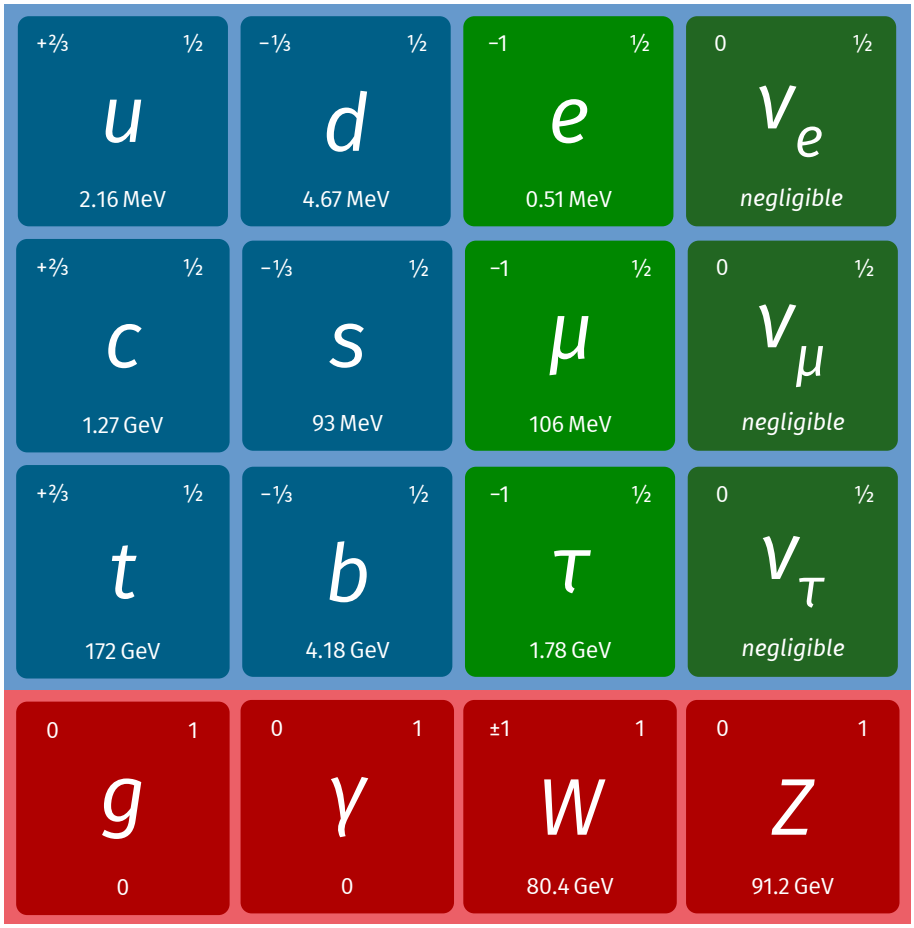


Figure 1.1: The particles of the Standard Model, apart from the Higgs boson. Those in the red region are bosons while those in the blue region are fermions. Of the fermions, those in blue are quarks while those in green are leptons. The darker green leptons are neutrinos. The charge is given in the upper left corner, in units of the elementary charge e, while the spin is given in the upper right corner, in units of $\hbar$. The mass is given below the symbol.

Gravity alone is not mediated by one of these bosons, being the result of the curvature of space-time itself in General Relativity. If quantized it would be mediated by a (hypothetical) massless spin-2 *graviton*.

However, there is a missing piece in this model. The Standard Model is a gauge theory, with a Lagrangian invariant under transformations described by its gauge group $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$. The $\mathrm{SU}(3)$ subgroup gives us the strong force, while $\mathrm{SU}(2) \times \mathrm{U}(1)$ produces the electromagnetic and weak forces, which can be unified into the electroweak force. In order to conserve gauge symmetry, the force carrying bosons of the electroweak force must be massless. However, we can see in Figure 1.1 that, unlike the photon (and the gluon), the W and Z bosons have significant masses. This can be achieved through spontaneous breaking of the $\mathrm{SU}(2) \times \mathrm{U}(1)$ electroweak symmetry.

In 1964 François Englert and Robert Brout [1], Peter Higgs [2], and later Gerald Guralnik, Richard Hagen, and Tom Kibble [3] published a mechanism for this. This mechanism is called the Brout-Englert-Higgs (BEH) Mechanism, or henceforth the Higgs Mechanism. A full derivation (summarized below) can be found in [4].

We can start with the terms in the Standard Model Lagrangian describing the electroweak gauge bosons [5]:

$$\mathcal{L}_{\text{EW-gauge}} = -\frac{1}{4} W^{i\mu\nu} W^{i}_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} \tag{1.1}$$

These describe three W bosons (with i taking the values 1, 2, 3) and one B boson, all of which are massless.

We then introduce two new complex scalar fields combined into a doublet (which we now call a *Higgs doublet*) ϕ. We then have the Higgs Lagrangian:

$$\begin{aligned}\mathcal{L}_{\text{Higgs}} &= \left(\mathrm{D}_\mu \phi\right)^\dagger \left(\mathrm{D}^\mu \phi\right) - \mathrm{V}(\phi) \\ &= \left(\mathrm{D}_\mu \phi\right)^\dagger \left(\mathrm{D}^\mu \phi\right) - \mu^2 \phi^\dagger \phi - \lambda\left(\phi^\dagger \phi\right)^2.\end{aligned} \tag{1.2}$$

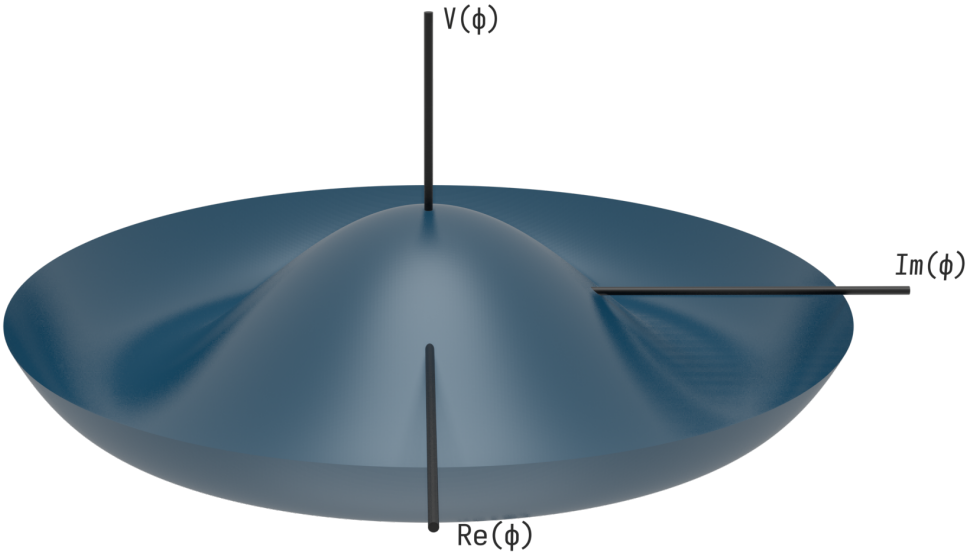


Figure 1.2: The Higgs potential $\mathrm{V}(\phi) = \mu^2 \phi^\dagger \phi + \lambda\left(\phi^\dagger \phi\right)^2$, with $\mu^2 < 0$. The minimum is at $\phi^2 = \frac{v^2}{2} = -\mu^2/2\lambda$.

The potential $\mathrm{V}(\phi) = \mu^2 \phi^\dagger \phi + \lambda\left(\phi^\dagger \phi\right)^2$ has the shape shown in Figure 1.2, where ϕ is a single complex field to enable visualization. The minimum is degenerate, so we can choose a convenient one to expand about:

$$\phi = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ v + h \end{bmatrix}. \tag{1.3}$$

Here, $v = \sqrt{-\mu^2/\lambda}$.

Among other terms [6],

$$\mathcal{L} \supset \overbrace{-\lambda v^2 h^2}^{\text{Massive Higgs boson}} + \overbrace{\frac{(gv)^2}{8}\left(W^{1\mu}W^1_\mu + W^{2\mu}W^2_\mu\right)}^{\text{Massive } W^\pm \text{ bosons}} + \underbrace{\frac{v^2}{8}\begin{bmatrix} W^{3\mu} & B^\mu \end{bmatrix}\begin{bmatrix} g^2 & -gg' \\ -gg' & g'^2 \end{bmatrix}\begin{bmatrix} W^3_\mu \\ B_\mu \end{bmatrix}}_{\text{Massive } Z \text{ boson and massless photon}} + \underbrace{-\lambda v h^3}_{\text{Higgs self-interaction}} \tag{1.4}$$

The orange and green terms appear from expanding $\left(\mathrm{D}_\mu \phi\right)^\dagger (\mathrm{D}^\mu \phi)$. The g and g' here are electroweak couplings.

Mass terms in Lagrangians take the form $-\frac{m^2}{2}\varphi^2$, for a field φ. Therefore, the new Higgs doublet produces mass terms for the $W^\pm$ and Z boson.[1] The orange term gives us $m_W = \sqrt{2\frac{\{gv\}^2}{8}} = \frac{gv}{2}$. The eigenvalues of the matrix in the green term are 0 and $g^2 + g'^2$. This gives us a massive Z boson, with mass $m_Z = \frac{v\sqrt{g^2+g'^2}}{2}$, and a massless photon. The Z boson field and photon field are a mixture of $W^{3\mu}$ and B^μ (specifically, rotated to lie along the eigenvectors of the matrix in green).

It also produces a new massive real scalar (i.e. spin 0) boson, which we call the *Higgs boson*, H. This has mass $m_H = v\sqrt{2\lambda}$. Its entry, in the same style as Figure 1.1, is shown in Figure 1.4. In a similar way, it produces mass terms for the various

1. $W^\pm = \frac{W^1 \mp i W^2}{\sqrt{2}}$

fermions, and every fermion interacts with the Higgs boson through a vertex like that shown in Figure 1.3, with a coupling proportional to its mass.

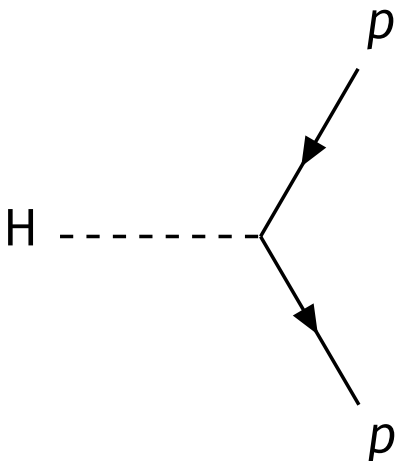


Figure 1.3: A generic Higgs – fermion interaction vertex.

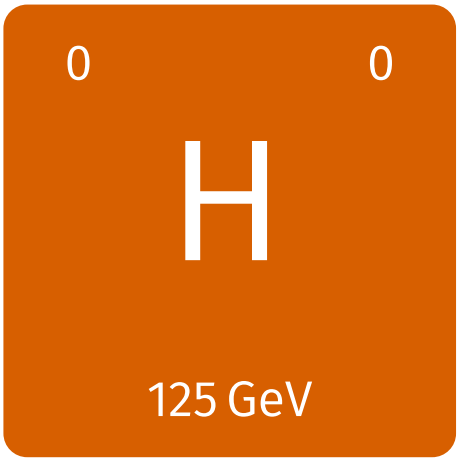


Figure 1.4: The Higgs boson, in the same style as in Figure 1.1.

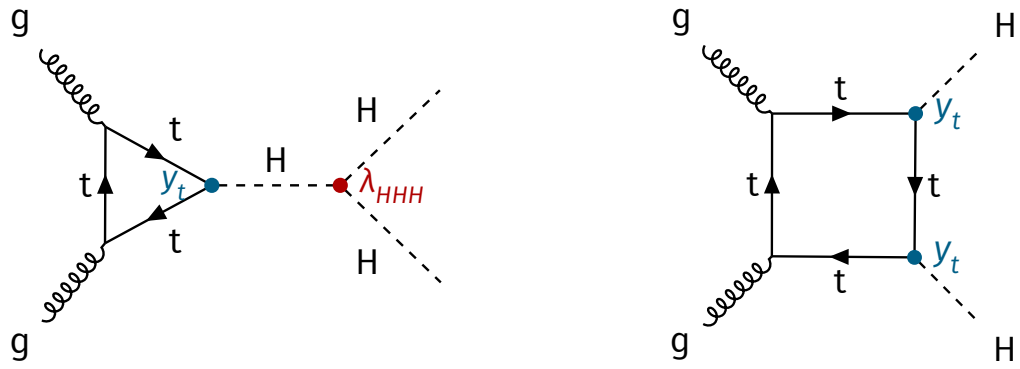
The measured value of the Higgs boson mass, m_H, is 125.10(14) GeV [7]. Given the measured values of other parameters, the Higgs vacuum expectation value, v, is 246 GeV [7]. We can also see in Eq. 1.4 that there is a trilinear Higgs self-interaction, with a coupling constant λv. By rearranging in terms of the Higgs boson mass m_H and vacuum expectation value v, we get

$$\lambda_{HHH} = \lambda v = \frac{m_H^2}{2v}. \tag{1.5}$$

By probing this self-coupling, we can test whether the relation in Eq. 1.5 is satisfied, and therefore test the Standard Model. The only direct way to probe this self-coupling is through measuring di-Higgs production.

1.2 HH Production in the Standard Model

At the LHC, the dominant route through which Higgs boson pairs can be produced in the Standard Model is through the two *gluon-gluon fusion* (ggF) diagrams shown in Figure 1.5.



(**a**) Triangle Diagram (**b**) Box Diagram

Figure 1.5: The two Standard Model gluon-gluon fusion HH production diagrams.

Observing and measuring this process provides the only direct way to measure λ_{HHH}. Limits are often reported in terms of $\kappa_\lambda = \lambda_{HHH}/\lambda_{HHH}^{\text{SM}}$ and $\kappa_t = y_t/y_t^{\text{SM}}$. With $\kappa_\lambda = 1$ and $\kappa_t = 1$, the cross-section of this process at $\sqrt{s} = 13\,\text{TeV}$ is 31.05 fb, at next-to-next-to-leading order, with the `FTapprox` top mass correction [8]. The best experimental limits on κ_λ are provided by the ATLAS Early Run 2 HH combination, which set observed 95% confidence level limits of $-5.0 < \kappa_\lambda < 12.0$ [9]. A combination of CMS analyses has constrained κ_t to $-0.9 < \kappa_t < 1.6$ at 95% confidence level [7].

1.2.1 *Variation with* κ_λ

As κ_λ varies, both the total cross-section of di-Higgs production and the differential cross-section with respect to the di-Higgs invariant mass (among other variables)

κ_λ	NLO X-Sec / fb
−1	116.71
0	62.51
1	27.84
2	12.42
2.4	11.65
3	16.28
5	81.74

Table 1.1: Next-to-leading order di-Higgs production cross-sections with varying κ_λ, at $\sqrt{s} = 13\,\text{TeV}$ [10].

vary. Due to destructive interference between the two tree-level production diagrams (Figure 1.5), this variation is non-trivial.

Table 1.1 [10] shows the total (NLO) cross-section for the di-Higgs process at various values of κ_λ. The total cross-section alone provides a reasonable level of discrimination between different values of κ_λ. However, there is a degeneracy between large positive and negative κ_λ, and discrimination power is low around $\kappa_\lambda = 2.4$

This can be resolved by looking at the differential cross-section as a function of the di-Higgs invariant mass, m_{HH}. Figure 1.6 shows the matrix element level leading order m_{HH} distribution for various κ_λ around 2.5. This figure was produced using samples simulated using MadGraph [11] and Pythia 8 [12]. The NNPDF 3.0 [13] parton distribution function set was used to generate this simulation.

At $\kappa_\lambda = 2.5$, for example, we see two peaks. The first, at the kinematic threshold ($m_{HH} = 250\,\text{GeV}$), is produced primarily by the process shown in the triangle diagram (Figure 1.5a) mediated by a single Higgs boson. This Higgs boson *must* be off-shell so the peak begins at the minimum m_{HH} required to produce two Higgs bosons: $250\,\text{GeV}$. It is this process that allows the measurement of κ_λ.

The second peak, at around 400 GeV, is produced by the process shown in the box diagram (Figure 1.5b), where the two Higgs bosons are produced directly from the top quark loop. Between the two peaks, we see a dip. This dip is caused by destructive interference between the box and triangle diagrams, which occurs because the triangle diagram contains three fermion lines, while the box diagram contains four. This leads to an overall minus sign on the triangle diagram, but not on the box diagram [14]. The point of maximum destructive interference is $\kappa_\lambda = 2.4$.

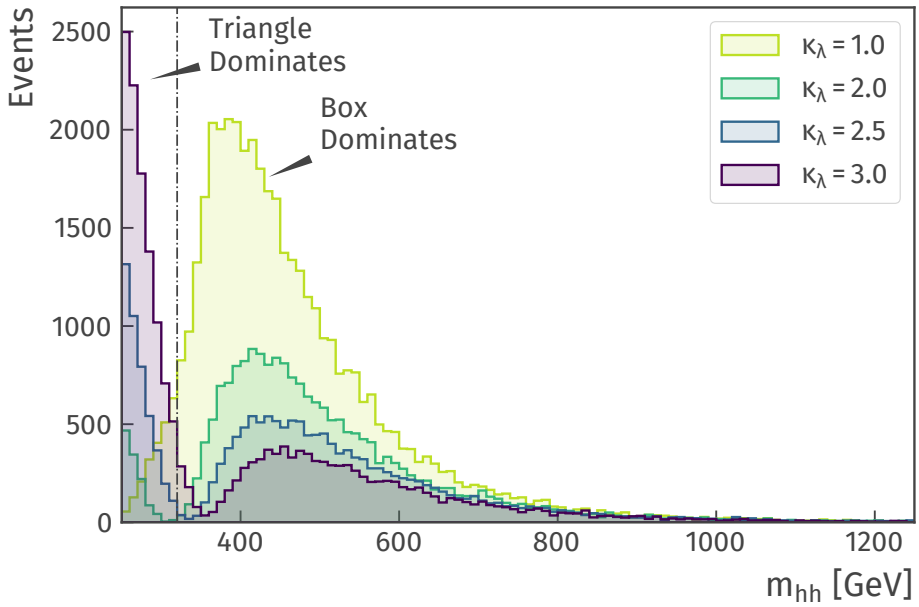


Figure 1.6: Matrix element level leading order m_{HH} distribution at various κ_λ around 2.5. Note the dip between the low and high m_{HH} peak due to destructive interference.

As a result, different κ_λ around 2.4 can be distinguished (at matrix element level) using the m_{HH} distribution, from the relative sizes of the low m_{HH} and high m_{HH} peaks and the position of the minimum between these peaks.

1.2.2 *Resolving m_{HH} with Detector Effects*

While we can see in Figure 1.6 that the matrix element level m_{HH} distributions allow us to determine κ_λ, this may not be possible with the m_{HH} distribution obtained by performing an analysis.

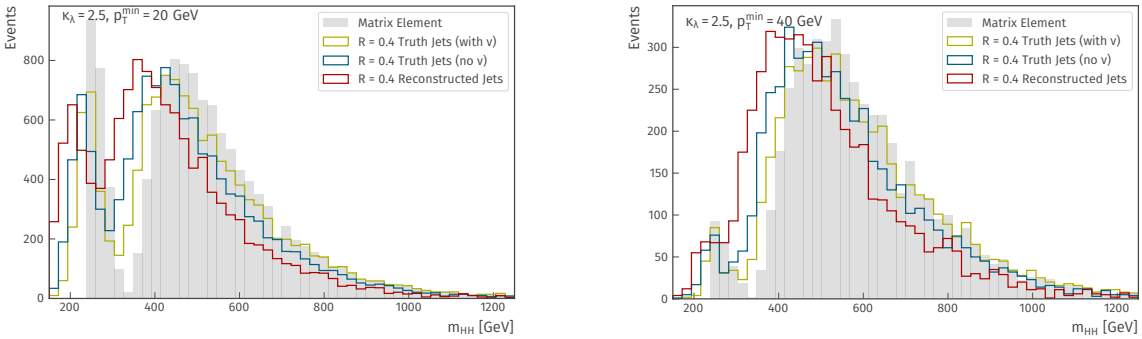
The primary way this happens is of course through the addition of background, which must then be mitigated. As a result, the m_{HH} distribution obtained from an analysis will suffer from significant background contamination, while some signal events will simultaneously be incorrectly removed by the background mitigation and estimation procedure.

Prior to this, however, reconstruction and detector effects may degrade m_{HH} resolution even in the absence of background, reducing the m_{HH} distribution's power to indicate the true value of κ_λ. I studied this effect as part of [14].

Figure 1.7 shows the m_{HH} distribution produced at three stages of resolution degradation, alongside the matrix element level distribution. The quarks produced in the final state decay into showers of hadrons. These showers are called jets, and the three stages are jets reconstructed from all final state particles produced by the simulation (truth jets, including neutrinos), all detectable final state particles produced by the simulation (truth jets, not including neutrinos), or the simulated detector signals produced by these final state particles (reconstructed jets). Jets are further described in Chapter 3.

The experiment's data acquisition system requires a trigger to control data rates, which enforces a minimum p_T requirement on jets. The effect of such triggers, enforcing a minimum p_T of 20 GeV or 40 GeV, were simulated by requiring these minimum transverse momenta on the b-quarks. This requirement is applied to

all three curves, and the matrix element level reference.[2] Figure 1.7 was made using a $\kappa_\lambda = 2.5$ MC sample produced using MadGraph and Pythia 8, with the detector simulated using Delphes [15] (emulating the ATLAS detector), and jet reconstruction using FastJet [16]. Jets were assigned to b-quarks through ghost association [17], and the plot is made using the four associated jets of each given type for each event.



(**a**) $p_T^{\text{min}} = 20\,\text{GeV}$ (**b**) $p_T^{\text{min}} = 40\,\text{GeV}$

Figure 1.7: Matrix element level m_{HH} distribution and m_{HH} distribution calculated using (1) truth jets, including neutrinos, (2) truth jets, not including neutrinos, and (3) reconstructed jets. A minimum p_T requirement is applied on the four assigned reconstructed jets to emulate a trigger.

From Figure 1.7 we can see that the main reason m_{HH} resolution is degraded (disregarding background contamination) is the removal of low m_{HH} events by the trigger. Since reconstructing the low m_{HH} peak is crucial to determining κ_λ, reducing trigger thresholds is extremely important in the $HH \to b\bar{b}b\bar{b}$ channel.

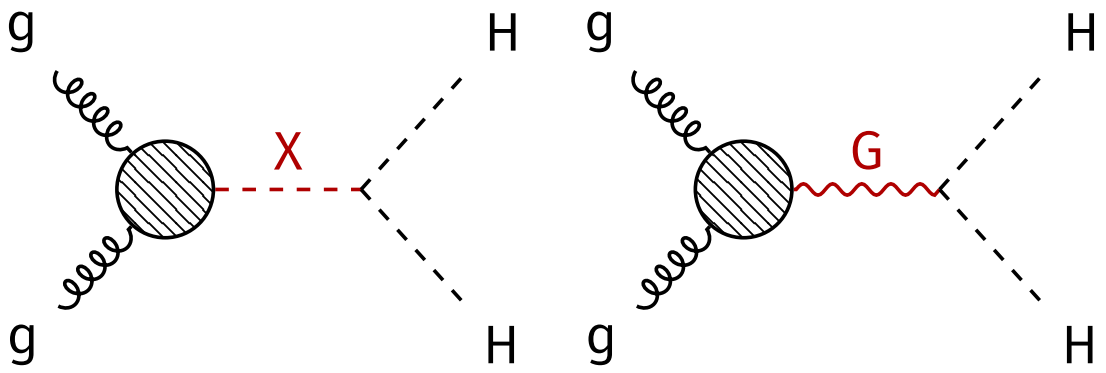
We can also see that the curves corresponding to all jet types produce an m_{HH} distribution that is shifted to lower m_{HH} compared to the matrix element level

2. The p_{T} requirements were applied to b-quarks rather than b-jets in order to keep the curves comparable.

distribution. This is caused by the $R = 0.4$ jets missing radiation that falls out of this jet cone. Such radiation is more prevalent at lower m_{HH}.

Modulo this effect, truth jets including neutrinos (and, to a lesser extent, not including neutrinos) can reproduce the matrix element level m_{HH} spectrum fairly well. However, with a 40 GeV minimum jet p_{T}, the low m_{HH} peak is almost entirely suppressed when using reconstructed jets.

1.3 Beyond the Standard Model HH Production



(**a**) Scalar intermediary (**b**) Graviton intermediary

Figure 1.8: Di-Higgs production through a heavy scalar or graviton (spin-2) intermediary.

Pairs of Higgs bosons can also be produced through the decay of heavy resonances in a Beyond the Standard Model (BSM) framework.

In order to decay to a pair of Higgs bosons such a resonance must have a mass of at least twice the mass of a Higgs boson, i.e. the resonance mass must be at least 250 GeV. The heavier the resonance mass, the higher the transverse momenta of the Higgs bosons produced, and therefore the closer together the two b-quarks produced by the decay of each Higgs boson. If the resonance mass is high enough

the two b-quarks produced by each Higgs boson will be close enough together to be captured by a single jet. The resolved analysis that is the subject of this thesis, which considers events where each Higgs boson decays to two b-quarks, each of which producing an individually resolvable jet, considers resonance masses from 251 GeV to 1200 GeV.

In the ATLAS $HH \to b\bar{b}b\bar{b}$ analysis two types of resonances in particular are considered: scalars (Figure 1.8a) and Kaluza-Klein gravitons (Figure 1.8b).[3] Simulations of graviton decays were used for some early studies in the analysis, but the limits shown in the results are set on scalars.

The scalar search conducted is a generic (model-independent) one, and the scalars simulated are given a width much smaller than detector resolution. An upper limit can then be set on the $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section, and the predictions of models can later be compared to this limit.

Such scalar resonances can be found in a number of BSM theories. Most notably, in a two Higgs doublet model (2HDM) [18] there will be *two* neutral scalar Higgs bosons. If the heavier Higgs boson has more than twice the mass of the lighter (in this analysis, the Standard Model Higgs boson) it can decay to a pair of light Higgs bosons.

Kaluza-Klein gravitons are found in Randall-Sundrum models [19][20] with a warped extra dimension [21]. The addition of this warped extra dimension leads to massive spin-2 particles called Kaluza-Klein gravitons since gravitons are the only particles which can propagate in the extra dimension. These Kaluza-Klein gravitons would have masses in the TeV range and one of their dominant decay

3. The terms Kaluza-Klein graviton and Randall-Sundrum (RS) graviton are used interchangeably in this thesis.

modes is to pairs of Higgs bosons [21]. Di-Higgs searches therefore provide a potential route for detecting these resonances.

1.4 The $HH \to b\bar{b}b\bar{b}$ Decay Channel

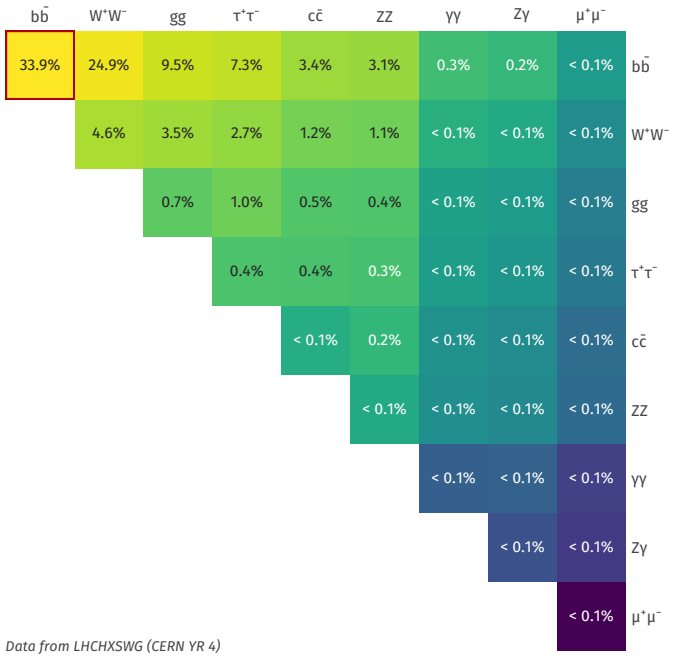


Figure 1.9: Di-Higgs branching ratios [22]. The $HH \to b\bar{b}b\bar{b}$ channel has the highest branching ratio.

While a pair of Higgs bosons may be produced by both Standard Model and Beyond the Standard Model processes, the subsequent decay of these Higgs bosons is well predicted by the Standard Model. These branching ratios are shown in Figure 1.9 [22]. The $HH \to b\bar{b}b\bar{b}$ channel has by far the largest branching ratio and is therefore a well motivated channel for studying di-Higgs production. The other two well-performing channels, $b\bar{b}\tau^-\tau^+$ and $b\bar{b}\gamma\gamma$, both also include a $H \to b\bar{b}$ decay, and the low branching ratio of the other Higgs boson decay is compensated for by the lower background levels in these channels.

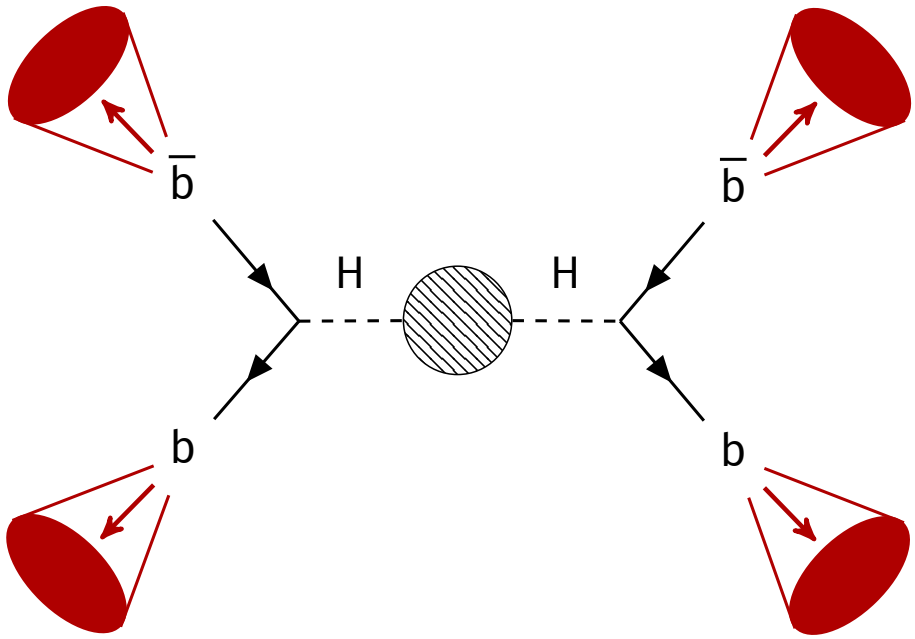


Figure 1.10: A resolved $HH \to b\bar{b}b\bar{b}$ event.

As shown in Figure 1.10, the resolved $HH \to b\bar{b}b\bar{b}$ analysis studies di-Higgs events where each Higgs boson decays to two b-quarks. QCD prohibits free coloured states [23]. As the b-quark and $\bar{b}$-antiquark fly apart, a "tube" of colour field forms between them which eventually breaks producing new quarks and antiquarks. These combine with each other and the original quarks and antiquarks to form sprays of hadrons. These hadrons will be concentrated around the direction of travel of each original quark or antiquark. If the four-vectors of all hadrons so produced were known, it would be possible to apply the anti-k_t [24] algorithm (with a suitably chosen R parameter) to recover two jets with axes in the approximate travel directions of the b-quark and $\bar{b}$-antiquark.

The hadrons formed by the *hadronization* process described above can also themselves decay. The b-quark and $\bar{b}$-antiquark themselves would have combined with one or two other (anti)quarks to form B hadrons. These have decay lengths on the

order of a few millimetres. This enables the flavour tagging procedure described in Section 3.3, which allows determination of the flavour of the quark from which a jet originates.

As a result, despite the large QCD background, the $HH \to b\bar{b}b\bar{b}$ channel is tractable. Combined with the large branching ratio, this makes $HH \to b\bar{b}b\bar{b}$ one of the best channels for a di-Higgs search.

2 The ATLAS Detector

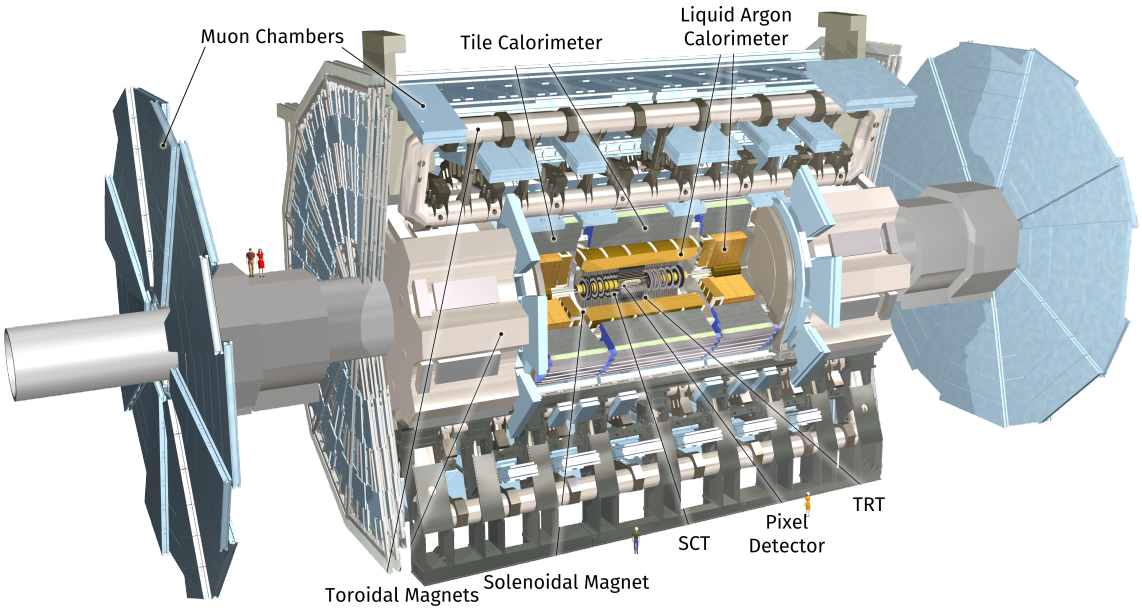


Figure 2.1: Diagram of the ATLAS detector [25].

The ATLAS detector (illustrated in Figure 2.1) is one of two general-purpose detectors at the LHC. It is a hermetic detector used to study a wide range of physics in both proton and heavy ion collisions. A full description of the design of this detector can be found in [26], which should be considered a general reference for this chapter.

In a broad sense, the ATLAS detector consists of

- A tracker (often called the *inner detector*, since it is the innermost part of the detector) which records the trajectories of charged particles, without stopping them.
- A calorimeter, which measures the energy of most particles (which interact with matter electromagnetically, or through the strong force). In doing so,

the energy carried by the particle is deposited in the calorimeter, and the particle is therefore stopped.

- Muon chambers, which record the trajectory and momentum of muons. Due to a combination of their mass, long lifetime, and lack of colour charge (which implies a lack of nuclear interactions), muons with at least moderate energy ($> 3\,\mathrm{GeV}$) pass through the calorimeter.

Each of these is further expanded upon later in this chapter.

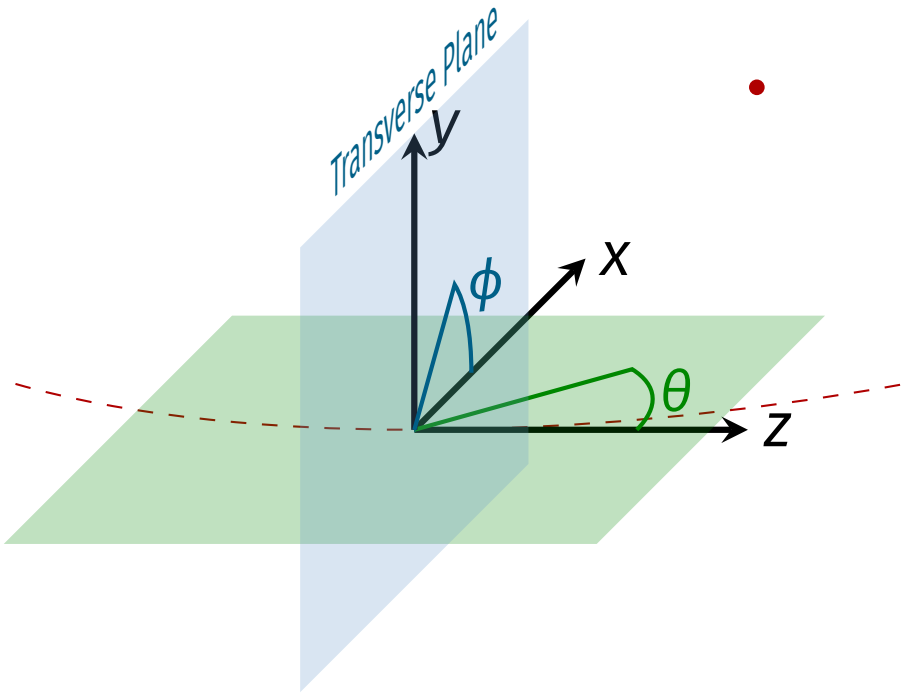


Figure 2.2: The ATLAS coordinate system. The red dashed line represents the LHC ring, and the red dot its centre. The x-axis points toward the centre of the LHC ring, while the y-axis points toward the surface of the Earth. The pseudo-rapidity $\eta = -\ln\tan(\theta/2)$. When polar coordinates are used, r is the magnitude of the displacement in the x-y plane.

Positions and object momenta in ATLAS are expressed in terms of a right-handed coordinate system, illustrated in Figure 2.2. The x and y coordinates are defined in the transverse plane (orthogonal to the beam axis), with x pointing toward the centre of the ring, and y pointing toward the surface of the Earth. These then define the positive z direction. Angles in the longitudinal plane are generally given as a pseudo-rapidity $\eta = -\ln\tan(\theta/2)$.

2.1 Inner Detector

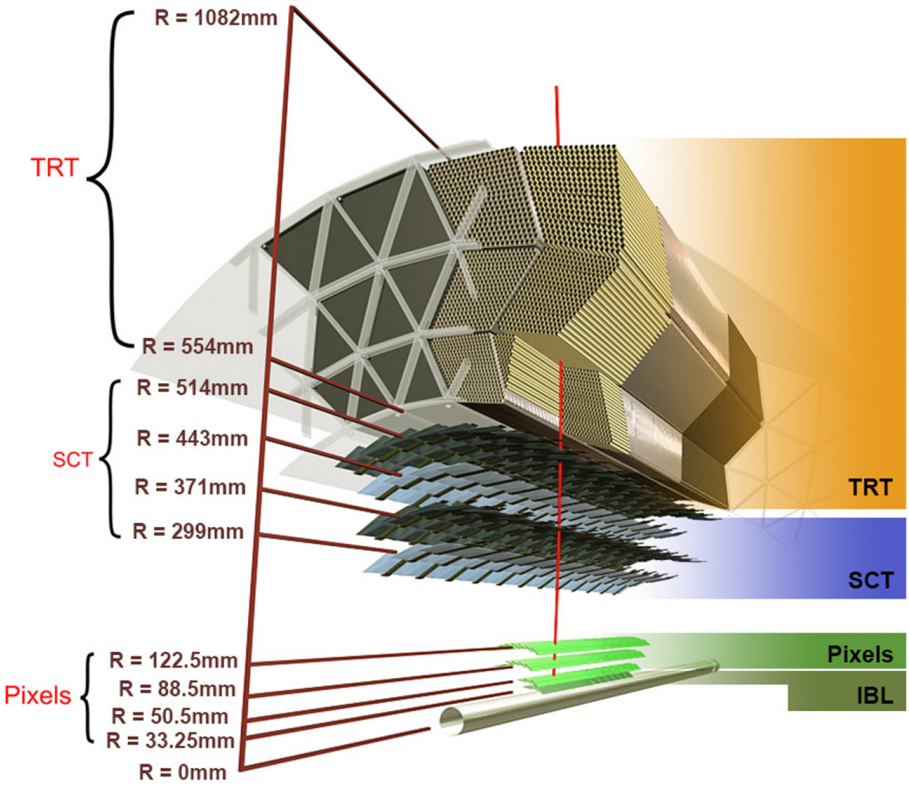
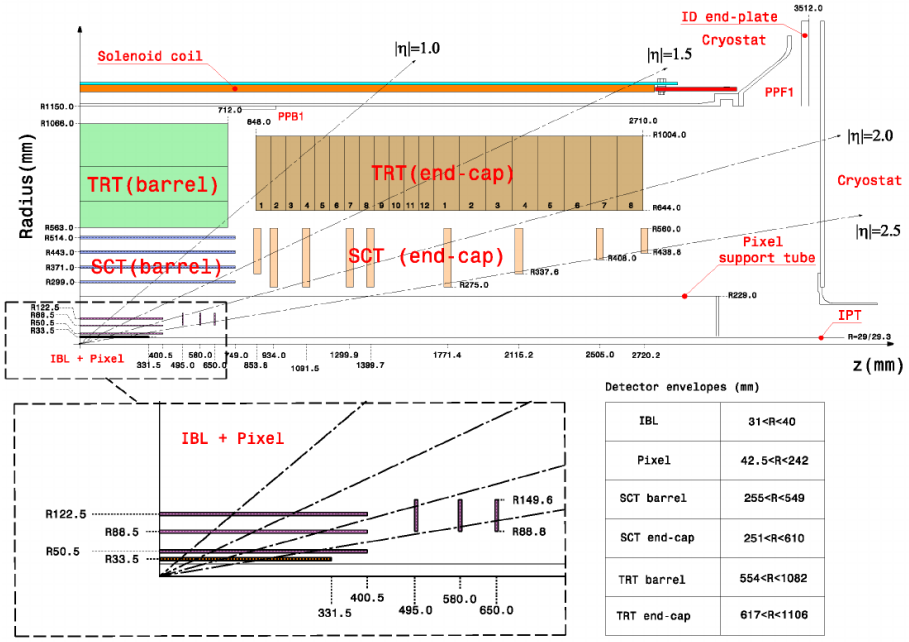


Figure 2.3: Layout of the ATLAS inner detector (tracker) [27].

The ATLAS tracker, or inner detector, illustrated in Figure 2.3, is divided into four major components. Figure 2.4 shows a more detailed view of the layout of these four components.



IBL	31<R<40
Pixel	42.5<R<242
SCT barrel	255<R<549
SCT end-cap	251<R<610
TRT barrel	554<R<1082
TRT end-cap	617<R<1106

Figure 2.4: Schematic of the ATLAS inner detector layout [28].

The insertable B layer (IBL), pixel detector and SCT (SemiConductor Tracker) use silicon detectors. The IBL is made up of 50 µm × 250 µm [29] pixels. This layer was added just before the start of Run 2 to improve b-tagging performance. As shown in Table 2.1, it has an $r\phi$ resolution of 8 µm, which translates to a ϕ resolution of 2.42×10^{-4} rad. Figure 2.5 from the IBL Technical Design Report [30] shows how the IBL leads to improved b-tagging performance.

The pixel detector then uses 50 µm × 400 µm pixels, while the SCT uses 80 µm × 64 mm strip detectors (in the barrel region), with the higher resolution (smaller dimension) in the transverse direction (the direction perpendicular to the beam axis).

The SCT barrel extends up to 80.5 cm from the centre of the detector in the z direction. Past this point, the SCT end-caps provide tracking coverage. These consist of strip sensors of varying sizes oriented in planes perpendicular to the

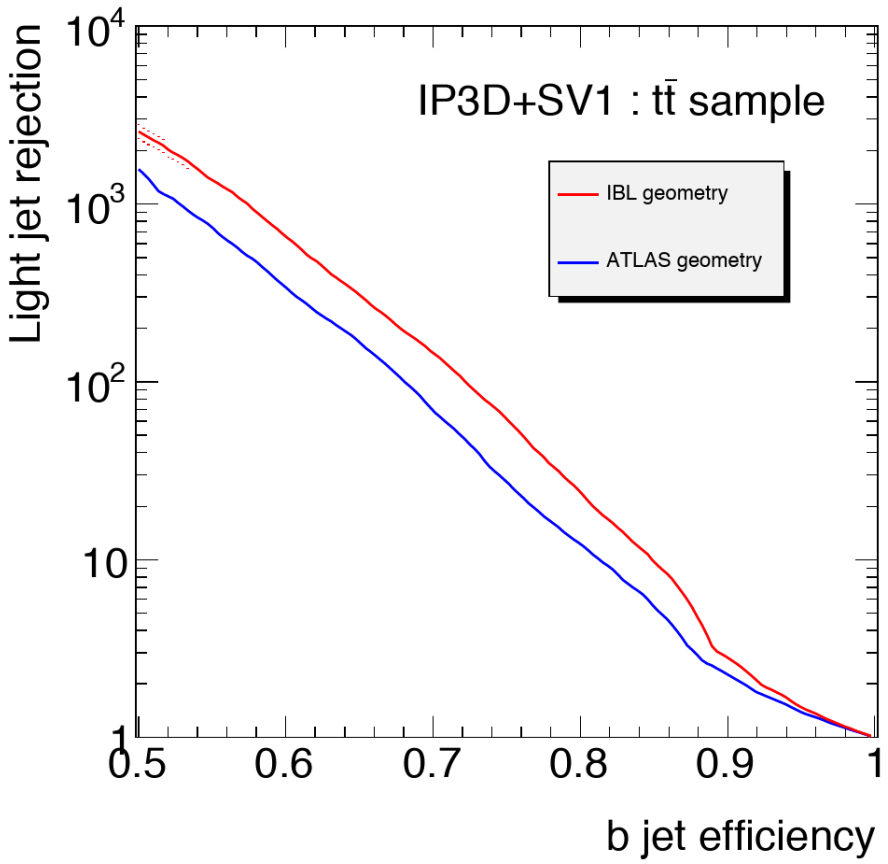


Figure 2.5: Improvement in b-tagging performance due to the inclusion of the insertable B-layer. At any given b-jet efficiency, the basic b-tagging algorithms (IP3D and SV1) reject more light jets when the IBL is available. Plot taken from the IBL Technical Design Report [30].

beam direction, with the long side in the r direction. Similarly, the pixel barrel extends up to 40 cm from the centre of the detector in the z direction, with disks perpendicular to the beam direction providing high resolution tracking up to 65 cm from the centre of the detector.

The longitudinal resolution of the SCT is improved by constructing each SCT module from two layers of microstrip detectors sandwiched back-to-back, with an angle of 40 mrad between the long (64 mm) axes. This gives a longitudinal resolution of about 0.5 mm.

The TRT (Transition Radiation Tracker) makes use of 4 mm diameter gas straws containing wires [26], fully described later in Section 2.1.2. It was designed to contribute to tracking through producing a large number of hits per track, while also contributing to particle ID by exploiting transition radiation.

The entire inner detector is surrounded by a solenoid producing a 2 T magnetic field in the z direction. The trajectory of any charged particle moving through the inner detector therefore follows an arc whose radius follows the relation

$$r = \frac{p_{\mathrm{T}}}{qB} \tag{2.1}$$

By joining the hits left by particles in each layer into tracks, taking into account the circular shape these tracks should have, the arc radius and hence the transverse momentum of the particle can be determined.

The pixel detector and SCT covers the full ϕ range (0 to 2π), and $\left|\eta\right| < 2.5$, with sensors in a barrel around the centre of the detector, and end-caps at the two ends.

Component	Resolution / μm $r\phi$	r	z
IBL [32]	8	–	40
Pixels (Barrel)	10	–	115
Pixels (End-cap)	10	115	–
SCT (Barrel)	17	–	580
SCT (Endcap)	17	580	–
TRT	130	130	–

Table 2.1: Spatial resolution of each component of the inner detector [26]. $r\phi$ resolution is given here since ϕ resolution changes with layer radius r.

They provide high resolution tracking information in the 2D r-ϕ plane, and lower resolution information in the z-axis, which is translated to a pseudo-rapidity.

The TRT covers $|\eta| < 2.0$. The TRT provides both tracking information (in r-ϕ only), with lower resolution but much higher hit rate (around 36 per track) than the silicon trackers [26], and aids in discriminating between hadrons and electrons through the production and detection of transition radiation [31].

Table 2.1 shows the spatial resolution of each of these components.

2.1.1 *Silicon Detectors*

The silicon microstrip detectors used in the SCT consist of p-n junctions held under a reverse-bias voltage, creating a wide depletion region. The microstrips consist of strips of p-type silicon deposited in an n-type substrate [33]. As a charged particle passes through the depletion region, it produces new electron-hole pairs. These electrons and holes drift under the reverse-bias voltage, producing a signal.

Due to their proximity to the interaction, the pixel sensors are subject to considerably more extreme radiation conditions. Under these conditions acceptors (holes)

are produced in the n-type substrate. This increases the depletion voltage, and eventually causes the n-type silicon to type invert and become p-type silicon [33].

In order to improve the radiation tolerance of the pixel sensors, a variation on the SCT design is used [26]. The pixels are formed of n+ doped silicon implants in the n-type substrate, with the p-n junctions being produced by corresponding p+ implants on the opposite side of the sensor [34]. This design is not used for the SCT sensors due to the higher cost of production [33].

2.1.2 *Transition Radiation Tracker*

The Transition Radiation Tracker consists of 4 mm diameter hollow Kapton tubes (referred to as *straws*), with the straw wall covered in a 0.2 μm layer of aluminium (itself covered in a thin graphite-Kapton layer to protect it) on each side [35]. These form the cathodes of a drift chamber, while 30 μm gold plated tungsten [35] wires within each straw form the anodes. The TRT is filled with a 70% xenon, 27% carbon dioxide, 3% oxygen mixture [26][1]. As a charged particle passes through a TRT straw, it ionizes the gas mixture, producing an electrical signal due to the high voltage. The position of the hit can be determined accurately by measuring the drift time of the electrons, which gives a spatial resolution of 130 μm [26].

The straws are also interleaved with polypropylene fibres (in the barrel) or foils (in the end-caps) [26]. As particles pass between polypropylene and the gas mixture, they produce transition radiation, which can also be detected by the wires. There are two signal thresholds, a high threshold for transition radiation hits, and a low threshold for track hits. Electrons can be identified because they produce much

1. In practice, parts of the TRT leak. For cost reasons, these parts of the TRT were filled with an argon dominated gas mixture in Run 2 [36].

more transition radiation, and hence a much larger number of high-threshold hits, than pions [36].

2.2 Calorimeter

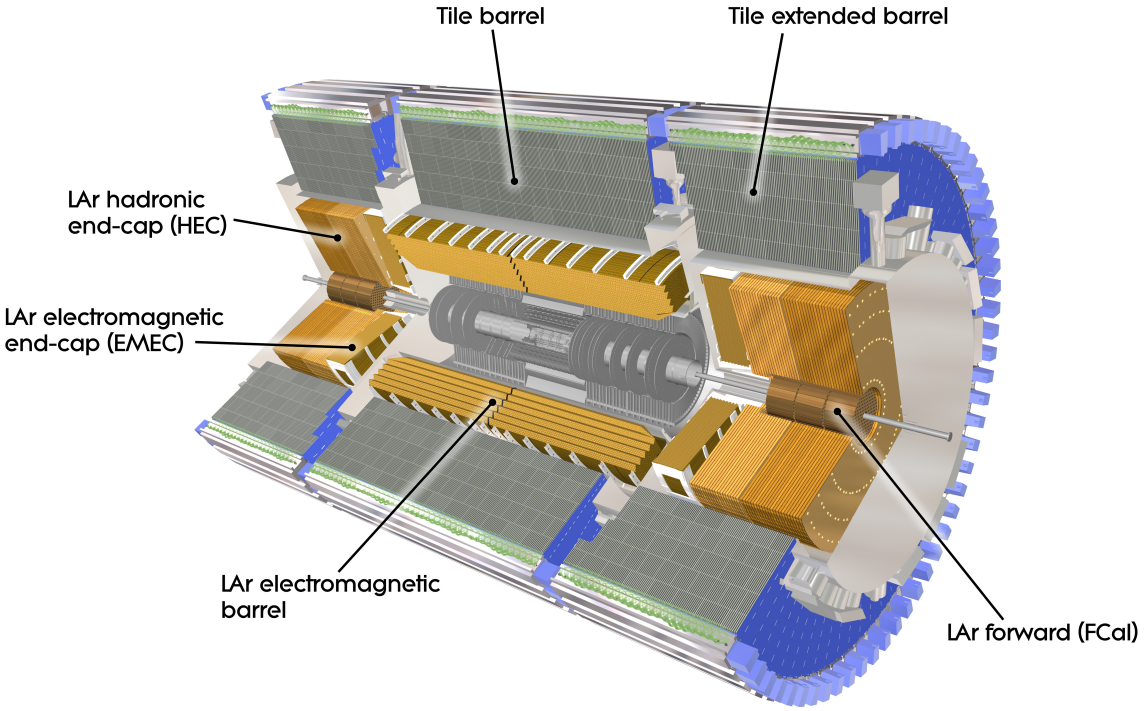


Figure 2.6: Schematic of the ATLAS calorimeter [37]. The LAr electromagnetic barrel and end-cap make up the Electromagnetic Calorimeter, while the other parts make up the Hadronic Calorimeter.

The ATLAS calorimeter, illustrated in Figure 2.6, measures the energy of photons, electrons, and charged and neutral hadrons. It is the most important detector for the measurement of jets, the central objects in the $HH \to b\bar{b}b\bar{b}$ analysis.

The calorimeter can be separated into the Electromagnetic Calorimeter (LAr Electromagnetic Barrel and End-Cap in Figure 2.6) which stops and captures the energy of electrons and photons, and the Hadronic Calorimeter (LAr Hadronic End-Cap, Tile Barrel and Tile Extended Barrel in Figure 2.6) which stops and captures the

remaining energy of hadrons which are not stopped by the electromagnetic calorimeter. The Forward Calorimeter (FCal, LAr Forward in Figure 2.6) provides electromagnetic and hadronic calorimetry in the forward region.

An electron (positron) or photon passing through the electromagnetic calorimeter will deposit its energy through an electromagnetic shower, consisting of alternating emission of photons through bremsstrahlung by electrons or positrons, and pair-production of electron-positron pairs by photons, as illustrated in Figure 2.7. When the particles have low enough energy, they rapidly lose that energy by ionizing the active material filling the calorimeter. This allows the deposited energy to be measured.

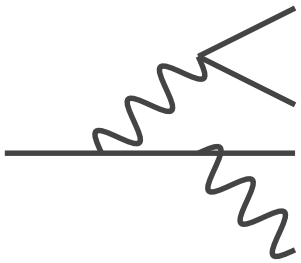


Figure 2.7: An EM shower.

The ATLAS electromagnetic calorimeter is a sampling calorimeter (that is, the calorimeter contains passive as well as active material, so only a fraction of the total energy deposited is measured) using liquid argon as the active material, and lead as the passive material. The layers of passive material are folded into an "accordion" configuration to create multiple effective layers of active material (see [26] Section 5.2.1 for a full description). The liquid argon active material produces an electrical signal proportional to deposited energy, which is read out by electrodes between the passive layers.

The electromagnetic calorimeter covers the full ϕ range, and $|\eta| < 3.2$ with higher precision at $|\eta| < 2.5$.

While hadrons also deposit energy through electromagnetic interactions, they also deposit energy through nuclear interactions. Since they are much more massive

they are not stopped by the electromagnetic calorimeter, so additional depth is required in the calorimeter which is provided by the hadronic calorimeter.

In the central region ($|\eta| < 1.7$), hadronic calorimetry is provided by the tile calorimeter, which consists of passive steel layers, and plastic scintillator tiles as the active material. The scintillation light is read out by wavelength-shifting optical fibres on the tile edges, connected to photomultiplier tubes. In the hadronic end-caps ($1.5 < |\eta| < 3.2$), the active material is once again liquid argon, and the passive material is copper. Finally, at very large $|\eta|$ ($3.1 < |\eta| < 4.9$), the FCal consists of active liquid argon and passive copper-tungsten layers.

Both the liquid argon calorimeters and the tile calorimeter require calibration. This is performed through a combination of simulations, test-beam measurements prior to installation, and using a charge-injection and (for the tile calorimeter) laser system after installation. The calibration system is fully described in the ATLAS detector paper [38].

2.2.1 *Resolution and Clustering*

All the ATLAS calorimeters are divided into calorimeter cells, whose size varies depending on the depth (r) and $|\eta|$ region. For instance, in the central barrel region, the electromagnetic calorimeter uses 0.0031×0.098 ($\eta \times \phi$) cells in the first layer and 0.025×0.0245 cells in the second and third layer.

These cells are combined into *towers* to form the input to the hardware (level 1) trigger. In this central region, these towers are 0.1×0.1. Each tower consists of the total energy deposited in that region in the full depth of the calorimeter.

The software or high-level trigger (HLT) and the offline analysis, on the other hand, make use of topological clusters (topo-clusters or just clusters) [39]. Here, calorimeter cells are combined into three dimensional topologically connected clusters which would correspond to a single deposit. They are also corrected to remove noise, and calibrated to account for uncaptured or mismeasured energy.

The ATLAS calorimeter provides sufficient resolution (both in direction, and energy) to allow these topo-clusters to form the basic jet input for the calorimeter jets used in Run 1 and early Run 2 ATLAS analyses. This analysis, however, uses *particle-flow* jets formed from particle-flow objects which combine tracks and calorimeter clusters to exploit the improved directional and transverse momentum resolution tracks can provide for charged particles. This particle-flow technique is described further in Chapter 3.

2.3 Muon Spectrometer

Muons are in essence heavier electrons, with a mass of 106 MeV. Compared to most particles, they are relatively long-lived with a lifetime of 2.2 μs in their rest frame. As a result, with sufficient energy they can pass through both the electromagnetic and hadronic calorimeters while retaining most of their energy. In order to measure the trajectories of these muons, ATLAS includes a muon spectrometer system as its outermost component. The muon spectrometers are placed at 5 m, 7.5 m and 10 m from the beam in a barrel, and in wheels 7.4 m, 10.8 m, 14 m and 21.5 m from the interaction point in the $\pm z$ directions. Precision measurement is performed by drift tube chambers (similar to the TRT) called Monitored Drift Tubes (MDTs), up to $|\eta| = 2.7$. In the forward region ($2.0 < |\eta| < 2.7$) these are augmented with multi-wire proportional chambers (Cathode Strip Chambers or CSCs), which

better time resolution and can handle higher rates [38]. In addition, Resistive Plate Chambers (RPCs) and Thin Gap Chambers (TGCs) are used to provide fast track information for triggering. These muon triggers are not used by this $HH \to b\bar{b}b\bar{b}$ analysis. The muon spectrometer is illustrated in Figure 2.8.

The magnetic field for the muon spectrometer is provided by toroidal magnets. The muon spectrometers alone can provide a reasonable measurement of muon p_{T} (roughly 10% uncertainty at 1 TeV). At lower p_{T} improved precision is available by combining the muon spectrometer track with an inner detector track. The muon spectrometer covers the range $|\eta| < 2.7$, though only the $|\eta| < 2.4$ region is used for triggering. The muon spectrometer is illustrated in Figure 2.8.

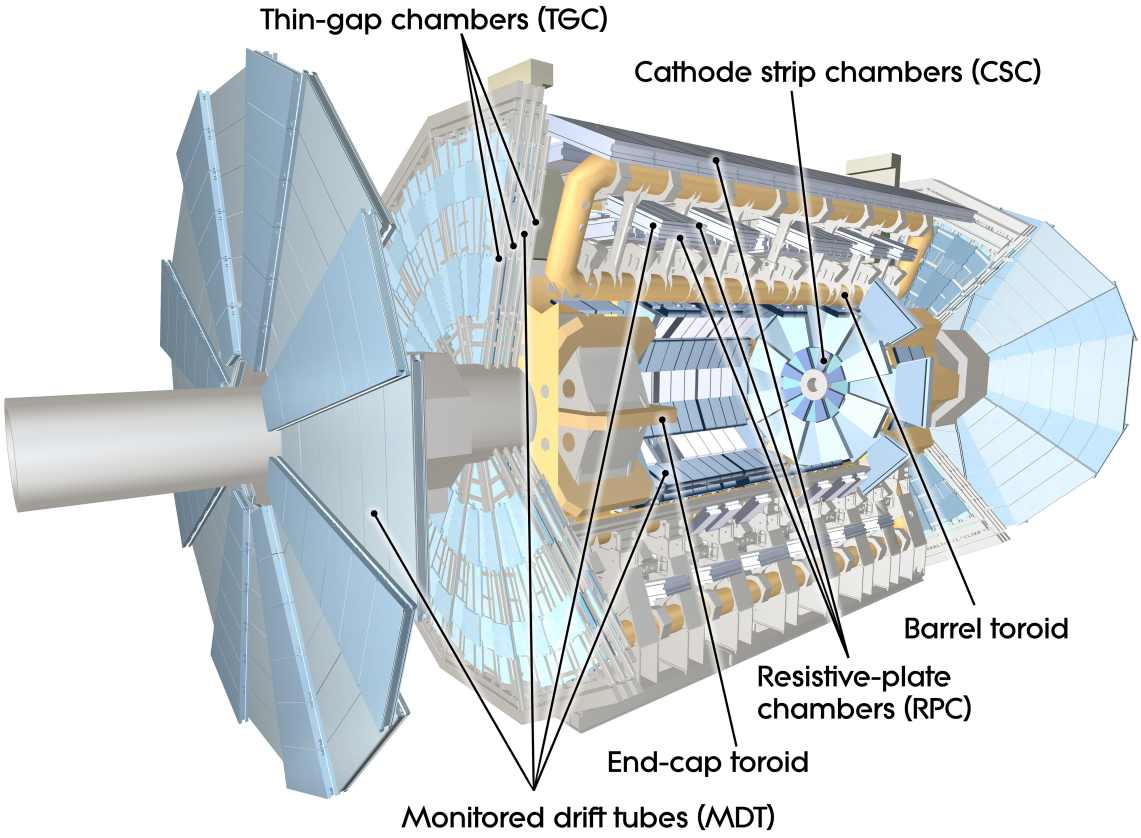


Figure 2.8: The ATLAS muon spectrometer [40].

3 Jets, Muons, and *b*-Tagging

The resolved $HH \to b\bar{b}b\bar{b}$ analysis searches for events with a final state containing four well-separated b-quarks. The list of reconstructed objects required by this analysis is therefore short:

Jets Quarks shower and hadronize into jets of hadrons. Reconstructed jets can be *b-tagged* to indicate whether they originated from a b-quark.

Muons B hadrons can decay semi-leptonically, with part of their energy being carried by a muon.[1] Reconstructed muons are required to correct jets for the energy carried away by such muons.

3.1 Jets

QCD does not allow for the existence of free coloured states. As a result, any quarks in the final state of an interaction undergo showering and hadronization to form sprays of charged and neutral hadrons. These sprays are called jets, and are reconstructed by clustering jet input objects such as tracks, calorimeter deposits, or, as in this analysis, particle flow objects. When interactions are simulated, these input objects may be final state particles from the simulation, in which case the jets can be referred to as truth jets.

3.1.1 *Particle Flow*

Neutral hadrons leave only calorimeter deposits. Charged hadrons leave both calorimeter deposits and tracks. The particle flow [41] algorithm combines calori-

1. Where this lepton is an electron, the electron energy would be fully captured by the jet.

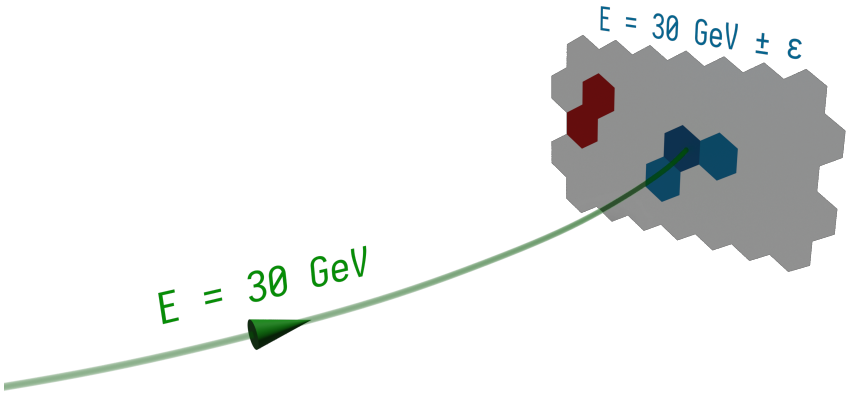


Figure 3.1: Illustration of the geometric matching procedure in the particle flow algorithm. The 30 GeV track is geometrically matched to the dark blue cluster. This cluster has less than 30 GeV so the two lighter blue clusters are included to make up the full energy. Additionally, the two red clusters have significant energy deposits but are not matched to any track. The final particle flow objects are the track (representing a charged hadron) and the two red clusters (representing deposits from neutral hadrons), since the remnant of the blue clusters is consistent with a fluctuation.

meter deposits with transverse momenta from tracks where these are available to improve the angular and energy resolution of jets.

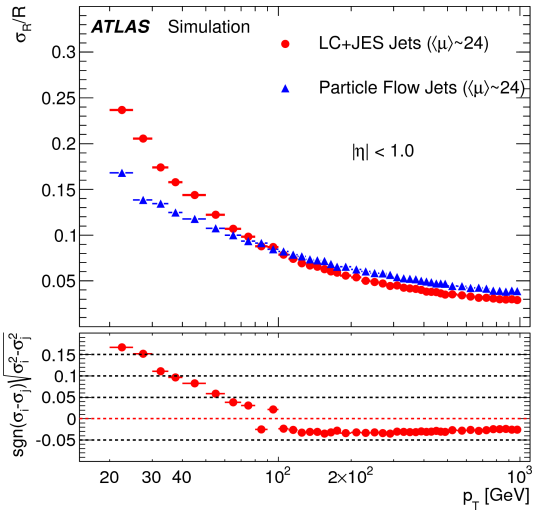
Calorimeter deposits enter the algorithm in the form of calorimeter clusters [39], which are topologically connected groups of cells with energy deposits around cells with energy deposits significantly higher than the noise floor.

Each high-quality[2] track is geometrically matched to a calorimeter cluster. If the energy the track would be expected to have deposited exceeds the energy actually deposited in the cluster, additional clusters are included until this is no longer the case. This procedure is illustrated in Figure 3.1.

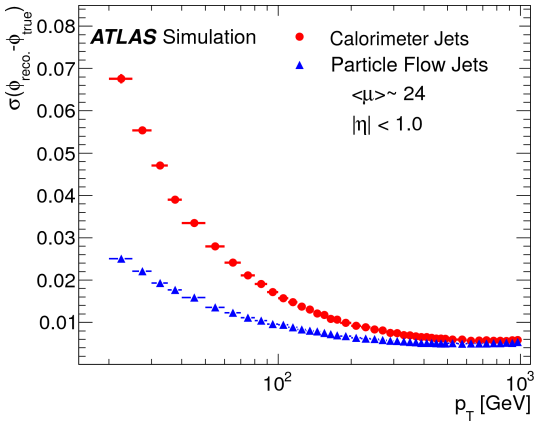
Calorimeter cells are then subtracted from the clusters to account for the energy the track would be expected to have deposited (to avoid double counting). The tracks, remnant calorimeter clusters (where they have sufficient energy not to be consistent with a fluctuation), and any clusters that were not matched to any track, are now considered *particle flow objects*. In Figure 3.1, once the track energy is subtracted from the group of blue clusters, what remains is consistent with a fluctuation so these remnant clusters are removed.

The improvement in p_{T} and ϕ resolution when particle flow jets are used is illustrated in Figure 3.2. Here it can be observed that ϕ resolution is improved throughout the p_{T} range, and p_{T} resolution is improved at low p_{T}. At higher p_{T}, tracks are straighter, which reduces p_{T} resolution, while the relative uncertainty on calorimeter energy resolution is smaller, since this scales as $E^{-\frac{1}{2}}$.

2. at least nine pixel / SCT hits, and no missing pixel hits [41]



(**a**) p_{T} Resolution



(**b**) ϕ Resolution

Figure 3.2: p_{T} and ϕ resolution improvement from the use of particle flow jets. Figures from [41].

3.1.2 *Anti-k_t Jet Clustering Algorithm*

These particle flow objects are clustered into jets using the anti-k_t algorithm [24], with the R parameter set to 0.4.

The core of the anti-k_t algorithm is the definition of the two "distances" given in Eqs. 3.1 and 3.2: a jet-jet distance d_{ij} and a jet-beam distance d_{iB}. ΔR_{ij} is the angular separation between object i and object j: $\sqrt{\left(\eta_i - \eta_j\right)^2 + \Delta\phi\,(i,j)^2}$. The name anti-$k_t$ originates from the use of the symbol k_t instead of p_T to designate the transverse momentum of an object. The prefix *anti-* refers to the power of -2, as opposed to 2 in the k_t algorithm.

$$d_{ij} = \min(k_{t,i}^{-2}, k_{t,j}^{-2}) \frac{\Delta R_{ij}^2}{R^2} \tag{3.1}$$

$$d_{iB} = k_{t,i}^{-2} \tag{3.2}$$

At each step, if the smallest distance is a d_{iB}, object i is moved to the list of jets. If the smallest distance is a d_{ij}, objects i and j are combined. When there are no objects left (i.e. they have all been moved to the list of jets), the jet clustering is complete.

$\min(k_{t,i}^{-2}, k_{t,j}^{-2})$ will return k_t^{-2} of the *hardest* jet in the pair, so the anti-k_t algorithm starts with the hardest objects and clusters in neighbouring soft radiation.

If ΔR_{ij} exceeds R, $d_{ij} > k_{t,i}^{-2} = d_{iB}$ (where i is the harder object) so object i is a jet, and will not be further combined with any other object. As a result, the radius of a jet cannot exceed R.

3.1.3 *Jet Calibration*

Due to pile-up, inefficiencies in the calorimeter, and other effects, the energy of a reconstructed jet built from calorimeter clusters, or even particle flow objects, will generally not be equal to the true energy of the jet (where the "true energy of the jet" is the energy of the corresponding truth jet built from final state particles in the case of Monte Carlo simulation, or the theoretical energy of the corresponding jet clustered from the final state particles if there were no detector for them to interact with, in the case of real data).

In order to correct for this, a multi-step calibration process is applied to the jets. This is fully detailed in [42], and summarized here. These corrections generally correct jet p_{T}, and jet energy is corrected by applying the same ratio $\frac{\text{corrected } p_{\mathrm{T}}}{\text{uncorrected } p_{\mathrm{T}}}$, leaving the directional components (η and ϕ) unaffected.

Jet Area Pile-up Correction The first correction accounts for energy added to jets due to pile-up (i.e. energy from other pp interactions in the same bunch crossing, or

residual energy in the detector from previous bunch crossings). A is defined to be the jet area, calculated using the ghost particle procedure [17]. A pile-up p_{T} density, ρ, is calculated by calculating the median p_{T}/A over jets with $\left|\eta\right| < 2.0$ reconstructed using the k_t algorithm, which is more sensitive to soft (low energy) activity. The k_t algorithm is similar to the anti-k_t algorithm described in Section 3.1.2, with k_t^{-2} replaced with k_t^2. Each anti-k_t jet in the event then has its pile-up component ρA subtracted, using the area of that jet.

Residual Pile-up Correction After the jet area dependent correction above, there is still some pile-up dependence of jet energies. These are corrected by applying a residual p_{T} correction, subtracting $\alpha(N_{\mathrm{PV}} - 1) + \beta\mu$, where N_{PV} is the number of primary vertices in the event, and μ is the number of pp interactions. The coefficients α and β are determined using MC simulated events.

Absolute MC-based Calibration After correcting for pile-up, a calibration called the "Absolute MC-based Calibration" is applied to correct for detector effects. This calibration is derived using an MC simulation of dijet events, which corrects reconstructed η and jet energy to true η and jet energy. When deriving this calibration, reconstructed jets are matched to truth jets geometrically, requiring $\Delta R < 0.3$.

Global Sequential Calibration Further discrepancies due to detector effects are corrected for through the Global Sequential Calibration. Six multiplicative corrections are applied to the jet energy and p_{T}. In addition to being a function of jet p_{T} and η, the six corrections are a function of each of six variables characterizing the interaction of the jet with the detector.

In Situ Calibration Finally, an *in situ* calibration is applied to data only, in order to correct for differences in detector response in data compared to detector response in MC simulation. In order to determine this calibration, particle-level jet transverse momenta are determined in data by balancing jets against well-calibrated objects or groups of objects. Jet response is here defined as $\langle p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{obj}} \rangle$. The *in situ* calibration includes the η intercalibration which corrects the response of forward ($|\eta| > 0.8$) jets to match that of central jets, Z+jet and γ+jet analyses where the hadrons in a Z/γ+jet event are balanced against a Z boson or photon, and the multi-jet balance analysis, where a high p_{T} jet is balanced against a number of low p_{T} (and therefore well-calibrated) jets.

3.2 Muons, and the Muon-in-Jet Correction

B mesons can decay semi-leptonically producing an electron or muon. While electrons produced in this manner are captured in the jet, since they deposit their energy in the calorimeter, muons will generally pass through the calorimeter, depositing only a small portion of their energy. It is this property of muons that allows the use of the muon spectrometers to detect virtually exclusively muons.

However, since muons are not stopped by the calorimeter, they can carry away a significant portion of the b-quark energy. In order to correct for this, we can apply a muon-in-jet correction.

First, muons are selected which

- have a $p_T > 4\,\mathrm{GeV}$,
- fall within the tracker acceptance ($|\eta| < 2.5$),
- and pass the `Medium` quality working point [43].

These muons are matched to a b-jet if they fall within its jet cone (i.e. within $\Delta R = 0.4$ of the jet axis). In this circumstance, the muon four-vector is added to the jet four-vector as a correction (with the energy deposited by the muon in the calorimeter subtracted to avoid double-counting.)

3.3 b-Tagging

In the $HH \to b\bar{b}b\bar{b}$ analysis, there are 4 b-quarks in the final state. We therefore make use of flavour tagging (or b-tagging) to determine which of the jets in our events originated from a b-quark.

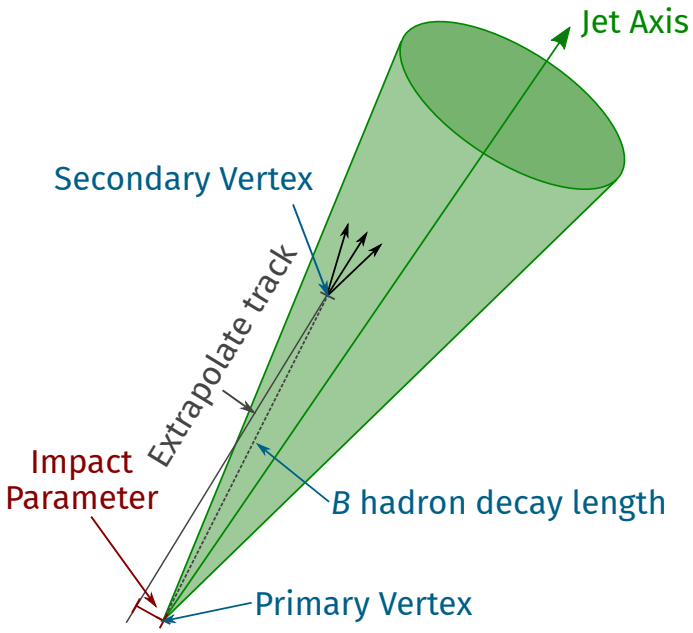


Figure 3.3: Diagram showing the impact parameter of one of a B hadron's decay products and the secondary vertex where the B hadron decays.

The mean lifetime of a B hadron (multiplied by the speed of light, c, to give length units), $c\tau$ is 0.45 mm to 0.5 mm. Its decay length is $\beta\gamma c\tau = {}^{pc\tau}/_{m}$. Even a B hadron with $p_{\mathrm{T}} = 40\,\mathrm{GeV}$ therefore has a decay length (in the transverse direction) of

around 4 mm. This is large enough to be discernible from studying the tracks of its decay products.

Figure 3.3 illustrates how a secondary vertex corresponding to the B hadron decay position can be found, and how extrapolating the tracks of the decay products allows impact parameters to be found (with respect to the primary vertex). Since the B hadron decay products were not produced at the primary vertex, these impact parameters are non-zero.

The flavour tagging procedure in ATLAS is founded on three basic low-level algorithms [44]:

- `IP2D/3D` using the 2D / 3D impact parameters
- `SV1` using the secondary vertex
- `JetFitter` fitting all displaced vertices

In addition, a fourth low-level algorithm, `RNNIP` [45], uses a recurrent neural network to identify b-jets using the tracks in the jet directly (with no prior vertex finding step).

These low level taggers produce a number of outputs which are used as inputs to the high-level taggers. The `MV2` [44] tagger is a boosted decision tree which uses inputs from `IP2D/3D` and `SV1`. This was the baseline tagger used in the Early Run 2 analysis. For this analysis, the `DL1r` tagger, a deep neural network making use of inputs from all four low-level algorithms listed above, is used.

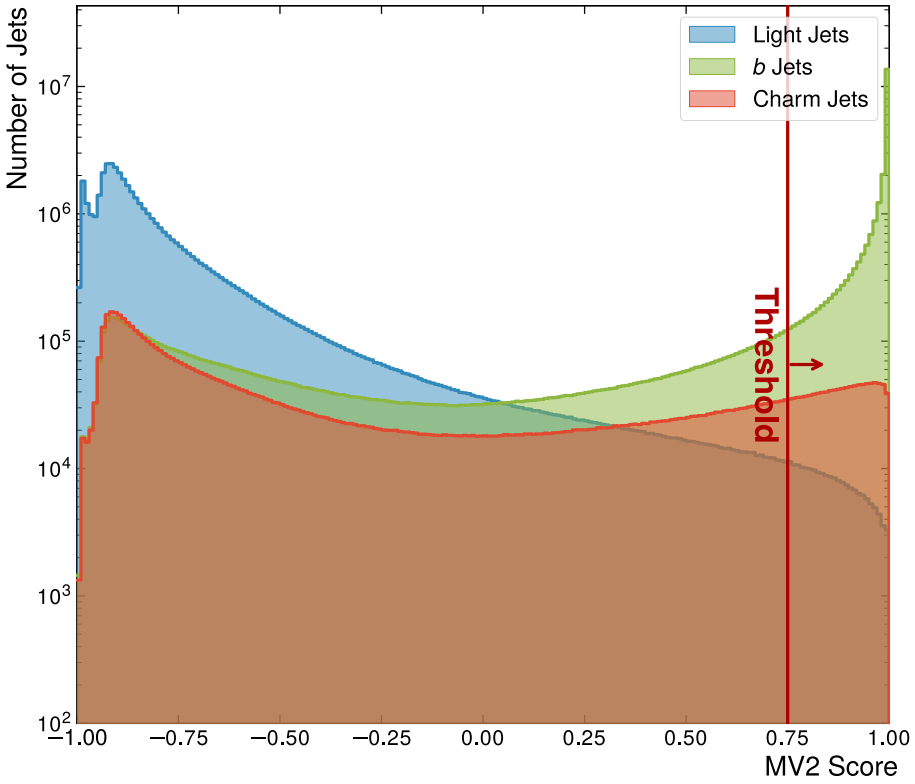


Figure 3.4: MV2 scores for a $t\bar{t}$ sample. A threshold is applied, and jets with a score above this threshold are considered tagged, while jets with a score below the threshold are untagged.

3.3.1 *Working Points*

Figure 3.4 shows how a threshold can be applied on the MV2 score distribution for a simulated $t\bar{t}$ sample to b-tag the jets in this sample. The high-level flavour tagging algorithms are used by determining the minimum score such that some predetermined percentage (the *efficiency*) of b-jets are selected. In practice, this efficiency is evaluated on a hybrid Monte Carlo sample consisting of $t\bar{t}$ and Z′ events at a range of masses (providing a flat p_{T} spectrum at high p_{T}). b-jets are then selected in the analysis if the b-tagging score is greater than that set by the working point.

These minimum scores are determined to give a certain efficiency across the hybrid Monte Carlo sample as a whole, but in fact the true efficiency depends on jet p_{T} and pseudo-rapidity, and differs between data and Monte Carlo. In order to correct for this, scale factors are applied to events [44]. Figure 3.5 shows these p_{T} and $|\eta|$ dependent efficiencies, and the scale factors applied to correct for differences between data and Monte Carlo, for the 77% (nominal) efficiency working point of the DL1r tagger. The best efficiencies for a given DL1r score threshold are achieved on central jets with a p_{T} between 40 GeV and 200 GeV.

The scale factors are determined by selecting dileptonic $t\bar{t}$ events in data, to obtain a high purity sample of b-jets. The b-tagging efficiency is then measured in both, and the ratio (for a given p_{T} bin) is the scale factor. These scale factors are generally close to one, but can compound when there are multiple jets in an event, as in the $HH \to b\bar{b}b\bar{b}$ analysis. The smooth curve in Figure 3.5 is the result of smoothing these scale factors. The uncertainties on this measurement are propagated through to become the flavour tagging uncertainties discussed in Section 6.2.1.

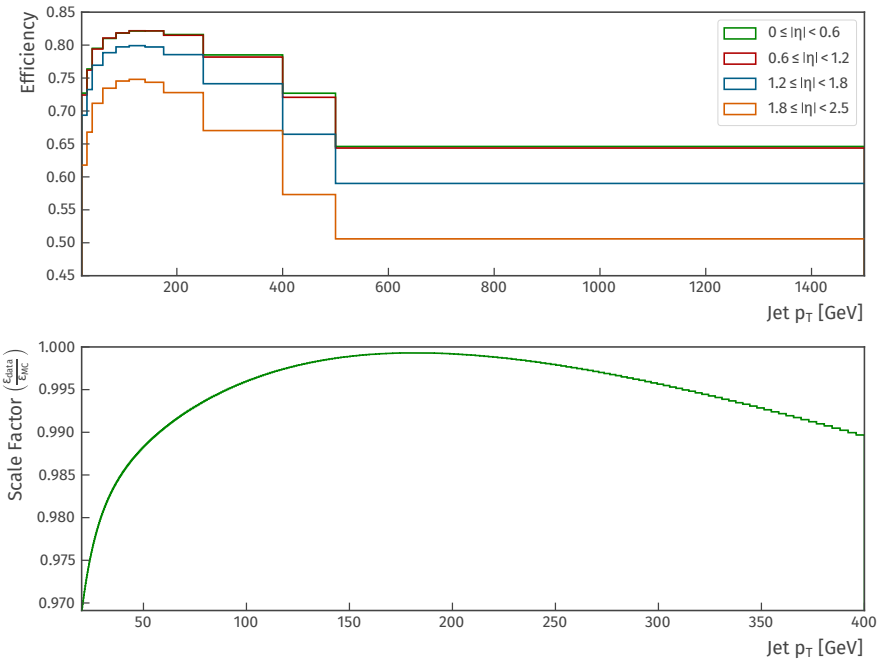


Figure 3.5: Flavour tagging efficiencies (in MC, as a function of jet p_T and $|\eta|$), and scale factors (as a function of jet p_T) for the DL1r 77% nominal efficiency working point.

The choice of working point is a trade-off between the b-jet selection efficiency and the rejection factor for light, charm, and τ lepton jets. These rejection factors are $\frac{\text{total jets of type } t}{b\text{-tagged jets of type } t}$, and the rejection factors for the DL1 tagger are shown in Table 3.1. This tagger is similar to the DL1r tagger used in this analysis.

In practice, the working point used in this analysis was selected by evaluating the expected limits produced with the various options, selecting the working point that

Working Point	Light Rejection	Charm Rejection	τ Rejection
60%	1300	27	220
70%	390	9.4	43
77%	130	4.9	14
85%	29	2.6	3.9

Table 3.1: Rejection factors for the DL1 tagger at each working point [44].

	Limit / fb @		
	300 GeV	500 GeV	800 GeV
MV2 @ $\varepsilon = 70\%$ WP	2167	97.6	19.9
DL1r @ $\varepsilon = 70\%$ WP	2050	99.3	16.2
DL1r @ $\varepsilon = 77\%$ WP	1921	96.2	16.0

Table 3.2: 95% CI upper limits obtained for scalar signals at 300 GeV, 500 GeV and 800 GeV, evaluated using 2016 data only. DL1r at the 77% efficiency working point consistently produces the best limit.

produced the tightest limits. The Early Run 2 analysis used the MV2 tagger at the 70% efficiency working point. For this analysis, I evaluated this combination, as well as the DL1r tagger at the 70% and 77% efficiency working points. This evaluation was done using only the $24.3\,\text{fb}^{-1}$ of data taken in 2016 and the corresponding simulated signals. The results of this are shown in Table 3.2, which shows that the 77% efficiency working point of the DL1r tagger consistently produced the tightest limits. As a result, this was the tagger and working point used in this analysis.

3.3.2 *b-jet Energy Correction*

In addition to the standard ATLAS jet calibration, and the muon-in-jet correction described in Section 3.2, an additional correction is applied to b-tagged jets to account for other differences in the reconstruction of b-jet energies as opposed to light jet energies. In particular, this accounts for energy loss due to out-of-cone radiation, and due to neutrinos in semi-leptonic b decays. The correction is derived by comparing the p_{T} of reconstructed jets to the geometrically matched ($\Delta R < 0.3$) truth jet in a $t\dot{t}$ simulation. This correction is derived separately for jets containing a muon and jets not containing a muon.

3.4 b-Tagging on Particle Flow Jets

The Early Run 2 analysis made use of jets clustered from calorimeter clusters (*calorimeter jets*), with no tracking information used. One of the key reasons for this was that the ATLAS b-tagging framework had not been configured to run b-tagging on particle flow jets.

In order to rectify this, I configured the ATLAS b-tagging framework to run b-tagging on particle flow jets, and compared the performance of `MV2` (the then-standard b-tagging algorithm) on particle flow jets to its performance on calorimeter jets. This work constituted my ATLAS authorship qualification task. Whilst my work here was not directly used in the $HH \to b\bar{b}b\bar{b}$ analysis (which uses the newer `DL1r` b-tagging algorithm), it helped illustrate that particle flow jets could be useful in a b-jet centric analysis like $HH \to b\bar{b}b\bar{b}$.

3.4.1 *Monte Carlo Sample and Preselection*

I conducted my evaluation of b-tagging performance on a simulated sample of approximately 60 million semi-leptonic and fully leptonic $t\bar{t}$ decays.

After applying the standard ATLAS jet calibrations, any jet within $\Delta R = 0.3$ of an electron from a $W^{\pm}$ or Z^0 decay (determined by examining the MC truth record) was removed.

A number of further selections were then applied to the jets to reject pile-up:

- $p_{\mathrm{T}} > 20\,\mathrm{GeV}$
- $\left|\eta\right| < 2.5$

- Where $p_{\mathrm{T}} < 60\,\mathrm{GeV}$ and $|\eta| < 2.4$,

$$\mathrm{JVT} > \begin{cases} 0.2 & \text{for particle flow jets} \\ 0.59 & \text{for calorimeter jets} \end{cases}$$

Here, JVT (Jet Vertex Tagger) is a composite variable designed to reject pile-up [46]. In essence, this variable compares the scalar sum of the transverse momenta of tracks associated to the jet and originating from the hard-scatter vertex (defined by calculating the scalar sum of the transverse momenta of all tracks originating from each vertex, and selecting the vertex where this is highest) to a) the p_{T} of the jet, and b) the scalar sum of the transverse momenta of all tracks associated to the jet. After these selections were applied, the sample contained 261 million calorimeter jets and 250 million particle flow jets.

Jets are labelled with their true flavours (truth labelled) [47] by matching each jet to the closest b or c containing hadron, or τ lepton, requiring that this hadron / lepton be within $\Delta R = 0.3$ of the jet axis. If more than one such hadron / lepton is found, priority is given to b, then c, then τ. If no such hadron / lepton is found, the jet is labelled as light flavour.

3.4.2 *Performance Evaluation*

In order to evaluate the performance of the `MV2` algorithm on the two types of jet, I separated the jets into p_{T} bins of width 10 GeV. In each bin, the appropriate working point[3] was applied. These working points were read from the CDI (calibration data interface) file provided by the ATLAS b-tagging group, and were derived by determining the threshold that would give the desired efficiency (70% or 77%) on the hybrid ($t\bar{t} + Z'$) MC sample.

3. See Section 3.3.1.

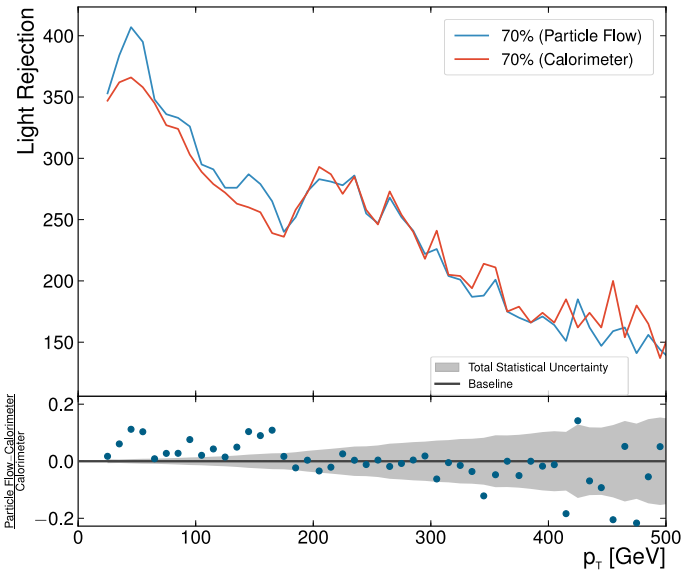
With N_i being the total number of jets labelled as flavour i (b, c, τ, or l if none of the above), and P_i the number of such jets selected upon applying the b-tagging working point, the light jet rejection and purity are calculated as given in Eqs. 3.3 and 3.4.

$$\text{light jet rejection} = {}^{N_l}/_{P_l} \tag{3.3}$$

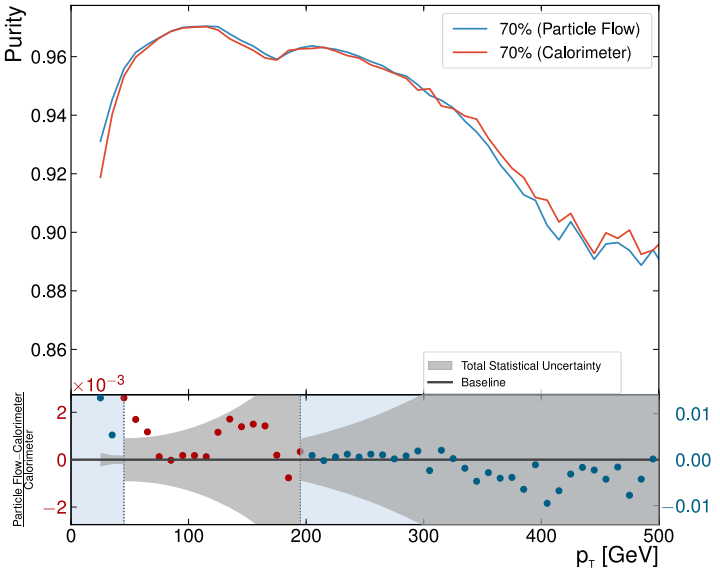
$$\text{purity} = \frac{P_b}{P_b + P_c + P_\tau + P_l} \tag{3.4}$$

Figure 3.6 shows the result of this performance comparison for the 70% efficiency working point of MV2. Despite MV2 having been trained on calorimeter jets, it produces a similar purity, and a slightly improved light jet rejection for lower p_T jets (particularly below 100 GeV), when running on particle flow jets. This reflects the improved resolution provided by particle flow jets at low p_T.

Figure 3.7 shows the result of the same performance comparison for the 77% efficiency working point of MV2.

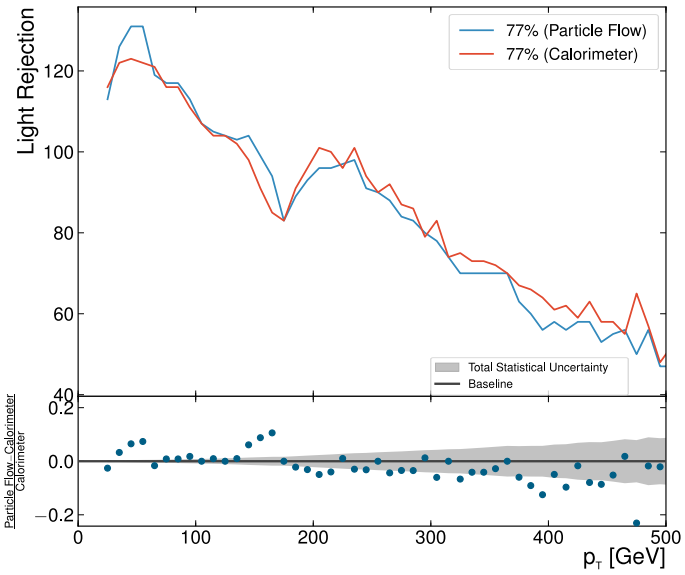


(**a**) Light rejection

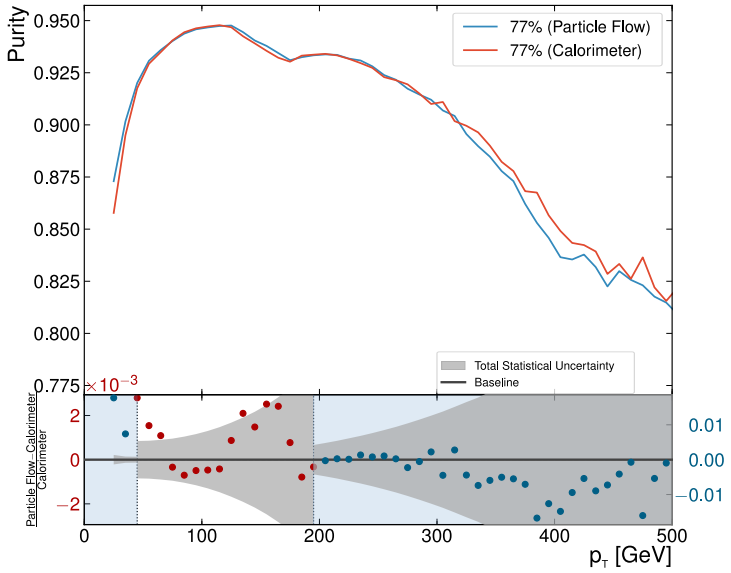


(**b**) Purity

Figure 3.6: Light rejection and purity for the MV2 70% efficiency working point. In plot (b) the red points (unshaded section) should be read against the red scale, while the blue points (shaded section) should be read against the blue scale.



(**a**) Light rejection



(**b**) Purity

Figure 3.7: Light rejection and purity for the MV2 77% efficiency working point. In plot (b) the red points (unshaded section) should be read against the red scale, while the blue points (shaded section) should be read against the blue scale.

4 The $HH \to b\bar{b}b\bar{b}$ Analysis

4.1 Trigger

Within the ATLAS detector there are 40 million bunch crossings every second, with each bunch crossing producing an average of 33.7 proton-proton collisions [48] over the course of Run 2. This value is called *pile-up* and is represented by the Greek letter μ, and Figure 4.1 shows the total integrated luminosity recorded against μ.

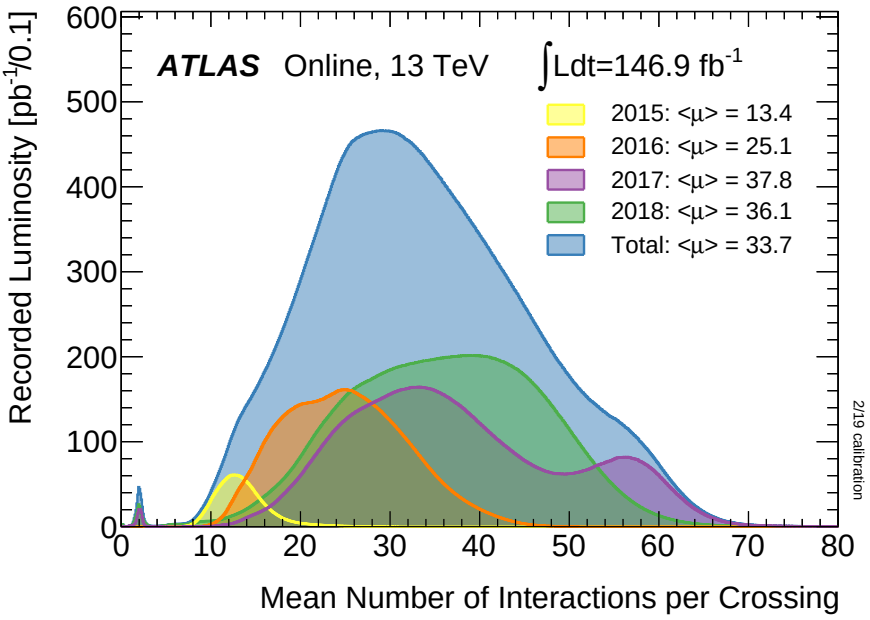


Figure 4.1: Total integrated luminosity recorded at various levels of pileup. Figure from [48].

It is infeasible even to read out the detector at a rate of 40 MHz, much less write all of this data to disk. In order to reduce the output rate, ATLAS employs a two-stage online trigger system [49].

The first stage is the Level 1 (L1) trigger, which is implemented in hardware, using custom-designed ASICs (Application Specific Integrated Circuits) and FPGAs (Field Programmable Gate Arrays) [50]. This trigger operates on a readout of the calorimeter and muon spectrometer, and applies simple conditions to identify potentially interesting events to be considered by the next stage. This reduces the event rate from 40 MHz to 100 kHz [50].

The $HH \to b\bar{b}b\bar{b}$ analysis uses jet and H_T (the scalar sum of jet transverse momenta) L1 triggers. In order to keep up with the 40 MHz beam crossing rate a pipeline is employed, allowing a maximum latency of about 2.5 μs [51]. Nevertheless, this is an extremely high event rate, and as a result the L1 jet and H_T triggers (together, the `L1Calo` trigger) operate on a coarse calorimeter readout (with *elements* of size 0.2×0.2 in $\eta \times \phi$). The combination of these into jets is also performed in a fast, approximate manner, using a sliding window algorithm in which a jet is a 4×4 block of elements, with the 2×2 element square at the centre being a local energy maximum [52]. Triggers then require n L1 jets meeting specified p_{T} and $|\eta|$ thresholds, and/or H_T exceeding a specified value.

The second stage is the High-Level Trigger (HLT), which is implemented in software, running on the HLT farm which is comprised of commodity x86_64 machines physically located close to the ATLAS detector (i.e. at LHC Point 1). This stage operates on a full readout of the detector, though in order to process events sufficiently quickly, events are usually only reconstructed in specific regions identified by the L1 trigger. The L1 trigger that initiates a specific HLT trigger is called the *seed*.

1	At least 4 jets with $p_T > 15\,\text{GeV}$, $\left\|\eta\right\| < 2.5$
2	At least 1 jet with $p_T > 85\,\text{GeV}$ and 2 with $p_T > 30\,\text{GeV}$
3	$H_T > 190\,\text{GeV}$, at least 1 jet with $p_T > 15\,\text{GeV}$, $\left\|\eta\right\| < 2.1$
4	At least 1 jet with $p_\text{T} > 100\,\text{GeV}$
5	At least 1 jet with $p_\text{T} > 75\,\text{GeV}$ and 2 with $p_\text{T} > 20\,\text{GeV}$

Table 4.1: Level 1 seeds for triggers in Table 4.2. The thresholds given are for L1 jets, not fully reconstructed (offline) jets.

Year	L1	Trigger	Category	Event must contain at least
2016	4	HLT_j225_bmv2c2060_split	1b	One b-jet (60% eff. WP) with $p_\text{T} > 225\,\text{GeV}$
	5	HLT_j100_2j55_bmv2c2060_split	2b1j	Two b-jets (60% eff. WP) with $p_T > 55\,\text{GeV}$ *and* one extra jet with $p_T > 100\,\text{GeV}$
	1	HLT_2j35_bmv2c2060_split_2j35_L14J15.0ETA25	2b2j	Two b-jets (60% eff. WP) with $p_T > 35\,\text{GeV}$ *and* two extra jets with $p_T > 35\,\text{GeV}$
2017	2	HLT_j110_gsc150_boffperf_split_2j35_gsc55_bmv2c1070_split_L1J85_3J30	2b1j	Two b-jets (70% eff. WP) with $p_T > 55\,\text{GeV}$ *and* one extra jet with $p_T > 150\,\text{GeV}$
	1	HLT_2j15_gsc35_bmv2c1040_split_2j15_gsc35_boffperf_split_L14J15.0ETA25	2b2j	Two b-jets (40% eff. WP) with $p_T > 35\,\text{GeV}$ *and* two extra jets with $p_T > 35\,\text{GeV}$
	3	HLT_2j35_gsc55_bmv2c1050_split_ht300_L1HT190-J15s5.ETA21	2bHT	Two b-jets (50% eff. WP) with $p_T > 55\,\text{GeV}$ *and* $H_T > 300\,\text{GeV}$
	4	HLT_j225_gsc300_bmv2c1070_split	1b	One b-jet (70% eff. WP) with $p_T > 400\,\text{GeV}$
2018	2	HLT_j110_gsc150_boffperf_split_2j45_gsc55_bmv2c1070_split_L1J85_3J30	2b1j	Two b-jets (70% eff. WP) with $p_T > 55\,\text{GeV}$ *and* one extra jet with $p_T > 150\,\text{GeV}$
	1	HLT_2j35_bmv2c1060_split_2j35_L14J15.0ETA25	2b2j	Two b-jets (60% eff. WP) with $p_T > 35\,\text{GeV}$ *and* two extra jets with $p_T > 35\,\text{GeV}$
	4	HLT_j225_gsc300_bmv2c1070_split	1b	One b-jet (70% eff. WP) with $p_T > 400\,\text{GeV}$
	3	HLT_2j45_gsc55_bmv2c1050_split_ht300_L1HT190-J15s5.ETA21	2bHT	Two b-jets (50% eff. WP) with $p_T > 55\,\text{GeV}$ *and* $H_T > 300\,\text{GeV}$

Table 4.2: Complete triggers used in the $HH \to b\bar{b}b\bar{b}$ analysis. The numbers in the second column are a cross-reference to Table 4.1.

At this stage, the trigger has access to the information from the trackers, and is therefore able to run b-tagging. Table 4.1 shows the L1 triggers, and Table 4.2 the HLT triggers used by the $HH \to b\bar{b}b\bar{b}$ analysis for each year of data taking. The categories here relate to the trigger exclusivity strategy described in Section 4.1.1. In Table 4.2 the HLT triggers are cross-referenced to their L1 seeds by the number preceding the trigger name.

4.1.1 *Trigger Exclusivity (Bucketing) Strategy*

In order to simplify the application of trigger scale factors[1] (which correct for differences in trigger efficiency between data and MC simulation) a trigger exclusivity strategy is applied. Under this strategy, an event is only considered if it passes the specific trigger assigned to it, based on fully reconstructed quantities only. It must also have fully reconstructed jets that can be geometrically matched to each HLT jet used in the decision. As a result of tuning the assignments for this analysis, this exclusivity strategy also led to a very small *increase* in sensitivity.

Each event is assigned to a trigger as follows.

1. If the p_{T} of the leading (i.e. highest p_{T}) fully reconstructed jet is over 325 GeV and this jet is b-tagged, the event is assigned to the `1b` trigger.
2. Else, if the leading jet is *not* b-tagged and its p_{T} is over 169 GeV, the event is assigned to the `2b1j` trigger.
3. Else, if the event was recorded in 2017 or 2018 (or if simulated, part of a simulation corresponding to data taken in 2017 or 2018), and the H_T calculated from fully reconstructed jets with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $\left|\eta\right| < 2.5$ is greater than 900 GeV, the event is assigned to the `2bHT` trigger.
4. If none of these conditions are satisfied, the event is assigned to the `2b2j` trigger.

1. In the event, the implementation of these scale factors was not ready, so they were not applied. Since they are small (close to 1) and affect only the simulated signal this does not have a significant impact on the results of this analysis.

4.2 Dataset

The $HH \to b\bar{b}b\bar{b}$ analysis makes use of data collected by ATLAS in the years 2016 through 2018. This amounts to a total of $127\,\text{fb}^{-1}$ of data, broken down into

2016 $24.6\,\text{fb}^{-1}$. This differs from the $32.9\,\text{fb}^{-1}$ integrated luminosity quoted by most analyses due to a misconfiguration of the HLT leading to a loss of $8.3\,\text{fb}^{-1}$ of data in analyses (like the $HH \to b\bar{b}b\bar{b}$ analysis) which use b-jet triggers.

2017 $43.7\,\text{fb}^{-1}$

2018 $58.5\,\text{fb}^{-1}$

The $3.2\,\text{fb}^{-1}$ of data collected in 2015 was not used because the necessary calibrations will not be produced.[2]

In addition, Monte Carlo simulations of narrow-width (much smaller than detector resolution) scalar resonances decaying to $HH \to b\bar{b}b\bar{b}$ are used to provide the reference signals being sought. These are generated using MadGraph [11], with Herwig7 [53] used for the parton shower, and EvtGen [54] simulating B hadron decays. The leading order NNPDF 2.3 [55] with QED corrections and $\alpha_S = 0.130$ parton distribution function (PDF) set was used for the matrix element level calculation. The parton shower used the leading order MMHT 2014 [56] PDF set. The simulations are then passed through the fast ATLAS detector simulation, AtlFast II. These signals are simulated at masses between 251 GeV and 900 GeV. Narrow-width scalar resonances are also simulated at 1 TeV to 1.2 TeV. These

2. Calibrations in ATLAS are provided by Combined Performance groups based on dedicated analyses. In this case, the b-jet trigger group will not produce calibrations for 2015 data since the integrated luminosity is so small.

higher mass simulations were passed through the full Geant4 [57] ATLAS simulation. This was because these higher mass samples are also used by the boosted $HH \to b\bar{b}b\bar{b}$ analysis, which makes use of jet substructure and therefore benefits from the full simulation. This resolved $HH \to b\bar{b}b\bar{b}$ analysis does not make use of jet substructure, and the fast simulation is therefore acceptable.

Most of the Kaluza-Klein graviton mass spectrum covered by the resolved analysis was excluded by the Early Run 2 analysis [58]. As a result, the narrow width scalars were the focus of this analysis. Nevertheless, gravitons were also simulated and used in early studies. For these, the reference simulations were generated using MadGraph, showered with Pythia [59], then passed through EvtGen and the full ATLAS detector simulation.

4.3 Event Selection

4.3.1 *Basic Requirements*

From the events selected by the triggers described in Section 4.1, the analysis first selects those that can plausibly be $HH \to b\bar{b}b\bar{b}$ events.

The key requirement here is that the event have at least four $R = 0.4$ anti-k_t jets, with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $\left|\eta\right| < 2.5$, in order to ensure:

1. The triggers are reasonably efficient. This also requires the application of the trigger exclusivity strategy described in Section 4.1.1.
2. The jets are within the region covered by the ATLAS tracker. This is necessary since tracking is required for b-tagging.

Henceforth, any jets that do not pass the above requirements are discarded entirely.

4.3.2 *Tag Categories*

To be considered b-tagged, a jet must pass the 77% efficiency working point of the DL1r tagger, as described in Section 3.3.

Events are separated into two categories:

Four tag Events with at least four b-tagged jets. In plots, this will be labeled "Data".

Two tag Events with precisely two b-tagged jets. These events are used for background estimation, as described in Section 4.5. In plots, this will be labeled "Two Tag Data", or "Background" if the background estimation procedure has been applied.

Events with fewer than two b-tagged jets are discarded entirely. Events with precisely three b-tagged jets are not used in the search. Nevertheless, they are not used for background estimation, such that they may be preserved for inclusion in a future, further improved iteration of this analysis.

These two categories are then, as far as possible, treated identically in the Higgs Reconstruction and Background Rejection phases of the analysis, such that the two tag events can be used to produce a good representation of the background present in the four tag events.

4.3.3 *Higgs Reconstruction*

All events passing the basic requirements are then processed through the jet pairing algorithm described in Section 4.4, which reconstructs the two Higgs bosons, i.e. they are all reconstructed as $HH \to b\bar{b}b\bar{b}$ events. This first requires the selection of four jets, which are henceforth considered our Higgs constituent (HC) jets.

This selection can be performed in a manner that treats two tag and four tag events identically by sorting the jets (in descending order) by b-tag (i.e. the pass/fail decision, *not* the DL1r score), then p_{T}. The first four of these jets are then selected. Where an event has four or more tagged jets, this leads to the four highest p_{T} tagged jets being chosen. Where an event has precisely two tagged jets, this leads to the two tagged jets and two highest p_{T} untagged jets being chosen.

4.3.4 *Background Rejection and Top Veto*

The largest source of background in this analysis is multijet QCD events, such as $2 \times \left(g \to b\bar{b}\right)$. In order to reduce this background, the separation between the two Higgs candidates in pseudo-rapidity is required to be less than 1.5. This cut is motivated by the expectation that a decay of a resonance would be more central than multijet background, where the two Higgs candidates are often unrelated. The effect of this cut is shown in Figure 4.2, where we see that eliminating events to the right of the dashed line removes far more data (noting that "data" is primarily background) than it does any of the signals shown. This threshold was taken from the Early Run 2 analysis [58].

The next largest source of background consists of $t\dot{t}$ events, and other top-containing final states. A top veto is applied to mitigate this background, by removing events containing products consistent with a $t \to b(W \to q_1\bar{q}_2)$ decay.

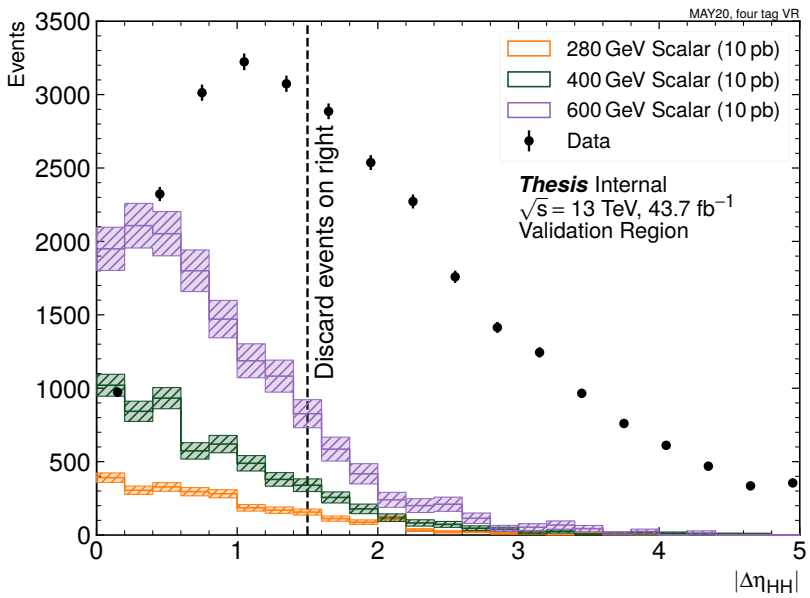


Figure 4.2: Effect of the $\left|\Delta\eta_{HH}\right|$ cut on simulated signals and (four tag) data, in the kinematic validation region (Section 4.3.5). This plot shows data taken in 2017, and the corresponding MC.

Top candidates are built from one HC jet (which is taken as the b-jet), and two other jets (which together form the W candidate). The jets forming the W candidate may themselves be HC jets. The quantity X_{Wt}, defined in Eq. 4.1 is then calculated for each top candidate, with m_t and m_W being the invariant mass of the top and W candidate respectively. X_{Wt} quantifies the three jet system's compatibility with being a decaying top quark. The event is discarded if $X_{Wt} < 1.5$ for any top candidate. The X_{Wt} definition and threshold were taken from the Early Run 2 analysis [58]. The effect of this cut is shown in Figure 4.3. Here we see that eliminating events to the left of the dashed line eliminates much of the $t\dot{t}$ simulation, while preserving most signal events. The 280 GeV signal suffers greater losses than the other signals shown here, which accounts for some of the increase in sensitivity with increasing mass that can be seen in this analysis.

$$X_{Wt} = 10\sqrt{\left(\frac{m_W - 80.4\,\text{GeV}}{m_W}\right)^2 + \left(\frac{m_t - 172.5\,\text{GeV}}{m_t}\right)^2} \tag{4.1}$$

Rarer backgrounds (e.g. single Higgs backgrounds) are negligible compared to these, and cuts to eliminate these were therefore considered unnecessary for this search for Beyond the Standard Model resonances. In a search for Standard Model di-Higgs production, these backgrounds should be considered, since they are far from negligible compared to the expected Standard Model di-Higgs cross-section.

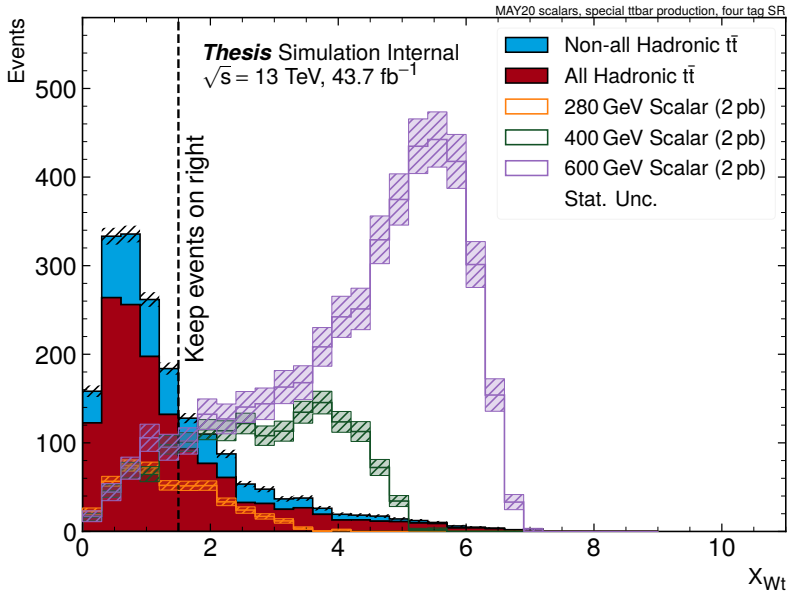


Figure 4.3: Effect of the X_{Wt} cut on simulated signals and $t\dot{t}$ events. The "all hadronic $t\dot{t}$" simulation contains those simulated $t\dot{t}$ events where both t-quarks decay fully hadronically, while "non-all hadronic $t\dot{t}$" contains those where at least one t-quark decays to a W boson which decays leptonically. This plot shows the MC simulation corresponding to data taken in 2017.

4.3.5 *Kinematic Region*

Finally, the events are separated into kinematic signal, validation, and control regions depending on the masses of the two Higgs candidates (m_{H1} and m_{H2}). These regions are illustrated in Figure 4.4.

The quantity X_{HH}, which quantifies an event's compatibility with the di-Higgs hypothesis, is defined in Eq. 4.2. The reference masses are 120 GeV and 110 GeV to account for energy losses in the detector. Events with $X_{HH} < 1.6$ are in the kinematic signal region.

The final signal selection consists of those four tag events that are in the kinematic signal region.

$$X_{HH} = 10\sqrt{\left(\frac{m_{H1} - 120\,\text{GeV}}{m_{H1}}\right)^2 + \left(\frac{m_{H2} - 110\,\text{GeV}}{m_{H2}}\right)^2} \tag{4.2}$$

The validation and control regions are defined similarly, using the quantities defined in Eqs. 4.3 and 4.4. The centre of the these regions is shifted to higher masses relative to the signal region by 3% for the validation region, and 5% for the control region. These shifts were optimized during the Early Run 2 analysis, such that the mean of the m_h 1 and m_h 2 distributions in the three regions would approximately coincide.

$$Y_{\text{Validation}} = \sqrt{(m_{H1} - 1.03 \times 120\,\text{GeV})^2 + (m_{H2} - 1.03 \times 110\,\text{GeV})^2} \tag{4.3}$$

$$Y_{\text{Control}} = \sqrt{(m_{H1} - 1.05 \times 120\,\text{GeV})^2 + (m_{H2} - 1.05 \times 110\,\text{GeV})^2} \tag{4.4}$$

Events not falling into the signal region, but with $Y_{\text{Validation}} < 30\,\text{GeV}$, are in the validation region. Events not falling into the signal or validation region, but with $Y_{\text{Control}} < 45\,\text{GeV}$, are in the control region.

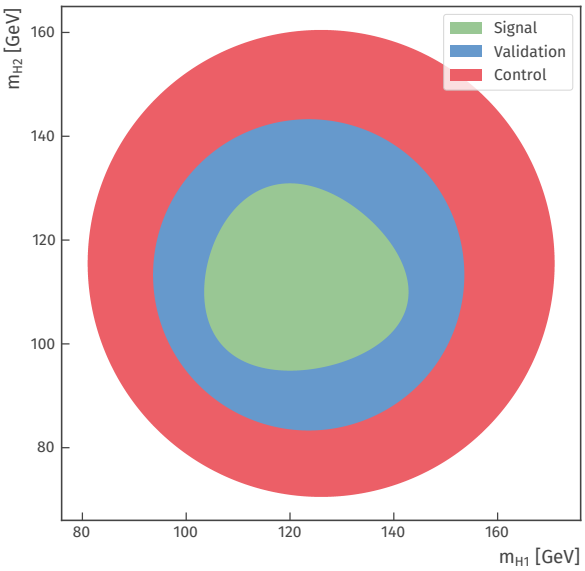


Figure 4.4: Kinematic signal, validation, and control regions in the m_{H1}-m_{H2} plane.

The kinematic region definitions were unchanged from the Early Run 2 analysis [58]. A re-optimization of the signal region yielded only a 5% improvement in sensitivity (as measured by $s/\sqrt{b}$), while modifying the definitions would present complications since this is a blind analysis.

4.4 Jet Pairing

A crucial step in the resolved $HH \to b\bar{b}b\bar{b}$ analysis is that of correctly pairing the four b-jets to form two Higgs candidates. With four b-jets, there are $\frac{\binom{4}{2}}{2} = 3$ ways to

do this (six ways to choose the two jets in the first Higgs candidate, then divide by two because there are two pairs in each pairing). This is illustrated in Figure 4.5.

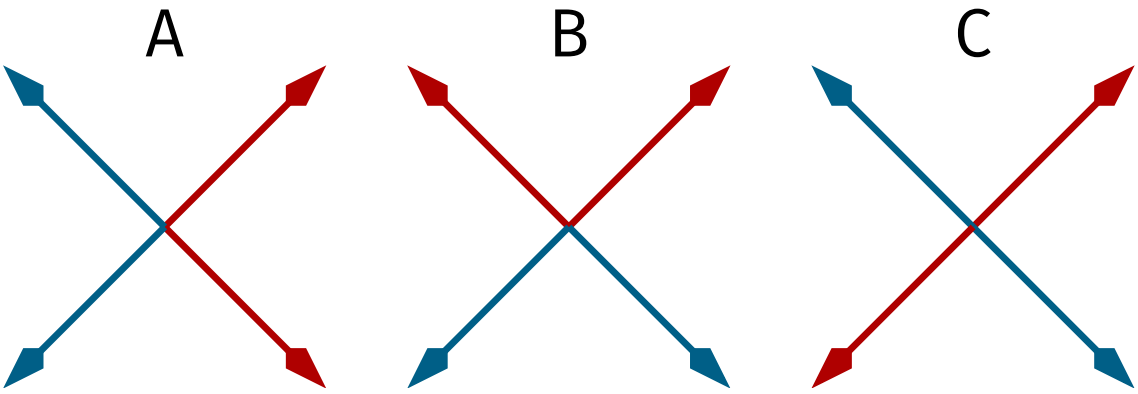


Figure 4.5: Three ways to pair four jets.

In the Early Run 2 analysis [58], this was done by first eliminating any pairing failing either of the conditions in Eqs. 4.5 and 4.6 (illustrated in Figure 4.6). These conditions were optimized to maximize signal sensitivity in the Early Run 2 analysis [58]. If no pairing passed these conditions, the event was eliminated entirely. If more than one pairing passed these conditions, the pairing used was selected by then minimizing the distance in the m_{H1}-m_{H2} plane from the pairing to the line joining (120 GeV, 110 GeV), as illustrated in Figure 4.7.

$$\frac{360}{m_{HH}/\text{GeV}} - 0.5 < \Delta R_{jj}^{\text{lead}} < \frac{653}{m_{HH}/\text{GeV}} + 0.47 \tag{4.5}$$

$$\frac{235}{m_{HH}/\text{GeV}} - 0.02 < \Delta R_{jj}^{\text{sublead}} < \frac{875}{m_{HH}/\text{GeV}} + 0.35 \tag{4.6}$$

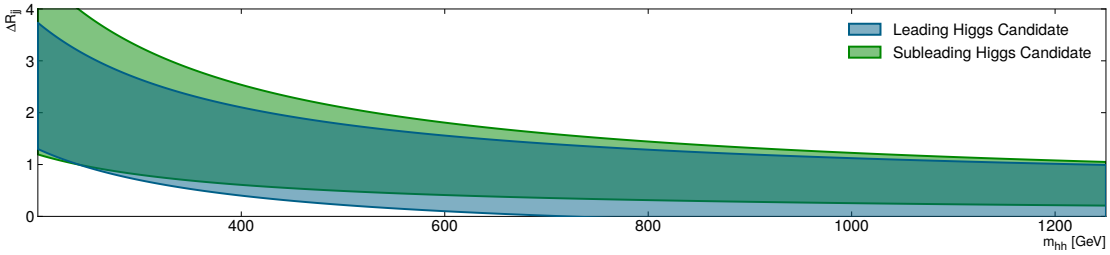


Figure 4.6: The ΔR conditions applied in the Early Run 2 analysis pairing strategy. A pairing passes these conditions only if *both* the leading and subleading Higgs candidate fall within the corresponding shaded region.

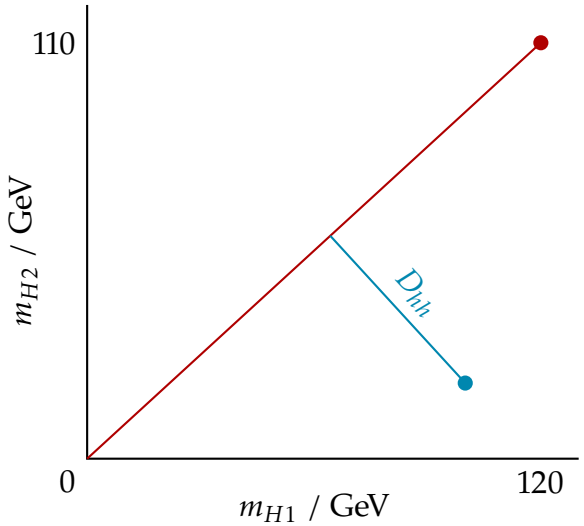


Figure 4.7: Definition of D_{hh}, the quantity minimized to break ambiguities in the Early Run 2 pairing strategy.

4.4.1 *Background Peak with Early Run 2 Pairing Strategy*

Figure 4.8 shows the result of applying the Early Run 2 analysis [58] to 2017 data. In the four tag data on the left we see, at the periphery of the blinded signal region, the tails of a peak. This peak is too prominent to expect it to be a true signal peak, and this can be confirmed by observing that such a peak is also present when the

same plot is made with two tag data (Figure 4.8 right). 2017 data is used purely for illustration here, and this effect can also be seen in data from 2016 or 2018.

This background peak in the signal region makes the task of predicting the background in the signal region given the known data in the other regions more difficult, and therefore increases the uncertainties on the background estimate.

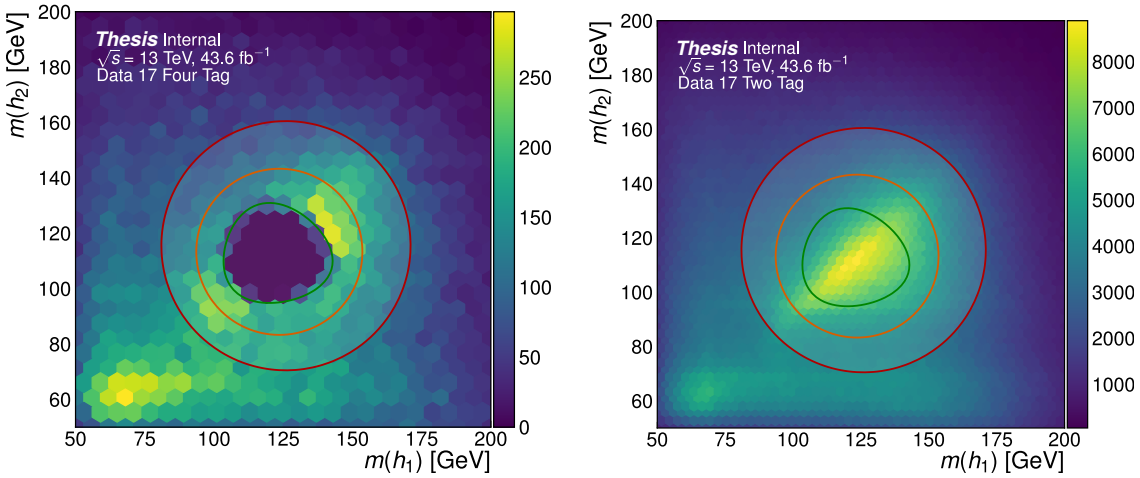


Figure 4.8: Result of applying the Early Run 2 analysis to 2017 data. The left figure shows the (blinded) four tag data, while the right figure shows the two tag data, which is not blinded and has many more events. A peak can be seen in the kinematic signal region (green) in both figures. The orange and red outlines delimit the boundary of the control and sideband regions respectively.

By removing the subsequent cuts, as in Figure 4.9, it can be seen that the peak lies along the line connecting the origin to (120 GeV, 110 GeV), suggesting that a major component is an artefact of the pairing strategy used in the Early Run 2 analysis, which preferentially chooses pairs that lie on this line, concentrating events into the kinematic signal region.

In this section I will describe how I developed an alternative pairing strategy which avoids the sculpting caused by the pairing strategy used in the Early Run 2 analysis.

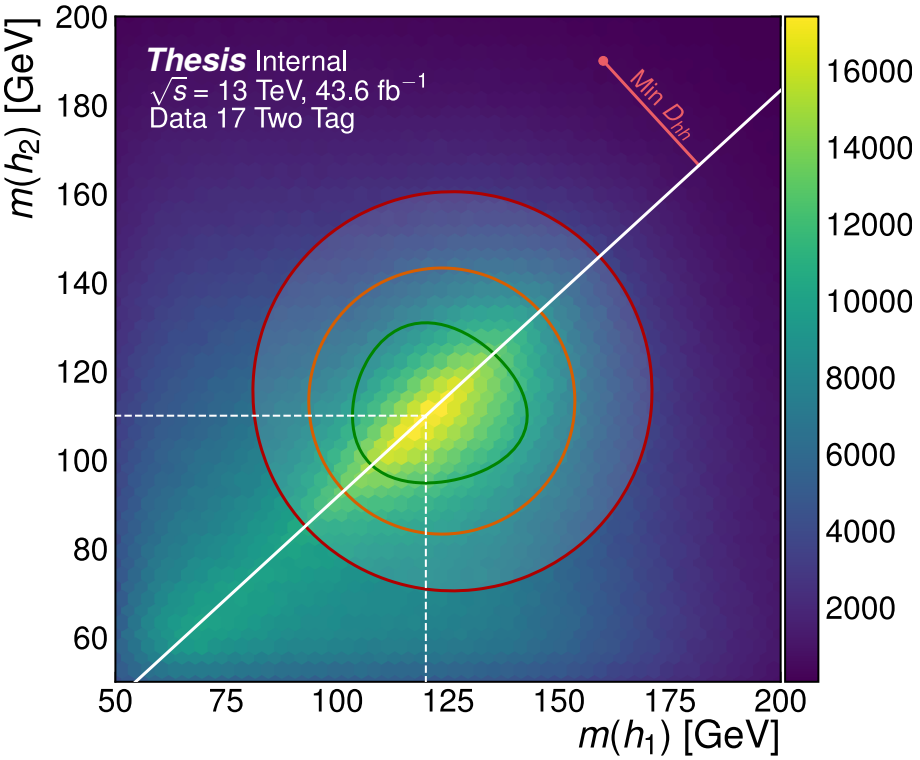


Figure 4.9: Result of applying the Early Run 2 analysis' pairing strategy to 2017 two tag data. In the Early Run 2 analysis, the four jets are paired into two Higgs candidates by first eliminating pairings that fail the ΔR cuts in Section 4.4, then choosing the remaining pairing closest to a line connecting the origin to $(120\,\text{GeV}, 110\,\text{GeV})$. The distance from a specific pairing to this line is called D_{hh}.

After adopting the new pairing strategy it can be seen (Figure 4.10) that though the background peak is not eliminated, it is less prominent, and no longer lies along the aforementioned line.

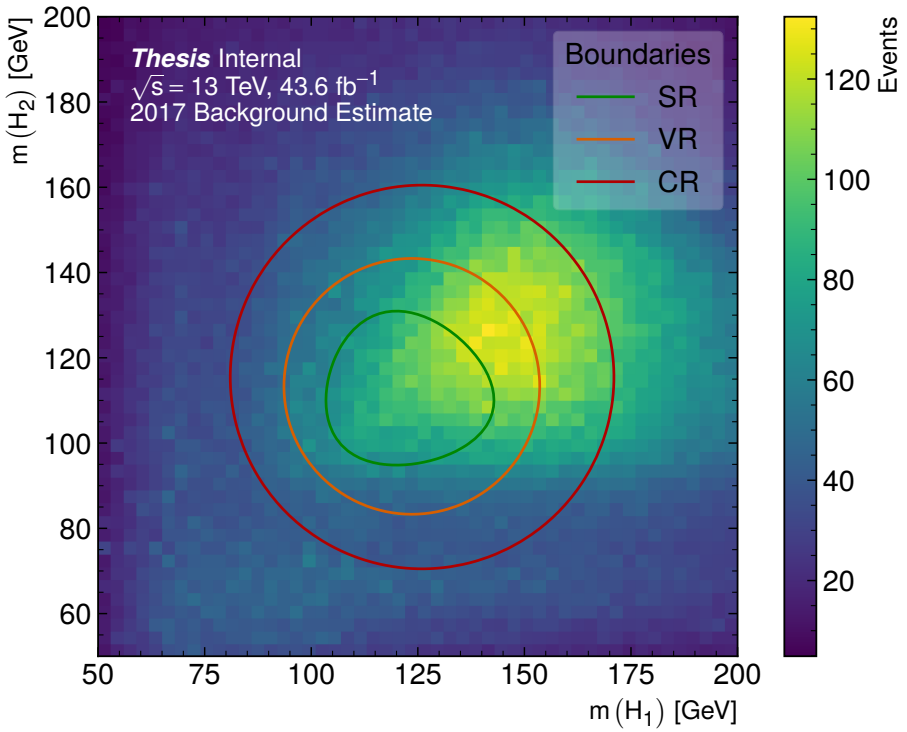


Figure 4.10: Result of replacing the Early Run 2 pairing strategy with that described in Section 4.4. The peak in the kinematic signal region is now much less prominent, and no longer lies over the line connecting the origin to $(120\,\text{GeV}, 110\,\text{GeV})$.

In Figure 4.11, separating the events into those where the invariant mass of the four-jet system as a whole (m_{HH}) is less than 400 GeV and those where $m_{HH} \geq 400\,\text{GeV}$ allows us to see that it is only the former set of events that peaks in the kinematic signal region.

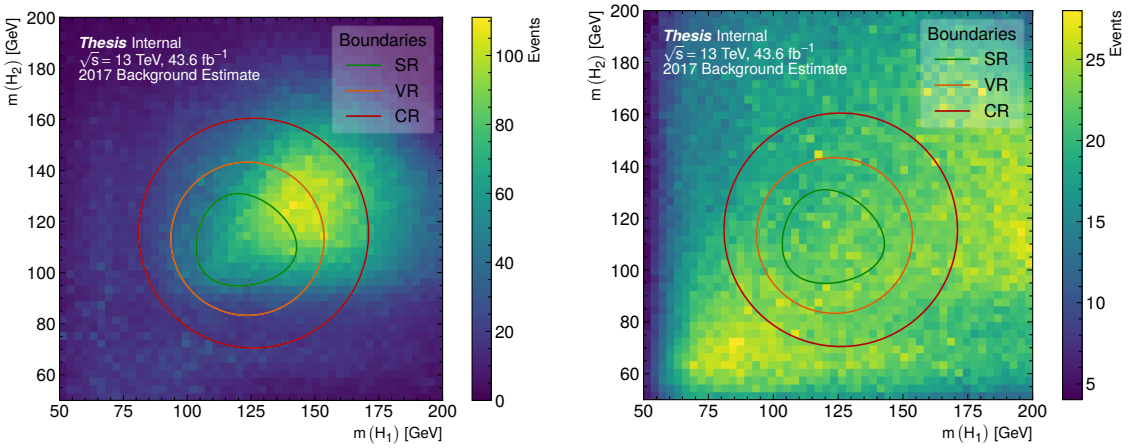


Figure 4.11: Result of replacing the Early Run 2 pairing strategy with that described in Section 4.4. The left figure shows those events where the four-jet invariant mass (m_{HH}) is less than 400 GeV while the right figure shows those events where this mass is greater than 400 GeV. We can observe that the background peak in the kinematic signal region consists of events in the former m_{HH} region.

4.4.2 *Monte Carlo Signal*

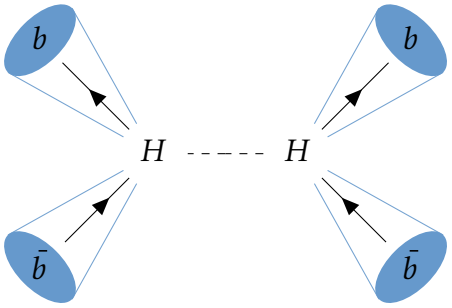
Monte Carlo simulated $HH \to b\bar{b}b\bar{b}$ signal events are necessary to evaluate the performance of pairing algorithms, and to train pairing algorithms that make use of machine learning.

When evaluating the performance of the old pairing, I used $G^*_{KK} \to HH \to b\bar{b}b\bar{b}$ events where the Randall-Sundrum graviton resonance mass was set at values between 300 GeV and 1400 GeV. These events were generated as described in Section 4.2, then passed through the full ATLAS detector simulation. A simulation of $HH \to b\bar{b}b\bar{b}$ events where the Higgs pair was produced through Standard Model processes was also used. These events were generated using aMC@NLO [11], Herwig 7 [53], and EvtGen, and again passed through the full ATLAS detector simulation.

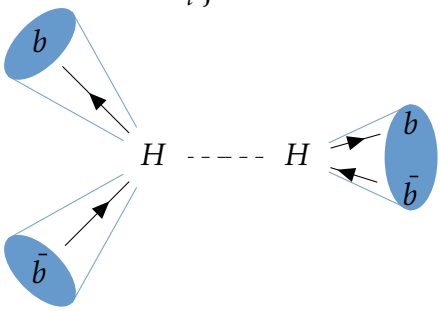
Events which did not have four b-tagged jets with $p_{\mathrm{T}} \geq 40\,\mathrm{GeV}$ and $|\eta| < 2.5$, and therefore would not enter the pairing stage of the analysis, were first dropped. The four b-quarks in each of the remaining events are then associated with the $R = 0.4$ anti-k_t jet whose cone it falls within. If multiple jet cones overlap in the vicinity of the b-quark, it is assigned to the jet whose axis is closest, in terms of ΔR. Here, only jets with $p_T > 40\,\mathrm{GeV}$ and $|\eta| < 2.5$ are considered, and they need not be b-tagged.

These events are then categorized into four exclusive categories:

Good — Events where each Higgs boson decays to two b-quarks in separate $R = 0.4$ anti-k_t jets



Merged 1 — Events where one Higgs boson decays to two b-quarks in separate $R = 0.4$ anti-k_t jets and the other decays to two b-quarks in the same $R = 0.4$ anti-k_t jet



Merged 2 Events where each Higgs bosons decays to two b-quarks in a single $R = 0.4$ anti-k_t jet



Incomplete Events where at least one b-quark falls outside any jet cone. For example, this can occur if one of the true b-jets is removed by the p_{T} or $|\eta|$.

Figure 4.12 shows the composition of the graviton and Standard Model Monte Carlo samples. It shows that the vast majority of events do fall into the "Good" category, at all mass points.

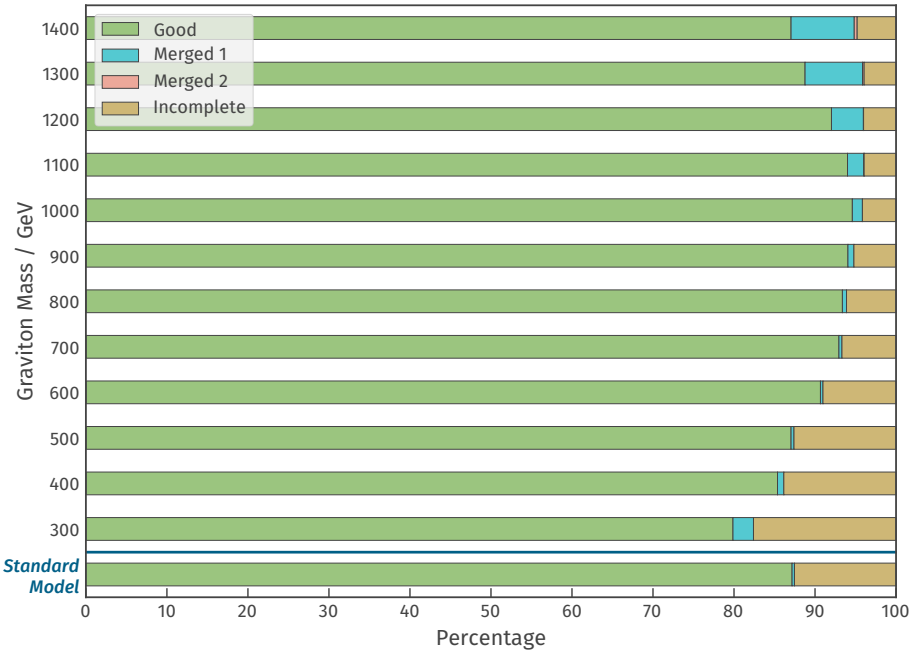


Figure 4.12: Composition of graviton and Standard Model samples. Most events fall into the "Good" category.

Only events categorized as "Good" are used for performance evaluation, since there is no "correct" pairing for merged or incomplete events.

For the purpose of training the new pairing algorithm, I used $X \to HH \to b\bar{b}b\bar{b}$ events. Here, the scalar resonance masses were set at values between 251 GeV and 900 GeV, with more events being generated at the lowest masses. These events were generated using MadGraph5_aMC@NLO 2.6.1 [11], Herwig 7.1.3, and EvtGen. They were then passed through the AtlFast II ATLAS fast detector simulation.

The Monte Carlo signal samples were prepared by first discarding any event which did not have four jets with $p_T > 35\,\text{GeV}$ and $\left|\eta\right| < 2.5$. This removes many events which would not pass the basic selections required to reach the pairing stage of the analysis, while preserving sufficient statistics to train the machine learning algorithm. A random $1/4$ of these events are selected to be used for training, since these events must later be excluded when these samples are used as reference signals in the statistical analysis.

As for the performance evaluation, I matched b-quarks to their closest jets, with the requirement that this closest jet be within $\Delta R = 0.4$ of the b-quark. Here, jets were considered down to 25 GeV to further increase the available statistics. Events where the four b-quarks were not all matched to four distinct jets were discarded.

4.4.3 *Performance Evaluation of Old Pairing*

Prior to investigating a new pairing strategy, I evaluated the performance of the strategy used in the Early Run 2 analysis. The results of this evaluation are shown in Figure 4.13. Above a graviton mass of around 500 GeV, the pairing algorithm is almost perfect, but for low mass gravitons and, to a lesser extent, the Standard Model sample, a significant number of events are incorrectly paired. The proportion of events where an incorrect jet is selected is fairly constant at 8%–10%.

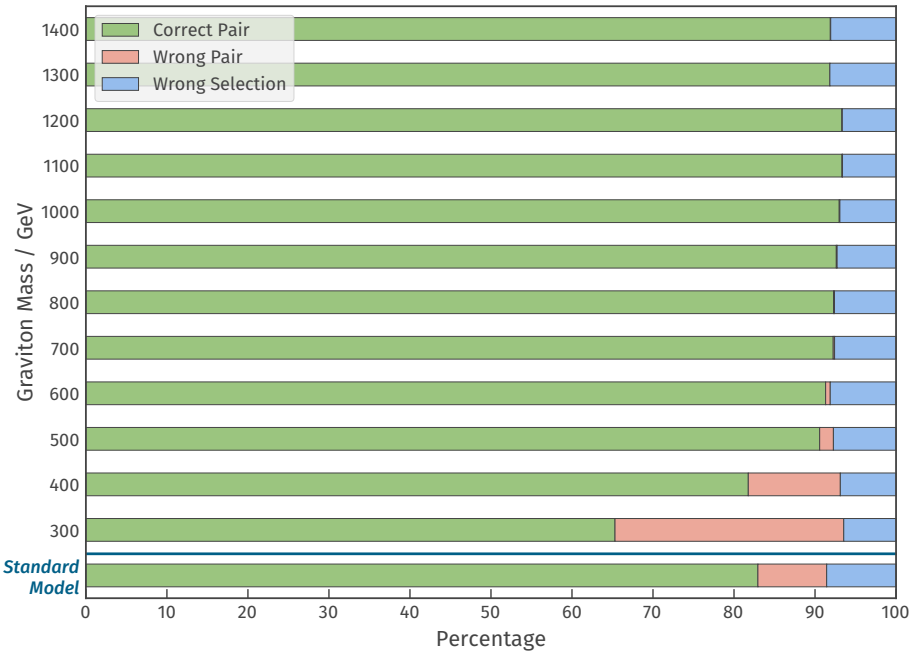


Figure 4.13: Efficiency of Early Run 2 analysis pairing strategy on graviton and Standard Model samples. Wrong Selection refers to those events where at least one of the four jets is selected incorrectly.

4.4.4 *New Pairing Strategy*

When designing a new pairing strategy, I had two key objectives:

1. To match or exceed the efficiency of the Early Run 2 strategy irrespective of resonance mass

2. To avoid sculpting background by avoiding use of Higgs candidate masses

In order to meet these objectives, I chose to train a gradient boosted tree [60], or Boosted Decision Tree (BDT), to distinguish correct pairings from incorrect pairings. A BDT uses a large number of decision trees to produce a model that can classify objects (here, pairings) based on a number of input variables, called *features*. For any event, the boosted decision tree assigns a score to each of the three possible pairings. This score is passed through a logit function, and is therefore in

essence the BDTs estimate of the probability that it is a correct pairing. For each event, the pairing with the highest score is then selected.

I constructed the boosted decision tree using LightGBM [61], and optimized the hyperparameters using a Bayesian search procedure [62]. The final hyperparameters returned by this procedure were:

- At least 50 events in each leaf (final) node in a tree
- No more than 180 leaves in each tree
- A `learning_rate` of 0.01

In order to meet the second objective, the variables used by the BDT are:

- ΔR_{jj} for the leading and subleading Higgs candidate
- $\Delta\eta_{jj}$ for the leading and subleading Higgs candidate
- $\Delta\phi_{jj}$ for the leading and subleading Higgs candidate

By comparing Figure 4.10 to the right hand plot of Figure 4.8 we can easily see that this second objective has been successfully met.

I also evaluated other variables, such as jet vertex charge (where track and secondary vertex information is used to determine the charge of jets) [63], jet pull (which measures the energy distribution inside a jet [64], where the energy in a pair of jets originating from the decay of a colour-neutral state would tend to fall between the two jet axes), and pseudo-continuous b-tagging categories (where the DL1r score is used to provide a more granular estimate of how b-like a jet is). These failed to improve the performance of the BDT.

Since the energy and momentum of the four jets is not available, the BDT is unable to reconstruct the Higgs candidate masses.

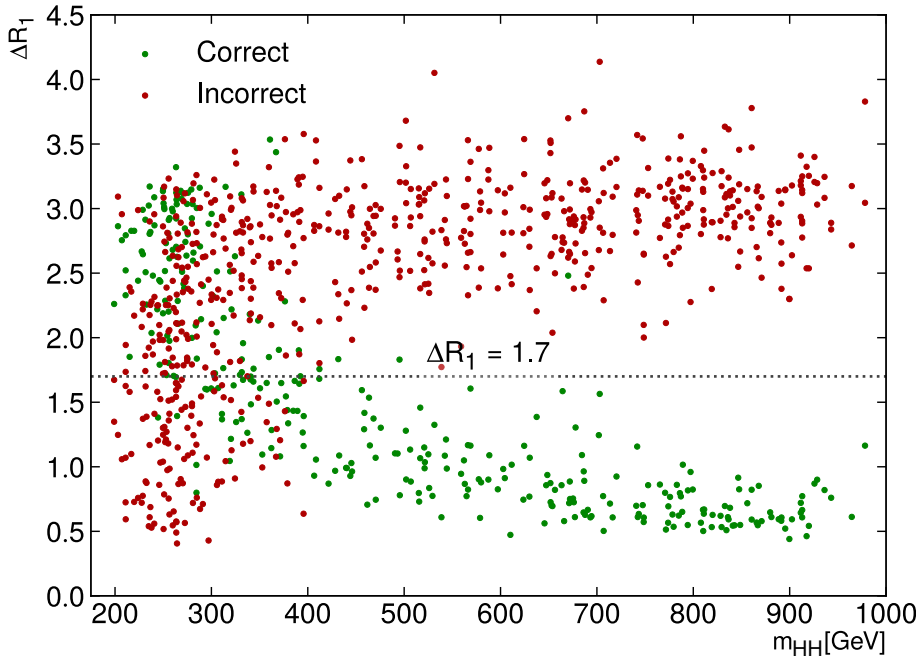


Figure 4.14: ΔR_{jj} for leading Higgs candidate for correct and incorrect pairings. Above 500 GeV these are very well separated.

Figure 4.14 shows how ΔR_1 (ΔR_{jj} for the leading Higgs candidate) changes for correct and incorrect pairings as a function of the di-Higgs mass. Above 500 GeV the correct and incorrect pairings are very well separated, since the Higgs bosons are sufficiently boosted that the decay products of each form a clear pair, close in ΔR. At lower masses the relationship gradually reverses, and ΔR_1 alone is insufficient to distinguish correct from incorrect pairings.

Since the angular behaviour of the jets produced by the Higgs decays changes so drastically, in order to be efficient across the m_{HH} range the BDT must be able to make use of the di-Higgs invariant mass. However, it is important that the BDT not make use of m_{HH} *as a discriminating variable*.

Fortunately, it is straightforward to parameterize the BDT on the di-Higgs invariant mass in this way. m_{HH} is independent of the pairing chosen, so by including all three pairings from each event in the training sample, I can ensure the proportion of correct to incorrect pairings is a constant 1:2 for each individual value of m_{HH}. The BDT can therefore use m_{HH} only as a parameter to guide the requirements it places on the angular separation variables.

4.4.5 *Performance Comparison*

Figure 4.15 shows an evaluation of the efficiency of the new pairing algorithm compared to the strategy used in the Early Run 2 analysis, and a third strategy where the pairing that minimizes ΔR_{jj} for the leading Higgs candidate is selected.

I conducted this evaluation by retraining the BDT 10 times, each time setting aside a different 25% of events. The efficiency is evaluated on this independent 25%, and the plot shows the mean efficiency from the 10 iterations. The plot also includes bands showing the standard deviation of these 10 iterations, but the statistical fluctuation in these efficiencies is small enough that this band is virtually invisible.

We see here that above $m_{HH} = 500\,\text{GeV}$, all three pairing strategies are extremely efficient, with the BDT and min ΔR_1 strategies being virtually 100% efficient. At low m_{HH}, however, the efficiency of the min ΔR_1 strategy drops off precipitously. The efficiency of the BDT strategy, on the other hand, drops off more gradually, and the BDT strategy has the best performance throughout the relevant m_{HH} range.

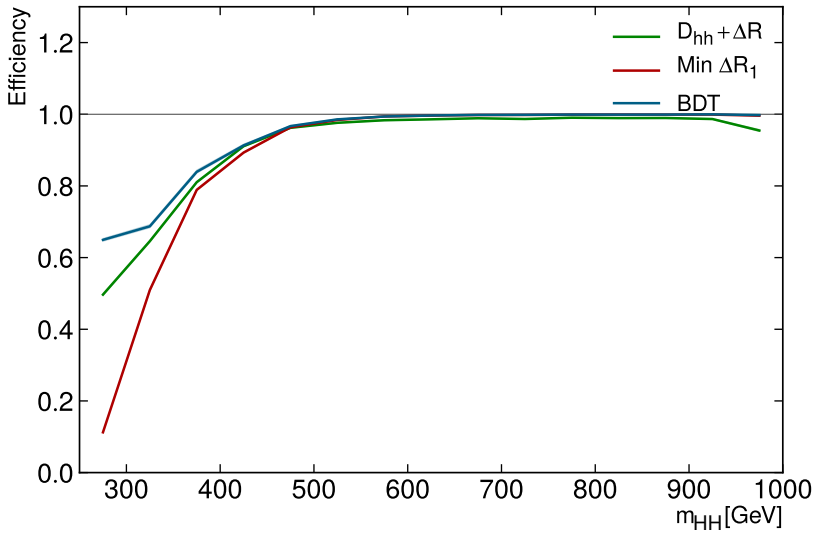


Figure 4.15: Performance of new (BDT) pairing algorithm compared to that used in the Early Run 2 analysis ($D_{hh} + \Delta R$) and a simple strategy picking the pairing where ΔR_{jj} for the leading Higgs candidate is smallest (Min ΔR_1).

4.4.6 *BDT Internals*

It is possible to gain some insight into how this boosted decision tree assigns scores to pairings by making use of SHAP values [65], which measure how much a given feature affected a machine learning models prediction for a given input item. Quoting [65],

> "SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$."

In the case of a boosted decision tree (as here), this can be done particularly efficiently [66] by the `TreeExplainer` in the Python `shap` package.

Figure 4.16 shows a summary plot. For each feature, each dot represents the impact that feature had on the prediction for a specific pairing. The dot colour represents the feature value.

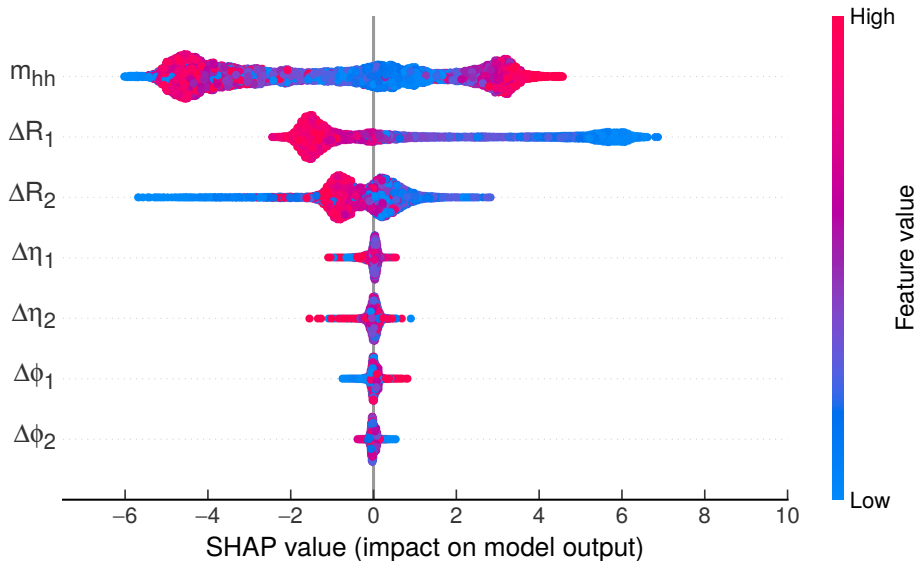


Figure 4.16: SHAP summary plot for pairing BDT. Features are ranked by impact, and each dot represents the impact that feature had on the prediction for that pairing. Dot colour represents feature value.

Features are ranked by average impact, so we can see that m_{HH} had the greatest impact. This indicates the pairing strategy learned by the BDT is strongly dependent on m_{HH}. We can also see that the BDT is most confident about high m_{HH} pairings. That is, on the m_{HH} row we see two pink (high m_{HH}) clusters at the two ends, while the low m_{HH} (blue) dots tend to be closer to the centre.

The other two high impact features are ΔR_1 and ΔR_2, the total separation between the jets in the leading and subleading Higgs candidate respectively. While the separations in η and ϕ alone do contribute, they have little impact overall.

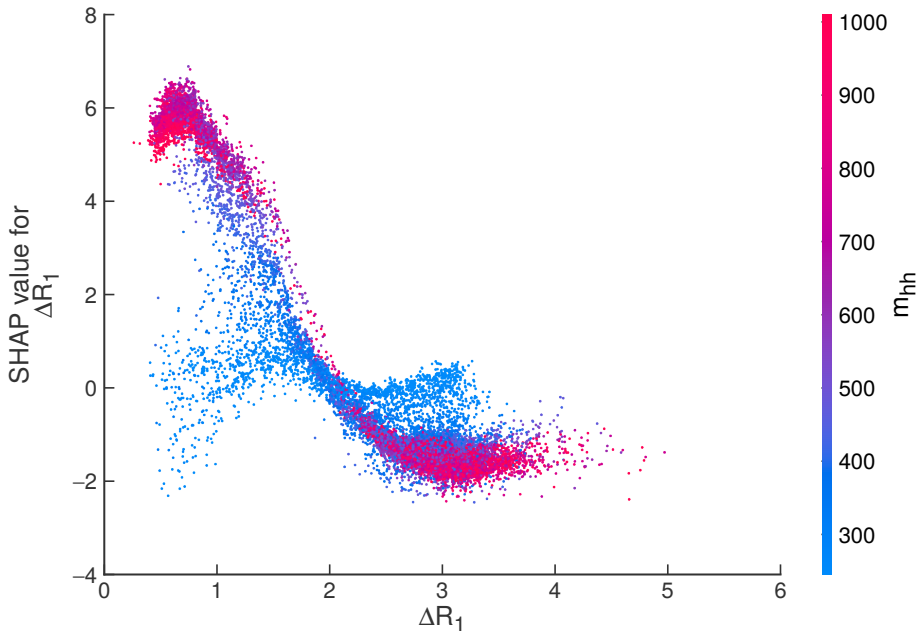


Figure 4.17: SHAP dependence plot for ΔR_1, coloured by m_{HH}. We can see that at high m_{HH}, low ΔR_1 indicates a high confidence that we have a correct pairing.

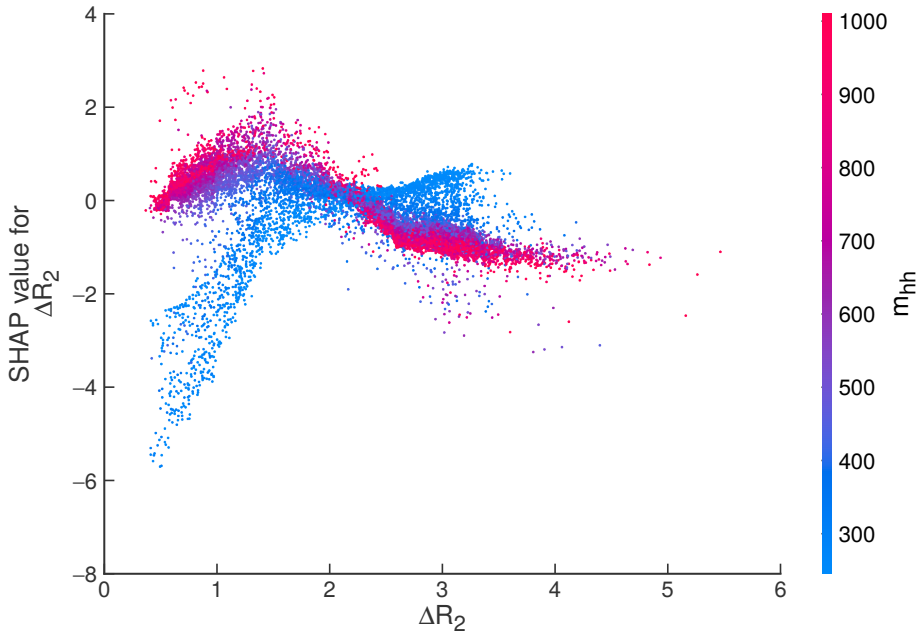


Figure 4.18: SHAP dependence plot for ΔR_2, coloured by m_{HH}. We can see that at low m_{HH}, low ΔR_2 indicates with moderate confidence that we have an incorrect pairing.

Figures 4.17 and 4.18 show SHAP dependence plots for ΔR_1 and ΔR_2, coloured by m_{HH}. We can now clearly see that at high m_{HH}, the BDT can easily distinguish correct pairings by the low separation between the jets in the leading Higgs candidate. At low m_{HH}, the task is more difficult. Here, the BDT can distinguish some incorrect pairings with moderate confidence based on low separation between the jets in the subleading (and to a lesser extent, leading) Higgs candidate. It can also make use of the limited additional information provided by the individual η and ϕ separations.

4.5 Background Estimation

The overwhelmingly dominant background in the $HH \to b\bar{b}b\bar{b}$ analysis is caused by multi-b QCD processes such as gluon splitting ($g \to b\bar{b}$). Where these events include fewer than four b-quarks, requiring four b-tagged jets can help eliminate much of this background. However due to the imperfection of the b-tagging algorithm, and multi-b events with four or more b-quarks, many of these events will pass the full event selection (including the four b-tagged jet requirement). It is therefore necessary to accurately estimate the background component in the final signal selection.

QCD backgrounds are notoriously difficult to accurately model in Monte Carlo. Due to this, the $HH \to b\bar{b}b\bar{b}$ analysis makes use of a fully data-driven strategy, which ultimately models both the QCD backgrounds, and smaller backgrounds such as the top backgrounds (e.g. $t\bar{t} \to b\bar{b}W^+W^-$), and single Higgs bosons decaying to a pair of b-quarks.

The background estimation strategy used in the $HH \to b\bar{b}b\bar{b}$ analysis is a variation on the classic "ABCD" method of producing a data driven background estimate.

As explained in Sections 4.3.2 and 4.3.5, events are divided into *tag categories*, and *kinematic regions*. The signal is expected to be concentrated in the four tag signal region. The background estimation strategy makes the basic assumption that the inherent variation of the background between the kinematic regions is uncorrelated with the inherent variation of the background between the tag categories, since these form the two orthogonal dimensions for the modified "ABCD" method.

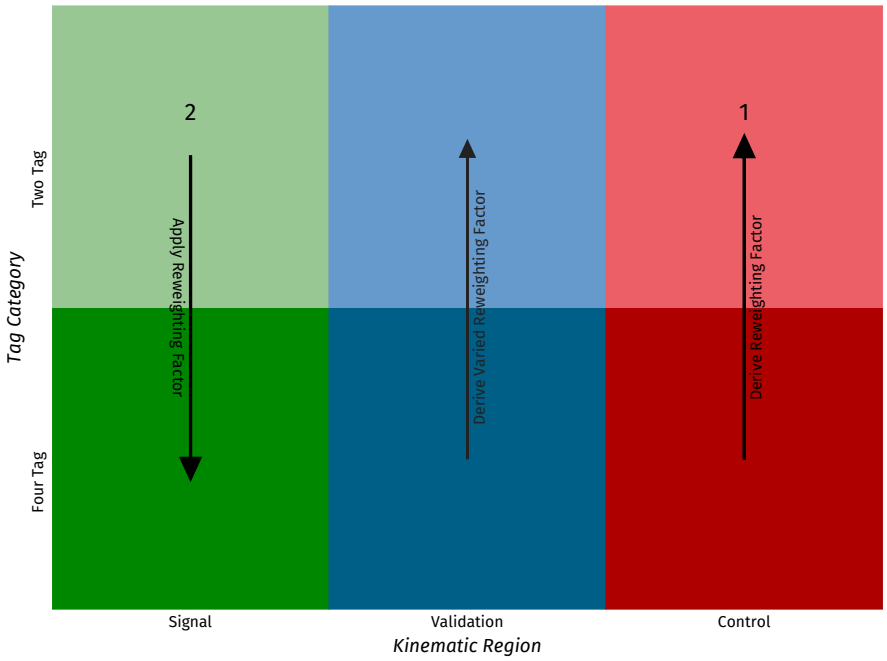


Figure 4.19: Schematic of basic background estimation strategy. The four regions for the modified "ABCD" method are Four Tag Signal Region, Four Tag Control Region, Two Tag Signal Region, and Two Tag Control Region. The validation region is used to measure the systematic uncertainty on the background model.

Figure 4.19 shows a schematic of the basic background estimation strategy. First, in the control region, reweighting factors are derived that, when applied to the two tag data, would yield a result resembling the four tag data. These reweighting

factors are then applied to the two tag data in the signal region to obtain an estimate of the background in the signal region.

These reweighting factors are a function that, given some properties of (i.e., variables describing) a two tag event, yield the weight that event must be given such that the distribution of the weighted two tag events in those variables is the same as the distribution of the four tag data in those variables.

In practice, this function consists of a neural net which returns the appropriate reweighting factor, given as input:

- the number of jets in the event
- the natural logarithm of the minimal X_{Wt} (calculated for the top veto)
- the natural logarithm of the transverse momentum of the second and fourth highest p_{T} HC jet,
- the natural logarithm of the transverse momentum of the complete four-jet system,
- the natural logarithm of the angular separation (ΔR) between the closest pair of HC jets,
- the natural logarithm of the angular separation between the other two HC jets,
- the angular separation (ΔR) between the Higgs candidates
- the separation in ϕ between the jets in each Higgs candidate
- and the average absolute pseudo-rapidity of the four HC jets.

Here it can be seen that in many of these cases, the natural logarithm of a variable is taken before it is used in the neural net. This reduces the range of the input variable, which improves the performance of the neural networks.

The neural network is trained to produce these weights by using a variation on the technique described in [67], since the reweighting function required *is*, in essence, the likelihood ratio between the four tag and two tag data.

A further correction is applied to the overall normalization, consisting of a normalization correction factor calculated by dividing the four tag control region yield by the reweighted two tag control region yield.

The statistical uncertainty on the background estimate is not accurately reflected by the standard procedure of applying a Poisson uncertainty, since this would reflect only the statistics of the two tag signal region, while the statistical uncertainty is driven by the statistics of the four tag background region (i.e. the statistical uncertainty on the reweighting factors). The statistical uncertainty is therefore determined using a bootstrapping procedure [68] where additional reweighting functions are derived using resamples (with replacement) of the control region. This is further described in Section 6.1.

The systematic uncertainty in this procedure (due to extrapolation in the m_{H1}-m_{H2} plane) is estimated by deriving an additional reweighting function in the validation region (as shown in Figure 4.19). The systematic uncertainty on a histogram bin for the background estimate is then the absolute difference between the bin value when using the nominal (control region derived) reweighting function, and the bin value when using the validation region derived reweighting function. Section 6.3 describes how this systematic uncertainty is decomposed into

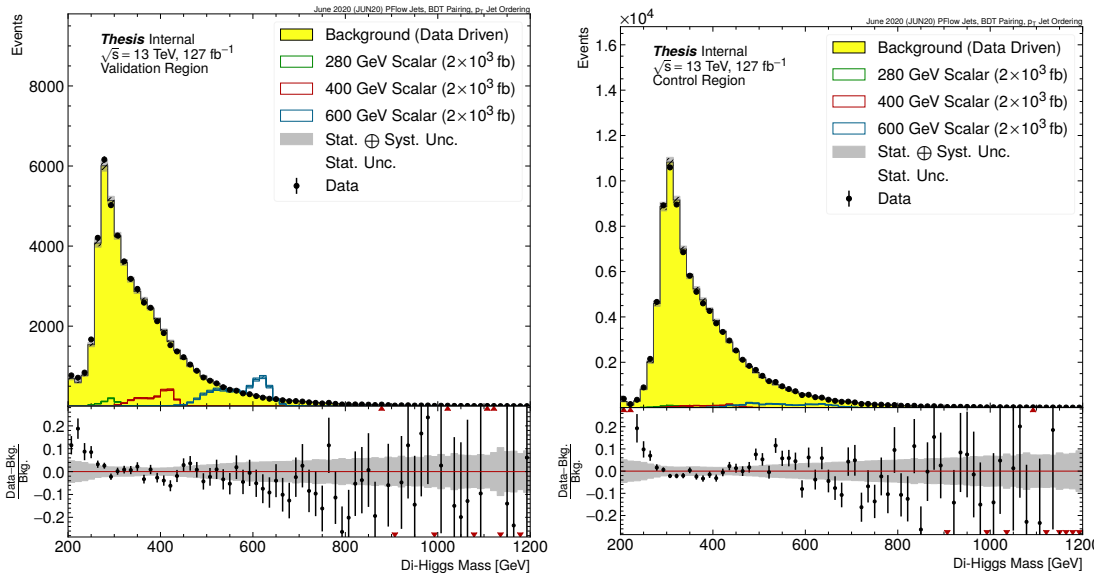
multiple nuisance parameters in the statistical analysis. In the validation plots below, however, this decomposition is not done and the raw difference between the two distributions produced with the two reweighting functions is shown as the systematic uncertainty.

4.5.1 *Validation*

The background estimation can be validated by comparing the distribution of some set of variables in the background estimate to the distribution in four tag data. This is done in Figures 4.20–4.22. Here, the background model has been applied to two tag data in the validation and control regions, and the comparison is made to four tag data in the same region. The control region plots demonstrate the level of closure of the background model (which is imperfect, demonstrating that the neural networks have not been over-trained) while those in the validation region demonstrate that the model performs satisfactorily.

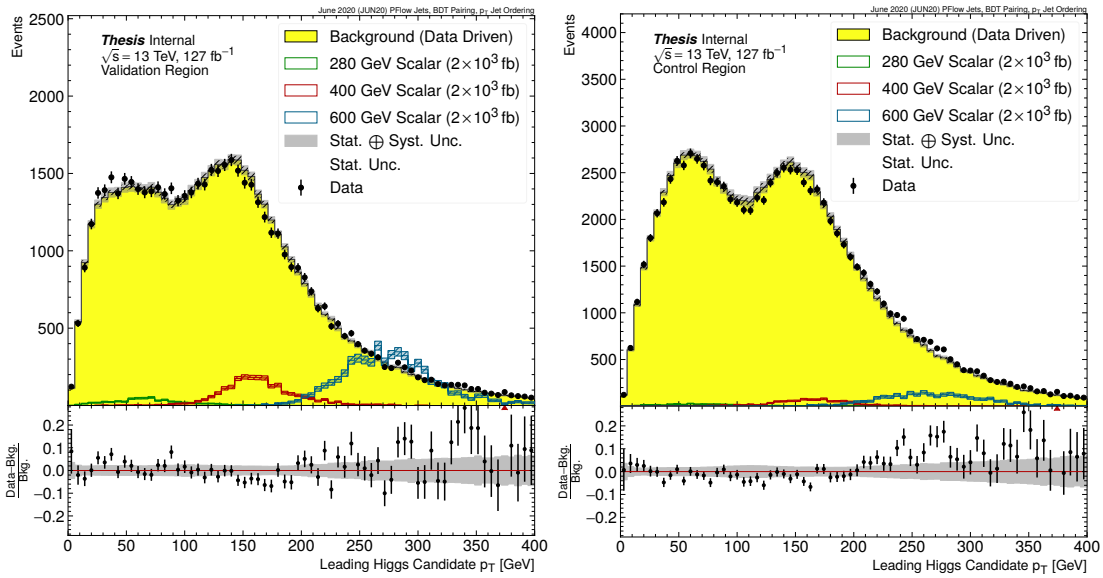
Modelling of the di-Higgs invariant mass, m_{HH}, is further improved after applying the Higgs mass correction, as can be seen in Figure 4.23d.

Figures 4.20 and 4.21 show variables that are not used in the reweighting procedure, demonstrating that the model is not solely valid for those variables that were part of the procedure.



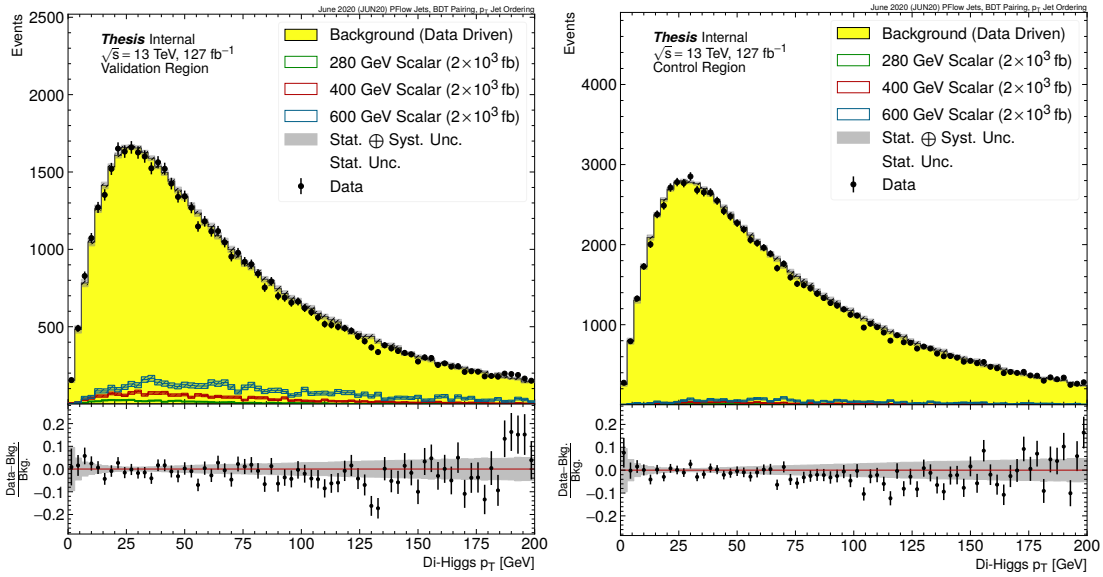
(**a**) Validation region (**b**) Control region

Figure 4.20: Di-Higgs invariant mass distribution in the validation and control regions. This variable is not used in the reweighting procedure. The signals illustrated are shown with a cross-section of 2 pb for visibility.



(**a**) Validation region (**b**) Control region

Figure 4.21: Leading Higgs boson transverse momentum distribution in the validation and control regions. This variable is not used in the reweighting procedure. The signals illustrated are shown with a cross-section of 2 pb for visibility.



(**a**) Validation region (**b**) Control region

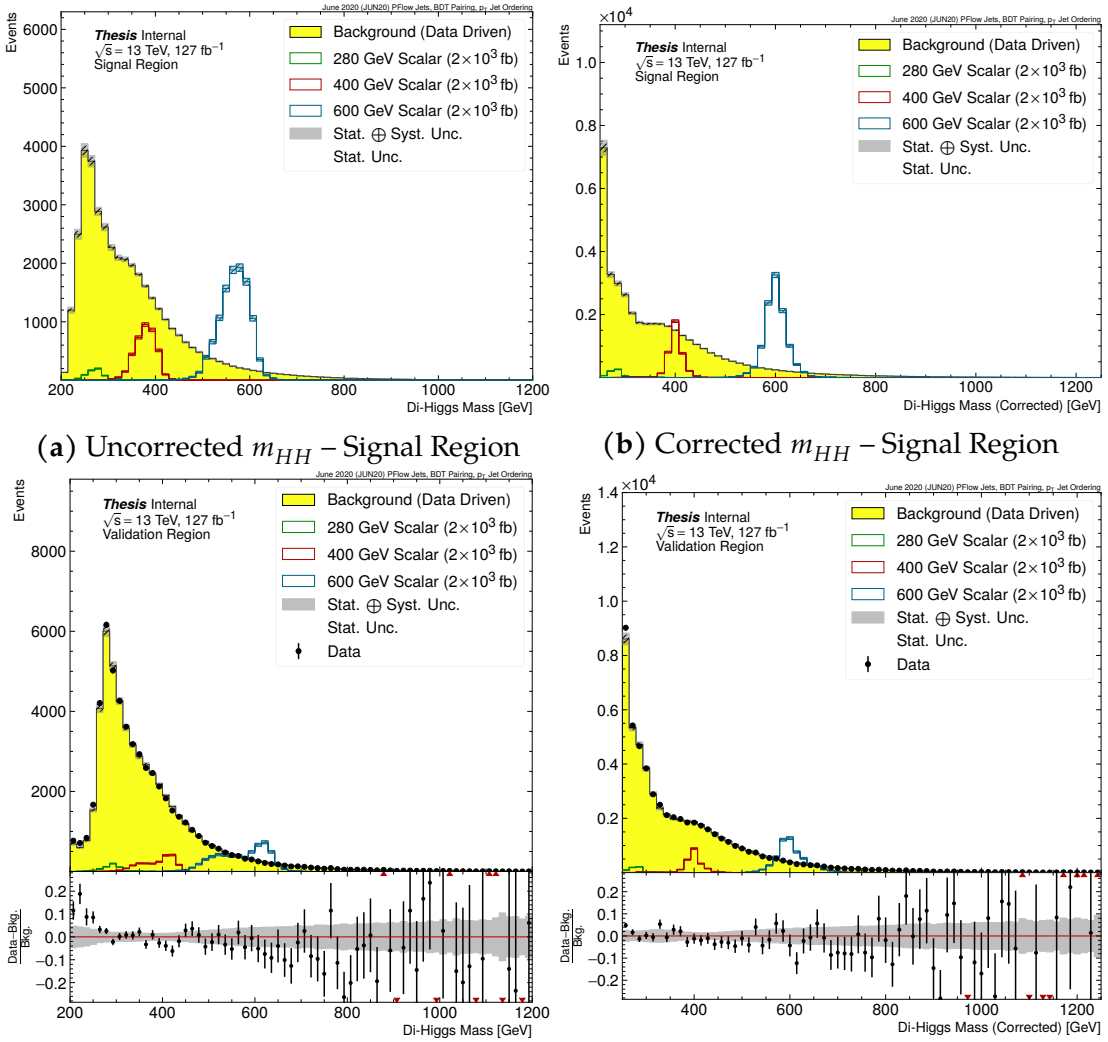
Figure 4.22: Di-Higgs transverse momentum distribution in the validation and control regions. The signals illustrated are shown with a cross-section of 2 pb for visibility.

4.6 Higgs Peak Sharpening

In order to compensate for energy losses shifting the di-Higgs invariant mass, a correction is applied to m_{HH} after the full event selection, to calculate a new variable called corrected m_{HH}. Each Higgs candidate four-vector is scaled such that its invariant mass is 125 GeV, and the corrected m_{HH} is then the invariant mass of the sum of the scaled four-vectors.

The effect of this correction on m_{HH} in the validation region can be seen in Figure 4.23. Each signal peak is centred at its nominal mass in the corrected plot, while in the uncorrected plot these peaks are shifted to lower masses due to energy losses in the detector and jet reconstruction. The widths of the signal peaks are also reduced. In addition, m_{HH} is now always greater than 250 GeV in the background estimate and data (shown in the validation region only as black dots), since the minimum energy of each Higgs candidate is 125 GeV.

In Figure 4.23 it is also evident that corrected m_{HH} is better modelled by the background estimation procedure than uncorrected m_{HH}. This is likely due to the correction reducing differences between two tag and four tag distributions, and further motivates the use of corrected m_{HH} in the statistical analysis.



(**a**) Uncorrected m_{HH} – Signal Region

(**b**) Corrected m_{HH} – Signal Region

(**c**) Uncorrected m_{HH} – Validation Region

(**d**) Corrected m_{HH} – Validation Region

Figure 4.23: Comparison between uncorrected and corrected m_{HH} in the signal and validation regions. The signals illustrated are shown with a cross-section of 2 pb for visibility. We can see here that the correction correctly centres the signal peaks, improves their mass resolution, and improves the quality of background modelling.

5 Statistical Analysis

The purpose of the statistical analysis is to take the raw data from the $HH \to b\bar{b}b\bar{b}$ analysis, in the form of histograms showing the (corrected) m_{HH} distribution in data and predicted background, and histograms giving the predicted m_{HH} shape for signals at various masses, and convert these to a 95% confidence level upper limit on the $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section, as a function of the mass of the scalar boson X. This is done taking into account both the statistical uncertainties on the various histograms, and the systematic uncertainties on the predicted background distribution and signal shapes, as described in Chapter 6.

The statistical analysis is conducted using RooStats [69], with a model constructed using HistFactory [70]. This section describes the mathematical technique used to set a limit on a single signal hypothesis (in other words, a single mass point in the resonant $HH \to b\bar{b}b\bar{b}$ analysis). Section 5.2 describes the practical implementation of the statistical analysis in the $HH \to b\bar{b}b\bar{b}$ analysis.

The limits are set using the CL_s technique [71]. Given some signal, the null hypothesis is the hypothesis in which there is only background, b, while the alternative hypothesis is that in which we have both signal and background, $s+b$. Given some data, CL_b is then the confidence level for the null hypothesis, and CL_{s+b} is the confidence level for the alternative hypothesis. Since signal and background are not well separated in this analysis, data highly compatible with $s+b$ will also be highly compatible with b. In such a situation, where there is insufficient statistical power to separate the hypotheses, CL_{s+b} can falsely exclude a signal hypothesis where observed data fluctuates below the expected background [72]. We therefore

define CL_s, which gives a better measure of the compatibility of the data with the signal hypothesis:

$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} \tag{5.1}$$

In order to set limits on the $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section, a signal strength parameter μ is defined, giving a continuous alternative hypothesis $s(\mu) + b = \mu s_0 + b$ where s_0 is a reference signal strength. In order to obtain a limit on the cross-section in fb, this reference cross-section is set to 1 fb.

The systematic uncertainties are accounted for by defining nuisance parameters (denoted θ), which are allowed to vary within constraints set by the nominal uncertainties. The variation of each nuisance parameter parameterizes the variation of the m_{HH} distribution due to a systematic uncertainty. In a process called *profiling*, a maximum likelihood fit is conducted, fitting both μ and the nuisance parameters simultaneously, determining best-fit values for the nuisance parameters, which modify the background estimate and simulated signal shape given the additional information learned from the data, and uncertainties on these, which provide a new estimate for the magnitudes of these uncertainties, given the additional information available from the data.

The analysis is conducted using a test statistic, q_μ.[1] This is defined in Eq. 5.2, where a single hat ($\hat{\cdot}$) represents the maximum likelihood estimate of the parameter $\cdot$, and a double hat ($\hat{\hat{\cdot}}(\mu_0)$) represents the conditional maximum likelihood estimate

1. The analysis was planned to be conducted using $\tilde{q}_\mu$, which encodes the physical condition that $\mu > 0$. It was later learned that $\tilde{q}_\mu$ should not be used without bounding μ when calculating the best-fit [73].

of ⬚ if μ is fixed to μ_0.

$$q_\mu = \begin{cases} -2\ln\left(\frac{\mathcal{L}\left(\mu,\hat{\hat{\theta}}(\mu)\right)}{\mathcal{L}\left(\hat{\mu},\hat{\theta}\right)}\right) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{5.2}$$

This test statistic is defined in terms of $\mathcal{L}(\mu, \boldsymbol{\theta})$, the likelihood of a specified signal strength μ and nuisance parameters $\boldsymbol{\theta}$, given some data. The canonical definition (summarized below) of the likelihood function generated by HistFactory can be found in its documentation [70], or in the documentation for pyhf [73], a Python re-implementation of the HistFactory model.

A likelihood given a certain hypothesis (μ and $\boldsymbol{\theta}$) is calculated by obtaining a probability distribution function $f(m_{HH})$ for a single event to have a given m_{HH} under that hypothesis. The likelihood is then the probability of the data under that hypothesis [70]:

$$\begin{aligned} \mathcal{L}(f(m_{HH}) \mid \text{data}) &= \text{P}\left(\text{data} \mid f(m_{HH})\right) && (5.3) \\ &= \text{Pois}\left(n \mid \lambda = n_{\text{exp}}\right) \prod_{i \in \text{events}} f(m_{HH}(i)) \prod_{\theta \in \boldsymbol{\theta}} \text{G}\left(\theta' \mid \theta\right) && (5.4) \\ &\propto \prod_{b \in \text{bins}} \text{Pois}\left(n_b \mid \lambda = n_{b,\text{exp}}\right) \prod_{\theta \in \boldsymbol{\theta}} \text{G}\left(\theta' \mid \theta\right), && (5.5) \end{aligned}$$

where n_b is the number of events in bin b, $n_{b,\text{exp}}$ is the expected number of events in that bin, n is the total number of events, and n_{exp} is the expected total number of events. Here, the second line makes use of the fact that, given the expected histogram generated by the hypothesis (the expected bin values, $n_{b,\text{exp}}$), the probability a particular bin will have n_b events can be determined by assuming the actual bin

value for each bin will be distributed according to the Poisson distribution. Eq. 5.4 gives an unbinned likelihood, while Eq. 5.5 gives a binned likelihood for a given histogram binning. Since likelihoods are used in ratios, both likelihoods in a ratio should be calculated with the same binning.

Each nuisance parameter has a Gaussian constraint term in this likelihood function, $\mathrm{G}\left(\theta' \mid \theta\right)$, where θ' is an *auxiliary measurement* representing the information about θ that can be obtained from data. This auxiliary measurement may, in the case of modelling and flavour tagging nuisance parameters, be from a calibration analysis or, in the case of the nuisance parameters representing the systematic uncertainty on the background estimate, may be an additional measurement made in this analysis.

Given this, CL_s is given by [71]

$$\mathrm{CL}_s\left(\mu\right) = \mathrm{CL}_{s(\mu)} = \frac{\mathrm{CL}_{s(\mu)+b}}{\mathrm{CL}_b} = \frac{\mathrm{p}\left(q_\mu \geq q_{\mu,\mathrm{obs}} \mid s(\mu)+b\right)}{\mathrm{p}\left(q_\mu \geq q_{\mu,\mathrm{obs}} \mid b\right)}. \tag{5.6}$$

Here, it should be noted that due to statistical uncertainties, the observed data under a fixed μ varies, and therefore so does q_μ. As a result q_μ follows a distribution under a fixed hypothesis. These p-values are calculated using the asymptotic approximation [74], which is valid because we are in the large sample limit, due to the large background.[2] In this limit, the distributions of q_μ can be approximated by analytic expressions given in [74]. The limit is then set by determining the signal strength that gives a $\mathrm{CL}_s = 0.05$. Larger signal strengths are excluded at the 95% confidence level.

2. Even where the sample size is more modest, the asymptotic approximation provides accurate results [74].

To summarize, the procedure of setting upper limits conceptually proceeds as follows. A scan is conducted over some range of μ. At each point, q_μ is calculated using the observed data to give $q_{\mu,\text{obs}}$. $\text{CL}_{s(\mu)+b}$ and CL_b are calculated using the asymptotic approximation. Alternatively, it would also be possible to generate many "toy" datasets under the b and $s(\mu)+b$ hypotheses, to empirically determine the distributions of q_μ. The value $\text{CL}_s(\mu)$ is then calculated as the ratio of $\text{CL}_{s(\mu)+b}$ and CL_b. If $\text{CL}_s(\mu) < 0.05$ the point μ is excluded at the 95% confidence level. In other words, the probability of producing the data observed if the true signal strength were μ is less than 5%.

5.1 Nuisance Parameters

The nuisance parameters θ are assigned to represent the variation of m_{HH} distributions (background estimate or signal simulation) due to systematic uncertainties. These parameterize the variation from a nominal distribution, given two variations, a $+1\sigma$ and -1σ variation.

In the case of a normalization uncertainty, the parameterized m_{HH} distribution $m_{HH}(i,\alpha)$, where i indexes over bins and α is the nuisance parameter value, is given by

$$m_{HH}(i,\alpha) = m_{HH}(i) \times \begin{cases} v_+^{|\alpha|} & \alpha > 0 \\ v_-^{|\alpha|} & \alpha < 0 \\ 1 & \alpha = 0 \end{cases} \tag{5.7}$$

where $v_\pm$ are the $\pm 1\sigma$ variations.

In the case of a shape uncertainty, the $\pm 1\sigma$ variations are themselves m_{HH} histograms, and will be denoted $v_\pm(i)$. The parameterized m_{HH} distribution is then

given by

$$m_{HH}(i,\alpha) = m_{HH}(i) + |\alpha| \times \begin{cases} v_+(i) - m_{HH}(i) & \alpha \geq 0 \\ v_-(i) - m_{HH}(i) & \alpha < 0. \end{cases} \tag{5.8}$$

These make up the bulk of the systematic uncertainties.

5.2 Concrete Implementation

This section will describe the concrete implementation of the statistical analysis (i.e. limit setting) procedure for the resonant $HH \to b\bar{b}b\bar{b}$ analysis.

A limit is set for each resonant mass point separately. The limits are produced using histograms of corrected m_{HH} (with 70 bins between 250 GeV and 1250 GeV). The following histograms are produced for each data-taking year:

- A background histogram, generated using two tag data in the kinematic signal region. Weights are applied to events to implement the background estimation ("reweighting") procedure. The statistical uncertainty on each bin in this histogram is replaced with one derived using bootstrapping, as described in Section 6.1.
- A data histogram, generated using four tag data in the kinematic signal region. Before unblinding, this is replaced with a copy of the background histogram.
- For each mass point, a signal histogram is generated using four tag signal simulation, in the kinematic signal region. The signal normalization is set to 1 fb, such that the signal strength μ is equal to the $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section limit in fb.

In addition, histograms are generated to represent the "up" and "down" variations for each nuisance parameter. For the signal uncertainties, these are generated using jet variations (for jet-related systematic uncertainties), or by applying a varied weight (for flavour tagging and trigger-related systematic uncertainties).

For the background shape uncertainties, these are generated using the low / high H_T decomposition, described in Section 6.3.1. The background shape uncertainty is separated into two nuisance parameters, one for events where the scalar sum of the transverse momenta of the four selected jets (called H_T for the purpose of this decomposition) is less than 300 GeV, and one for events where this $H_T > 300\,\text{GeV}$. For the low and high H_T nuisance parameters, the "up" variation is generated in the same way as the nominal background histogram, but for events with a low (or high, respectively) H_T, the weight applied is from the alternate background model. The "down" variation is generated by inverting the difference between the "up" variation and the nominal histogram.

These histograms are used to generate one workspace for each mass point. Each workspace consists of three channels, one for each year of data taking. Each channel consists of two samples (a background sample containing the predicted background, and a signal sample containing the simulated signal) in addition to the observed data. RooStats will set a limit in all three channels simultaneously, and combine these to give an overall limit.

The background sample is based on the background histogram, and statistical errors are activated on this sample. `SetNormalizeByTheory` is set to false, since the background sample is derived from data. This prevents the integrated luminosity uncertainty from being applied to the background prediction. A `HistoSys` is used to

implement each of the low and high H_T nuisance parameters, while an `OverallSys` implements the normalization nuisance parameter. The up and down variation of this parameter are set to $1 \pm |n_{\text{nom}} - n_{\text{alt}}|/n_{\text{nom}}$, where n_{nom} and n_{alt} are the normalization corrections derived in the kinematic control and validation region, respectively (see Section 6.3.1 for full details).

The signal sample is based on the histogram for the mass point in question. `HistoSyses` are used to implement each of the detector modelling and reconstruction uncertainty nuisance parameters, as described in Section 6.2. `SetNormalizeByTheory` is set to true, since these samples are simulated. This allows the normalization of these samples to be varied to implement the integrated luminosity uncertainty.

RooStats is then used to set a limit for each workspace. The `AsymptoticCalculator` is used to compute CL_s in the asymptotic approximation for each μ tested. RooStats is configured to set an upper limit by enabling `SetOneSided`. The expected limit is then set by computing CL_s (and the $\pm 1\sigma$ and $\pm 2\sigma$ CL_s bounds) at various test values in a root-finding procedure, using a background-only Asimov dataset [74].

The nominal limit is set at the μ value whose $\text{CL}_s = 0.05$, while the uncertainty bands are set at the μ values whose $\pm 1\sigma$ and $\pm 2\sigma$ CL_s bounds are equal to 0.05. The root-finding procedure is configured to calculate each of these to a precision of at least 1%.

5.3 Expected (Pre-Unblinding) Limits

Figure 5.1 shows the expected $pp \to X \to HH$ cross-section limit (at 95% confidence level), before unblinding. Since this is a pre-unblinding expected limit, it is conducted on an Asimov dataset [74] where μ and all nuisance parameters are set

to zero. In other words, the Asimov dataset is equal to the predicted background. The $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section limit was converted to a $pp \to X \to HH$ cross-section limit by dividing by the branching ratio, 33.9%.

Here we see that the limit is stronger (smaller) at 251 GeV than at immediately higher masses, despite the limit shrinking with increasing mass in general. This is the result of the signal efficiency of the analysis being stronger for the 251 GeV mass point.

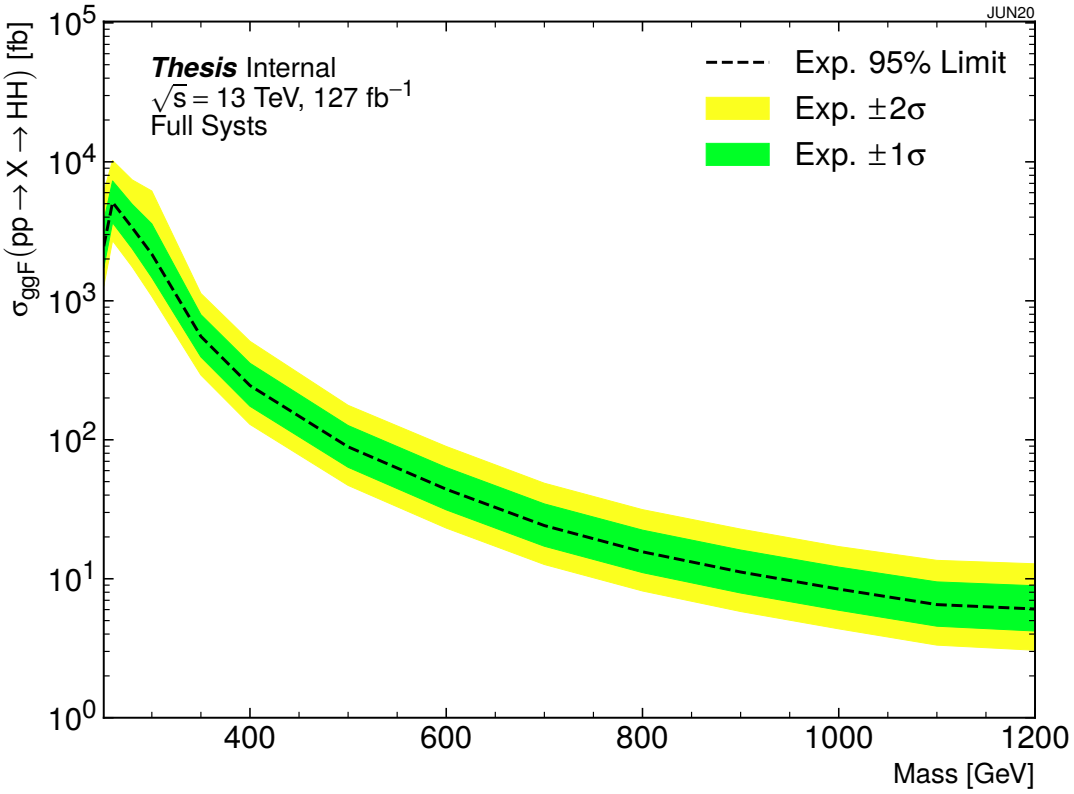


Figure 5.1: Pre-unblinding expected $pp \to X \to HH$ cross-section limit, at the 95% confidence level. The $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section limit was converted to a $pp \to X \to HH$ cross-section limit by dividing by the branching ratio, 33.9%.

6 Sources of Uncertainty

The uncertainties in this $HH \to b\bar{b}b\bar{b}$ analysis can be broken down into statistical uncertainties, and systematic uncertainties from various sources. Each of these will be described here.

6.1 Statistical Uncertainties

The term "statistical uncertainty" refers to the uncertainty stemming from the analysis drawing from a finite pool of data. As is common in High Energy Physics, the foundation of this analysis is the histogram, which describes the distribution of a variable in terms of the fraction of events encountered where that variable fell into each of a set of finite-width bins. Since the total number of events is finite, the number of events falling into each bin follows a Poisson distribution, illustrated in Figure 6.1. The uncertainty on the bin should therefore be the square root of the number of events the bin should ideally have contained. Since this number is unknown, it is estimated by the square root of the number of events the bin actually contains.

This treatment is sufficient to determine the statistical uncertainty on four tag data, since this data is not weighted. It does not suffice for simulated signal distributions or for the background estimate, since in both cases, events are weighted. That is, the events falling into a bin do not contribute equally to the value of that bin in the histogram.

In the case of simulated signal distributions, the uncertainties on the weights themselves are dealt with separately, as systematic uncertainties. The statistical

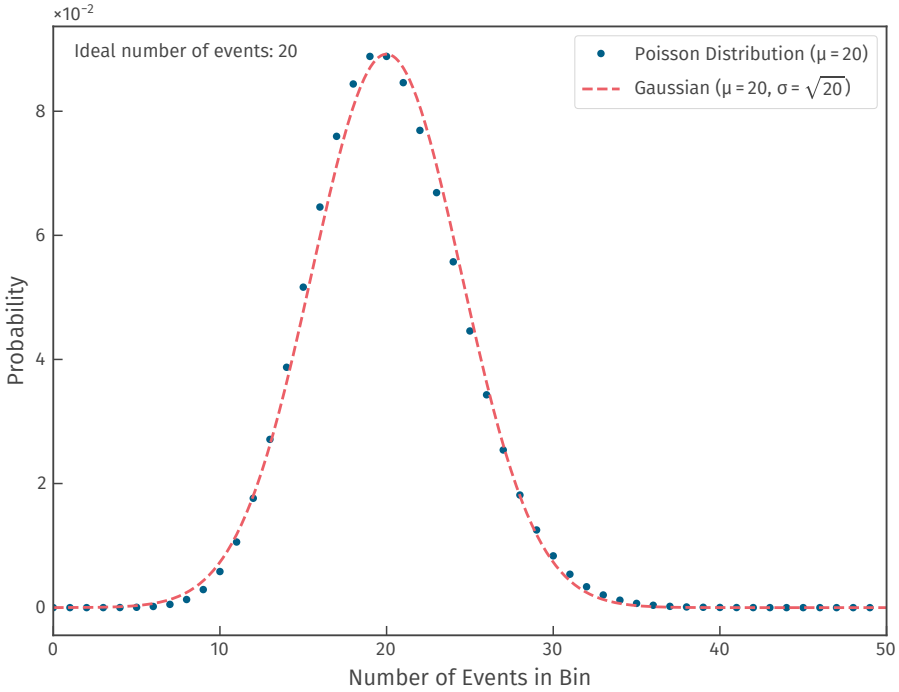


Figure 6.1: The actual number of events falling into a bin that should, in theory, have 20 events, will follow a Poisson distribution (shown in blue) due to random fluctuations. The standard deviation of a Poisson distribution is the square root of its mean, and this can be seen by noting that the Gaussian distribution with mean 20, and standard deviation $\sqrt{20}$ (dashed red line) falls very close to the Poisson distribution with mean 20.

uncertainty therefore only needs to account for the fact that it would be determined by the raw number of events used, not the weighted number. This is done by using the square root of the sum of *squares* of weights. Eq. 6.1 considers the case of K distinct weights w_i carried by N_i events respectively. In the first term, these events are considered together as N events, with weights w_j. The equation shows that this procedure correctly yields the sum in quadrature of the individual uncertainty on each set of events with a particular weight.

$$\begin{aligned}\sqrt{\sum_{j=0}^{N} w_j^2} &= \sqrt{\sum_{i=0}^{K} w_i^2 N_i} \\ &= \sqrt{\sum_{i=0}^{K} \left(w_i \sqrt{N_i}\right)^2}\end{aligned} \tag{6.1}$$

The statistical uncertainty on the background estimate is more difficult to calculate. The background estimate is determined by weighting two tag events in the kinematic signal region, using weights produced by a neural network trained by comparing two tag and four tag events in the control region. Since any region contains about $100\times$ more two tag events than four tag events, the limiting factor (and hence the driving statistical uncertainty) is the number of four tag control region events. Simply using the squared sum of weights method would yield an uncertainty that reflected the number of two tag signal region events.

Naively, this uncertainty might be obtained by taking the reciprocal of the square-root of the number of four tag control region events falling into a di-Higgs mass bin as the fractional uncertainty for that bin. However, since the reweighting neural network makes use of a large set of variables (and in fact, m_{HH} is not one of these), this would not yield the correct uncertainty. In the face of this, this

analysis calculates the statistical uncertainty on the background estimate using the bootstrapping [68] technique.

The events in the kinematic control region are resampled with replacement, to produce 100 replicas, each with the same number of events as were originally in the pool. An ensemble of 100 neural networks is then trained, one with each of these replicas. The statistical uncertainty on each weight is then taken to be $1/2$ the interquartile range of the weights produced by the 100 neural networks. Similarly, the uncertainty on the normalization factor is $1/2$ the interquartile range of the normalization factors derived from each replica. In order to reduce fluctuations in the nominal weights (normalization factor), these are replaced with the median of the 100 weights (normalization factors) produced by the replica-trained neural networks.

These weights must then be combined to produce statistical uncertainties on histogram bins. Rigorously, this should be done by generating 100 histograms, one with each replica-trained neural network, which would produce a distribution of bin values for each bin. Unfortunately, this was found to be infeasible due to the computational and storage requirements.

A conservative approach is therefore adopted to combine the weights with statistical uncertainties. An alternate histogram is generated using nominal weight + interquartile range/2 as the weight for each event. In order to reduce the overestimation of the uncertainty, this histogram is scaled to the same yield as the nominal histogram before the nominal + interquartile range/2 normalization factor is applied. The statistical uncertainty on each bin of the nominal histogram is then set to the difference between the nominal value of that bin and the alternate value of that bin.

6.2 Systematic Uncertainties — Modelling

Systematic uncertainties, on the other hand, are those uncertainties that exist due to imperfections in techniques used in this analysis, or imperfect modelling. This section describes the process of correcting for, and calculating uncertainties due to, imperfect modelling in simulations, while the one after describes the uncertainty due to imperfection in the data-driven background model, which is the dominant systematic uncertainty in this analysis.

Systematic uncertainties are accounted for by introducing nuisance parameters, each of which have an up and down variation. Each of these nuisance parameters may represent a single uncertainty, a component of a single uncertainty, or a mixture of related (and partially correlated) uncertainties.

6.2.1 *Jet and Flavour Tagging Uncertainties*

The jet and flavour tagging uncertainties are the result of imperfect knowledge of the detector, leading to jet reconstruction and flavour tagging exhibiting differing performance on Monte Carlo simulations as opposed to on real data. These are corrected through calibrations derived from dedicated measurements, and the uncertainties in these calibrations are propagated to this analysis.

The results of the $HH \to b\bar{b}b\bar{b}$ analysis will later be combined with results from other di-Higgs channels, such as $HH \to b\bar{b}\gamma\gamma$ and $HH \to b\bar{b}\tau^-\tau^+$, in order to set a more accurate constraint on inclusive (all decay channel) HH production. In order to accurately perform this combination, information about correlations between jet uncertainties must be preserved even where those uncertainties are not dominant in this analysis. There are therefore a total of forty jet uncertainty

nuisance parameters in this analysis. These can be divided into two types: Jet Energy Scale (JES) and Jet Energy Resolution (JER).

Jet Energy Scale uncertainties are related to the jet calibration process. This process corrects reconstructed jet energies to what the corresponding truth jet energy would be, had the event been simulated. The calibration procedure is determined using Monte Carlo simulation, as described in [42]. This procedure corrects for additional energy from pile-up, energy missed in the detection process, and other differences between the measured jet energy and the true (particle-level) jet energy. It is a function of both jet properties and global data-taking conditions (in particular, pile-up conditions). An *in situ* calibration is also applied to data, using measurements taken by comparing data and simulation, to account for differences between the two.

The JES uncertainties originate from uncertainties in this calibration procedure due to statistical and theoretical uncertainties in the simulations used to derive them, and uncertainties in the measurements (in data) used to derive the in situ calibration. In total, this leads to 125 nuisance parameters [42], which would be evaluated in the $HH \to b\bar{b}b\bar{b}$ analysis by varying the nuisance parameter up or down by one standard deviation, and re-running the jet calibration process (and the remainder of the analysis) with each variation. This would lead to 250 variations, and 125 nuisance parameters in the statistical analysis. Though these variations are only applied to MC simulations, this would be computationally expensive, and is unnecessary for an analysis which is not a precision measurement. This analysis will, however, be statistically combined with other di-Higgs analyses (and the respective CMS analysis), which requires retaining sufficient information about correlations between these nuisance parameters. We therefore make use of the

category reduction, where these have been reduced to 30 orthogonal components which preserve the physical meaning necessary to allow statistical combinations.

JER, meanwhile, measures the precision with which jet energies can be measured. Where simulated jets have better resolution than jets found in data, the simulated jets are smeared to account for this difference. There are around ten JER nuisance parameters in the parameterization used in the $HH \to b\bar{b}b\bar{b}$ analysis, which are applied by varying the severity of this smearing, and re-running the remainder of the analysis for each variation.

Figure 6.2 shows the pulls and impacts of the ten most important jet systematics for the 500 GeV scalar mass point, evaluated on an $s+b$ Asimov dataset with $\mu = 1$. *Pulls* refer to the best fit value of each nuisance parameter $(\hat{\theta})$ and the uncertainty on this value. *Impacts* are the change in the best fit value of the signal strength μ when the nuisance parameter is fixed at $\pm 1\sigma$ (post-fit). The impact is shown as $\hat{\mu} - \mu_{\text{var}}$, the difference between the best-fit signal strength, and the varied best-fit signal strength.

It can be noted that these are almost all JER systematics, demonstrating that the jet energy resolution is the dominant systematic uncertainty on jets in this analysis. Figure 6.3 then shows the pre- and post-fit plots of the simulated signal for the five most important jet systematics, while Figure 6.4 shows the correlations between all jet systematics. Since these systematics are not constrained by the data, the pre- and post-fit variations are virtually identical. We can see from the correlation plot that these systematics are almost entirely uncorrelated.

It is also notable that the JER systematics appear one-sided. This is a result of these being inherently one-sided, since they are applied by smearing jet energies.

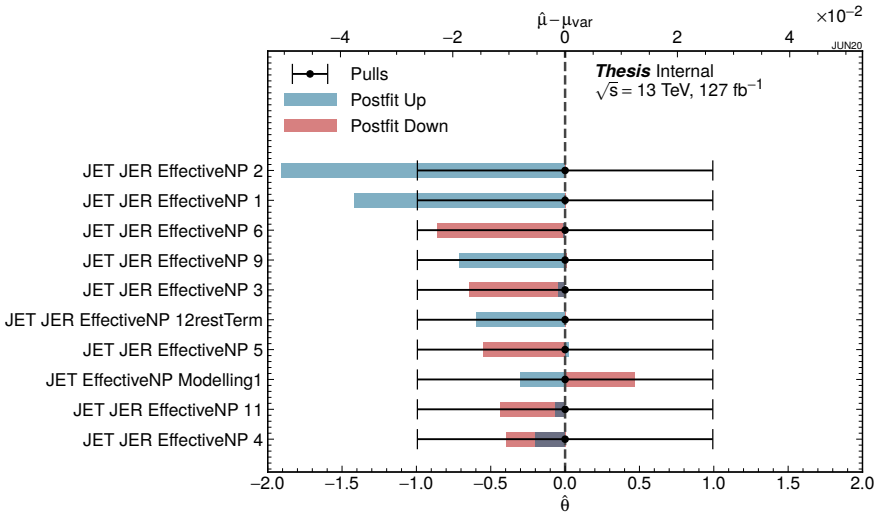
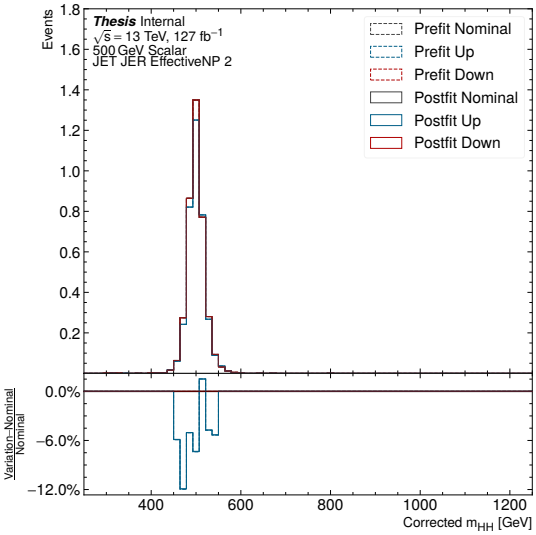
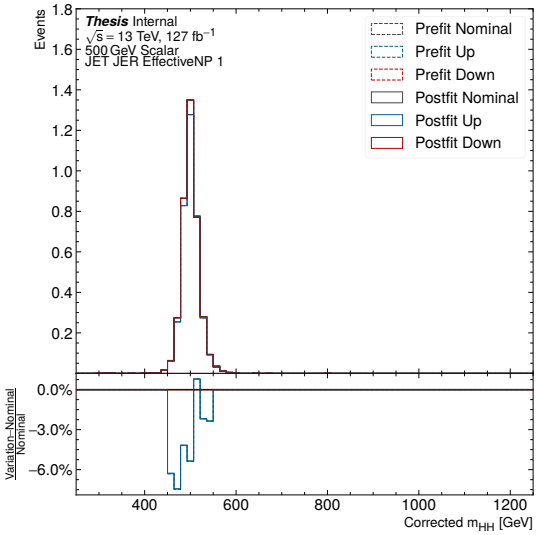


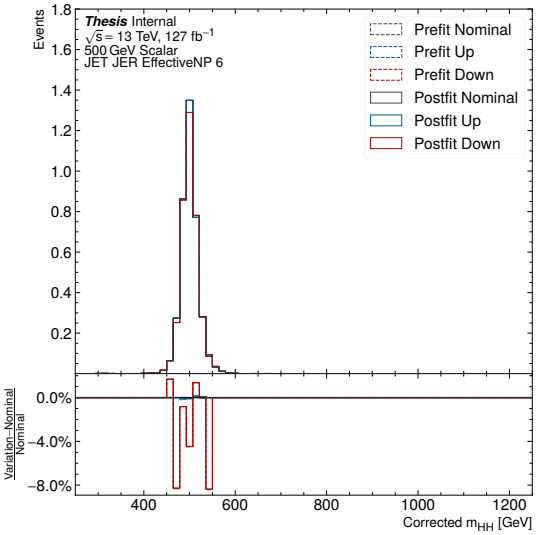
Figure 6.2: Pulls (black dots, with uncertainties) and impacts (pink and blue bars) of the ten most important (i.e. highest impact) jet systematics. The 500 GeV narrow-width scalar simulation is used as a benchmark. These are evaluated using an Asimov dataset generated with $\mu = 1$ (i.e. an $s + b$ Asimov dataset), where the reference signal cross-section is 1 fb. This is done in order to return meaningful impacts. We can note that almost all of these are JER systematics.



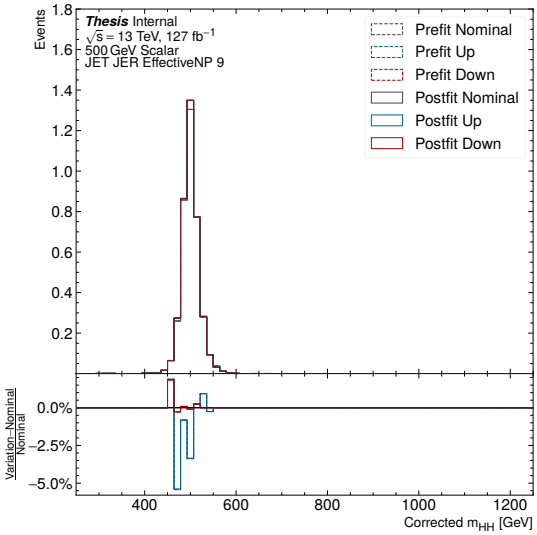
(**a**) JER EffectiveNP 2



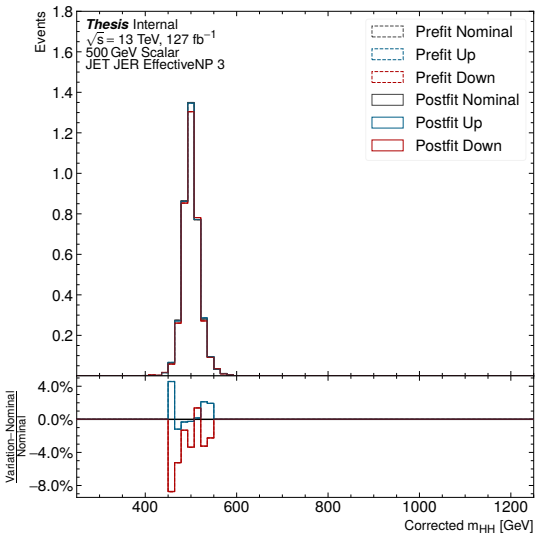
(**b**) JER EffectiveNP 1



(**c**) JER EffectiveNP 6



(**d**) JER EffectiveNP 9



(**e**) JER EffectiveNP 3

Figure 6.3: Pre- and post-fit plots of the 500 GeV narrow-width scalar signal, showing the five most important jet systematics. Bins where the nominal histogram contains fewer than 0.05 (weighted) events are suppressed in the ratio to avoid division by zero. The pre- and post-fit variations are virtually identical.

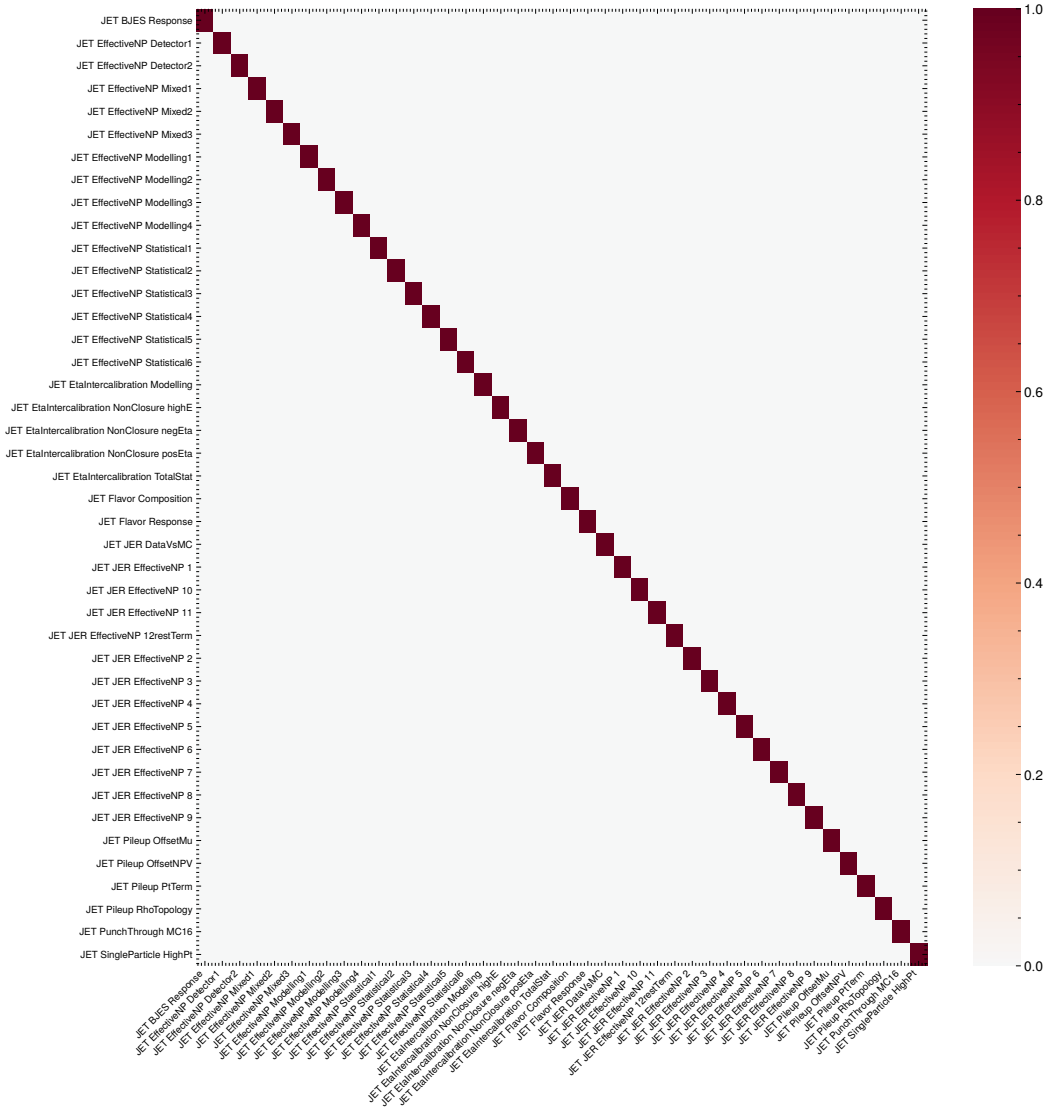


Figure 6.4: Correlations between jet systematics, evaluated on a 500 GeV narrow-width scalar simulation. It can be noted that these systematics are almost perfectly uncorrelated.

Flavour tagging is applied by checking whether the DL1r score for a jet exceeds the threshold set by the working point (in the $HH \to b\bar{b}b\bar{b}$ analysis, the 77% efficiency working point). If it does, the jet is tagged, and if it does not, the jet is not tagged. The efficiency of this working point is measured using the b-jets in a simulated $t\dot{t}$ sample [44], where parton showering and hadronization is done using Pythia. This is not necessarily the same as the efficiency in data, or in a simulation where parton showering and hadronization is done using a different program (e.g. Herwig). Similarly, the mistag rate of charm and light jets will not necessarily match between data and Monte-Carlo simulation, or between simulations with different parton showering programs.

These differences are accounted for by applying a scale factor to the event weight of simulated events. This event-level scale factor is calculated by finding the product of the per-jet tag (for tagged jets) or anti-tag (for untagged jets) scale factor over every jet considered for b-tagging. In the $HH \to b\bar{b}b\bar{b}$ analysis, this is every jet with $p_{\mathrm{T}} \geq 40\,\mathrm{GeV}$, $\left|\eta\right| < 2.5$, and p_{T} greater than or equal to that of the fourth hardest (i.e. fourth highest p_{T}) tagged jet.

The per-jet scale factors include an Monte-Carlo–Monte-Carlo component for simulations where the parton showering program used was not Pythia, and a data-MC component calculated using a highly $t\dot{t}$ enriched sample selected in data. The calibration procedure for Run 2 is fully detailed in [44].

The flavour tagging uncertainties, then, originate from the uncertainties in the measurements used to derive these scale factors. These are propagated to the $HH \to b\bar{b}b\bar{b}$ analysis as variations on the per-jet tag and anti-tag scale factors. Different variations lead to nine nuisance parameters related to correctly tagging true b-jets,

four related to mistagging of true charm jets, and four related to mistagging true light jets. Each of these sets of nuisance parameters is determined through an eigenvector decomposition of the uncertainties on the calibration analysis. There are also two nuisance parameters relating to extrapolation of these scale factors outside the calibrated p_{T} range.

Figure 6.5 shows the pulls and impacts of the ten most important flavour tagging systematics for the 500 GeV scalar mass point, evaluated on an $s+b$ ($\mu = 1$) Asimov dataset. It is notable that the most of these are eigenvectors of the b-jet efficiency related uncertainties.

Figure 6.6 then shows the pre- and post-fit plots of the simulated signal for the five most important systematics, while Figure 6.7 shows the correlations between all flavour tagging systematics. Figure 6.6 shows that the pre- and post-fit variations are virtually identical, while Figure 6.7 shows that these systematics are again almost entirely uncorrelated.

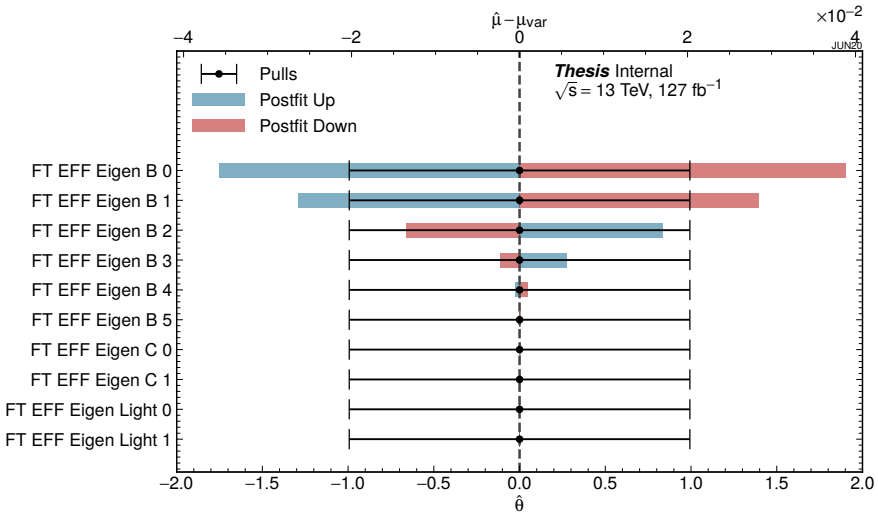
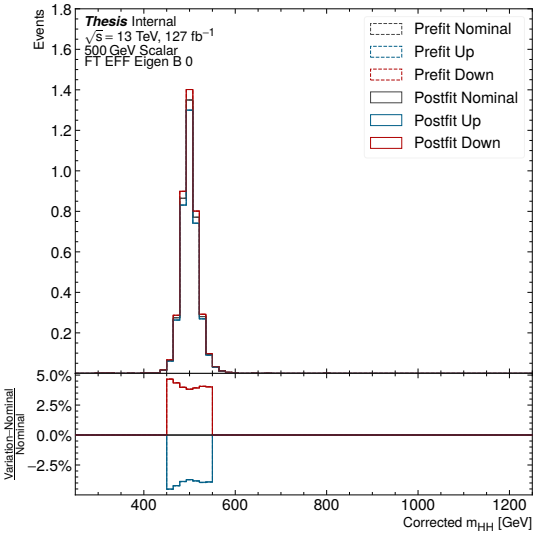
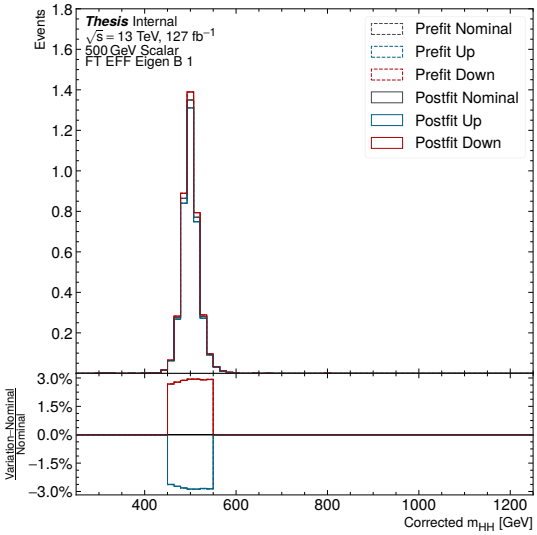


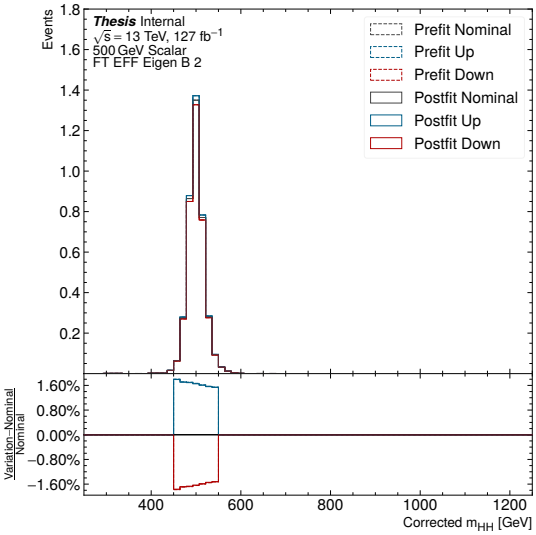
Figure 6.5: Pulls and impacts of the ten most important (i.e. highest impact) flavour tagging systematics. The 500 GeV narrow-width scalar simulation is used as a benchmark. These are evaluated using an Asimov dataset generated with $\mu = 1$ (i.e. an $s + b$ Asimov dataset), where the reference signal cross-section is 1 fb. This is done in order to return meaningful impacts. Here, it is notable that most of these are eigenvectors of the b-jet efficiency related uncertainty.



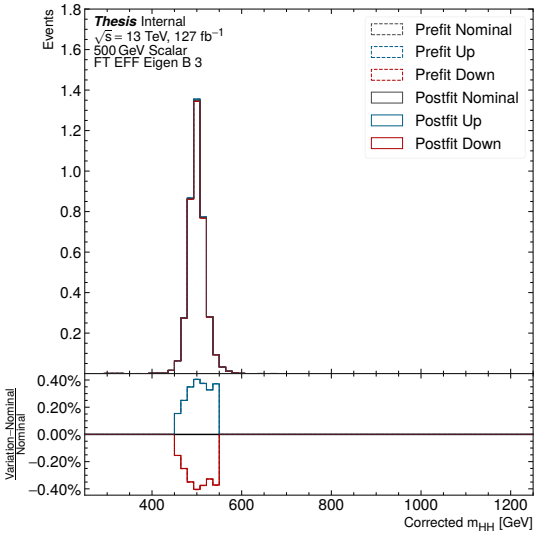
(**a**) Flavour Tagging B EigenNP 0



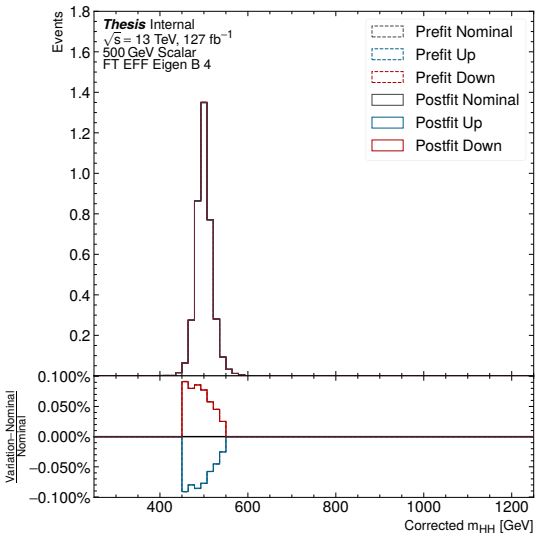
(**b**) Flavour Tagging B EigenNP 1



(**c**) Flavour Tagging B EigenNP 2



(**d**) Flavour Tagging B EigenNP 3



(**e**) Flavour Tagging B EigenNP 4

Figure 6.6: Pre- and post-fit plots of the 500 GeV narrow-width scalar signal, showing the five most important flavour tagging systematics. Bins where the nominal histogram contains fewer than 0.05 (weighted) events are suppressed in the ratio to avoid division by zero. The pre- and post-fit variations are virtually identical.

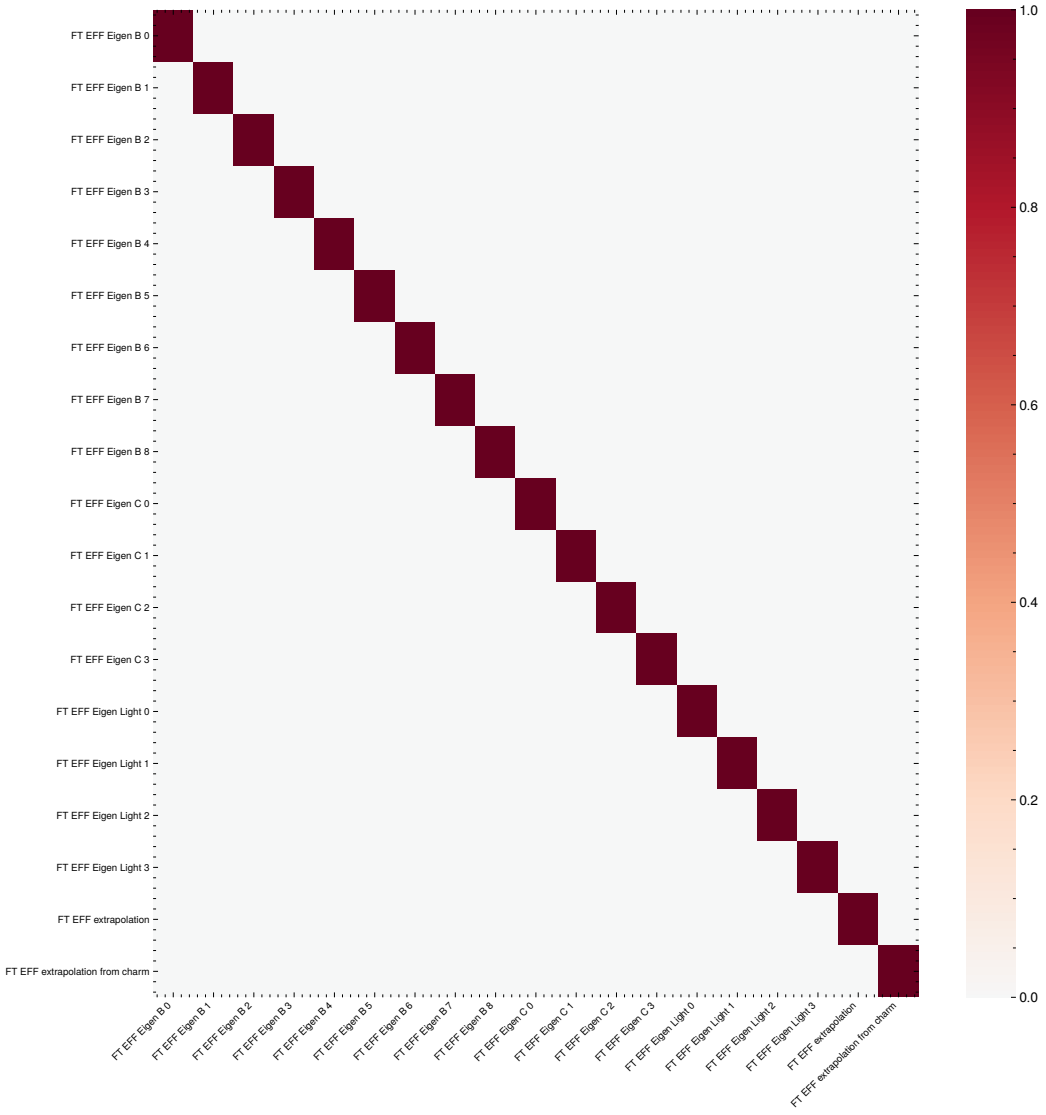


Figure 6.7: Correlations between flavour tagging systematics, evaluated on a 500 GeV narrow-width scalar simulation. It can be noted that these systematics are almost perfectly uncorrelated.

6.2.2 *Theory Uncertainties*

Finally, there is a modelling uncertainty due to imperfect knowledge of simulated signal process, and in particular the QCD aspects of this process. The uncertainty due to imperfect knowledge of the proton parton distribution function is accounted for by varying this parton distribution function, while the uncertainty due to imperfect knowledge of the showering and hadronization of the b-quarks in the final state is accounted for using alternate signal simulations where this step is calculated using Pythia in place of Herwig. There are a number of differences [75] between these generators, notably that Pythia uses p_{T}-ordered parton showers and a string [76] hadronization model, while Herwig uses angular-ordered parton showers and a cluster [77] hadronization model. As a result, comparing these simulations can give a qualitative measure of the theoretical uncertainties associated with these choices. Following studies of these alternate simulations, a 5% normalization uncertainty was applied to the signal simulations to account for these theory uncertainties.

An additional 1.7% normalization uncertainty was also applied to the signal simulations to account for the uncertainty in the measurement of the integrated luminosity of the collected data.

6.3 Systematic Uncertainties — Background Estimation

This section is also included in the ATLAS internal documentation for this analysis (which I co-edited) as an appendix [78].

The dominant systematic uncertainty in the $HH \to b\bar{b}b\bar{b}$ analysis, however, stems from the fully data-driven background estimation procedure. The uncertainty due to this procedure can be broken down into:

- Non-closure due to limited statistics in the data used to train the neural networks. This is smoothed and accounted for through the bootstrapping procedure used to derive the statistical uncertainty on the background estimate, as described in Section 6.1.
- An uncertainty on the normalization of the background estimate.
- The uncertainty on the shape of the background estimate.

There is no *a priori* "correct" way to measure the systematic uncertainty on the normalization and shape of the background estimate. In the $HH \to b\bar{b}b\bar{b}$ analysis, this is done by building an alternative background model, where the neural network is trained in the kinematic validation region.

Simply applying a two-point systematic to account for the total systematic uncertainty in the normalization and shape of the background estimate would severely limit the space of deviations from the nominal estimate that can be modelled. The uncertainty on the background estimate for each year is therefore decomposed into multiple nuisance parameters. Two ways of conducting this decomposition were studied.

In addition, separate sets of background estimate uncertainty nuisance parameters are used for each year of data-taking, since the background estimation is itself conducted separately on each year. This is because the data-taking conditions and triggers varied significantly between the three years.

6.3.1 *Low / High H_T Decomposition*

One of the decompositions studied was that adopted in the Early Run 2 analysis.

The background shape uncertainty is separated into a low H_T and high H_T region. For the purpose of defining this uncertainty, H_T is defined as the scalar sum of the transverse momenta of the four Higgs constituent jets. Events with $H_T \leq 300\,\text{GeV}$ are considered *low* H_T events, while those with $H_T > 300\,\text{GeV}$ are considered *high* H_T events.

In the pre-fit one sigma variation of the low H_T nuisance parameter, the alternate (validation region derived) background weight is used for low H_T two tag events while the nominal (control region derived) weight is used for high H_T two tag events.

The pre-fit one sigma variation of the *high* H_T nuisance parameter is generated similarly, but here the alternate weight is used for high H_T two tag events, while the nominal weight is used for low H_T two tag events.

This split on H_T effectively allows the background shape uncertainty on low m_{HH} events to be decoupled from that on high m_{HH} events, without causing a discontinuity in this uncertainty. The split arose after such a coupling across the m_{HH} range was found to be an issue in the Early Run 2 analysis.

In addition to these shape uncertainty nuisance parameters, an additional overall normalization uncertainty is added to account for additional uncertainty in the background normalization. The pre-fit one sigma variation of this uncertainty is set to the ratio between the residual normalization correction for the alternate background model and that for the nominal background model.

Figure 6.8 shows the pulls and impacts of the background estimation nuisance parameters in the low / high H_T decomposition scheme. As before, for the purposes of illustration, these are evaluated using the 500 GeV narrow-width scalar as a benchmark. Here we see striking differences compared to the modelling uncertainties.

The impacts are orders of magnitude larger than the impacts of the modelling uncertainties. This is because the background cross-section is itself orders of magnitude larger than the reference signal cross-section (1 fb). As a result, an x% uncertainty on the background estimate can be much larger than the reference signal cross-section, and therefore have a large impact on the best-fit μ. Meanwhile, an x% uncertainty on the signal will only have an impact of about x% on the best-fit μ. As a result, the dominant uncertainties in this analysis are those on the background estimate.

Since this is a pre-unblinding evaluation of these pulls on background-only Asimov data, these nuisance parameters are not pulled. They are, however, significantly constrained by the fit in the signal region. Unlike the modelling uncertainties, where the one sigma deviations have been determined through calibration analyses, and are well understood, the background estimation uncertainties are evaluated in a somewhat *ad hoc* manner. The constraints here reflect an overestimation of the nominal one sigma deviations of these nuisance parameters.

Figure 6.9 shows the correlations between the background estimation nuisance parameters in this decomposition scheme. Here we see that there are non-negligible correlations between these nuisance parameters, where they relate to the same data-taking year. This is a result of their being raw nuisance parameters, with

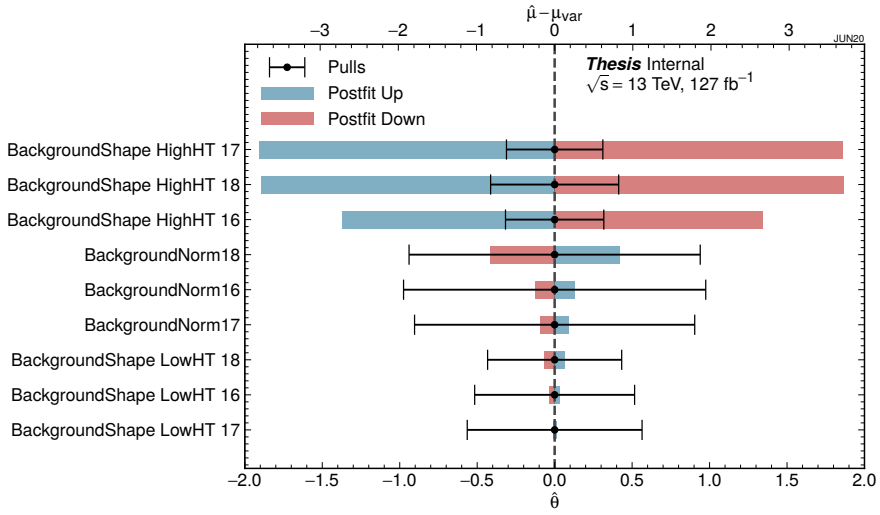


Figure 6.8: Pulls and impacts of the nine background estimation nuisance parameters in the low / high H_T decomposition scheme. For the 500 GeV mass point, the dominant uncertainties are the high H_T uncertainties.

no eigenvector decomposition performed. This is done in order to preserve the physical meaning of these nuisance parameters.

Figure 6.10 then shows the pre- and post-fit distributions of m_{HH} for each of these nuisance parameters.

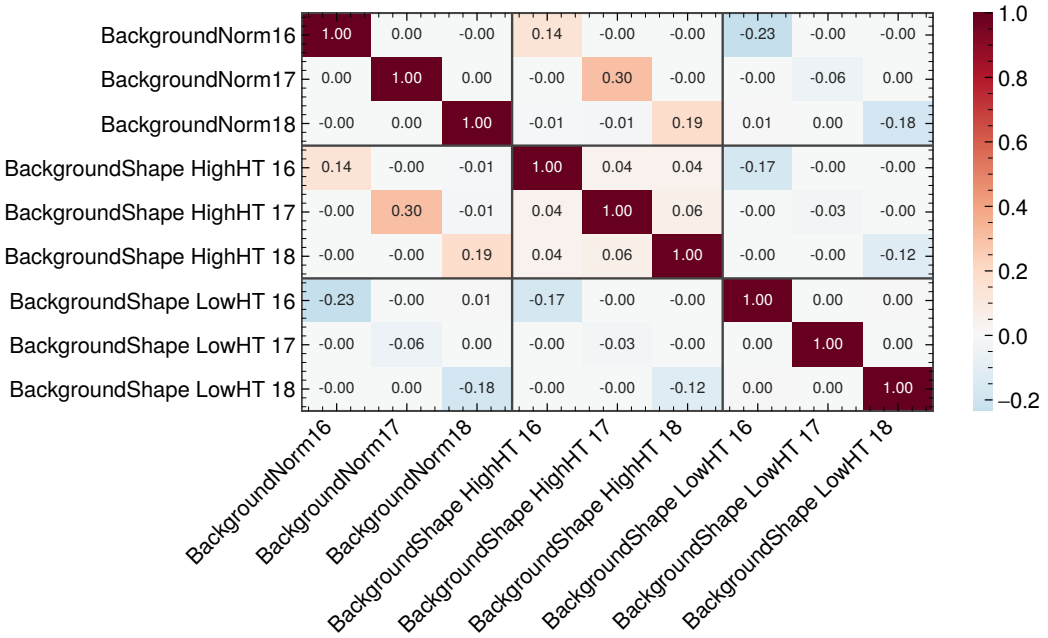
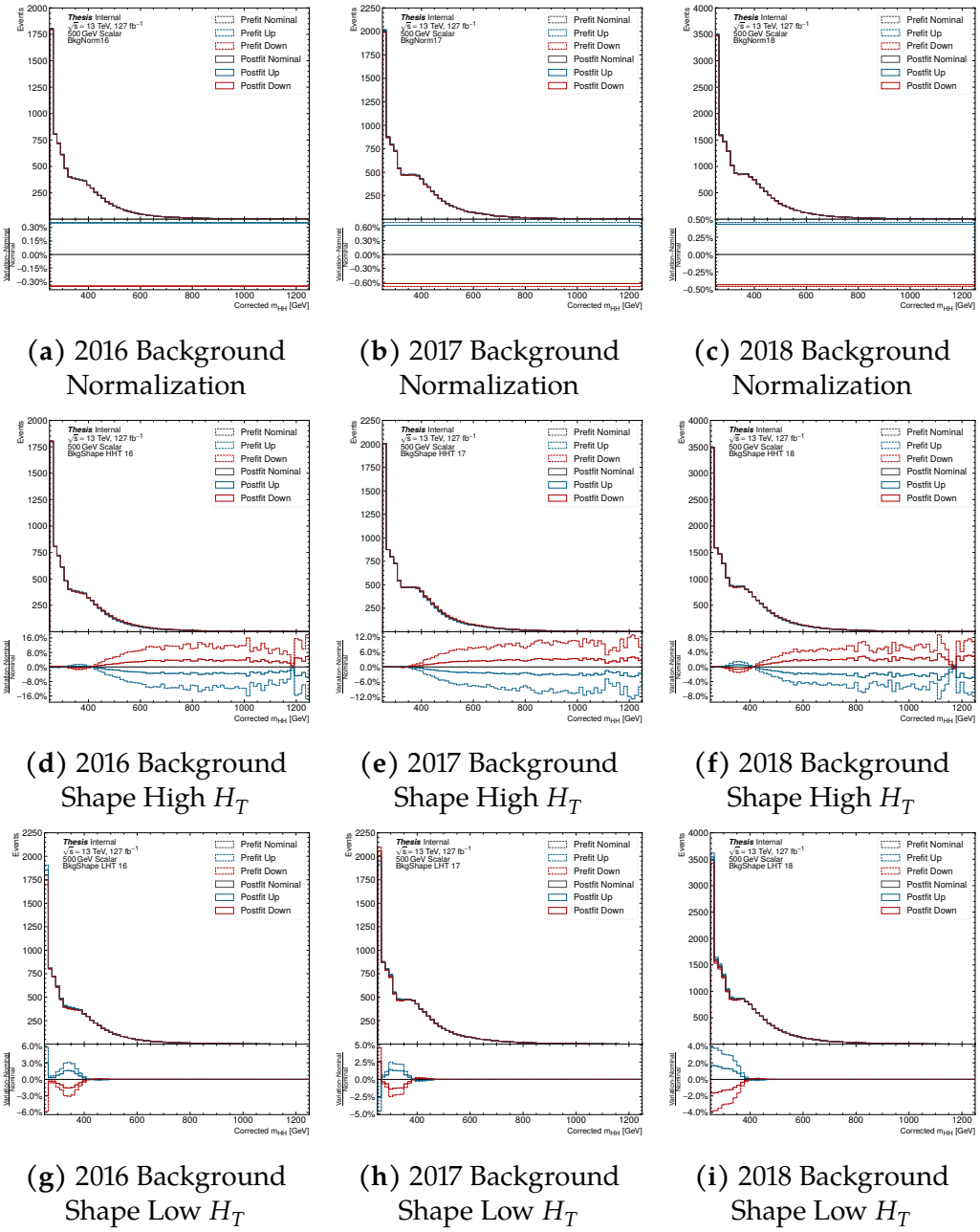


Figure 6.9: Correlations between the background estimation nuisance parameters in the low / high H_T decomposition scheme. Here we find that there are non-negligible correlations between nuisance parameters relating to the same year. This is because no eigenvector decomposition is performed on these nuisance parameters, in order to preserve their physical meaning.



(**a**) 2016 Background Normalization

(**b**) 2017 Background Normalization

(**c**) 2018 Background Normalization

(**d**) 2016 Background Shape High H_T

(**e**) 2017 Background Shape High H_T

(**f**) 2018 Background Shape High H_T

(**g**) 2016 Background Shape Low H_T

(**h**) 2017 Background Shape Low H_T

(**i**) 2018 Background Shape Low H_T

Figure 6.10: Pre- and post-fit m_{HH} distributions for the background estimation nuisance parameters in the low / high H_T decomposition scheme.

6.3.2 *Orthogonal Polynomial Decomposition*

While the separation of the background estimation uncertainty into a normalization term, and separate low and high H_T shape terms does allow the modelling of a wider range of deviations from the nominal model than a single nuisance parameter, within each of these H_T regions, a strong assumption is still being made about the precise shape of the deviation.

In order to weaken the assumption made about the shape of this deviation, I evaluated modelling this uncertainty using a polynomial decomposition.

This decomposition requires the m_{HH} distribution from the nominal background model, and an alternate m_{HH} distribution. While developing the strategy, I worked in the kinematic validation region, using four tag data as the alternate distribution. Development was also done using only data taken in 2017.

Later, when implementing the strategy, I returned to using the neural network background model derived in the kinematic validation region as the alternate distribution, in order to work in the kinematic signal region. The strategy was also implemented for all years of data-taking.

With the nominal m_{HH} distribution denoted f_i, and the alternate m_{HH} distribution denoted g_i, where i is an index over the bins, the objective is to model

$$\frac{g_i - f_i}{f_i} \tag{6.2}$$

using a polynomial. If a degree d polynomial is used, there will be $d + 1$ terms $p_{j\in[0,d]}(m_{HH})$, with coefficients c_j. These coefficients can be determined by carrying out a least squares fit to the ratios described by Eq. 6.2. Each of these terms gives

rise to a nuisance parameter, and the pre-fit one sigma variation on this nuisance parameter is given (for bin i) by

$$\Big(1 + c_j p_j(m_{HH})\Big) f_i \tag{6.3}$$

The f_i factor in Eq. 6.3 converts a variation on the ratio to a variation on the background estimate itself.

This decomposition interacts well with HistFactory's implementation of systematic uncertainties, where the nuisance parameter α is a multiplicative factor on the difference between the pre-fit one sigma variation and the nominal value of the background estimate. As a result, the RooStats fit in the signal region will, in essence, adjust the coefficients of the polynomial within the standard Gaussian constraints.

There are an infinite number of options for the specific basis used for this decomposition (that is, the individual terms p_j). The simplest is to use

$$\Big\{1, x, x^2, x^3, \ldots\Big\} \tag{6.4}$$

where x could be i, or m_{HH} at the bin centres. However, these terms do not form an orthogonal basis. As a result, they can produce extremely large uncertainties at the extremes of the m_{HH} range. This can be seen in Figure 6.11 (though here, the polynomial is fit to the difference $g_i - f_i$), where we see that the individual terms (with the exception of the constant term P_0) grow very large (though with large cancellations) at high m_{HH}.

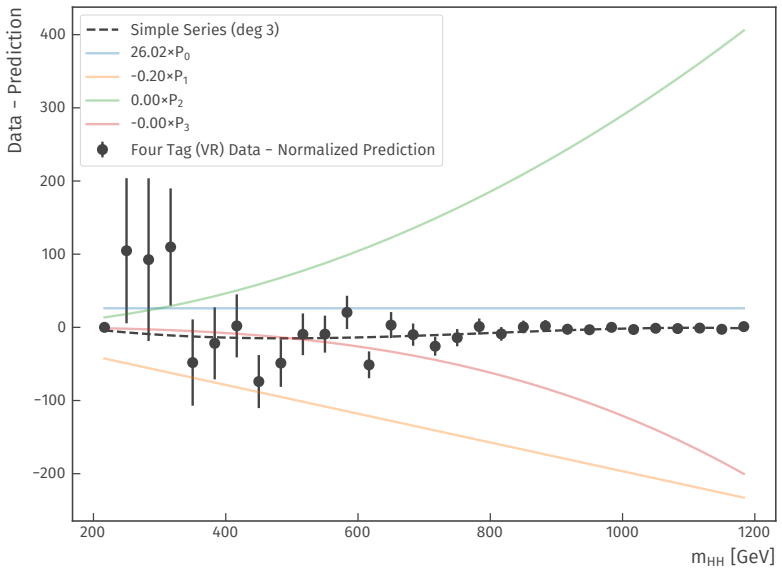


Figure 6.11: Individual terms when a degree three polynomial is fit to the difference between the nominal background estimate (prediction) and the alternate m_{HH} distribution (four tag (VR) data). Note that the individual terms are very large at high m_{HH}.

This property can be avoided by making use of an orthogonal set of polynomials. Since the m_{HH} distributions are represented by histograms, the appropriate set of polynomials would be the relevant Gram [79] polynomials (also known as the discrete Chebyshev polynomials). Since we only intend to use a low degree polynomial series (much smaller than the number of bins), the Legendre polynomials are a good approximation to these. The Legendre polynomials have the advantage of not depending on the number of bins. These are defined [80] such that:

$$\int_{-1}^{1} p_m(x) p_n(x)\,\mathrm{d}x = \frac{2}{2n+1}\delta_{mn}$$

Since the Legendre polynomials are defined to be orthogonal on the range (-1, 1), I remapped the bin centres to this range. That is, in place of $P_j(m_{HH})$, I used $P_j(ax+b)$, where the scale a and offset b are determined such that the centre of the first bin maps to -1 and the centre of the last bin maps to 1.

Figure 6.12 shows the individual terms, along with the total prediction from the series, when Legendre series of degree three and five are fit to the ratio in Eq. 6.2. Even with a degree three polynomial, most data points are within the statistical error bars of the prediction from the series. This demonstrates that a degree three polynomial is sufficient to model the systematic components in this ratio. By using the lowest degree polynomial that is sufficient, I can reduce the ability of the polynomial to fit away a real signal.

This is confirmed in Figure 6.13, where I have injected small Gaussian (width 50 GeV) signals at 400 GeV and 900 GeV. Comparing to Figure 6.12a, we see that the points where signal has been injected are further from the series total (centre line in the ratio plot) than the corresponding points where signal has not been

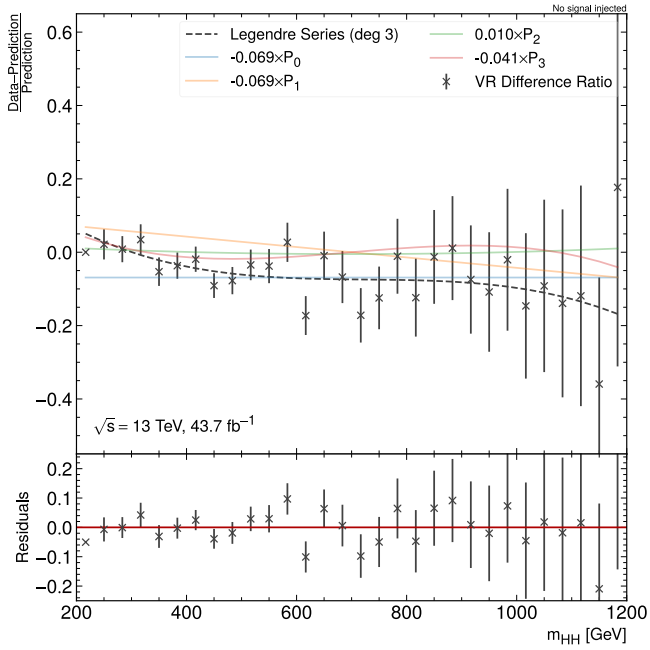
injected. This demonstrates that these signals are not entirely fit away, even when the full polynomial fit is performed with the signals injected. When the polynomial components are varied as nuisance parameters, it is even less likely that a signal would be fit away, since the nuisance parameters are subject to Gaussian constraints, and the fit is performed simultaneously with the signal strength parameter μ.

I then implemented this strategy for use in the $HH \to b\bar{b}b\bar{b}$ analysis. Since the limit setting procedure is done using C++ code, I implemented the algorithm in NumPy's [81] `legfit` function in C++ using the Eigen [82] library. This is used to determine the coefficients to be used by fitting a degree three Legendre series to the ratio in Eq. 6.2, with the nominal m_{HH} distribution supplied by the nominal control region derived background model, and the alternate distribution supplied by the validation region derived background model. The one sigma variations on each nuisance parameter supplied to HistFactory are then the product in Eq. 6.3, with $p_j(m_{HH})$ replaced with $P_j(ax + b)$.

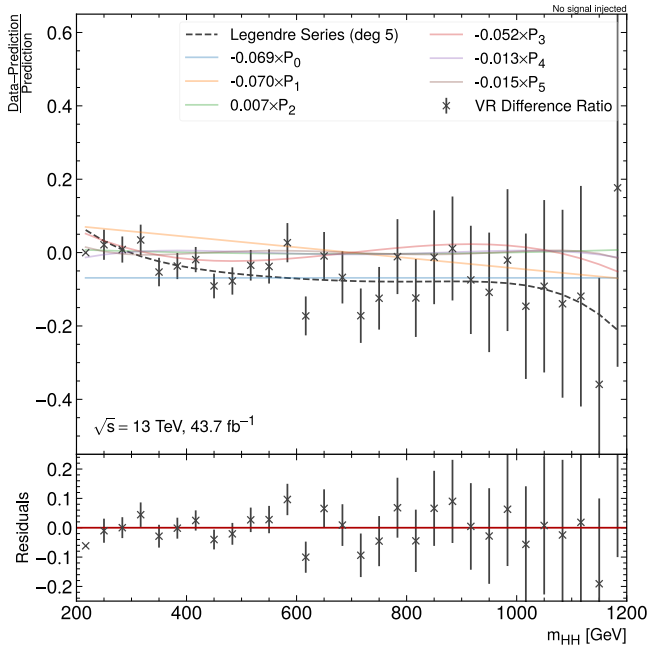
No normalization nuisance parameter is used with this decomposition since the 0th nuisance parameter itself acts as a normalization term.

Figure 6.14 shows the pulls and impacts associated with these nuisance parameters. Again, these are evaluated using a background-only Asimov dataset using the 500 GeV narrow-width scalar as a benchmark. The nuisance parameters are less constrained than with the low / high H_T decomposition strategy.

It is notable that NP0, the nuisance parameter corresponding to the constant term, is strongly constrained. While this nuisance parameter acts as a normalization term, its nominal one sigma variation is not the normalization ratio between the validation derived and control region derived background models. It cannot,



(**a**) Degree 3 Legendre Series



(**b**) Degree 5 Legendre Series

Figure 6.12: Individual terms when a Legendre series of degree three (a) or five (b) is fit to the ratio in Eq. 6.2. The dashed line shows the total prediction from this series, and the ratio demonstrates that, in both cases, the remaining variation fits within the statistical error bars.

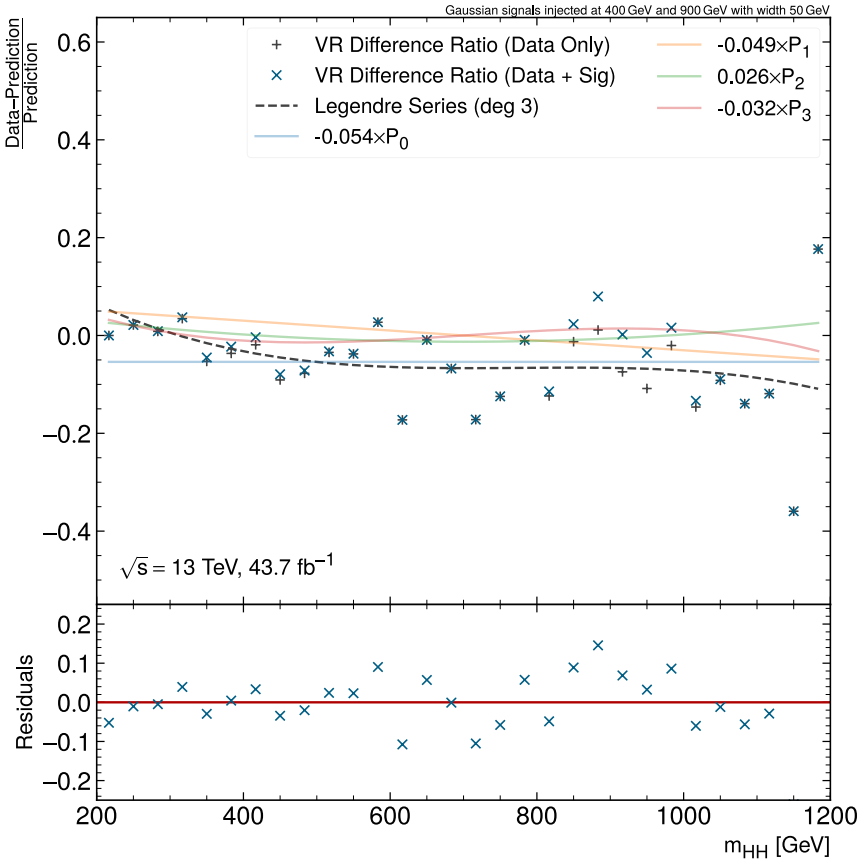


Figure 6.13: A Legendre series of degree three fit to the ratio from Eq. 6.2, with small signals injected at 400 GeV and 900 GeV. Comparison with Figure 6.12a shows that those signals are not fit away by the polynomial.

therefore, be considered equivalent to the normalization nuisance parameter in the low / high H_T decomposition.

NP0 is, in fact, the mean ratio of bin values over the bins in the histogram. The normalization ratio would be the mean ratio of bin values weighted by the number of events in each bin. The higher terms in the polynomial (in particular the term correcting for an overall trend, NP1) adjust for any discrepancy this may cause. In Figure 6.12a, we see that NP1 cancels NP0 at low m_{HH}, and adds to NP0 at high m_{HH}. It is for this reason that the correlation matrix shown in Figure 6.15 shows significant correlations, particularly between NP0 and NP1, despite these polynomials being globally orthogonal.

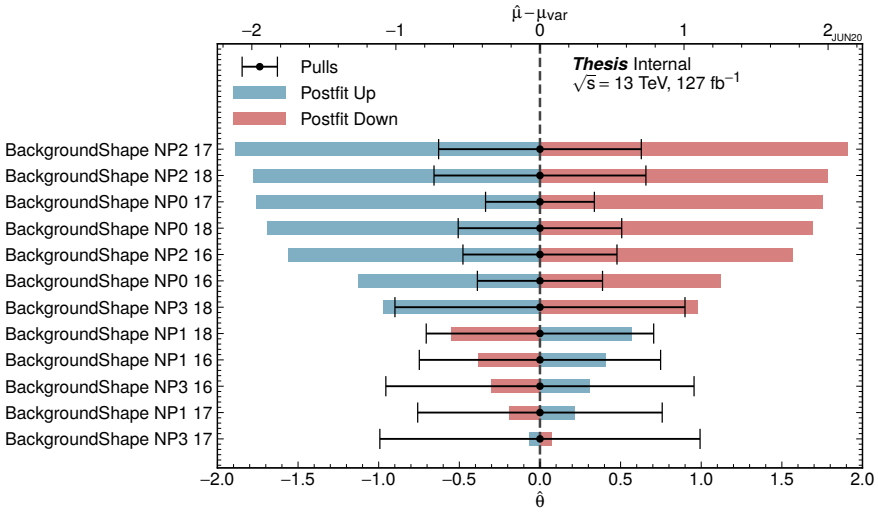


Figure 6.14: Pulls and impacts of the background estimation nuisance parameters in the orthogonal polynomial decomposition. Evaluated using a b-only Asimov dataset on the 500 GeV mass point scalar.

Figure 6.16 then shows the pre- and post-fit m_{HH} distributions with each nuisance parameter.

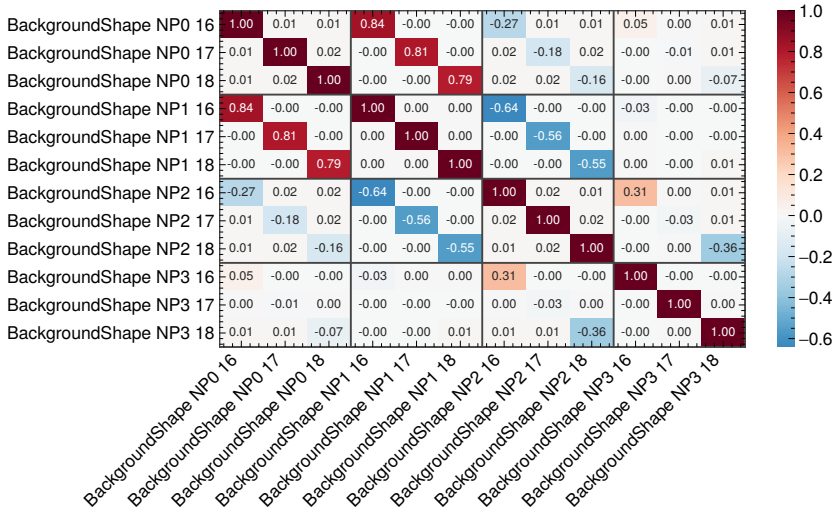
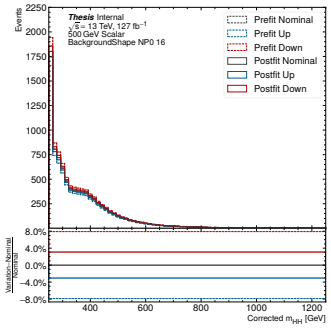
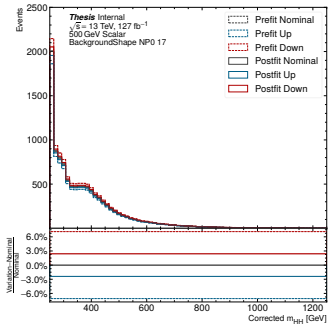


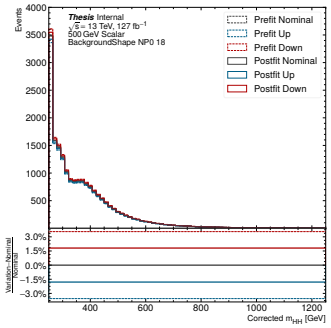
Figure 6.15: Correlations between the background estimation nuisance parameters in the orthogonal polynomial decomposition. Evaluated using a *b*-only Asimov dataset on the 500 GeV mass point scalar.



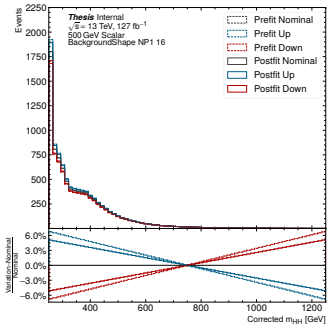
(**a**) 2016 Background Shape NP0



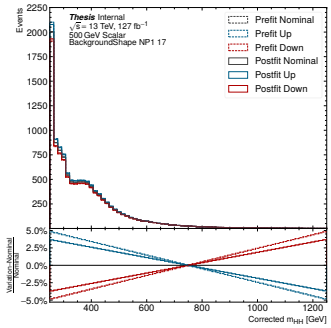
(**b**) 2017 Background Shape NP0



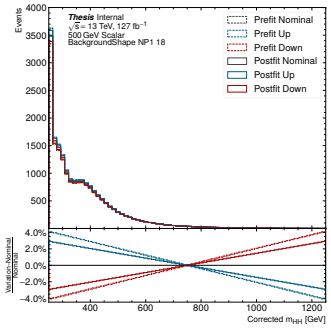
(**c**) 2018 Background Shape NP0



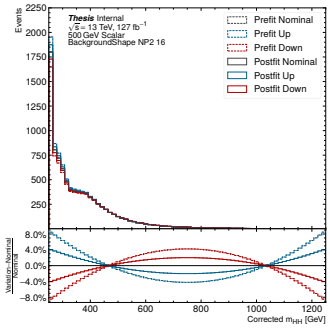
(**d**) 2016 Background Shape NP1



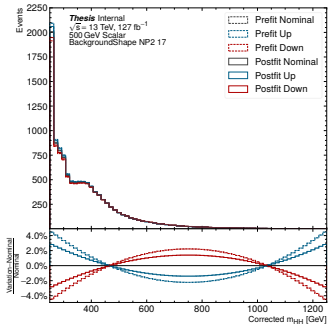
(**e**) 2017 Background Shape NP1



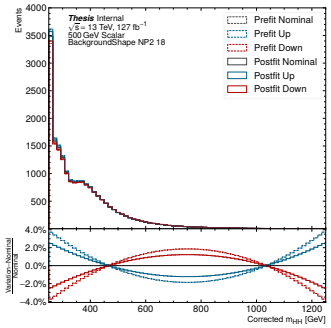
(**f**) 2018 Background Shape NP1



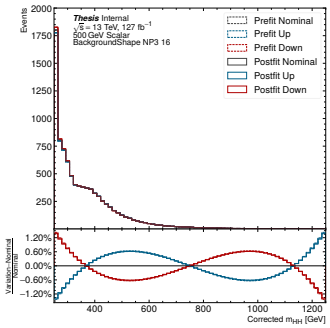
(**g**) 2016 Background Shape NP2



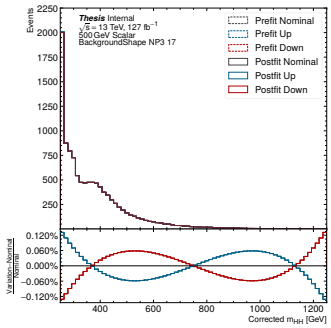
(**h**) 2017 Background Shape NP2



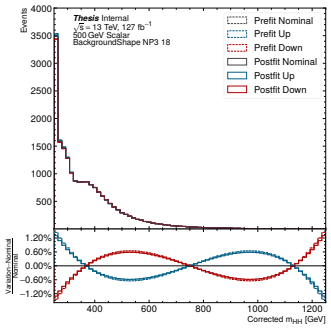
(**i**) 2018 Background Shape NP2



(**j**) 2016 Background Shape NP3



(**k**) 2017 Background Shape NP3



(**l**) 2018 Background Shape NP3

Figure 6.16: Pre- and post-fit m_{HH} distributions with uncertainties due to each nuisance parameter in the orthogonal polynomial decomposition scheme for background estimation uncertainties.

6.3.3 *Conclusion*

With two possible decompositions for the systematic uncertainty on the shape of the background estimate, it is necessary to choose one to be used in the analysis. In order to do this, I produced expected limits using background-only Asimov data. The results of this are given in Figure 6.17. Since the polynomial systematic decomposition is more flexible than the low / high H_T decomposition, it carries the risk that uncertainties may be seriously overestimated. As this figure shows, the orthogonal polynomial scheme causes no significant deterioration in the expected limit.

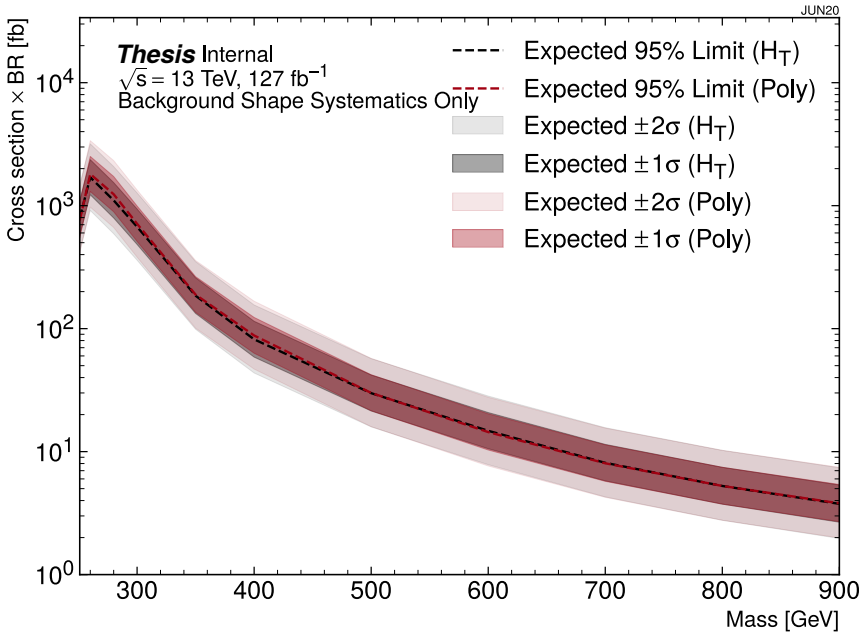
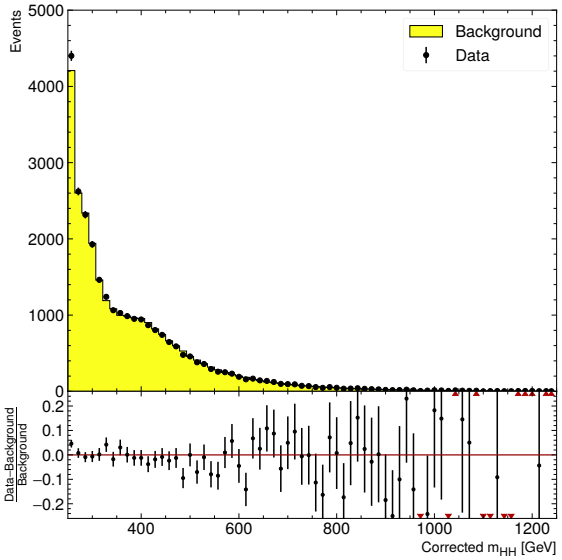


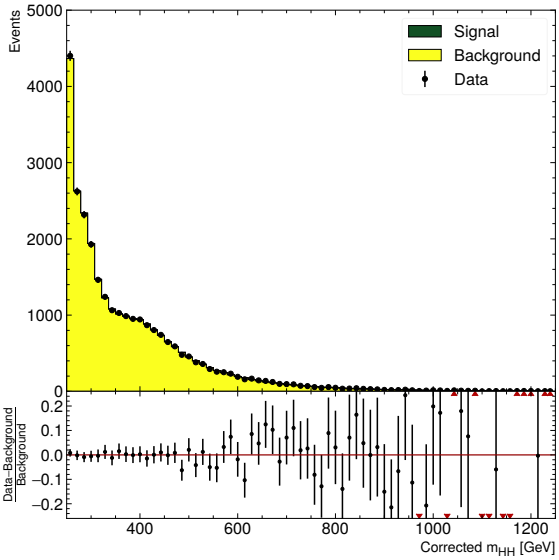
Figure 6.17: 95% confidence level limits with only statistical and background estimation uncertainties included, using the low / high H_T and orthogonal polynomial decomposition schemes. The orthogonal polynomial scheme do not cause a significant deterioration in the expected limit.

In order to make the final decision, I produced post-fit plots of the background and predicted signal distribution in the kinematic validation region, under the two decomposition schemes. These show the final best-fit distribution given the four tag data in the validation region, with the signal strength μ and nuisance parameters set to their best-fit values. This was done with the 280 GeV scalar signal simulation, since it was at this mass point that the Early Run 2 analysis found a 3.6σ (local significance) excess [58].

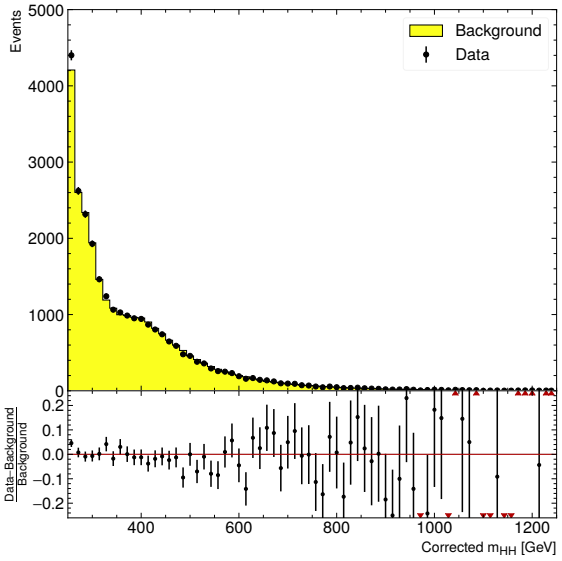
Figure 6.18 shows the result, which demonstrates that both schemes appear to be able to cover the systematic uncertainty due to the background estimate. As a result, at this (pre-unblinding) stage, I decided to use the low / high H_T scheme (which is simpler) as the nominal decomposition. In the event that a deviation like that found in the Early Run 2 analysis appears again in this analysis, the orthogonal polynomial scheme can then also be used to verify the deviation.



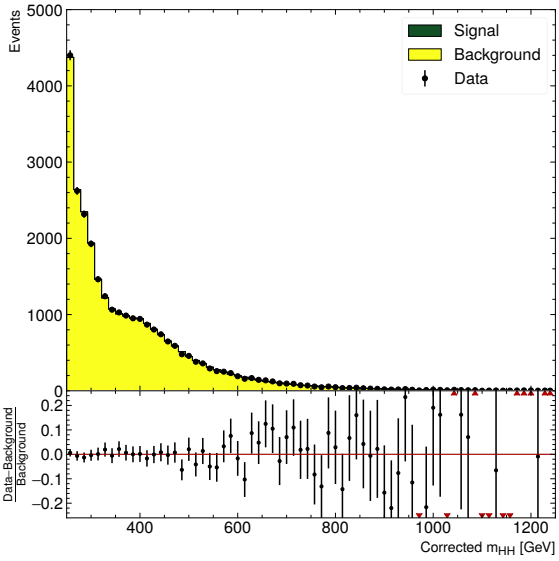
(**a**) 2018 data pre-fit corrected m_{HH} distribution, using the low / high H_T decomposition scheme.



(**b**) 2018 data post-fit corrected m_{HH} distribution, using the low / high H_T decomposition scheme.



(**c**) 2018 data pre-fit corrected m_{HH} distribution, using the orthogonal polynomial decomposition scheme.

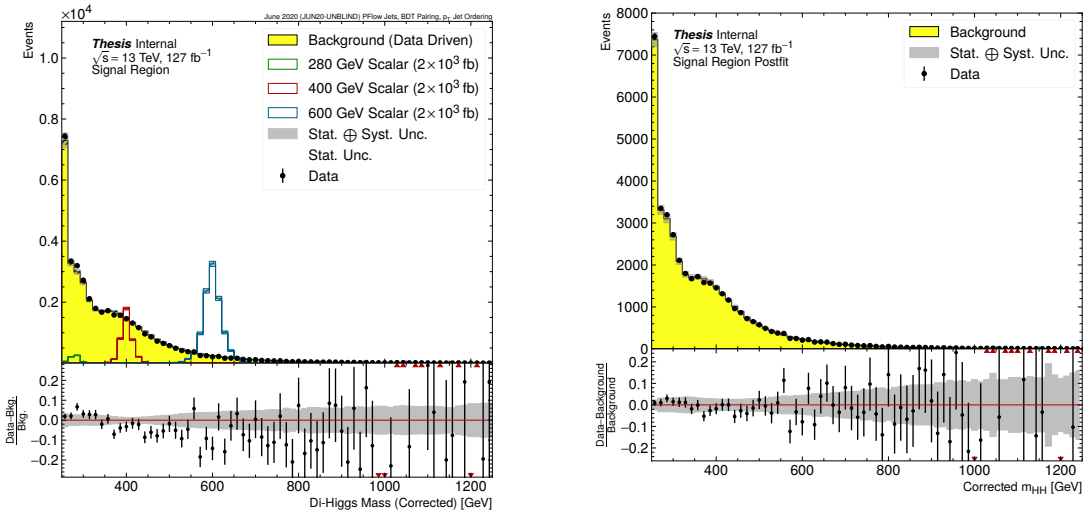


(**d**) 2018 data post-fit corrected m_{HH} distribution, using the orthogonal polynomial decomposition scheme.

Figure 6.18: Pre- and post-fit corrected m_{HH} distributions for 2018 data in the validation region, using the two background estimate systematic uncertainty decomposition schemes. These fits were produced using the 280 GeV reference signal. Both schemes are able to cover the systematic uncertainties.

7 Unblinded Results

Figure 7.1 shows the pre- and post-fit corrected m_{HH} distribution in the signal region, now with the data shown. Comparing these, we can see that the nominal background estimate underestimates the data at low masses, and overestimates at high masses. The fit corrected for these by "pulling" the background modelling nuisance parameters (i.e. setting these to non-zero values).



(**a**) Pre-fit (**b**) Post-fit

Figure 7.1: Pre- and post-fit corrected m_{HH} distributions in the signal region. A selection of signals is illustrated on the pre-fit plot as a guide.

These pulls can be seen (for the 500 GeV signal fit) in the background modelling pull and impact plot in Figure 7.2. The background shape nuisance parameters are pulled toward 1.0. This is understandable because these uncertainties are quantified by the difference between the background model from neural networks trained in the control region and that from neural networks trained in the validation region. Since the signal region is immediately surrounded by the validation region

(which is itself surrounded by the control region), it would be more similar to the validation region, leading to these pulls.

The impact plot shows that, as in the Asimov fit prior to unblinding (Figure 6.8), the high H_T nuisance parameters have by far the largest impact. The constraints on the nuisance parameters are also consistent with what was observed on the Asimov dataset.

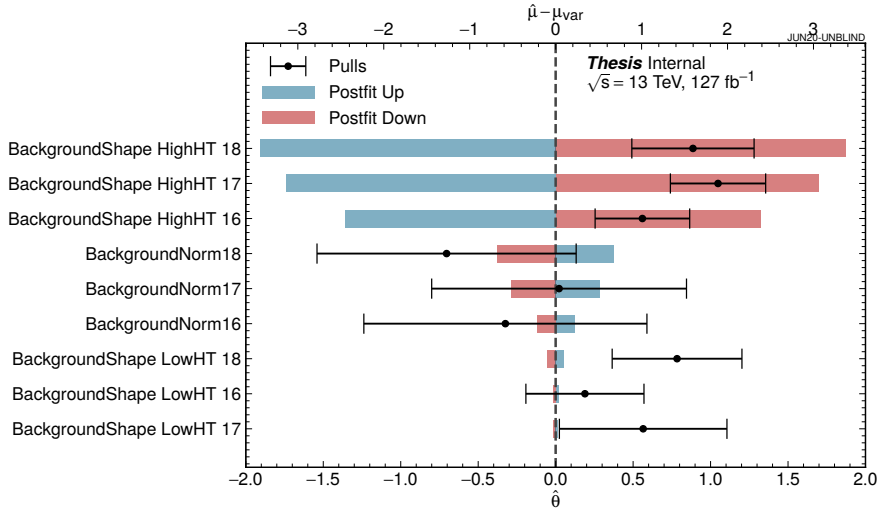


Figure 7.2: Background modelling nuisance parameter pulls and impacts for the 500 GeV signal fit. The shape nuisance parameters are pulled toward one.

I then proceeded to produce a new limit using the observed dataset. This is shown in Figure 7.3. Here, the Asimov dataset used to set the expected limit and uncertainty bands is generated by setting all nuisance parameters to their best-fit values from a conditional maximum likelihood fit performed with μ fixed at 0.

At the 280 GeV mass point where the Early Run 2 analysis observed a large excess, the observed limit is now within the 2σ uncertainty band of the expected limit. The best-fit signal ($pp \to X \to HH \to b\bar{b}b\bar{b}$) cross-section here is 1159 ± 572 fb.

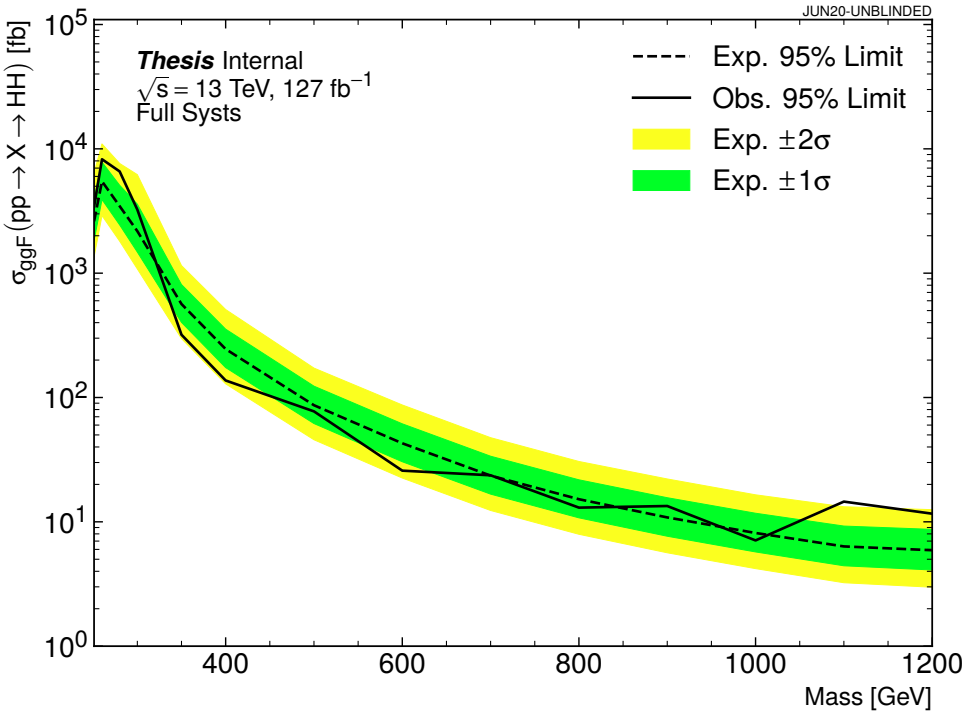
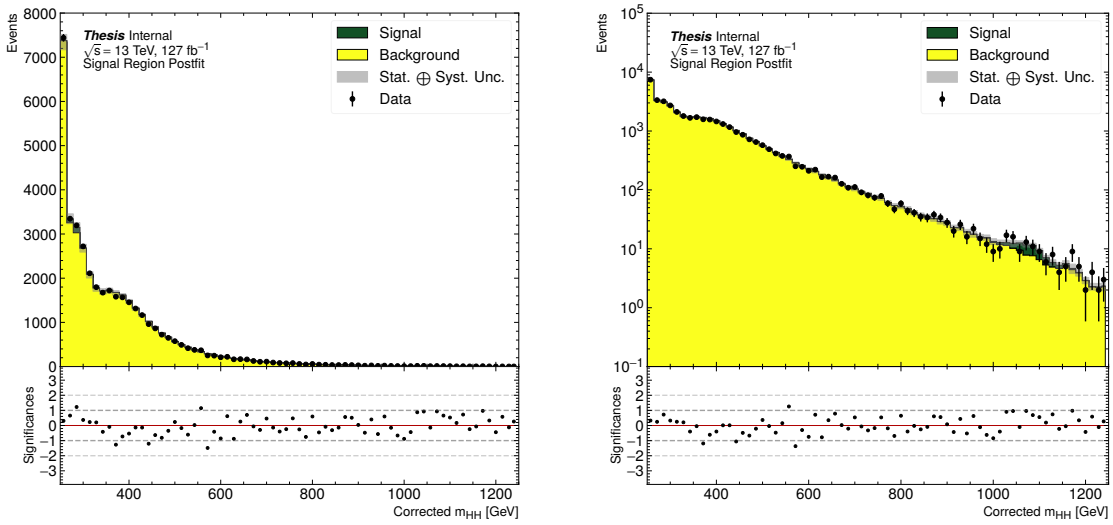


Figure 7.3: Unblinded limit of the $pp \to X \to HH$ cross-section, calculated in the $HH \to b\bar{b}b\bar{b}$ decay channel. The observed limit is within the $\pm 2\sigma$ band of the expected limit at all mass points except 1100 GeV.

A spike can also be seen in the observed limit at 1100 GeV, where the observed limit is just above the 2σ band of the expected limit. The best-fit signal ($pp \to X \to HH \to b\bar{b}b\bar{b}$) cross-section at this point is 2.72 ± 1.15 fb.

These excesses are illustrated on Figure 7.4. The 1100 GeV post-fit distribution is shown with a log-scale y-axis to make the behaviour at high m_{HH} more visible. The bin-by-bin significances were calculated using the recommendation given in [83], which takes into account the full (statistical + systematic) uncertainty on the predicted background. These are with respect to the background distribution only.



(**a**) 280 GeV signal fit.

(**b**) 1100 GeV signal fit. The y-axis is shown with a log scale.

Figure 7.4: Post-fit corrected m_{HH} distribution showing excesses at 280 GeV and 1100 GeV.

Discovery significances can be assigned to these excesses using the q_0 test statistic [74]

$$q_0 = \begin{cases} -2\ln\left(\dfrac{\mathcal{L}\left(0,\hat{\hat{\theta}}(0)\right)}{\mathcal{L}\left(\hat{\mu},\hat{\theta}\right)}\right) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} \tag{7.1}$$

The excess at 1100 GeV has a local significance of 2.6σ, while the excess at 280 GeV has a local significance of 2σ. In order to account for the look-elsewhere effect, these must be converted to global significances. Following the ATLAS recommendation [84], this can be done by first calculating the number of "downcrossings", N_{dc}, from 0.5 in a plot of local p_0 (the probability of achieving an excess of equal or greater magnitude due to a random fluctuation under the background-only hypothesis). This "downcrossings" method is a restatement of the trial factor

method shown by Gross and Vitells [85]. The local p_0 plot is shown in Figure 7.5, which reveals that for the results of this analysis, this number is 3.

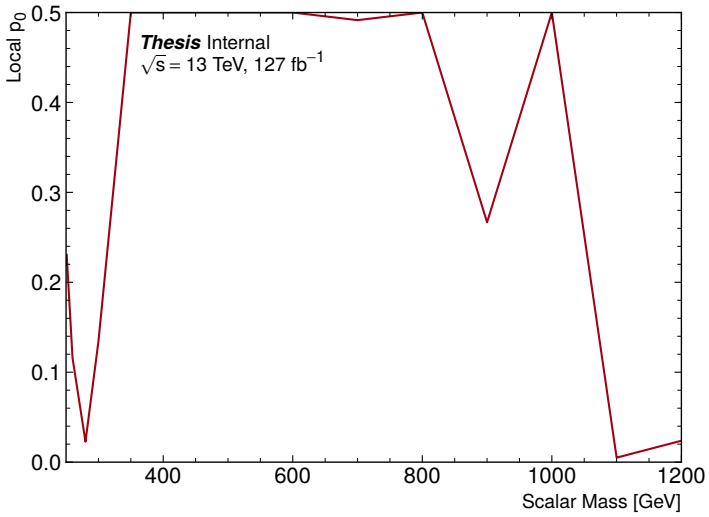


Figure 7.5: Local p_0 at each tested mass point. The number of downcrossings from 0.5 is 3.

Local significances can then be converted to global significances using the formula:

$$p_{0,\text{global}} = p_{0,\text{local}} + \left(N_{\text{dc}} \pm \sqrt{N_{\text{dc}}}\right) e^{-\frac{1}{2}Z^2_{\text{local}}} \tag{7.2}$$

The excess at 1100 GeV then has a global significance of $1.2^{+0.4}_{-0.3}\sigma$ while the 280 GeV excess has a global significance of only $0.2^{+0.7}_{-0.6}\sigma$. Therefore, neither excess is statistically significant.

The limit is also stronger than expected at 350 GeV and 400 GeV. This was due to the observed number of events in this region being lower than the background prediction. This can be seen in Figure 7.6 which shows the same post-fit corrected m_{HH} distribution as Figure 7.1b with bin-by-bin significances in the lower axes.

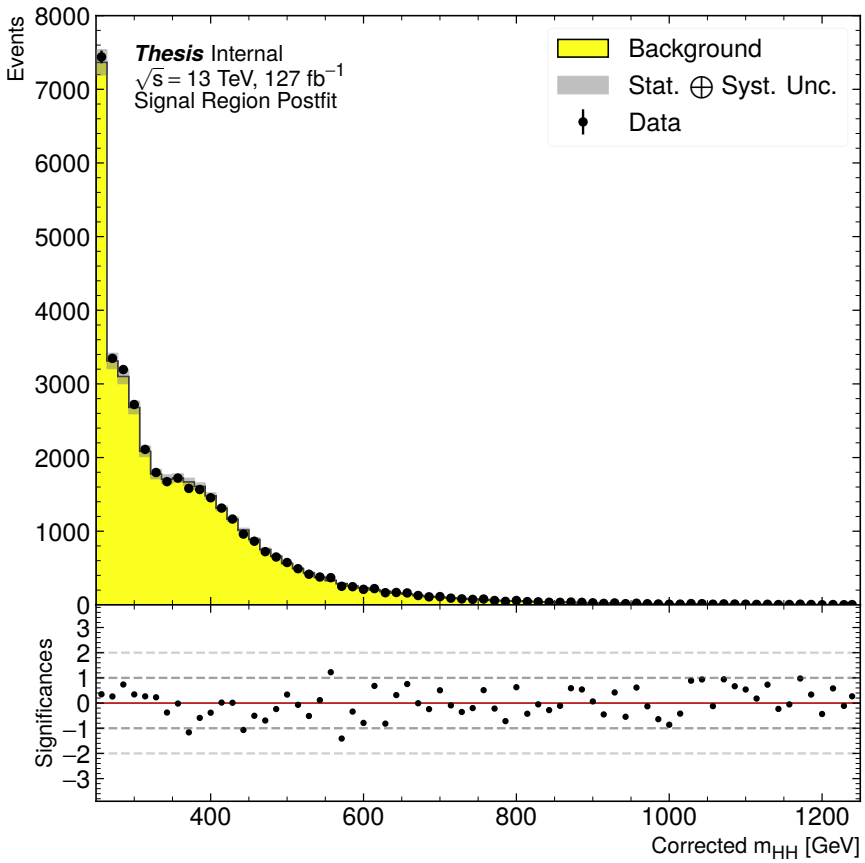
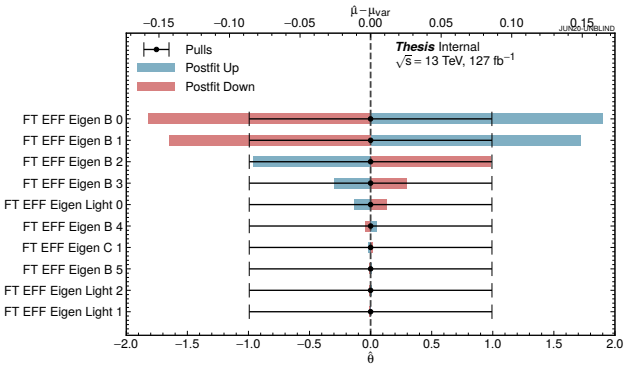
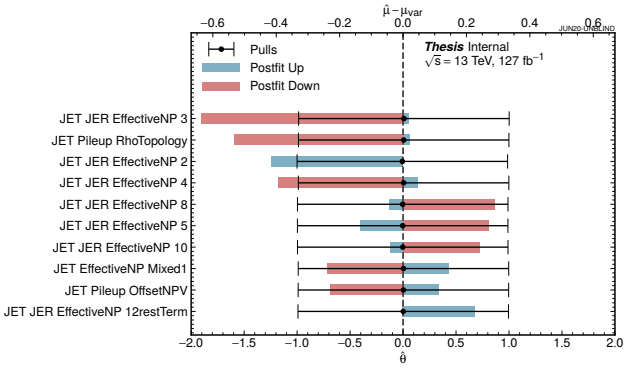


Figure 7.6: Comparison of the corrected m_{HH} distribution in the signal region with the post-fit background prediction.

Figure 7.7 then shows the 10 highest impact jet and flavour tagging nuisance parameters, calculated for the 500 GeV signal fit. These show no pulls or constraints from the signal region data. Again, this is consistent with what was observed on the Asimov fit prior to unblinding (see Figures 6.2 and 6.5).



(**a**) Flavour Tagging NPs



(**b**) Jet NPs

Figure 7.7: The 10 highest impact jet and flavour tagging nuisance parameters. These show no pulls or constraints post unblinding. These plots were made for the 500 GeV signal fit.

8 Conclusions

In summary, the resolved resonant $HH \to b\bar{b}b\bar{b}$ analysis has set stringent limits on Beyond the Standard Model models which predict a moderate to high mass (300 GeV to 1200 GeV) narrow resonance decaying to a pair of Higgs bosons. Above 350 GeV these limits are stronger than those of the combination of all ATLAS di-Higgs searches with 27.5 fb^{-1} to 36.1 fb^{-1} of data [9].[1] No statistically significant excesses have been found and, in particular, the excess seen at 280 GeV in the Early Run 2 ATLAS analysis is not confirmed.

The results of this resonant analysis are also promising for the non-resonant search for Standard Model di-Higgs production. The non-resonant $HH \to b\bar{b}b\bar{b}$ analysis will consist of incremental improvements on this analysis, with no additional data added. The stringent limits set by this analysis between 300 GeV and 900 GeV suggest the non-resonant analysis will also be able to set stringent limits on Standard Model di-Higgs production.

In the more distant future, the results of this analysis suggest it may be prudent to rethink the choice of control region made in this analysis, calculating the nominal background estimate in a region closer to the signal region. Care must be taken when doing this, however, to avoid selecting a region which may itself contain significant signal.

Di-Higgs searches, particularly searches for the Standard Model signal and measurements of the Higgs self-coupling, are some of the most important analyses

1. In this combination, $HH \to b\bar{b}b\bar{b}$ becomes the highest sensitivity channel around a mass of 400 GeV. At lower masses, the sensitivity is driven by the $HH \to b\bar{b}\gamma\gamma$ channel due to the low threshold photon triggers this analysis is able to exploit.

in Run 3 and HL-LHC (Runs 4 and 5). Work is ongoing to design triggers for Run 3 that will be useful for Di-Higgs analyses. Nevertheless it is likely that trigger thresholds will need to be higher than at present, which would adversely affect the $HH \to b\bar{b}b\bar{b}$ analysis. Looking further ahead to HL-LHC, by combining the $b\bar{b}b\bar{b}$, $b\bar{b}\tau\tau$, and $b\bar{b}\gamma\gamma$ channels, ATLAS projects that the Standard Model HH signal could be observed with a signal strength of 3.0σ [86].

A Unblinded Cutflow

Table A.1 shows the yields in data and three representative simulated signals following each cut. "Background" refers to unreweighted two tag data since the weights are derived after most cuts. With the weights applied, the background prediction is 3.949×10^4 events in the kinematic signal region, 5.437×10^4 in the validation region, and 9.378×10^4 in the control region.

Table A.1: Signal and data yields following each cut. "Background" refers to unreweighted two tag data.

	Data	**Background**		**Scalar Signal**	
	–	–	280 GeV	400 GeV	600 GeV
Initial (Unweighted for MC)	34 239 471 653	34 239 471 653	240 566 598	131 193 650	10 847 624
Reweighted to ($127\,\text{fb}^{-1}$ × Gen. x-sec × BR)	–	–	$1.325\,48 \times 10^7$	$1.445\,86 \times 10^7$	$2.453\,62 \times 10^6$
Pass NTuple Preselection	1 326 516 376	1 326 516 376	$1.752\,53 \times 10^6$	$4.711\,26 \times 10^6$	$1.488\,09 \times 10^6$
Pass Trigger	607 880 977	607 880 977	$1.700\,74 \times 10^6$	$4.592\,91 \times 10^6$	$1.460\,46 \times 10^6$
At least four jets with $p_T \geq 40\,\text{GeV}, \lvert\eta\rvert < 2.5$	426 515 496	426 515 496	$1.317\,95 \times 10^6$	$3.633\,60 \times 10^6$	$1.191\,70 \times 10^6$
Four (two for background)					
b-tagged jets	4 249 588	373 988 622	$2.673\,73 \times 10^5$	$7.594\,03 \times 10^5$	$3.103\,10 \times 10^5$
Trigger Bucket Cut	3 810 038	326 947 555	$2.659\,84 \times 10^5$	$7.237\,46 \times 10^5$	$2.809\,84 \times 10^5$
$\Delta\eta_{HH} < 1.5$	1 966 344	168 807 819	$2.002\,23 \times 10^5$	$5.999\,36 \times 10^5$	$2.302\,85 \times 10^5$
Top Veto	1 457 530	140 446 079	$9.378\,15 \times 10^4$	$4.630\,57 \times 10^5$	$2.094\,23 \times 10^5$
Signal Region	39 300	6 228 418	$3.393\,87 \times 10^4$	$2.213\,70 \times 10^5$	$1.064\,33 \times 10^5$
Validation Region	109 180	8 530 358	$3.043\,39 \times 10^4$	$1.231\,90 \times 10^5$	$5.459\,20 \times 10^4$
Control Region	187 644	14 424 562	$1.885\,26 \times 10^4$	$5.922\,90 \times 10^4$	$2.475\,40 \times 10^4$

A.1 Signal Contamination in the Validation Region

In Table A.1, we see that for the 280 GeV signal, the kinematic validation region contains almost as many signal events as the signal region. This could be a cause

for concern if these signal events led the background shape uncertainty to be artificially increased at low m_{HH}, reducing sensitivity to that signal. In order to investigate this, a signal at 280 GeV was injected into

1. the Asimov dataset for the kinematic signal region only
2. the Asimov dataset for the kinematic signal region, and the validation and control region data. The nominal and alternative background estimates would then include the effect of this additional signal contamination. In the validation and control region, the injected signal would be additional to any preexisting signal.

The injected signal cross-section in both cases was 286.86 fb, the current best upper limit (at 95% confidence level) on a signal of this mass, which was set by the combination of all ATLAS Early Run 2 di-Higgs analyses [9]. RooStats was then used to determine the best-fit $pp \to X \to HH \to b\bar{b}b\bar{b}$ cross-section in each case. Since the signal was injected over Asimov data in the signal region, we would expect to recover the injected value, and this was indeed the case (modulo numerical stability) with a best-fit cross-section of 286.850 ± 564.173 fb when the signal was injected only over the Asimov dataset, and 286.850 ± 592.354 fb when the signal was injected over both the Asimov dataset in the signal region, and the validation and control region data. With only a 5% increase on the uncertainty on the best-fit cross-section, and no change in the best-fit cross-section itself, this demonstrates that signal contamination in the validation region does not lead to an appreciable loss of sensitivity to low mass signals.

Bibliography

[1] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Physical Review Letters* 13.9 (Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[2] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13.16 (Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[3] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Physical Review Letters* 13.20 (Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[4] Alessandro De Angelis and Mário João Martins Pimenta. "The Higgs Mechanism and the Standard Model of Particle Physics". In: *Introduction to Particle and Astroparticle Physics: Questions to the Universe*. Milano: Springer Milan, 2015, pp. 361–420. ISBN: 978-88-470-2688-9. DOI: 10.1007/978-88-470-2688-9_7.

[5] Nikos Konstantinidis. *Lecture Notes for the 2017 School for Experimental High Energy Physics Students*. RAL Technical Report RAL-TR-2017-009. STFC, 2017.

[6] Mark Thomson. *Modern Particle Physics*. English. Cambridge, United Kingdom ; New York: Cambridge University Press, Sept. 2013. ISBN: 978-1-107-03426-6.

[7] M. Tanabashi et al. "Review of Particle Physics". In: *Physical Review D* 98.3 (Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001.

[8] M. Grazzini et al. "Higgs Boson Pair Production at NNLO with Top Quark Mass Effects". In: *Journal of High Energy Physics* 2018.5 (May 2018). DOI: 10.1007/jhep05(2018)059.

[9] G. Aad et al. "Combination of Searches for Higgs Boson Pairs in Pp Collisions at S=13TeV with the ATLAS Detector". en. In: *Physics Letters B* 800 (Jan. 2020), p. 135103. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2019.135103.

[10] G. Heinrich et al. "Probing the Trilinear Higgs Boson Coupling in Di-Higgs Production at NLO QCD Including Parton Shower Effects". en. In: *Journal of High Energy Physics* 2019.6 (June 2019), p. 66. ISSN: 1029-8479. DOI: 10.1007/JHEP06(2019)066.

[11] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[12] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". en. In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024.

[13] The NNPDF Collaboration et al. "Parton Distributions for the LHC Run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015), p. 40. ISSN: 1029-8479. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849.

[14] Jacob Amacker et al. "Higgs Self-Coupling Measurements Using Deep Learning in the $b\bar{b}b\bar{b}$ Final State". In: *arXiv:2004.04240 [hep-ex, physics:hep-ph]* (July 2020). arXiv: 2004.04240 [hep-ex, physics:hep-ph].

[15] J. de Favereau et al. "DELPHES 3: A Modular Framework for Fast Simulation of a Generic Collider Experiment". en. In: *Journal of High Energy Physics* 2014.2 (Feb. 2014), p. 57. ISSN: 1029-8479. DOI: 10.1007/JHEP02(2014)057.

[16] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. "FastJet user manual". In: *Eur. Phys. J. C* 72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[17] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. "The Catchment Area of Jets". en. In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 005–005. ISSN: 1126-6708. DOI: 10.1088/1126-6708/2008/04/005.

[18] G.C. Branco et al. "Theory and Phenomenology of Two-Higgs-Doublet Models". In: *Physics Reports* 516.1 (2012), pp. 1–102. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2012.02.002.

[19] Lisa Randall and Raman Sundrum. "Large Mass Hierarchy from a Small Extra Dimension". In: *Physical Review Letters* 83.17 (Oct. 1999), pp. 3370–3373. DOI: 10.1103/PhysRevLett.83.3370.

[20] Lisa Randall and Raman Sundrum. "An Alternative to Compactification". In: *Physical Review Letters* 83.23 (Dec. 1999), pp. 4690–4693. DOI: 10.1103/PhysRevLett.83.4690.

[21] Kaustubh Agashe et al. "Warped Gravitons at the CERN LHC and Beyond". In: *Physical Review D* 76.3 (Aug. 2007), p. 36006. DOI: 10.1103/PhysRevD.76.036006.

[22] CERN. "CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". English. In: (2017). DOI: 10.23731/CYRM-2017-002.

[23] Mark Thomson. "Colour Confinement". In: *Modern Particle Physics*. Cambridge: Cambridge University Press, 2019, pp. 248–253. ISBN: 978-1-107-03426-6.

[24] Matteo Cacciari, Gavin P Salam and Gregory Soyez. "The Anti-Ktjet Clustering Algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. DOI: 10.1088/1126-6708/2008/04/063.

[25] Joao Pequenao. "Computer Generated Image of the Whole ATLAS Detector". Mar. 2008.

[26] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[27] Atlas Collaboration. *Experiment Briefing: Keeping the ATLAS Inner Detector in Perfect Alignment*. en. https://cds.cern.ch/record/2723878. July 2020.

[28] Claudia Gemme, Allan G. Clark and Alexandre Rozanov. *Production and Integration of the ATLAS Insertable B-Layer*. en. https://cds.cern.ch/record/2306002. Feb. 2018.

[29] ATLAS Collaboration. *ATLAS Insertable B-Layer Technical Design Report*. ATLAS-TDR-19. 2010. URL: https://cds.cern.ch/record/1291633.

[30] M Capeans et al. *ATLAS Insertable B-Layer Technical Design Report*. Tech. rep. CERN-LHCC-2010-013. ATLAS-TDR-19. Sept. 2010.

[31] A Vogel. *ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates*. Tech. rep. ATL-INDET-PROC-2013-005. Geneva: CERN, Apr. 2013.

[32] Giulia Ripellino. *The Alignment of the ATLAS Inner Detector in Run 2*. Tech. rep. ATL-INDET-PROC-2016-003. Geneva: CERN, Sept. 2016. DOI: 10.22323/1.276.0196.

[33] A. Ahmad et al. "The Silicon Microstrip Sensors of the ATLAS Semiconductor Tracker". en. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 578.1 (July 2007), pp. 98–118. ISSN: 0168-9002. DOI: 10.1016/j.nima.2007.04.157.

[34] G. Aad et al. "ATLAS Pixel Detector Electronics and Sensors". en. In: *Journal of Instrumentation* 3.07 (July 2008), P07007–P07007. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/07/P07007.

[35] The ATLAS TRT collaboration et al. "The ATLAS Transition Radiation Tracker (TRT) Proportional Drift Tube: Design and Performance". en. In: *Journal of Instrumentation* 3.02 (Feb. 2008), P02013–P02013. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/02/P02013.

[36] Bartosz Mindur. "ATLAS Transition Radiation Tracker (TRT): Straw Tubes for Tracking and Particle Identification at the Large Hadron Collider". en. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 845 (Feb. 2017), pp. 257–261. ISSN: 01689002. DOI: 10.1016/j.nima.2016.04.026.

[37] Joao Pequenao. "Computer Generated Image of the ATLAS Calorimeter". Mar. 2008.

[38] The ATLAS Collaboration et al. "The ATLAS Experiment at the CERN Large Hadron Collider". en. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/08/S08003.

[39] ATLAS Collaboration. "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1". In: *Eur. Phys. J. C* 77 (2017), p. 490. DOI: 10.1140/epjc/s10052-017-5004-5. arXiv: 1603.02934 [hep-ex].

[40] Joao Pequenao. "Computer Generated Image of the ATLAS Muons Subsystem". Mar. 2008.

[41] ATLAS Collaboration. "Jet reconstruction and performance using particle flow with the ATLAS Detector". In: *Eur. Phys. J. C* 77 (2017), p. 466. DOI: 10.1140/epjc/s10052-017-5031-2. arXiv: 1703.10485 [hep-ex].

[42] Georges Aad et al. "Jet Energy Scale and Resolution Measured in Proton-Proton Collisions at $\sqrt{s}$=13 TeV with the ATLAS Detector". In: CERN-EP-2020-083 (July 2020). arXiv: 2007.02645 [hep-ex].

[43] ATLAS Collaboration. "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13\,\text{TeV}$". In: *Eur. Phys. J. C* 76 (2016), p. 292. DOI: 10.1140/epjc/s10052-016-4120-y. arXiv: 1603.05598 [hep-ex].

[44] ATLAS Collaboration. "ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13\,\text{TeV}$". In: (2019). arXiv: 1907.05120 [hep-ex].

[45] ATLAS Collaboration. *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*. ATL-PHYS-PUB-2017-003. 2017. URL: https://cds.cern.ch/record/2255226.

[46] ATLAS Collaboration. *Tagging and suppression of pileup jets with the ATLAS detector*. ATLAS-CONF-2014-018. 2014. URL: https://cds.cern.ch/record/1700870.

[47] ATLAS Collaboration. *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*. ATL-PHYS-PUB-2017-013. 2017. URL: https://cds.cern.ch/record/2273281.

[48] ATLAS Collaboration. "Luminosity Public Results Run 2". In: *ATLAS Public TWiki* ().

[49] ATLAS Collaboration. "Performance of the ATLAS trigger system in 2015". In: *Eur. Phys. J. C* 77 (2017), p. 317. DOI: 10.1140/epjc/s10052-017-4852-3. arXiv: 1611.09661 [hep-ex].

[50] Imma Riu and ATLAS Collaboration. "The ATLAS Level-1 Topological Trigger Performance in Run 2". en. In: *Journal of Physics: Conference Series* 898 (Oct. 2017), p. 032037. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/898/3/032037.

[51] Yu Nakahama. "The ATLAS Trigger System: Ready for Run-2". en. In: *Journal of Physics: Conference Series* 664.8 (Dec. 2015), p. 082037. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/664/8/082037.

[52] Theodor Herwig and ATLAS collaboration. "ATLAS Jet Trigger Performance in 2016 Data". en. In: *Proceedings of 38th International Conference on High Energy Physics — PoS(ICHEP2016)*. Chicago, USA: Sissa Medialab, Feb. 2017, p. 854. DOI: 10.22323/1.282.0854.

[53] Johannes Bellm et al. "Herwig 7.0/Herwig++ 3.0 release note". In: *Eur. Phys. J. C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. arXiv: 1512.01178 [hep-ph].

[54] D. J. Lange. "The EvtGen particle decay simulation package". In: *Nucl. Instrum. Meth. A* 462 (2001), p. 152. DOI: 10.1016/S0168-9002(01)00089-4.

[55] The NNPDF Collaboration et al. "Parton Distributions with QED Corrections". In: *arXiv:1308.0598 [hep-ph]* (Nov. 2013). arXiv: 1308.0598 [hep-ph].

[56] L. A. Harland-Lang et al. "Parton Distributions in the LHC Era: MMHT 2014 PDFs". In: *The European Physical Journal C* 75.5 (May 2015), p. 204. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-015-3397-6. arXiv: 1412.3989.

[57] S. Agostinelli et al. "GEANT4 – a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), p. 250. DOI: 10.1016/S0168-9002(03)01368-8.

[58] ATLAS Collaboration. "Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *JHEP* 01 (2019), p. 030. DOI: 10.1007/JHEP01(2019)030. arXiv: 1804.06174 [hep-ex].

[59] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[60] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. ISSN: 0090-5364.

[61] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3146–3154.

[62] Tim Head et al. *Scikit-Optimize/Scikit-Optimize: V0.5.2*. Version v0.5.2. Mar. 2018. DOI: 10.5281/zenodo.1207017.

[63] ATLAS Collaboration. *A new tagger for the charge identification of b-jets*. ATL-PHYS-PUB-2015-040. 2015. URL: https://cds.cern.ch/record/2048132.

[64] Jason Gallicchio and Matthew D. Schwartz. "Seeing in Color: Jet Superstructure". In: *Physical Review Letters* 105.2 (July 2010), p. 022001. DOI: 10.1103/PhysRevLett.105.022001.

[65] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774.

[66] Scott M Lundberg et al. "Explainable AI for Trees: From Local Explanations to Global Understanding". In: *arXiv preprint arXiv:1905.04610* (2019). arXiv: 1905.04610.

[67] Gilles Louppe. "Likelihood Ratio Estimation for Statistical Inference in Physical Sciences". In: (2019).

[68] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (Jan. 1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[69] Lorenzo Moneta et al. "The RooStats Project". In: *Proceedings, 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2010): Jaipur, India, February 22-27, 2010*. Vol. ACAT2010. PoS. 2010, p. 57. DOI: 10.22323/1.093.0057. arXiv: 1009.1003 [physics.data-an].

[70] Kyle Cranmer et al. "HistFactory: A Tool for Creating Statistical Models for Use with RooFit and RooStats". In: (2012).

[71] A L Read. "Presentation of Search Results: The CLs Technique". In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), pp. 2693–2704. ISSN: 0954-3899. DOI: 10.1088/0954-3899/28/10/313.

[72] Eilam Gross. "LHC Statistics for Pedestrians". In: (2008). DOI: 10.5170/CERN-2008-001.205.

[73] Heinrich, Lukas and Feickert, Matthew and Stark, Giordon. *Pyhf: V0.4.4*. DOI: 10.5281/zenodo.1169739.

[74] Glen Cowan et al. "Asymptotic Formulae for Likelihood-Based Tests of New Physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0.

[75] *Impact of Fragmentation Modelling on the Top Quark Mass Measurement Using the ATLAS Detector*. Tech. rep. ATL-PHYS-PUB-2015-042. Geneva: CERN, Sept. 2015.

[76] B. Andersson et al. "Parton Fragmentation and String Dynamics". en. In: *Physics Reports* 97.2-3 (July 1983), pp. 31–145. ISSN: 03701573. DOI: 10.1016/0370-1573(83)90080-7.

[77] B. R. Webber. "A QCD Model for Jet Fragmentation Including Soft Gluon Interference". en. In: *Nuclear Physics B* 238.3 (June 1984), pp. 492–528. ISSN: 0550-3213. DOI: 10.1016/0550-3213(84)90333-X.

[78] Dale Charles Abbott et al. *Supporting Document: The Search for Resonant HH Production Decaying to the 4b Final State Using the Full Run-2 Data and the Resolved Analysis Channel*. Tech. rep. ATL-COM-PHYS-2020-084. Geneva: CERN, Feb. 2020.

[79] R. W Barnard et al. "Gram Polynomials and the Kummer Function". en. In: *Journal of Approximation Theory* 94.1 (July 1998), pp. 128–143. ISSN: 0021-9045. DOI: 10.1006/jath.1998.3181.

[80] Eric W. Weisstein. *Legendre Polynomial*. https://mathworld.wolfram.com/LegendrePolynomial.html.

[81] Stefan van der Walt, S. Chris Colbert and Gael Varoquaux. "The NumPy Array: A Structure for Efficient Numerical Computation". In: *Computing in Science Engineering* 13.2 (Mar. 2011), pp. 22–30. ISSN: 1558-366X. DOI: 10.1109/MCSE.2011.37.

[82] Gaël Guennebaud, Benoît Jacob et al. "Eigen V3". In: (2010).

[83] William Buttinger and Michel Lefebvre. *Formulae for Estimating Significance*. en. https://cds.cern.ch/record/2643488. Oct. 2018.

[84] ATLAS Collaboration. *GlobalSignificanceComputations < AtlasProtected < TWiki*. https://twiki.cern.ch/twiki/bin/view/AtlasProtected/GlobalSignificanceComputations.

[85] Eilam Gross and Ofer Vitells. "Trial Factors for the Look Elsewhere Effect in High Energy Physics". In: *The European Physical Journal C* 70.1-2 (Nov. 2010), pp. 525–530. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-010-1470-8. arXiv: 1005.1891.

[86] *Report on the Physics at the HL-LHC,and Perspectives for the HE-LHC - INSPIRE*. https://inspirehep.net/literature/1768569.