Graph Neural Network Flavour Tagging and Boosted Higgs Measurements at the LHC

Samuel John Van Stroud
University College London

Submitted to University College London in fulfilment
of the requirements for the award of the
degree of **Doctor of Philosophy**

June 21, 2022

Declaration

I, Samuel John Van Stroud confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Samuel Van Stroud

Abstract

Here some useful packages are demonstrated. In particular, the `hepunit` package which adds additional units to `SIUnit`. A variety of jet measurements are made using data collected during the first year of 7 TeV proton-proton collisions from the general-purpose ATLAS experiment at the LHC. no more than 300 words

Impact Statement

impact statement 500 words link to ucl info

Acknowledgements

Here is an example of how to declare commands for use in a single file that will not be needed elsewhere. Additionally, it serves to illustrate the chapter referencing system.

Perhaps you might want to point out that Peter Higgs provided helpful advice for Chapter 1.

Preface

blah this is 300 TeV in text mode. this is $300\,\mathrm{TeV}$ in math mode.

Contents

1

Chapter 1

Theoretical Framework

1.1 The Standard Model

1.2 The Higgs Mechanism

Chapter 2

The Large Hadron Collider and the ATLAS Detector

2.1 Overview

The Large Hadron Collider (LHC) at CERN has extended the frontiers of particle physics through its unprecedented energy and luminosity. In 2010, the LHC collided proton bunches, each containing more than 10^{11} particles, 20 million times per second, providing 7 TeV proton-proton collisions at instantaneous luminosities of up to $2.1 \times 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

2.2 Trigger system

An LHCb trigger table borrowed from `hepthesis` is shown in Table 2.1:

	L0	L1	HLT
Input rate	40 MHz	1 MHz	40 kHz
Output rate	1 MHz	40 kHz	2 kHz
Location	On detector	Counting room	Counting room

Table 2.1: Characteristics of the trigger levels and offline analysis.

2.3 Reconstructed Physics Objects

2.3.1 Tracks

2.3.2 Jets

- Jet finding algorithms

2.3.3 Leptons

Chapter 3

Investigating Tracking Improvements

Todo:

- Check all info wrt to this PDG review

3.1 *b*-hadron Reconstruction

3.1.1 *b*-hadron Decay Topology

b-hadrons are quasi-stable bound states of quarks, where one of the quarks is a bottom quark (b quark). The proper lifetimes τ of the various b-hadrons are similar and relatively long, with $\tau \sim 10^{-12}$s. This lifetime corresponds to a proper decay length $c\tau \sim 300$ μm. In the rest frame of the detector, the typical b-hadron travels a distance $d = \beta\gamma c\tau$ before decaying, where at high energies $\gamma \sim E_B/m_B$. For a 1 TeV b-hadron, this gives $d \sim 60$ mm - well beyond the radius of the first pixel layer (IBL) at 33 mm. At the LHC, b quarks are generated in the hard scattering of proton-proton (pp) collisions. They quickly hadronize into a b-hadron, which is often initially in an excited state due to the high energies of the pp collisions at the LHC ($\sqrt{s} = 13$ TeV). The hadronisation process is hard - around 70-80% of the b quark's momentum goes into the b-hadron, with the rest being radiated as other particles. The excited b-hadron will quickly fragment (i.e. de-excite) by radiating particles, which are prompt (they are formed closed to the primary vertex). These fragmentation particles have an increasing multiplicity and collimation to the b-hadron axis as the p_{T} of the b-hadron increases. The de-excited b-hadron subsequently weakly decays to on average 4 or 5 particles (the multiplicity of the decay products of the weak decay of the b-hadron is unaffected by increases in the b-hadron p_{T}.).

Due to their lifetimes, energetic b-hadrons can travel a significant distance from the primary pp interaction point before decaying to a spray of collimated stable particles. This signature is registered in the detector as a displaced jet. Due to the elements of the CKM matrix, b-hadrons decay with a high probability to D hadrons (which contain a c quark), which also have significant lifetimes - this can lead to reconstructed tertiary vertices in the jet core. The typical features of a b-jet, and in particular the large track impact parameter d_0 which can result from displaced decays, are shown in fig. 3.1. Many ATLAS analyses rely on a method of tagging jets instantiated by b quarks and rejecting jets created from other quarks (c and light flavours u, d, s). These "b-tagging" algorithms work by discriminating against the unique signatures of b-jets discussed above. b-tagging relies on the efficient and accurate reconstruction the tracks corresponding to the b-hadron decay products. These tracks are then used as inputs to vertex reconstruction algorithms and jet making algorithms.

3.1.2 *b*-hadron Decay Track Reconstruction

A necessary requirement for successful jet b-tagging is the efficient and accurate reconstruction of the charged particle trajectories in the jet. For high p_{T} jets (p_{T}

chapters/3.tracking/figs/b-jet-diagram.png

Figure 3.1: Diagram of a typical b-jet (blue) which has been produced along with two light jets (grey). The b-hadron has travelled a significant distance from the primary interaction point (pink dot) before its decay. The large transverse impact parameter d_0 is a characteristic property of the trajectories of b-hadron decay products.

chapters/3.tracking/figs/high-pt-b-tracks.png

Figure 3.2: As b-hadron p_{T} increases, the time of flight of the B increases, so tracks will have less room to diverge before reaching detector elements. To compound the problem, the collimation of the tracks increases. The detector may then be unable to resolve individual tracks.

$>$ 200 GeV) this task becomes difficult due to a combination of effects. As the jet energy increases, the track multiplicity of the jet increases due to the presence of additional fragmentation tracks. Tracks in the jet also become increasingly collimated as their inherited transverse momentum increases. Together, these two effects lead to a very high density of charged particles in the jet core, making reconstruction difficult. At high energies, the increased decay length of B (and D) hadrons means that decay products have less of an opportunity to diverge before reaching the first tracking layers of the detector. If the decay takes place very close to a detector layer, or if the decays are sufficiently collimated, hits left by nearby particles may not be resolved individually, leading to merged clusters (shown in fig. 3.2). Shared hits generally predict bad tracks. As such, shared hits are heavily penalised during reconstruction (and in particular as part of ambiguity solving). However, in the core of high p_{T} b-jets, where decay particles are displaced from the primary vertex and are highly collimated, the density of particles is high enough that the probability of clusters being merged increases dramatically. The presence of merged clusters requires that the corresponding tracks share hits (if they are to be reconstructed successfully), which may end up impairing the successfully reconstruction of the track. Furthermore, decays may also take place inside the tracking detectors themselves, which can lead to missing or wrong innermost cluster assignment. The combination of effects described above makes reconstructing tracks in the core of high p_{T} b-jets particularly challenging.

chapters/3.tracking/figs/overlay_po_nHitsOnIBL_Frac_pT.pdf chapters/3.tracking/figs/overlay_po_nHitsOnPix_F

a **b**

Figure 3.3: Hit multiplicities on the IBL (fig. 3.3a) and the all pixel layers (fig. 3.3b) as a function of the transverse momentum p_{T} of the reconstructed track. Tracks from the weak decay of the b-hadron are shown in red, while fragmentation tracks (which are prompt) are in blue. For each of these, standard tracks and pseudo-tracks are plotted. Hit multiplicities on the pseudo-tracks at high p_{T} due to the increased flight of the b-hadron. The baseline tracks have more hits than the pseudo-tracks, indicating that they are being incorrectly assigned additional hits.

Figure 3.4: Track reconstruction efficiency from b-hadron decay products for baseline ATLAS tracking (black), Bcut+Refit procedures applied (green), pseudo-tracking (blue), and for tracking where the ambiguity solver has been manually removed (orange).

Figure 3.5: The total number of pixel hits on tracks from b-hadron decays as a function of the production radius of the decay product. An excess of hits is assigned to the standard tracks in comparison to the ideal pseudo-tracks.

Concretely, then, the issues relating to high p_{T} b-hadron tracking can be factorised into two parts. The first part is a drop in track reconstruction efficiency. As mentioned, tracks originating from high energy b-hadron decay products can have a high rate of shared hits due to the number of particles present in a high

p_{T} b-jet and their relative collimation. Additionally, tracks may be missing hits on the inner layers of the detector. This occurs primarily when the decay b-hadron decays inside the detector. These features of can make it difficult for B decay tracks to meet the ambiguity solver's stringent track quality requirements. As a result, many B decay tracks are rejected in the ambiguity solving stage, leading to a severe drop in tracking reconstruction efficiency. This is shown by the severe decrease in reconstruction efficiency visible when comparing baseline tracking with the ideal pseudo-tracks in fig. 3.4. This situation presents a problem: relaxing cuts on shared hits significantly degrades the ambiguity solver's power to reject bad tracks. However for b-hadron decay tracks it seems these same restrictions on shared hits are seriously impairing the reconstruction efficiency of good tracks. The second part of the problem is that, due to the high density of clusters available for assignment in the vicinity of the typical high energy b-hadron decay track, and also given the strong positive bias of the ambiguity solver towards those tracks with precise pixel measurements (especially the innermost IBL measurement), many b-hadron decay tracks are assigned incorrect inner layer hits. This is only a problem for those decay products which were produced inside the pixel detector as a result of a long-flying b-hadron, and so do not have a correct hit available for assignment (evidenced in fig. 3.8b). The incorrect hits may skew the parameters of the track, which can in turn mislead b-tagging algorithms. In particular, b-tagging algorithms rely heavily on the transverse impact parameter significance $d_0/\sigma(d_0)$ of the track. The quality of this measurement is expected to be adversely affected by wrong inner-layer hits on the track. This combination of reduced reconstruction efficiency and incorrectly assigned hits is thought to be the cause of the observed drop in b-tagging efficiency at high energies , although it is not clear which effect may dominate.

3.2 Pseudotracks and Ideal Tracks

Pseudotracking and ideal tracking are used as benchmarks of the best tracking possible given the ATLAS detector. Both pseudotracks and ideal tracks are constructed using truth information to group combinations of hits that have been left by the same truth particle. As a result, hit-to-track association and track reconstruction efficiency are both ideal (given the ATLAS detector). Ideal tracks represent a yet more idealised tracking scenario by correcting the cluster positions based on truth information, and smearing the cluster position based on the detector resolution.

When pseudotracking is run alongside standard tracking, those clusters which are shared on the reconstructed tracks run through the cluster splitting machinery. If a cluster is found to be compatible with being split, its definition is changed, and the pseudotracks use this definition too. As a result, pseudotracks can have split clusters.

3.3 Investigating Improvements for High p_{T} B Tracking

An investigation into

3.3.1 Looser Track Cuts & Track Refit Procedure

A solution for the problem of wrong inner-layer hits on B tracks had previously previously been developed. This solution selects tracks which pass a b-jet Region of Interest (ROI) selection, and then removes the innermost hits on these tracks based on the result of a "refit" procedure. The refit procedure runs as follows. Each track is refitted without the innermost hit, and if there is a significant improvement in the fit quality (the χ^2 of the track fit divided by the number of degrees of freedom on the track n), the innermost hit is rejected and the new track is replaces the old. If the fit quality does not improve by a certain amount, the initial track is kept. This procedure is recursively applied. The b-jet ROI selection selects tracks that are matched within $dR < 0.14$ ($|\eta| < 0.1$, $|\phi| < 0.1$) of a CaloCluster with $E_T > 150$ GeV. The track itself must also pass a transverse momentum cut with $p_{\mathrm{T}} > 15$ GeV. The refit procedure was previously shown to lead to a reduction in the rate of wrongly assigned IBL hits on B decay tracks (see fig. 3.8b). However, this apparent improvement did not lead to an increase in b-tagging performance. It was found that the refit procedure also removed unacceptable numbers of good hits, degrading the quality of un-problematic tracks, shown in fig. 3.8a. This is likely the cause of the underwhelming b-tagging performance improvement.

The performance of both the ROI, and the hit removal using track fit information, is examined, and an attempt at improving the performance of the refit procedure is made. Results are discussed in the following two sections.

3.3.2 Region of Interest Optimisation

Selection cuts for the b-jet ROI were determined on a largely ad-hoc basis. An effort was made to systematically optimise the selection cuts. The decay tracks of B hadrons are tightly collimated with the B itself, with most decay products satisfying $dR(B, \text{track}) < 0.02$, as shown in fig. 3.6a. Meanwhile, calorimeter clusters relating to the B hadrons are generally found within $dR < 0.05$ of the B fig. 3.6b. In total, then, B decay tracks will usually be found within $dR < 0.07$ of the relevant calorimeter cluster, which suggests that the current $dR < 0.14$ is loose by a factor of two. Similar analysis of cluster and track energy distributions found that the related cuts were also loose, and so they were modified from $E_T > 150$ GeV to $E_T > 300$ GeV, and from $p_T > 15$ GeV to $p_T > 30$ GeV.

Additionally examined in the course of this work was the fake rate of the b-jet ROI. The distributions in fig. 3.7a demonstrate that most of clusters passing the $E_T > 150$ GeV selection were unable to be matched to a nearby B hadron using truth information. Clusters that pass the selection but do not correspond to energy depositions from B hadrons lead to fake ROIs. As a consequence of these distributions, tracks selected by the ROI are largely impure in the desired B hadron tracks.

The modified ROI was used to re-run the refit procedure. A comparison of of "standard" and "optimised" (using the optimised b-jet ROI) refit procedures is found in fig. 3.8. These results show that whilst tighter selection cuts did lead to a recovery of some good hits (fig. 3.8a), performance with respect to the baseline is still significantly degraded.

3.3.3 Fit Quality as a Discriminant for Wrong Hits

As mentioned, tracks selected by the ROI are refitted without their innermost hit, and, if an improvement in fit quality is observed, the hit is rejected. In order to test a b

Figure 3.6: Distributions of angular distance dR between B hadrons and their weak decays and other fragmentation tracks (fig. 3.6a), and the distribution of angular distance dR between B hadrons and the calorimeter clusters in the hadronic calorimeter (fig. 3.6b). In fig. 3.6a, the tracks from the weak decay of the B are significantly more collimated to the B than the other fragmentation tracks.

the effectiveness of this procedure, a dataset of two sets of tracks was produced. The first set contained unmodified baseline-reconstructed tracks. The second contained the same tracks as the first, but modifications made during reconstruction removed the innermost hit on each track. Then, using Monte Carlo (MC) truth information, a track-by-track fit quality comparison was made for tracks with good and wrong innermost hits.

It is clear from the distributions in fig. 3.7b that the fit quality improvement (measured by fractional change in χ^2/n of the track before and after the innermost hit is removed) is not a discriminating variable for wrong hits, and indeed attempted optimisations of the of the refit procedure based on these distributions were found to be ineffectual. While wrong hits are likely to degrade the track fit, it is also true that any additional measurement, good or wrong, constrains the track, and therefore removal of that measurement will be likely to lead to an increase in the χ^2/n of the track. Removing hits in this way is therefore problematic.

a b

Figure 3.7: The distribution of cluster transverse momentum, in fig. 3.7a for both clusters that were able (orange) and unable (blue) to be matched to a B hadron using MC truth information. The normalisation shows that the majority of clusters are not matched to B hadrons, resulting in fake ROIs. In fig. 3.7b, the fractional improvement in track fit quality (χ^2/n) is shown for all track (blue), tracks with good IBL hits (green), and tracks with wrong IBL hits (orange). The distributions are overlapping, suggesting that the χ^2/n improvement is not a good discriminator of good and wrong hits.

3.3.4 Conclusion

The work outlined in the two preceding sections has uncovered issues with both the b-jet ROI, and the methodology of identification and removal of wrong hits on tracks inside a given ROI. Attempts were made to optimise the selection cuts of the ROI, however the large background of energetic phenomena produced in collisions that are not B hadron related means that the ROI is largely unsuccessful in selecting a pure sample of likely B hadron candidates. An additional effort was made to improve the removal of wrong hits using other information in addition to the track fit improvement. Information such as the type and locations of its, and track d_0 were considered. While progress here was not insignificant, without substantial overhaul of the ROI to improve B purity, the results were not strong enough to demonstrate any viable solutions that would successful target and then improve B hadron decay tracks. Alongside the refit procedure, a “Bcut” cut scheme was suggested in order

a **b**

Figure 3.8: Distributions of good (fig. 3.8a) and wrong (fig:refit optimisation results sub2) hit assignment rates on the IBL for tracks using baseline tracking (black), the original unmodified refit procedure (green), and the refit procedure with an optimise set of ROI selection cuts (blue). The IBL lies at a radius of 33 mm from the beam pipe. Hence, particles produced with a production radius greater than this cannot leave good hits on the IBL.

to improve reconstruction performance. This consisted primarily of loosening the shared hit cuts in the ambiguity solver. While this did lead to a measurement increase in track reconstruction efficiency (see fig. 3.4), it was determined that the corresponding increase in fake tracks (i.e. those tracks for which the majority of hits do not come from a single truth particle) was too large to justify the implementation of the “Bcut” scheme. In conclusion, then, a different approach is required to address the problems discussed.

3.4 Global χ^2 Fitter Outlier Removal

This section documents ongoing progress into improving hit assignments using the Global χ^2 Fitter (GX2F) to prevent wrong hits from being assigned to tracks during the track fit. This is in contrast to the approach discussed in cref sec:refit, which attempts to identify and remove wrong hits after the reconstruction of the track (of which the track fit is a part). As part of the track fit, an outlier removal procedure is run, in which suspicious hits are indentified and removed. The GX2F code, as a relatively low-level component of track reconstruction, has not undergone significant modification for several years. During this time, a new tracking sub-detector, the IBL, was installed, and subsequently precise detector alignments have been derived. The motivation for looking at the GX2F is that these changes may require re-optimisation of the GX2F code, and in particular the outlier removal procedures. Further motivation for this approach comes from the low rate of labelled outliers in baseline tracking. For example, while approximately 15% of B hadron decay tracks have a wrong IBL hit (a value which only increases with the p_{T} of the B), less than 1% of this tracks have had their IBL hit labelled and removed as an outlier.

Implementation

The outlier removal procedure for the pixel detector is described in this section. The states (also called measurements, or hits) on the track are looped over in order of increasing radial distance to the beam pipe. For each state, errors $\sigma(m_i)$ on the measurement of the transverse and longitudinal coordinates are calculated. These errors are dependent on the sub-detector which recorded the measurement (as some sub-detectors are more precise than others). Additionally, a residual displacement r_i between the predicted position of the track x_i (inclusive of the current measurement), and the position of the measurement itself, m_i, is calculated. The pull p_i on the track state due to the current measurement is calculated according to

$$p_i = \frac{m_i - x_i}{\sqrt{\sigma(m_i)^2 - \sigma(x_i)^2}} \;, \qquad r_i = m_i - x_i. \tag{3.1}$$

This pull is computed for the transverse and longitudinal coordinates of the measurement, and the maximum of the two is selected and checked to see if it exceeds a certain threshold. If it does, the hit will be removed, after some additional checks are made to confirm or deny the presence of the outlier. The threshold is set as a member variable `m_outlcut`. The results of varying this cut are described in section 3.4.1.

3.4.1 Cut Optimisation

A systematic variation of the cut point `m_outlcut` has been carried out. The results, demonstrating a reduction in wrong hit assignment whist keeping virtually all good hits assigned to tracks, are shown in fig. 3.9. The rate of wrong hits assigned to tracks decreases from 0.32 to 0.28 at the highest energies (12.5% reduction). Moreover, this result is obtained looking at all tracks inclusively, and the demonstrated improvement removes the need for a specific b-jet ROI (a requirement which led to problems outlined in section 3.3.2). These results hold when looking exclusively at B decay tracks. The fact that, as shown in fig. 3.8a, virtually all correctly assigned

a **b**

Figure 3.9: Profiles, as a function of parent B hadron p_{T}, of good (fig. 3.9a) and wrong (fig. 3.9b) hit assignment rates on the IBL for tracks using baseline tracking (black), and various looser values of the outlier cut.

hits are retained suggests that it may possible to relax this cut further. Tests are ongoing which will confirm this. The current GX2F treats all layers in the pixel detector in the same way - applying the same cut to each. While fig. 3.8a shows no adverse affects for hits on the IBL, when relaxing `m_outlcut` to a value of 1, some small reduction in good hit assignment efficiency was observed in other layers of the pixel detector, which are less precise. This difference in precision motivates the need to treat different layers in the pixel detector differently. To this end, layer-specific cutting capabilities for the GX2F are under development, which will allow each pixel layer to have their own cut point for outlier removal. Layer specific cuts will then be optimised to see if greater numbers of wrong hits can be successfully identified as outliers and removed, while maintaining high good hit assignment efficiency.

3.5 Tracking software validation

- tracking validation
- qspi validation

Chapter 4

Track Classification MVA

4.1 Track Truth Origin Labelling

4.2 Multivariate Track Selection for *b*-tagging

4.2.1 Fake Tracks Identification Tool

Probably talk about this model as a stepping stone to the general classifier

4.2.2 *b*-hadron Decay Track Identification Tool

Maybe don't need this section since it was talked about less

4.3 General Track Origin Classifier Tool

Culmination of this work in the general tool Martino has implemented

Applications:

- Frack to jet association
- Fake track studies (removal and for recommendations)

4.4 Conclusion

Chapter 5

Graph Neural Network Flavour Tagger

- import note

Chapter 6

VHbb Analysis Preamble

6.1 Overview

The Higgs boson, discovered at the LHC in 2012, is predicted by the standard model to decay primarily to two b quarks, with a branching factor of 0.582 ± 0.007 [1]. Observation of this decay mode was recently reported by ATLAS [2]. Whilst the dominant Higgs production mode at the LHC is gluon-gluon fusion, this mode has an overwhelming QCD multijet background and so sensitivity to the Higgs is low. The $\mathrm{H} \rightarrow b\bar{b}$ observation therefore searched for Higgs bosons produced in association with a vector boson (W or Z). This production mechanism results in leptonic final states from the decay of the vector boson, allowing for leptonic triggering, whilst at the same time significantly reducing the multi-jet background.

A closely related analyses now searches for the $\mathrm{H} \rightarrow b\bar{b}$ decay of the Higgs boson, produced in association with a vector boson, when the vector boson and Higgs are highly boosted. The full Run-2 dataset is used for a total integrated luminosity of 139 fb^{-1}. The analysis is split into 0-, 1- and 2-lepton channels depending on the number of selected electrons and muons, to target the $\mathrm{ZH} \rightarrow \nu\nu bb$, $\mathrm{WH} \rightarrow \ell\nu bb$, $\mathrm{ZH} \rightarrow \ell\ell bb$ processes, respectively, where ℓ is an electron or muon. In all channels, events are required to have exactly two b-tagged jets, which form the Higgs boson candidate. At least one of the b-tagged jets is required to have p_{T} greater than 45 GeV. Events are further split into 2-jet or 3-jet categories depending on whether additional, untagged jets are present.

In the 0- and 1-lepton channels, the analysis is further split into signal and control regions. To leading order, there are no additional b-jets in the event other than the two coming from the reconstructed Higgs candidate. For this reason, there is a signal region veto (i.e. events are not accepted into the signal region) for events with additional b-tagged jets in the event. Events with additional b-tagged jets are included in the control region, which is highly pure in $t\bar{t}$ events. The control region is used to constrain the normalisation of the $t\bar{t}$ background.

Chapter 7

VHbb Boosted Analysis

7.1 Overview

7.2 Modelling Work

7.2.1 Background

Source of Uncertainty	Implementation
Renormalisation scale (μ_R)	Internal weights
Factorisation scale (μ_F)	Internal weights
PDF set	Internal weights
α_S value	Internal weights
Parton Shower (PS) models	Alternative samples
Underlying Event (UE) models	Alternative samples
Resummation scale (QSF)	Parameterisation
CKKW merging scale	Parameterisation

Table 7.1: Different sources of uncertainty (i.e. variations in the model) considered for V+jets background, and the corresponding implementation. For each uncertainty, acceptance and shape uncertainties are derived.

Alternative Samples

As mentioned, alternative samples of V+jets events was generated using MadGraph5_aMC@NLO+Pythia8, and the results are compared with the nominal Sherpa 2.2.1 samples. This allows for a comparison of different parton showering and underlying event models, and derivation of the systematic uncertainties on the nominal choice of models.

Internal Weight Variations

Nominal signal samples generated with Sherpa 2.2.1 include systematic variations of certain modelling parameters which are stored as alternative event weights. The samples contain event weight variations which correspond to variations of renormalisation scale μ_R, and factorisation scale μ_F, of 0.5 and 2 times the nominal value. Additionally stored is event weight variations corresponding to 30 different variations on the PDF and two variations of the strong coupling constant α_S. Variations of α_S were found to have negligible impact on the results of the analysis, and are not discussed further.

Parameterisation Methods

While the inclusion of internal weight variation in MC event generators has decreased simulation times and increased available statistics, there are in Sherpa 2.2.1 currently some sources of systematic uncertainty that are unable to be stored as internal weight variations due to technical limitations. Two such systematics relate to the choice of CKKW matrix element merging scale, and resummation scale (QSF). The generation of high statistics alternative samples is a time consuming process, as is typically not done for all samples for every new generator release. A method to parameterise the systematic variation using one sample, and to then apply this parameterisation to another sample, has been developed by the ATLAS SUSY group [3]. This method was used to derive CKKW and QSF uncertainties for the nominal Sherpa 2.2.1 sample, using a previous (lower statistic) Sherpa 2.1 alternative sample. The resulting uncertainties were studied and found to be negligible in comparison with systemics from other sources.

Shape Uncertainties

In order to derive shape uncertainties (which as the name suggests affect shapes but not overall normalisations of distributions), the following procedure is carried out. Normalised distributions of the reconstructed Higgs candidate mass m_J are compared for the nominal sample and variations. For each variation, the ratio of the variation to nominal is calculated, and an analytic function is fit to those sources of variation which have a ratio deviating from unity. If different analysis regions or channels show the same pattern of variation, a common uncertainty is assigned. An example of a significant source of uncertainty, arising from choice of factorisation scale μ_R is shown in fig. 7.1. An exponential function has been fitted to the ratio of the normalised distributions. Two different analysis regions (medium and high $p_{\mathrm{T}}{}^{V}$ bins) are shown. The difference of the shape of the variation means that two separate uncertainties have to be added in the fit, and applied individually in each $p_{\mathrm{T}}{}^{V}$ region.

chapters/6.vhbb_boosted/figs/1L_Whf

chapters/6.vhbb_boosted/figs/

a

b

Figure 7.1: Normalised distributions of leading fat jet mass m_J for medium (fig. 7.1a) and high (fig. 7.1b) $p_{\mathrm{T}}{}^{V}$ analysis regions for W+heavy-flavour-jets (merged in heavy flavours, high and low purity signal regions) in the 0 lepton channel. The renormalisation scale μ_R has been varied by a factor of 2 ("1up") and 0.5 ("1down"). An exponential function has been fit to the ratio.

Acceptance Uncertainties

Several different types of acceptance uncertainties have been calculated. These are implemented as nuisance parameters in the fit and for the most part account for the migration of events between different analysis regions. The list acceptance uncertainties relevant to the V+jets processes are given summarised below.

- **Overall normalisation:** only relevant where normalisation cannot be left floating (i.e. determined in the fit).
- **SR-to-CR relative acceptance:** the uncertainty on the normalisation of the signal region due to events migrating between the signal and control regions.
- **HP-to-LP relative acceptance:** the uncertainty on the normalisation of the high-purity (HP) signal region due to events migrating between the high- and low-purity signal regions.
- **Medium-to-high p_{T^V} relative acceptance:** describes any 'shape' effect in p_{T^V} distribution, given that the analysis only uses two p_{T^V} bins (medium and high).
- **Flavour relative acceptance:** for each flavour Vxx, where $xx \in \{\mathrm{bc,bl,cc}\}$ the ratio of Vxx/Vbb events is calculated. This corresponds to the uncertainty of Vbb events due to the miss-tagging of other flavours Vxx.

The uncertainties on different systematics are summed in quadrature to give a total uncertainty on each region. A summary of the different acceptance uncertainties that were derived in this way and subsequently applied in the fit are given in table 7.2. An effort has been made, wherever possible, to harmonise similar uncertainties across different analysis regions and channels.

7.2.2 Vector Boson + Jets Modelling

The background processes involving W or Z boson decays into leptons (including those in which the W boson arises from a top-quark decay) are collectively referred to as electroweak (EW), or V+jets, backgrounds. W+jets events are most relevant to the 1-lepton channel via the leptonic decay of W$\rightarrow \ell\nu$. In the event of W$\rightarrow \tau\nu$, and subsequent decay of the τ, or the lack of the successful reconstruction of the

e or μ, W+jets can also contribute to the 0-lepton channel. Meanwhile, Z+jets contributes primarily to the 0- and 2-lepton channels via the processes Z$\rightarrow \nu\nu$ and Z$\rightarrow \ell\ell$ respectively.

Modelling is used to predict the outcomes of the analysis and to assess the impact of sources of different systematic uncertainty. Signal and background modelling has has primarily consisted of using Monte Carlo (MC) generators to produce simulated events. The uncertainties on the simulated output must be well understood to perform a successful analysis. To achieve this, a set of "nominal" samples are first defined as a reference to which different variations can be compared. The nominal samples are chosen as the best possible representation of the underlying physical process. "Alternative" samples are used to understand the systematic uncertainties on the nominal samples. To generate an alternative sample, some aspect of the model is varied, and the simulation is re-run. A comparison back to the nominal sample gives a handle on the systematic uncertainty associated with the model parameter which was changed. Detailed information can be found in [4]. In order to access uncertainties associated with the use of MC generators, variations of the data are produced using alternative generators or variation of nominal generator parameters. The variation of nominal generator parameters can in certain cases be implemented using internal weight variations stored alongside the nominal events, and in other cases a new independent sample must be generated. The nominal generator used for V+jets events is SHERPA 2.2.1, while MADGRAPH5_aMC@NLO+PYTHIA8 (which uses different parton showering models) is used as an alternative generator. As production of large MC samples is computationally expensive, a feature of state of the art simulation packages is to store some sources of variation as internal event weights, which can be generated alongside the nominal samples, saving computation time. Several sources of uncertainty, summarised in table 7.1, have been assessed.

V+jets Acceptance Uncertainties				
Boson	**W**		**Z**	
Channel	0L	1L	0L	2L
Vbb Norm.	30%	-	-	-
SR/CR	90%†	40%†	40%	-
HP/LP	18%		18%	-
High/Medium p_T^V	30%	10%*	10%	
Channel Extrap.	20%	-	16%	-
Vbc/Vbb	30%			
Vbl/Vbb	30%			
Vcc/Vbb	20%			
Vcl Norm.	30%			
Vl Norm.	30%			

Table 7.2: V+jets acceptance uncertainties. W+jets SR/CR uncertainties marked by † are correlated. The 1L W+jets H/M uncertainty marked by $*$ is applied as independent and uncorrelated NPs in both HP and LP signal regions. The 0L W+jets Wbb Norm uncertainty is only applied when a floating normalisation for Wbb cannot be obtained from the 1L channel. A 30% uncertainty for Zbb norm is applied in the 1L channel when a floating normalisation for Zbb cannot be obtained from the 0L or 2L channels.

7.2.3 Diboson Modelling

7.3 Fit Studies

7.3.1 Fit Model

A global profile likelihood fit is used to extract the signal strength μ and its significance from the data. This statistical setup treats each bin as a Poisson counting experiment. The combined likelihood over N bins, without considering sources of systematic uncertainty, is given by

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} \exp\left[-(\mu s_i + b_i)\right], \tag{7.1}$$

where s_i (b_i) is the expected number of signal (background) events in bin i, and n_i is the number of events observed in data in bin i. The presence of systematic uncertainties which can affect the expected numbers of signal and background events necessitates the addition of nuisance parameters (NPs), θ, to the likelihood. Each source of systematic uncertainty for V+jets samples discussed in the previous section was implemented as a NP θ_j in the fit. The presence of NPs modifies the likelihood as

$$\mathcal{L}(\mu) \rightarrow \mathcal{L}(\mu, \theta) = \mathcal{L}(\mu) \times \mathcal{L}(\theta) \ , \quad s_i \rightarrow s_i(\theta) \ , \quad b_i \rightarrow b_i(\theta), \tag{7.2}$$

where

$$\mathcal{L}(\theta) = \prod_{\theta_j \in \theta} \frac{\exp\left[-\theta_j^2/2\right]}{\sqrt{2\pi}}. \tag{7.3}$$

Post-fit m_J distributions in the high-purity medium p_{T^V} regions for the 0- and 2-lepton channels are shown in fig. 7.2. The plots show large falling backgrounds, predominantly made up of W+jets and Z+jets events, and a signal distribution corresponding to the Standard Model Higgs boson peaking around $m_H = 125$ GeV.

a

b

Figure 7.2: Post-fit distributions for the 0-lepton (fig. 7.2a) and 2-lepton (fig. 7.2b) channels in the high purity medium p_{T^V} region, obtained in the combined conditional $\mu = 1$ fit to data. The last bin of each plot is an overflow bin.

7.4 Conclusion

Work has been carried out as part of the boosted VHbb analysis group to understand, and implement in the global profile likelihood fit, systematic uncertainties on V+jets samples. This background modelling work is an essential part of the success of the analysis. So far the fit has proved stable with the inclusion of the V+jets uncertainties, and detailed studies are now underway to determine the causes behind any observed pulls of the added NPs. Additional work is ongoing to help with the derivation of uncertainties on diboson samples, another important background. The analysis is already advanced, and is now progressing into its final stages. Publication is expected in the new year.

Chapter 8

VHbb Legacy Analysis

8.1 Overview

Chapter 9

Conclusion

Appendix A

Combining Multiple Triggers

Colophon

This thesis was made in LaTeX 2_ε using the “hepthesis” class [5].

Bibliography

[1] D. de Florian *et al.*, arXiv e-prints , arXiv:1610.07922 (2016), 1610.07922.

[2] M. Aaboud *et al.*, Physics Letters B 786, 59–86 (2018).

[3] J. K. Anders and M. D'Onofrio, CERN Report No. ATL-COM-PHYS-2016-044, 2016 (unpublished).

[4] A. S. Bell and F. Lo Sterzo, CERN Report No. ATL-COM-PHYS-2018-505, 2018 (unpublished).

[5] A. Buckley, *A class for typesetting academic theses*, 2010.