

UNIVERSITY OF OXFORD

DOCTORATE OF PHILOSOPHY

PARTICLE PHYSICS

DPhil Thesis

**Advanced Machine Learning Applications for the Study
of the Dance of the Higgs and Heavy Flavour Quarks at
ATLAS**

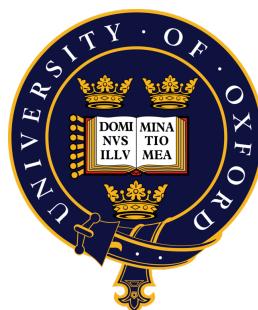
Candidate

Maxence DRAGUET

Supervisor

Daniela BORTOLETTO

2020-2024



Contents

Abstract

This confirmation of status report summarises some of the work I have carried out as part of the $VH(H \rightarrow c\bar{c})$ combined analysis of Run 2, with data from 2016 to 2018, as well as the development of the new jet flavour tagger called *DLLd*, designed to identify b - and c -jets with high efficiency.

Thesis plan

Writing from VS Code In my thesis, I will present a coherent and connected story of the research I have carried out during my DPhil. Starting with two chapters introducing respectively the required theoretical backgrounds and the necessary details of the A Toroidal LHC ApparatuS (ATLAS) experiment, I will then present my work in three main parts. The thesis will be structured around the following set of chapters:

Chapter 1 Introduction.

Chapter 2 Theoretical context: the Standard Model and the Brout-Englert-Higgs mechanism.

Chapter 3 Experimental context: the Large Hadron Collider and the ATLAS experiment, covering details of both apparatuses, object reconstruction, and other topics relevant to flavour tagging in ATLAS.

Chapter 4 Flavour tagging: this part will address the challenge of identifying the flavour of jets in the ATLAS experiment and will present my work on the development of an upgraded tagger called *DL1d* compared to what was then the reference, called *DL1r*. For future considerations, I will also introduce and make references to the development of a new graph neural network tagger called *GN1*, as my work on DL1d served as a basis of comparison for the public note, to which I have contributed, introducing this new algorithm to the experiment [**ATL-PHYS-PUB-2022-027**].

Chapter 5 Development of a *Xbb* tagger: contribution to the development of a $X \rightarrow q\bar{q}$ tagger capable of identifying pairs of *b*- and *c*-jets. Due to the importance of searches targeting events with a final state made of a $b\bar{b}$ or a $c\bar{c}$ pair, a special task force has been set up in ATLAS at the end of 2022 to deliver a tagger capable of efficiently tagging these states. Such a tagger, called *Xbb*, is naturally of great interest for the $VH(H \rightarrow b\bar{b}/c\bar{c})$ analyses. Our team in Oxford joined the effort in the development of this tagger, which is taking place throughout 2023. The first contribution to the project was the retraining of the DIPS and DL1d algorithm with variable-radius track jets, as this particular training of these two algorithms is required to train the benchmark model to which the new tagger will be compared. The next contribution will be to include neutral information in the new graph neural network being developed, thereby complementing the charged information obtained by analysing the tracks.

Chapter 6 This chapter will contain the core of the work carried out during the DPhil on the $VH(H \rightarrow c\bar{c})$ and the $VH(H \rightarrow b\bar{b}/c\bar{c})$ combined analyses using data collected between 2015 and 2018. A coherent and complete overview of the experimental approach used will be presented. Some elements to be discussed are the analysis strategy, studies on the harmonisation of the $VH(H \rightarrow c\bar{c})$ and $VH(H \rightarrow b\bar{b})$, the definition of a common top control region, the full statistical analysis, and, naturally, the outcome of the analysis.

Chapter 7 Conclusion and outlook.

The objective of the proposed structure is to establish a logical chain connecting the different topics: starting from the theory, moving on to the experimental reality of ATLAS, and then going into the details of flavour tagging. The latter is indeed the most important tool for the $VH(H \rightarrow b\bar{b}/c\bar{c})$ analyses, which will be addressed last as it is the culmination of the work to be presented in the thesis. In the present report, only chapters 4 and 6 as above listed are addressed in further detail, as the work with the *Xbb* task force is only starting at this point. The $VH(H \rightarrow b\bar{b}/c\bar{c})$ combined analysis is still ongoing at the moment of writing and the status presented here is therefore not final. Due to the complexity of the combined analysis, this report will focus on the $VH(H \rightarrow c\bar{c})$ sub-analysis and, in particular, the study of the background from top-quark decays. In some regions of the analysis, the top process represents up to 80% of the background at signal-like values of the discriminant variable and is therefore a significant contribution.

1 Tagging Flavour at ATLAS

1.1 Introduction

A fundamental ingredient in any ATLAS analysis is the ability to correctly identify particles in the aftermath of a collision, from τ -leptons, to b - and c -quarks, and gluons g . Having well-calibrated and optimally performing b - and c -tagging tools is of primary importance in studies of the Higgs boson couplings to b - and c -quarks. It is also critical for top t -quark measurements and searches for extensions of the Standard Model (SM). As described by the theory of Quantum Chromodynamics (QCD), colour-charged objects, such as a b - or a c -quark, undergo hadronisation to form collections of colourless hadrons. These hadrons, mostly B for b -quark and D for c -quark, are unstable and further decay in the volume of the detector. Such a succession of decays leaves a collection of particles within a cone oriented in the direction of the original parton, an easily recognisable pattern referred to as a *jet*. From an analysis of the complicated structure of the jet, the flavour of the initially decaying particle can be reconstructed. This is the task of *flavour tagging*. In the ATLAS experiment, the tool to achieve this identification, called *tagger*, is developed and maintained centrally by the Flavour Tagging Group (FTAG).

Tagging b -jets benefits from a particularly advantageous configuration: the b is the lightest element of the third generation and must decay through a flavour-changing process. Because of the SM all value of the V_{bc} Cabibbo–Kobayashi–Maskawa (CKM) matrix element, this decay process is suppressed, giving B -hadrons a characteristically long-lifetime and decay length ($c\tau$) $_B \sim 400 \mu\text{m}$ [Tanabashi:2018oca]. When considering boosted objects with a Lorentz γ factor above unity, the location of the B -hadron decay, called Secondary Vertex (SV), can be reconstructed by the ATLAS pixel detector [Aad:2019aic]. Other important variables describing the decay of B -hadrons are the Impact Parameters (IPs) d_0 and z_0 of charged particles emanating from the SV. As shown in Figure ??, d_0 and z_0 are the transversal and longitudinal distances from the primary vertex to the perigee¹ of the track associated with the charged particle. For a b -jet, either or both of the IPs can be large thanks to the long lifetime of the associated hadron. Other characteristics of b -jets are the large number of particles in the final state following their hadronisation, a property described as *high multiplicity* which is due to the large mass of the b -quark, and the likely presence of leptons in the jet cone, as 40% of the decays of B -hadrons including either an e or a μ [Tanabashi:2018oca].

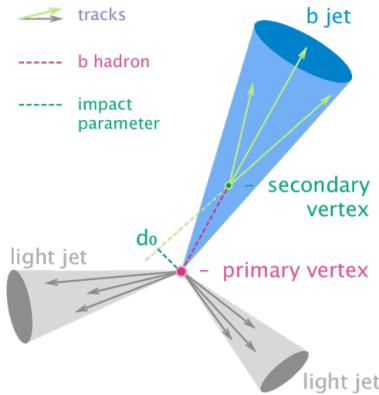


Figure 1: Representation of a b -jet.

On the contrary, tagging c -jets, being at an intermediate scale between light- (u , d , s , and gluons) and b -jets, is much more challenging. The decay length for charged (neutral) D -hadrons, $(c\tau)_D \sim 300$ (100) μm [Tanabashi:2018oca], is smaller than for B -hadrons and is difficult to resolve with the currently deployed tracker. The decay chain of B -hadrons often includes D -hadrons, making a clean separation of c -jets from b -jets harder. Compared to b -jets, c -jets have a lower multiplicity which leads to τ -jets being easily mistaken for c -jets, as these leptons can hadronically decay into a sufficiently large number of particles to mimic a jet in the detector. For all these reasons, less effort has been historically dedicated to constructing c -taggers in ATLAS. The task is however gaining particular attention due to the focus on the challenging $H \rightarrow c\bar{c}$ search [Aaboud:2018fhh]. This chapter presents the development of a novel b - and c -tagger for the ATLAS experiment.

¹The point of closest approach.

1.2 Flavour Tagging at ATLAS

In the ATLAS experiment, a choice was made to develop centrally a tagger to be used by the whole collaboration. It relies on a dedicated set of algorithms to perform simultaneously b - and c -tagging and is continuously improved to meet the requirements of the physics program. Currently, all adopted approaches rely on Deep Learning (DL) methods, given their vastly superior effectiveness. These DL methods rely on high-level features reconstructed by sub-algorithms based on physics variables, such as the tracks IPs, and the reconstruction of secondary vertices. The low-level algorithms consist of [Paganini:2289214]:

- Jet Fitter: a vertexing algorithm based on a Kalman filter to reconstruct the topology and fit the decay chain Primary Vertex (PV) $\rightarrow B \rightarrow D$ with the assumption that the vertices of the weakly decaying B/D-hadrons tend to align with the PV [ATL-PHYS-PUB-2018-025].
- Secondary Vertex method 1 (SV1): combining a secondary vertex finder and a tagger to offer flavour discrimination information. The former, based on the VKalVrt vertex reconstruction package [Kostyukhin:685551], returns a list of candidate secondary vertices with measured quantities assigned to each vertex. The latter derives jet weights based on discriminative variables and computes properties of the SV, such as the mass.
- IP likelihood: IP2D and IP3D are both likelihood-based methods to assign flavour-discriminating weights based, respectively, on the transversal and global impact parameters significance (corresponding to the reweighted IP variables by their respective uncertainties) of the tracks [ATLAS:2017bcq].
- Track collection analyser: either with Recurrent Neural Network Impact Parameter (RNNIP) [ATL-PHYS-PUB-2017-003] or Deep Impact Parameter Sets (DIPS) [ATL-PHYS-PUB-2020-014]. These are *Neural Network* (NN) approaches to extract discrimination information on the set of tracks associated with a jet. These taggers are further described later in this section.

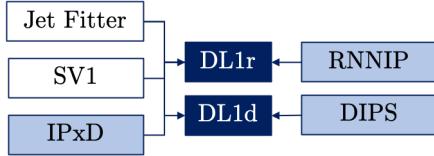


Figure 2: The algorithms for flavour tagging at ATLAS. High-level taggers are in dark blue, track-based taggers in light blue and vertex-related taggers in white.

The outputs of these low-level algorithms, as well as certain jet-related variables, such as p_T , are then combined as input to a high-level tagger consisting of a fully-connected called Deep Learner 1 Model with RNNIP input (DL1r) or Deep Learner 1 Model with DIPS input (DL1d), respectively if RNNIP or DIPS is used. The input vector is typically made of 44-45 features. This high-level tagger outputs three probabilities p_X for the analysed jet to correspond to a b -, c -, or light-flavour (indicated with the letter u) such that $p_b + p_c + p_u = 1$. A b -tagging score D_b is then derived by computing a scaled log-likelihood ratio:

$$D_b = \log \frac{p_b}{f_c^b \times p_c + (1 - f_c^b) \times p_u}, \quad (1)$$

where f_c^b is the charm fraction, a parameter that can be modified to tweak the importance of each flavour. The analogous c -tagging score D_c is:

$$D_c = \log \frac{p_c}{f_b^c \times p_b + (1 - f_b^c) \times p_u}. \quad (2)$$

A jet is X -tagged if the D_X discriminant score is above a set threshold constant c_{wp} , defining a *Working Point* (WP) with a unique configuration of signal and background (mis-tag) efficiencies. In this context, the efficiency ϵ_Y^X for Y -flavour jets to be X -tagged and the corresponding rejection \mathcal{R}_Y^X are respectively defined as:

$$\epsilon_Y^X = \frac{N_{Y-jets}^{X\text{-tagged}}}{N_{Y-jets}} \quad \text{and} \quad \mathcal{R}_Y^X = \frac{1}{\epsilon_Y^X}, \quad (3)$$

where $N_{Y-jets}^{X\text{-tagged}}$ and N_{Y-jets} are respectively the number of X -tagged Y -flavoured jets and the total number of Y -flavoured jets.

These high-level models are trained on Monte Carlo (MC) simulated data samples and need to be calibrated on real data to deliver an unbiased estimate, by deriving weights correcting the predictions for each jet. Uncertainties are derived on the predicted score and passed along to analyses using the tool. The novel algorithm introduced in this work is the DL1d tagger, which relies on the DIPS sub-tagger to extract correlations between the tracks.

1.2.1 RNNIP

The RNNIP tagger runs on arbitrary-length input sequences made of track features, as ordered by the absolute transverse IP significance, to extract tagging information from correlations between tracks [ATL-PHYS-PUB-2017-003]. The vector of track features includes the transverse and longitudinal impact parameter significances, the jet p_T fraction carried by the track, the distance between the track and the jet axis, and a learned 2D embedding of the track quality [Paganini:2289214]. It outputs a probability p_X for the jet to belong to flavour $X \in [b, c, \text{light}]$.

1.2.2 dips

The DIPS tagger, based on the Deep Sets architecture [NIPS2017_f22e4747], is a Graph Neural Network (GNN) approach to model the correlations between an arbitrary number of tracks. Two NNs are trained in this effort. A model Φ maps each individual track feature vector (similar to RNNIP) to a latent space (forming the nodes of the graph). The representations of each track in this latent space are then concatenated by a simple summation operation (the edges of the graph) and given as input to a secondary model F outputting the predicted probability p_X for the jet to belong to flavour $X \in [b, c, \text{light}]$ (the prediction of the graph). This approach has several advantages over RNNIP, mainly the physically motivated permutation-invariance of the input and the improved training/evaluation time thanks to a more parallelisable architecture. Furthermore, the performance delivered by DIPS is observed to globally outperform RNNIP [ATL-PHYS-PUB-2020-014].

1.3 Samples and Training Procedure

In preparation for the next run of the LHC, ATLAS has improved its reconstruction software. As such, important aspects of flavour tagging have changed, requiring to retrain the taggers to ensure optimal performance. This work presents the first ATLAS study of the retraining of DL1r on the new release (R22) and the inclusion of DIPS in the complete flavour tagging tools, a tagger called DL1d. Another novelty is the possible inclusion of τ -jets, to improve their rejection from c -jets. However, due to the widespread use of the FTAG algorithms and the difficulties arising in calibrating a tagger with exceptionally good rejection against τ -jets, these are not included in the default version of the tagger nor in the results shown here.

Two samples from proton-proton collisions at $\sqrt{s} = 13$ are simulated for this analysis:

- $t\bar{t}$ events, with at least one of the W -boson from the top-quark decay further decaying leptonically.
- Z' events, where an exotic boson Z' decays as $Z' \rightarrow q\bar{q}$ or $\tau\bar{\tau}$, with a variable Z' mass to generate a flat p_T spectrum extending the p_T -range of the jets studied up to 4 TeV.

For both samples, the ATLAS detector is simulated using GEANT4 [Agostinelli:602040] and jets are reconstructed with the anti-kT algorithm with a radius of $R = 0.4$. For RNNIP, the tracks considered must pass the following quality selection: ≥ 7 hits in the silicon layers, ≤ 2 missing hits in the silicon layers, ≥ 1 hit in the pixel detector, ≤ 1 hit shared by multiple tracks, $p_T > 1$ GeV, $|d_0| < 1$ mm, and $|z_0 \sin(\theta)| < 1.5$ mm. For DIPS, a looser track selection is applied to capture more tracks, modifying the nominal selection in the following way: $p_T > 0.5$ GeV, $|d_0| < 3.5$ mm, and $|z_0 \sin(\theta)| < 5$ mm [ATL-PHYS-PUB-2020-014]. Loosening the selection was observed to lead to a significant improvement in performance for jets with a $p_T < 250$ GeV for DIPS.

These two samples are combined into a hybrid sample to train the taggers, with 70% of the total number of jets coming from $t\bar{t}$ and the remaining from the Z' . The $t\bar{t}$ and Z' samples cover, respectively, a low- and high- p_T region based on a reconstructed B -hadron p_T separation threshold of 250 GeV for b -jets and a jet p_T of 250 GeV for non- b -jets. The evaluation of the performance of a trained tagger is performed on the separated sets and unfolded over the flavours.

ATLAS flavour tagging tools are widely used across the collaboration. It is therefore important for the taggers not to learn specific features of the simulations used for training but to focus on the inherent differences between the studied flavours. The hybrid sample is therefore downsampled in $[p_T - \eta]$ bins to have the same number of b -, c -, and light-jets in each 2D bin. The total statistics available is of 25×10^6 jets per flavour for training. The $t\bar{t}$ and Z' samples for validation and testing are each made of 1 million jets and are not downsampled.

Training is done with the Umami framework [Umami] based on TensorFlow [tensorflow2015-whitepaper] for 300 epochs with a variable learning rate schedule and the default network structure adopted in already released DL1r: 8 fully connected of smoothly-decreasing sizes in [256, 128, 60, 48, 36, 24, 12, 6] with ReLu activation leading to a final softmax layer producing the predicted probabilities for each flavour. While the DIPS probabilities used as inputs to DL1d come from a model trained on the new release, the RNNIP probabilities are still from a model trained on the previous one (R21)

[ATL-PHYS-PUB-2017-003, ATL-PHYS-PUB-2020-014]. Indeed, due to the large difference in performance, RNNIP is no longer supported in the new release and is included for sake of comparability to the previous techniques. The models at an epoch offering the best combined results in terms of b -tagging efficiency and rejection from b -jets on the validation set are selected for further analysis. Importantly, every training converged to a fixed set of performance values, with no overtraining occurring.

Several modifications to the model architecture, list of input variables, and preprocessing and training procedures have been explored, with no significant gain observed:

- The preprocessing steps were revised to reduce the size of the evaluation sets for the benefit of the training one. A dual approach, down-sampling light-jets and up-sampling c -jets to the b -jets [$p_T - \eta$], has also been implemented. This approach uses importance sampling with replacement to obtain the same fraction of the different flavours and the same p_T and $|\eta|$ distribution. While the performance of the majority classes was observed to improve, the efficiency at tagging the upsampled minority class (c -jets) was slightly lower. This trade-off can be controlled by modifying the flavour fractions.
- Several modifications to the list of input features have been attempted, with no clear advantage uncovered. Adding pile-up information (the actual number of interactions per crossing and the number of primary vertices were tested) was not observed to have an impact on the tagging efficiency. Adding other variables from SV1 or JetFitter was also not observed to improve performance. However, a positive observation is that the IP2D and IP3D taggers can both be safely removed without change to performance, as the information they add is in all likelihood now covered by the DIPS sub-tagger, thereby reducing the list of sub-taggers to maintain.
- The structure of the network and its training procedure. Using samples produced with an older release of the ATLAS software to pre-train the model was not observed to deliver a boost in performance. Changing the size of the network and the batch size was also not observed to have a positive effect.

The performance of the retrained DL1r tagger on the new release was found to be in good agreement with the currently recommended DL1r, despite using the same training of RNNIP on the previous release.

1.4 Analysis of Results

In order to establish a meaningful benchmark for the newly trained taggers, the performance of the recommended DL1r tagger, trained and evaluated on an analogous set of samples from the previous release, is included in the following results as benchmark under the label *Recom. DL1r*.

Figure ?? presents the Receiver Operating Characteristic (ROC) curves on the $t\bar{t}$ (left) and Z' (right) samples for b -tagging. These ROC plots show, on the x -axis, the b -tagging efficiency (ϵ_b^b) versus, on the y -axis, the rejection \mathcal{R}_Y^b for $Y \in [c, \text{light}]$. The two bottom sub-plots present the ratio of the c -jet rejection and light-jet rejection curves to the blue ones. This blue curve is the recommended DL1r performance and serves as the baseline of the comparison, while the new tagger DL1d is plotted in orange. Figure ?? shows the same plots for c -tagging, with respect to b - and light-jet rejection. The important observation is the clear gain obtained when replacing RNNIP with DIPS. Both the b - and c -tagging performance of DL1d clearly dominate the DL1r version, with a significant improvement in background flavour rejection for all tagging efficiency considered, as summarised in Table ???. The largest improvement in performance is obtained for b -tagging on the $t\bar{t}$ process, corresponding to a lower jet momentum.

In the light-rejection from b -jets ROC curves in Figure ??, there is an elbow in the curve at high b -jet efficiency. This effect is also present in the b -rejection from c -tagging, in Figure ???. Both correspond to a set of, respectively, light-jets and b -jets that do not overlap with the b -jets b -tagging and c -jets c -tagging discriminants distributions, as shown in Figures ?? and ???. These “background” jets are easily removed from the core set of “signal” jets due to internal differences between the flavours and the discrete nature of some sub-taggers used.

The background rejections of the various taggers for b -tagging (c -tagging) as a function of the jet transverse momentum at a fixed b -efficiency of 70% (c -efficiency of 30%) per region displayed are shown in Figure ?? (Figure ??). Throughout the p_T range considered, DL1d outperforms the DL1r tagger. The low p_T b -rejection from c -jets is noticeably better for the retrained tagger compared to DL1r.

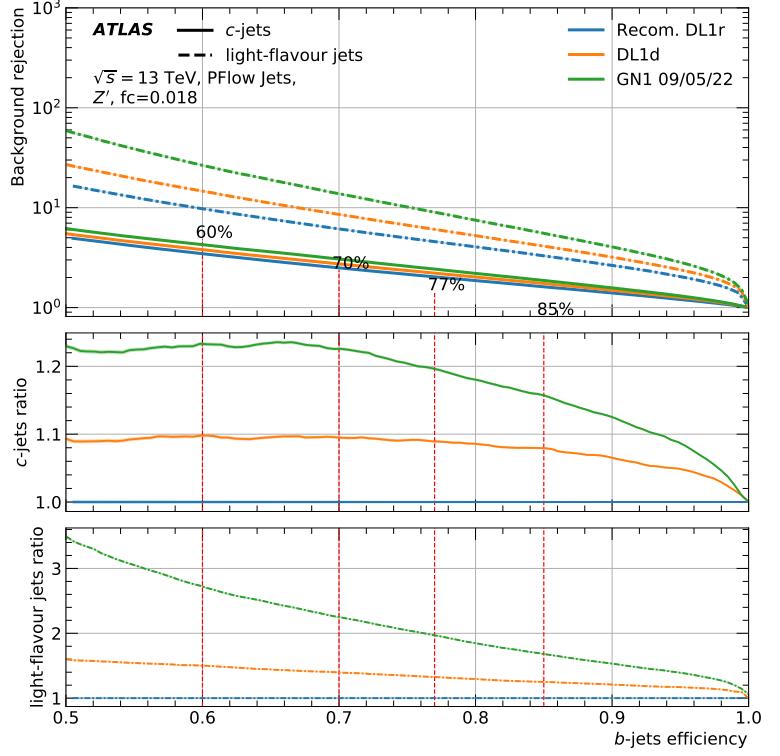
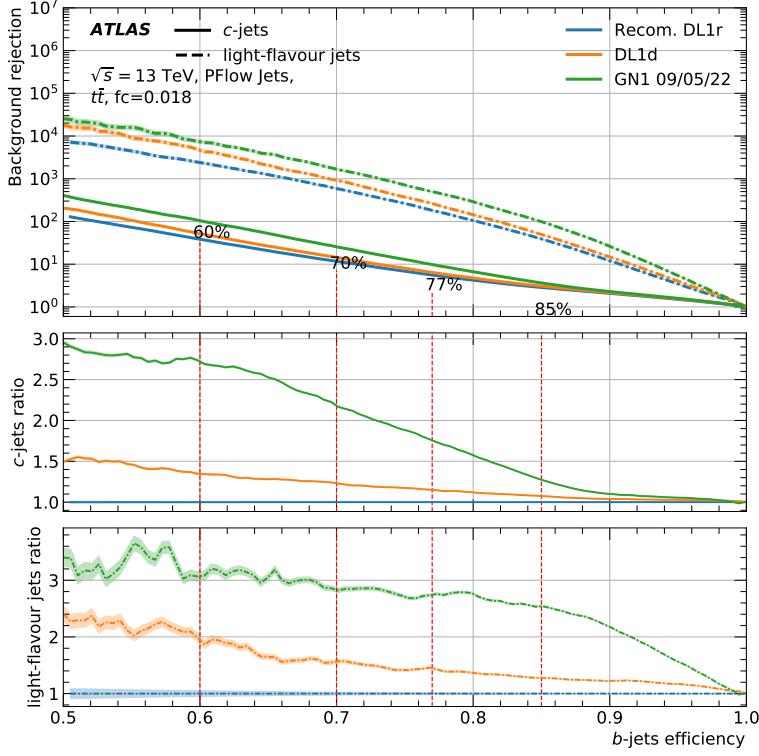


Figure 3: Performance for b -tagging with a flavour fraction of $f_c^b = 0.018$. Left: $t\bar{t}$; right: Z' . Top: ROC curves; centre: ratio of c -jets rejection from b -jets relative to the R22-retrained DL1r; bottom: same ratio for light-jets rejection. List of taggers: recommended DL1r from the previous release; DL1d trained on the new release; Graph Network 1 Model (GN1) test-model trained on the new release.

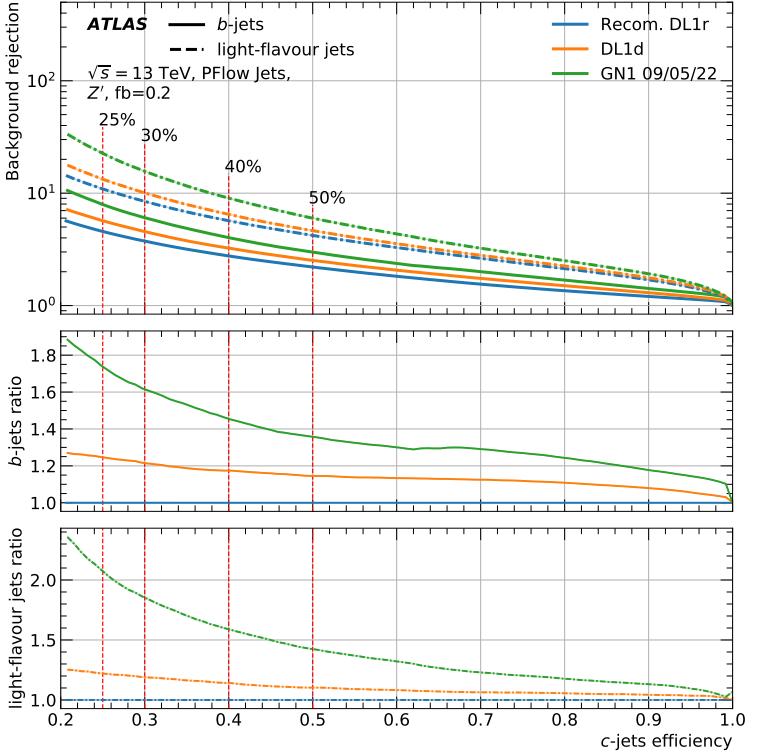
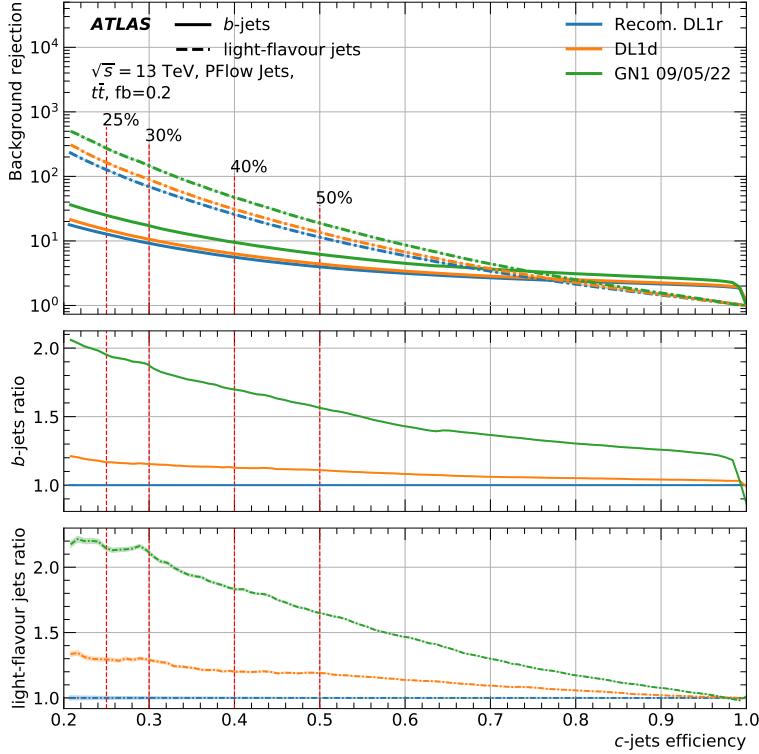


Figure 4: Performance for c -tagging with a flavour fraction of $f_b^c = 0.2$. Left: $t\bar{t}$; right: Z' . Top: ROC curves; centre: ratio of b -jets rejection from c -jets relative to the R22-retrained DL1r; bottom: same ratio for light-jets rejection. List of taggers: recommended DL1r from the previous release; DL1d trained on the new release; GN1 test-model trained on the new release.

| <i>b</i> -tagging on $t\bar{t}$ | | | <i>b</i> -tagging on Z' | | |
|---------------------------------|---------------------|-----------------|---------------------------|---------------------|-----------------|
| WP | <i>c</i> -rejection | light-rejection | WP | <i>c</i> -rejection | light-rejection |
| 60% | +26% | +73% | 60% | +19% | +43% |
| 70% | +19% | +56% | 70% | +10% | +32% |
| 77% | +12% | +41% | 77% | +9% | +26% |
| 85% | +7% | +32% | 85% | +6% | +19% |

| <i>c</i> -tagging on $t\bar{t}$ | | | <i>c</i> -tagging on Z' | | |
|---------------------------------|---------------------|-----------------|---------------------------|---------------------|-----------------|
| WP | <i>b</i> -rejection | light-rejection | WP | <i>b</i> -rejection | light-rejection |
| 25% | +26% | +5% | 25% | +12% | +22% |
| 30% | +25% | +9% | 30% | +11% | +19% |
| 40% | +22% | +12% | 40% | +8% | +14% |
| 50% | +18% | +15% | 50% | +7% | +10% |

Table 1: The ratio of background flavour rejection of the DL1d to DL1r at various tagging efficiencies, both trained on the new release. Top: *b*-tagging ($f_c^b = 0.018$); bottom: *c*-tagging ($f_c^c = 0.2$); left: $t\bar{t}$; right: Z' .

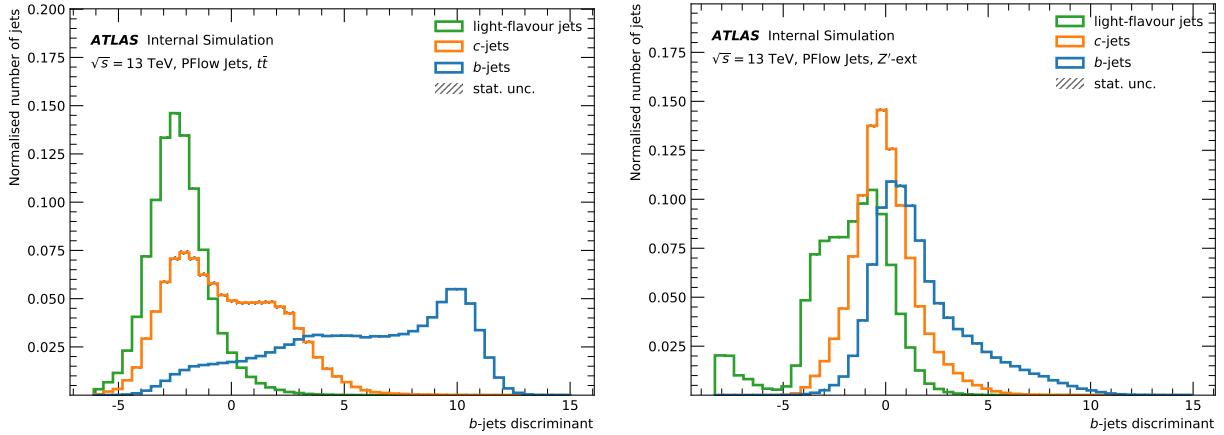


Figure 5: Distribution of DL1d *b*-tagging discriminant for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

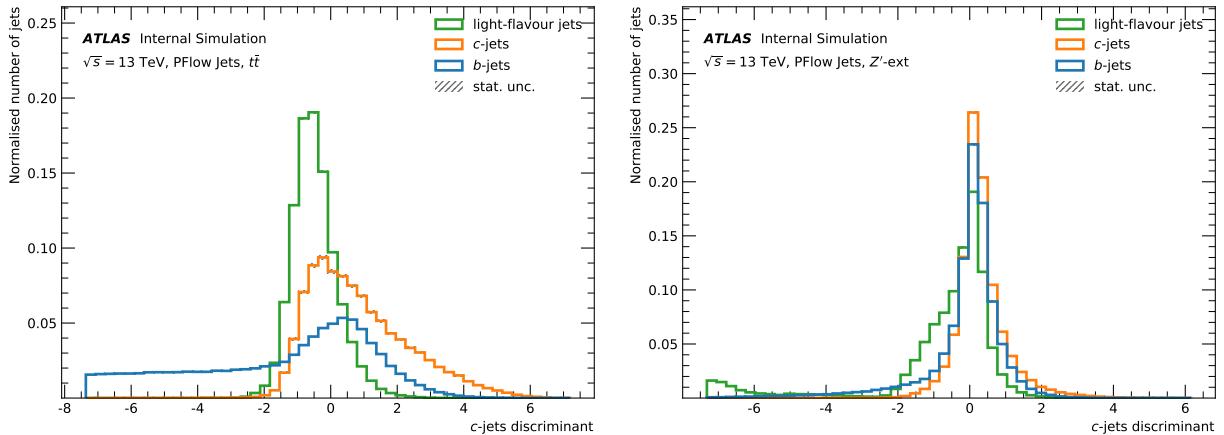


Figure 6: Distribution of DL1d *c*-tagging discriminant for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

In Figures ?? and ??, a new tagger currently in development is introduced: GN1 [ATL-PHYS-PUB-2022-027]. This model is based on a graph neural network directly processing low-level inputs, thereby diverging from the traditional ATLAS flavour tagging philosophy of combining several low-level sub-taggers into a high-level one, such as in DL1d. GN1 uses the information associated with charged tracks in a jet to directly output the flavour-tag probabilities, which are then combined into analogous discriminants to Equations ?? and ???. Alongside predicting the flavour of the jet flavour, two auxiliary objectives are also optimised for the training:

- the physics process that produced the different tracks and
- grouping the tracks into vertices.

These complementary objectives improve performance and provide useful information on the content of the jets. Thanks to this guided training procedure, GN1 is able to compete favourably with the higher-level taggers presented so far, with on average a remarkable increase in the background flavour rejection for all tagging efficiencies. Another non-negligible advantage of the GN1 approach is the simplification of the structure, removing the need to maintain and retune several low-level algorithms. At the time of compiling the results, this tagger is still in development and is included solely to offer an exciting suggestion of the future performance of the tools deployed by the ATLAS flavour tagging group.

1.5 Conclusion

This chapter introduces the novel tagger DL1d to the set of flavour tagging tools available in ATLAS. Based on the DIPS sub-tagger, DL1d is found to have improved background rejections at a fixed working point for both b - and c -tagging. In parallel to the development of this new tagger, the preprocessing pipeline has been revamped and several changes to the input features list and model architecture were explored in order to deliver the best-performing tagger to the collaboration.

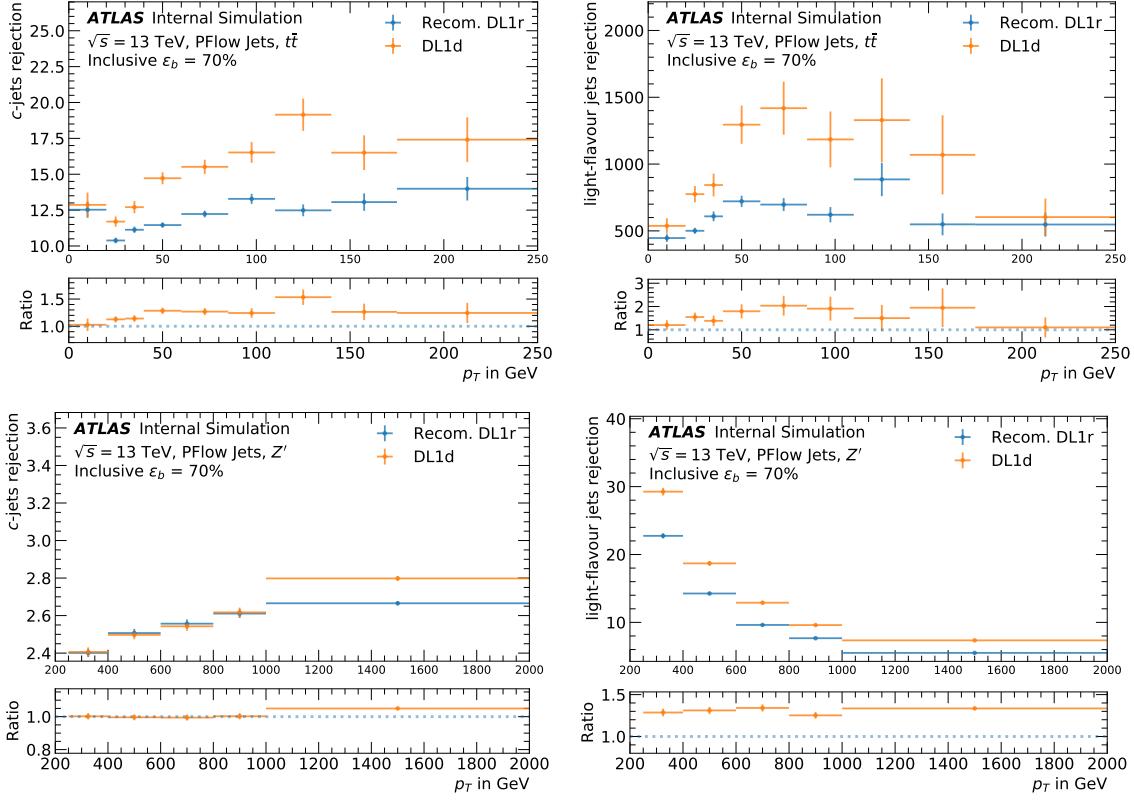


Figure 7: Background flavour rejections at a fixed b -tagging efficiency of 70% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z' ; left: c -rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

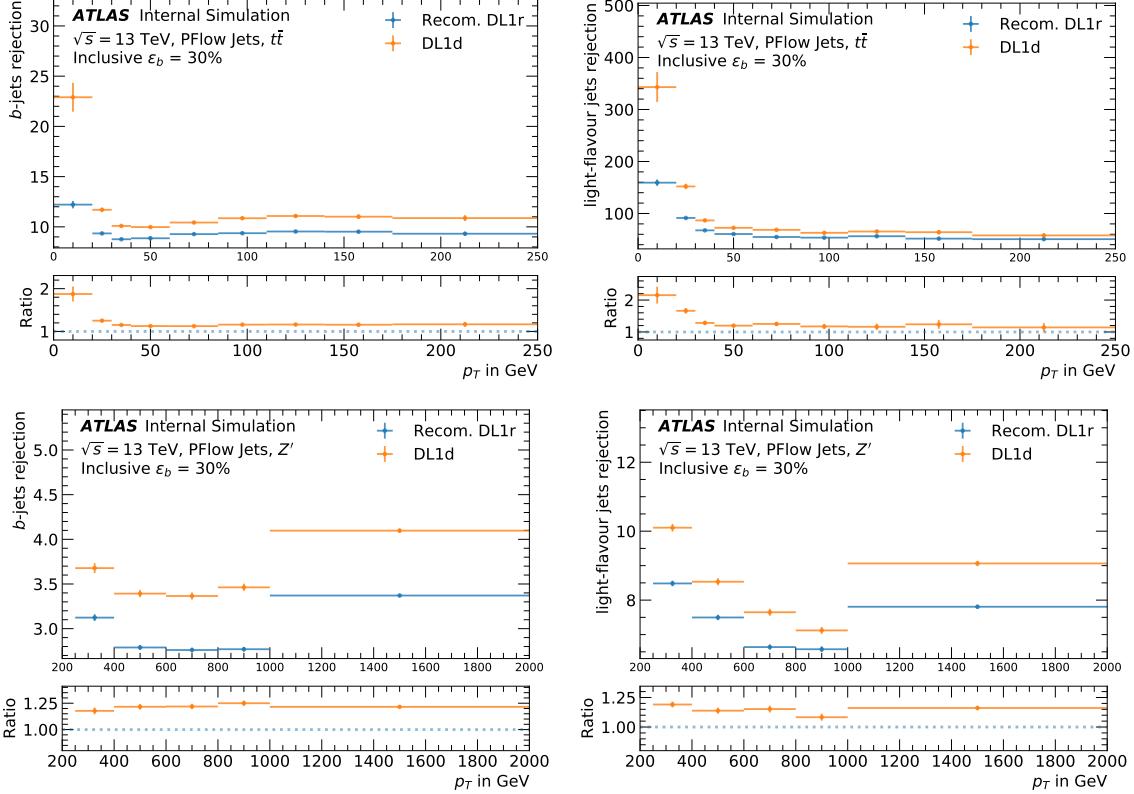


Figure 8: Background flavour rejections at a fixed c -tagging efficiency of 30% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z' ; left: b -rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

2 The $VH(H \rightarrow b\bar{b}/c\bar{c})$ Combined Analysis

2.1 Introduction

Perhaps the most important *raison d'être* of the *Large Hadron Collider* (LHC) was to discover the Brout-Englert-Higgs boson (Higgs - H), a feat achieved by the ATLAS and CMS Experiments in July 2012 [[ATLAS:2012yve](#), [CMS:2012qbp](#)]. Theorised in 1964 by two independent papers introducing the mechanism of spontaneous symmetry breaking to give mass to the gauge bosons [[Englert:1964et](#), [PhysRevLett.13.508](#)], its discovery almost forty years later marked one of the greatest achievements of the particle physics community. The Higgs boson is an essential part of the SM as it is tied to the mechanism through which particles acquire mass without breaking the electroweak gauge invariance.

Since then, both experiments have been studying the specific properties of the discovered particle, and in particular the different production processes and decay channels. During the LHC Run 2, corresponding to data taken from 2015 to 2018, the $t\bar{t}H$ production mechanism was observed for the first time [[ATLAS:2018mme](#), [CMS:2018uxb](#)]. The decay channel of the Higgs boson to a $b\bar{b}$ pair was observed [[ATLAS:2018kot](#), [CMS:2018nsn](#)] and there now is evidence of the decay to a $\mu^-\mu^+$ pair [[ATLAS:2020fzp](#), [CMS:2020xwi](#)]. The former decay channel is of significance since it has the largest predicted branching ratio of 58% for $m_H = 125$ GeV. The latter is the first evidence of the decay of the Higgs to second-generation fermions. Furthermore, constraints on the branching ratio of the H to another second-generation fermion, the c -quark, have been set by both collaborations studying the $H \rightarrow c\bar{c}$ decay [[Aaboud:2018fhh](#)]. This decay mode is the most common Higgs decay mode that has yet to be observed. It is indeed particularly challenging due to the small predicted branching ratio of 2.89% compared to $H \rightarrow b\bar{b}$, the large background rate, and the experimental difficulties in identifying c -jets. It is a fertile ground for new physics beyond the SM as well as an important test of the validity of the model. The fermion couplings in the SM were indeed added ad-hoc and there is a distinct mass hierarchy between the three generations of quarks that can be probed by studying their coupling strengths to the Higgs boson. In the $VH(H \rightarrow b\bar{b}/c\bar{c})$ analysis, the hierarchy of mass between the b - and c -quark, respectively a 3rd and 2nd generation quark, is probed.

2.2 The $VH(H \rightarrow b\bar{b}/c\bar{c})$ analyses in ATLAS

While $H \rightarrow b\bar{b}$ enjoys the largest branching ratio at the observed Higgs mass, the large multi-jet background in a hadron collider like the LHC makes this decay mode very challenging. The measurements for both the $b\bar{b}$ and $c\bar{c}$ decay modes are therefore performed in a so-called *associated production mode*, where the H is produced in addition with an extra vector boson V (W or Z) decaying leptonically, to electrons (e), muons (μ), neutrinos (ν), or a combination $e\nu$ or $\mu\nu$. Taus (τ) are not included at the moment though some tests are being run to add them, in particular, to migrate 0L-channel events with a hadronically decaying τ to the 1L channel. Despite the relatively small cross-section of the VH production mode ($\sigma_{VH} = 2.25$ pb compared to the total H production $\sigma_H \approx 51$ pb), the process benefits from experimentally favourable conditions thanks to the presence of leptons in the event signature: these allow for efficient triggering, the act of rapidly deciding whether to further process and store an event when collecting the data, and greatly reduce the contribution of the multi-jet background.

In the ATLAS collaboration, the $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$ analyses adopt very similar strategies, with the main ingredient being the ability to reliably tag the flavour of jets produced in an event, thereby reconstructing the heavy quark pair produced in the H decay. Using the full Run 2 dataset, the published ATLAS analysis obtained the following upper limits on the signal strength of the $VH(H \rightarrow c\bar{c})$ as predicted by the SM: an observed (expected) upper limit of $26 \times \text{SM}$ ($31 \times \text{SM}$) [[Collaboration:2721696](#)]. For comparison, CMS reported an observed (expected) upper limit of $14 \times \text{SM}$ ($7.6 \times \text{SM}$) [[arXiv:2205.05550](#)].

For the $VH(H \rightarrow b\bar{b})$, thanks to a larger expected signal, the analysis reaches a sensitivity of 6.7 standard deviations [[ATLAS:2020fcg](#)]. Therefore, more detailed measurements can be made, such as the cross-section as a function of momentum in the reduced Simplified Template Cross-Section (STXS) scheme. To probe larger p_T ranges, the analysis is now split into the *resolved* [[ATLAS:2020fcg](#)] and *boosted* [[ATLAS:2020jwz](#)] analyses, with the latter restricting to values of the transverse momentum of the associated vector boson p_T^V above 250 GeV. The name of these analyses comes from the ability to independently resolve the two b -jets into two distinct small cone radius (small- R) jets at low p_T^V . At high p_T^V , the Higgs p_T^H is highly Lorentz-boosted and a single large-radius ($R = 1$) jet, merging the two b -jets, can be reconstructed as a candidate for the Higgs. The measured signal strengths, the ratio of the measured yield to the SM predictions, are:

- For the resolved analysis in Run 2: a signal strength of $1.02^{+0.18}_{-0.17}$ corresponding to an observed (expected) significance of 6.7 (6.7) standard deviations [[ATLAS:2020fcg](#)]. Due to the good sensitivity of the analysis, the result is further detailed into the WH and ZH production processes with observed (expected) significances of, respectively, 4.0 (4.1)

and 5.3 (5.1) standard deviations. Furthermore, the VH cross-section times the $H \rightarrow b\bar{b}$ and $V \rightarrow$ leptons branchings fractions ($\sigma \times BR$) are reported in the reduced STXS scheme. Finally, limits are set on the coefficients of effective Lagrangian operators which can affect the VH production and the $H \rightarrow b\bar{b}$ decay.

- For the boosted analysis: a signal strength of $0.72^{+0.39}_{-0.36}$ corresponding to an observed (expected) significance of 2.1 (2.7) standard deviations [ATLAS:2020jwz].

It is interesting to note that the combination has already been tested for the resolved $VH(H \rightarrow b\bar{b}) + VH(H \rightarrow c\bar{c})$ [Collaboration:2721696] and the resolved + boosted $VH(H \rightarrow b\bar{b})$ [ATLAS:2021wqh]. However, these used the published analyses and the objective of the new Combined Analysis is to define a common analysis strategy for both Higgs decay modes, thereby improving the measurements of $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$ simultaneously. This combined measurement has several additional benefits:

- The Higgs-charm and -beauty coupling modifiers κ_c and κ_b can be measured directly, as well as their ratio κ_c/κ_b .
- The auxiliary measurements of background processes are shared, leading to a better knowledge of background processes that contribute to both phase spaces such as the $V+jets$ and top-quark processes.

The rest of this section focuses on the current state of the $VH(H \rightarrow c\bar{c})$ part of the Combined Analysis, as the analysis is not yet concluded.

2.3 Analysis strategy

The Combined Analysis is performed with the full ATLAS Run 2 proton-proton collision data, from 2015 to 2018, for a total integrated luminosity of 140 fb^{-1} at a centre of mass energy $\sqrt{s} = 13 \text{ TeV}$. The regions and boundaries between the different regimes of the analysis are illustrated in Figure ???. The separation is based on jet flavour tagging (for $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$) and on a p_T^V cut of 400 GeV (for the resolved - boosted $VH(H \rightarrow b\bar{b})$). The ($\rightarrow b\bar{b}$) - ($\rightarrow c\bar{c}$) analyses are separated by the required presence of two b -tagged jets or a c -tagged jets respectively. The p_T^V cut marks the difference between the Higgs candidate reconstruction scheme: two small radius ($R = 0.4$) jets below a p_T^V of 400 GeV and one large radius ($R = 1$) jets with two sub-jets above.

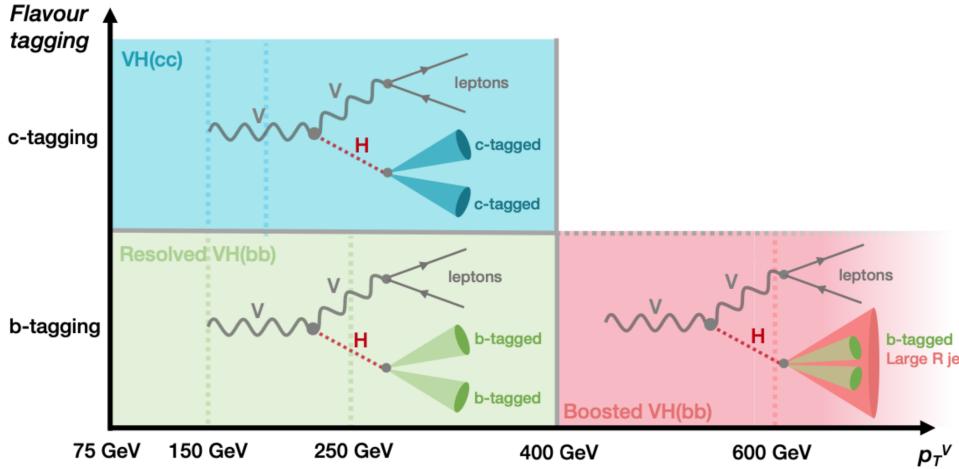


Figure 9: The analysis regimes considered in the $VH(H \rightarrow b\bar{b}/c\bar{c})$ analysis, from the internal documentation of the team [Chisholm:2743096].

For all analysis regions, three different channels are defined based on the decay mode of the vector boson V : $Z \rightarrow vv$ (0-lepton - 0L), $W \rightarrow lv$ (1-lepton - 1L), and $Z \rightarrow l^+l^-$ (2-lepton - 2L), where l refers to an electron or a muon and v to any type of neutrino. The signal considered are the $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$ processes. The SM diboson processes $VZ(Z \rightarrow b\bar{b})$ and $VZ(Z \rightarrow c\bar{c})$ are considered as signals in a cross-check analysis. Having a larger cross-section, these processes kinematically similar to the signals can be measured with good statistical significance, thereby offering a suitable test case to verify the performance of the strategy deployed. The main backgrounds are the $V+jets$ (mostly $Z+jets$ in 0L and 2L, $W+jets$ in 1L) and top-quark processes (predominantly the top-quark pair production $t\bar{t}$, with one of the t decaying leptonically, and a sub-leading contribution from single top-quark production, both in 0L and 1L). Minor backgrounds are the multi-jet (thanks to the required presence of leptons) and diboson pair productions (VV). All backgrounds are simulated

using Monte Carlo (MC) simulation packages, except for the multi-jet which is estimated from a data-driven method in the 1L channel and is negligible in other channels. To reproduce the conditions of the ATLAS detector, simulated samples are passed through the GEANT4 software [**Agostinelli:602040**] and the ATLAS reconstruction software.

2.3.1 Objects description

When a particle passes through the ATLAS detector, different types of electronic signals are collected by the various sub-detectors. This raw information is then processed in successive computational steps to reconstruct high-level physics objects that are then used by analyses. For the particular $VH(H \rightarrow c\bar{c})$ analysis, the most important objects are:

- Jets: calorimeter jets with a radius $R = 0.4$ reconstructed with the Particle Flow and anti- k_t algorithms [**Cacciari:2008gp**] are used. A jet is considered as *signal* if it has a $p_T > 20$ GeV and $|\eta| < 2.5$. It is considered as *forward* if $2.5 \leq |\eta| < 4.5$, with a $p_T > 30$ GeV. Because the Higgs candidates are selected based on the tagged flavour and the tagger works up to $|\eta| = 2.5$, due to its reliance on the tracking detectors, only signal jets are considered as candidates. To suppress pile-up jets, which is predominately made of light-jets, a 30 GeV p_T is applied to all non-Higgs candidate jets. For the Higgs candidates, the lower p_T threshold of 20 GeV is kept, as flavour tagging suppresses most of these light-jets.
- Electrons: identified with a likelihood-based method matching a deposit in the electromagnetic calorimeter with a track [**Aaboud:2657964**]. Requirements are summarised in Table ?? and vary depending on the lepton channel. A VH -Loose with loose likelihood identification is applied to electrons in all channels. Additionally, the ZH -Signal and WH -Signal criteria are applied in the 1L and 2L channels respectively, with a tighter p_T due to the trigger threshold. The 1L likelihood identification and isolation selections are tighter to suppress the multi-jet background.
- Muons: reconstructed by matching an energy deposit in the muon detector with information from the Inner Detector and Muon Spectrometer [**Aad:2746302**]. Requirements are summarised in Table ?? and vary depending on the lepton channel. The VH -Loose requirement is applied to muons in all channels. The ZH -Signal and WH -Signal are additionally applied to the 1L and 2L channels respectively.
- E_T^{miss} (MET): neutrinos are not detectable in ATLAS and their presence can be inferred from momentum imbalance in the transverse plane. E_T^{miss} is the negative vectorial sum of the p_T of physics objects (leptons, jets, ...) as well as a track-based *soft term*, introduced to include a contribution from good quality tracks not associated to a main physics object.
- Taus: hadronically decaying tau-leptons are identified and vetoed in 1L using a recurrent neural network (RNN) [**ATL-PHYS-PUB-2019-033**]. In 0L and 2L, if the jet passes the RNN requirement for hadronically decaying tau-leptons, it is no longer considered as a jet and cannot be considered as a Higgs candidate.

| Electron Selection | p_T | η | ID | d_0^{sig} w.r.t. BL | $ \Delta z_0 \sin \theta $ | Isolation |
|--------------------|-------------------|-----------------|----------|------------------------------|----------------------------|----------------|
| VH-Loose | >7 GeV | $ \eta < 2.47$ | LH Loose | < 5 | < 0.5 mm | Loose_VarRad |
| ZH-Signal | >27 GeV | | | Same as VH-Loose | | |
| WH-Signal | Same as ZH-Signal | | LH Tight | Same as ZH-Signal | | HighPtCaloOnly |

Table 2: Electron Selection

| Muon Selection | p_T | η | ID | d_0^{sig} w.r.t. BL | $ \Delta z_0 \sin \theta $ | Isolation |
|----------------|--|----------------|----------------|------------------------------|----------------------------|-----------------|
| VH-Loose | >7 GeV | $ \eta < 2.7$ | Loose quality | < 3 | < 0.5 mm | Loose_VarRad |
| ZH-Signal | >27 GeV | $ \eta < 2.5$ | | Same as VH-Loose | | |
| WH-Signal | >25 GeV when $p_T^V > 150$ GeV >27 GeV when $p_T^V < 150$ GeV | $ \eta < 2.5$ | Medium quality | < 3 | < 0.5 mm | HighPtTrackOnly |

Table 3: Muon Selection

2.3.2 Event categorisation

Events are categorised following a decomposition into regions of specific flavour-tags, p_T^V range, number of jets, and cuts on the ΔR between the Higgs candidate jets. As $VH(H \rightarrow b\bar{b})$ is completely orthogonal to $VH(H \rightarrow c\bar{c})$, the event categorisation of the boosted analysis is not described here.

Flavour tagging: Jet flavour tagging is perhaps the most important part of the analysis. The DL1r algorithm is used and calibrated for the analysis for both b - and c -tagging. At the time of the analysis, the superior DL1d and GN1 taggers of the previous section were not yet available and their calibration is an ongoing effort. The VH analyses require extensive flavour-tagging and an in-depth study of the numerous and important backgrounds and their flavour components. This takes a significant amount of work, making a switch to a new tagger not feasible from a practical point of view in the timing of the analysis. They represent, however, an exciting avenue for progress in future iterations of this search. The *Pseudo-Continuous Flavour Tagging* scheme (PCFT), shown in Figure ??, is employed for a coherent joint definition of b - and c -tagged jets. For a given jet, it is first decided whether the jet is b -tagged (B), based on a b -tagging working point with 70% efficiency. These working points are derived on a set of calibrating samples: dileptonic $t\bar{t}$ for b -jets, semileptonic $t\bar{t}$ for c -jets, and $Z+jet$ samples for light-jets. If the b -tagging requirement is not met, it is then considered whether the jet is c -tagged using the same DL1r tagger at a working point defined and calibrated in the context of the analysis. A *loose* (L) and *tight* (T) c -tagging requirements are defined, each at an exclusive c -tagging efficiency of 20%. Being made of the top tier c -jets, the b - and light-rejections of the tight c -tagging working point are improved compared to the looser working point. The c -jet (light-jet) efficiency in the b -tagged bins of the PCFT scheme is 7.85% (0.181%) and the b -jet (light-jet) efficiency in the loose c -tagged (L) bin is 11.5% (6.5%) while it is 4.5% (0.9%) in the tight c -tagged bin (T). Jets that are not tagged are ascribed the letter N .

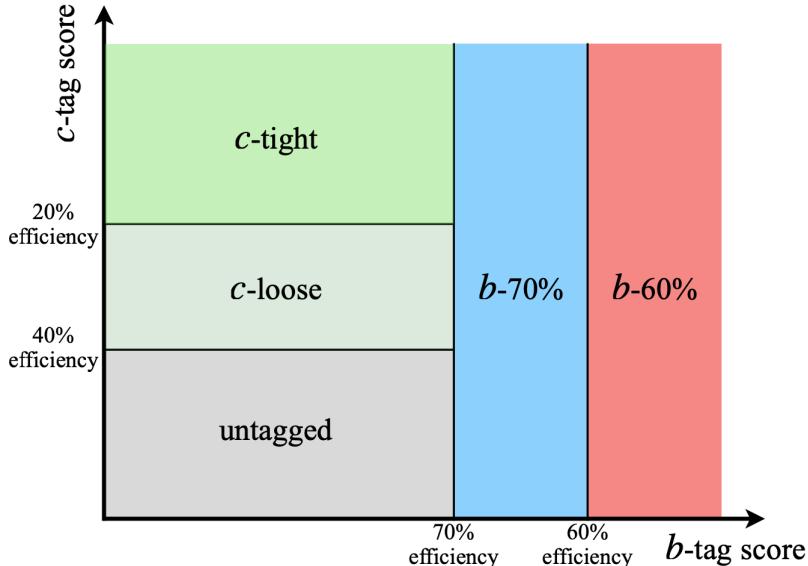


Figure 10: The pseudo-continuous tagging scheme, from the internal documentation of the team [Chisholm:2743096].

In an event with at least two signal jets, the candidate jets selected to reconstruct the Higgs define an event-tag by combining their individual tags. While in the past the so-called *Leading 2-jet* strategy was used, where the candidate jets were the two highest p_T signal jets, a new strategy called *All Signal Jets* is now preferred in the VH analysis as it improves the expected statistical significance. The strategy introduces a hierarchy of tags: $B > T > L > N$. The pair of candidates is made of the two signal jets having the highest tags, and, in case of ties, the highest p_T . The tagging requirements for an event to be in the signal regions (SR) are:

- For $VH(H \rightarrow b\bar{b})$, strictly two jets must be b -tagged and no tight c -tagged jets are allowed.
- For $VH(H \rightarrow c\bar{c})$, no b -tagged jets are allowed and at least one jet must be tight c -tagged. This defines three possible signal regions, where a second jet is either tight c -tagged (TT), loose c -tagged (LT), or not tagged (TN).

Two control regions are defined by modifying the conditions on the tags of the jets in the event:

- A combined $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$ top control region (topCR) is obtained by requiring at least 1 b -tagged jet and 1 c -tagged jet. The definition of this control region is the core subject of this work and will be further addressed later in this report.
- For $VH(H \rightarrow c\bar{c})$ only, an additional control region is defined by requiring no b - nor c -tagged jets in the event to constrain the significant $V+light-jet$ background.

Ranges of p_T^V : The categorisation is made of the following p_T^V regions in the non-boosted regime: [150, 250] GeV and [250, 400] GeV. For the 2L channel only, a low $p_T^V \in [75, 150]$ GeV is included - this is not feasible in 0L due to the trigger threshold on MET nor in 1L as the large amount of background at low p_T means this region does not improve the statistical significance.

Number of jets: Every region is split between a 2-jet and a 3-jet region, except for the 2L channel where events with 3 or more jets are included in the 3-jet region as there is almost no $t\bar{t}$ background thanks to the two leptons and c -jets requirements. In 0L and 1L, there is a large amount of $t\bar{t}$ background in the 4J region, hence the tighter requirement. Including this potential higher jets number region was observed to not improve the statistical significance.

ΔR between the jets: the angular separation between the two candidate jets $\Delta R(j_1, j_2)$, as defined in Equation ??, can be used to define a control region enriched in V +jets and $t\bar{t}$ backgrounds since these two processes give candidate jets with a flat angular spectrum while the signal peaks at low values of ΔR .

$$\Delta R(j_1, j_2) = \sqrt{(\eta_{j_1} - \eta_{j_2})^2 + (\phi_{j_1} - \phi_{j_2})^2} \quad (4)$$

A *high ΔR* control region (High ΔR CR) is defined using parametrised cuts on ΔR between the Higgs candidate jets as a function of p_T^V . An additional *low ΔR* control region (Low ΔR CR) for the 1L channel in $VH(H \rightarrow b\bar{b})$ resolved is also introduced (it is merged with the signal region for the $VH(H \rightarrow c\bar{c})$). The philosophy behind the parametrisation of this function is to adapt the cut on the expected angular separation between the two Higgs candidate jets as a function of how boosted they are, as described by the p_T^V variable. For signal events, we expect the H and V to be approximately back-to-back hence p_T^V is a good proxy for p_T^H while benefiting from better experimental resolution, as it is reconstructed from leptons p_T and/or E_T , depending on the channel. From physical principals, boosted candidate jets are indeed expected to have a lower angular separation. From the point of view of $VH(H \rightarrow c\bar{c})$, this modified cut represents a significant modification to the standalone analysis that relied on a simple fixed $\Delta R_{c\bar{c}}$ cut. The new cuts are defined by fitting a template function $c_1 \times e^{c_2 + c_3 \times p_T^V}$ so that:

- 95% (85%) of the $VH(H \rightarrow c\bar{c})$ signal is below the top limit for the 2-jet (3-jet) signal region,
- 90% of the diboson process is above the bottom limit in both signal regions.

The results of these fits for the 1L channel are displayed in Figure ??, showing the signal yield in a 2-dimensional histogram (p_T^V vs $\Delta R_{c\bar{c}}$) for different tags applied. The cut used in $VH(H \rightarrow b\bar{b})$, in yellow, shows a good agreement with the one derived on the TT -tagged (tight-tight) events in cyan and the LT -tagged (loose-tight) in green. The $VH(H \rightarrow b\bar{b})$ cuts is chosen so that the kinematic selection of the two analyses is harmonised. In $VH(H \rightarrow c\bar{c})$, the low ΔR CR does not improve the statistical significance hence this region is merged with the signal region. In $VH(H \rightarrow b\bar{b})$, these CRs are used to extract the normalisation of the backgrounds while in $VH(H \rightarrow c\bar{c})$ the shape of the $m_{c\bar{c}}$ spectrum is also used.

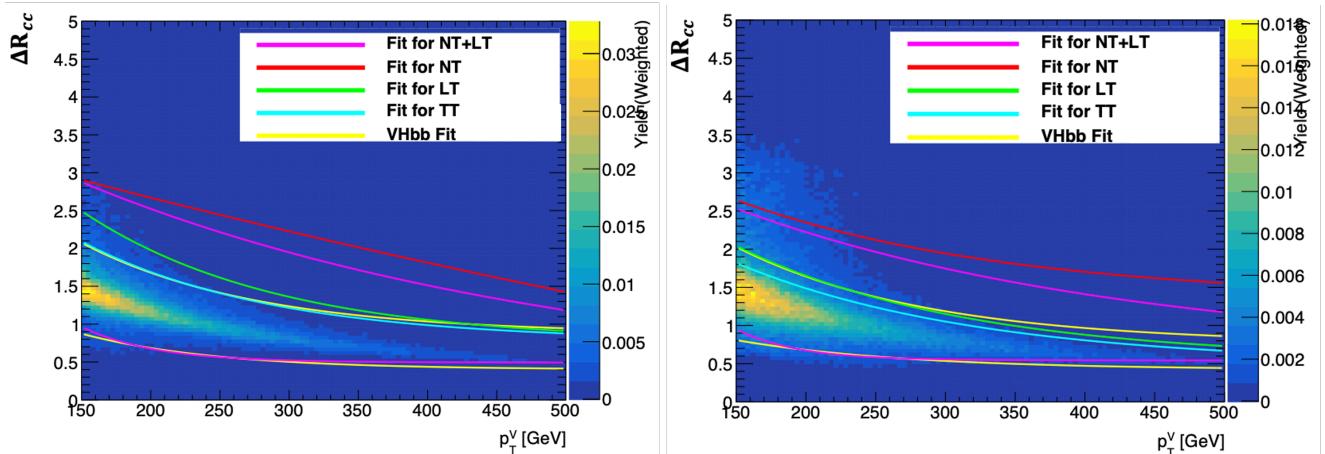


Figure 11: The p_T^V - $\Delta R_{c\bar{c}}$ 2D-histograms showing the signal yield of the 1-lepton $VH(H \rightarrow c\bar{c})$, for the 2-jet (left) and 3-jet (right) signal regions. The lines are the results of fitting the $\Delta R_{c\bar{c}} - p_T^V$ cuts for various signal tags, with the yellow curve showing the $VH(H \rightarrow b\bar{b})$ p_T^V - $\Delta R_{b\bar{b}}$ cut.

A summary of the full event categorisation and the regions thus defined at the moment of writing is shown in Figure ?? and a summary of the signal event selection is shown in Table ?? . As the analysis has not yet concluded, some modifications are still being investigated, such as merging the $L T$ - and $T T$ -tag regions. An important update to the $VH(H \rightarrow c\bar{c})$ analysis is the introduction of a *Boosted Decision Trees* (BDT) score distribution as discriminant variables instead of the invariant mass of the Higgs candidates jets $m_{c\bar{c}}$. The BDT is trained with kinematic and flavour information about the Higgs candidate jets as well as lepton information and higher-level variables such as angles between the objects, the sum of momenta, and the invariant mass. The algorithm provides a score in the $[-1, 1]$ range corresponding to low or high signal probability and significantly enhances the background-signal separation of the analysis.

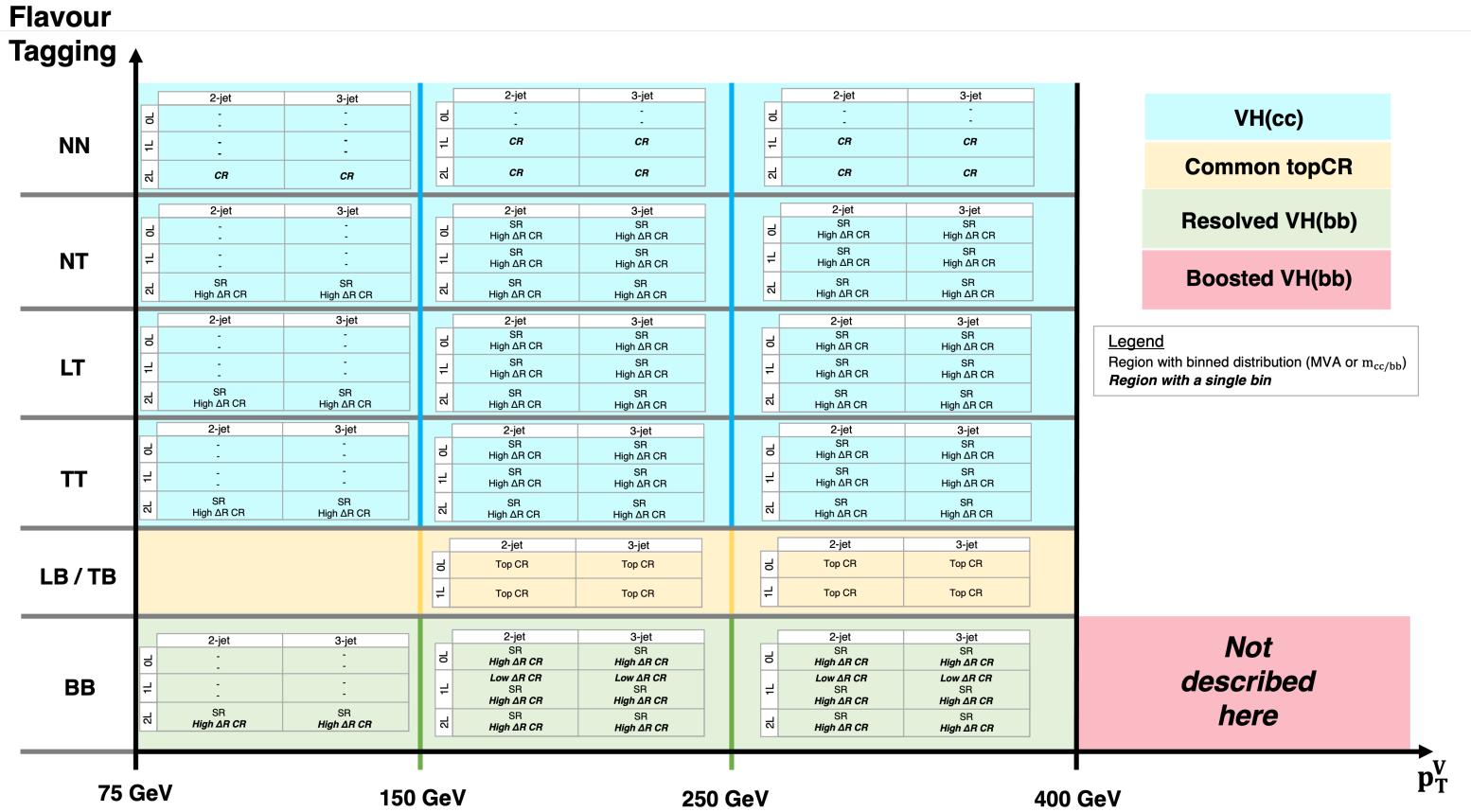


Figure 12: The current split of the analysis regions considered in the $VH(H \rightarrow b\bar{b}/c\bar{c})$ Combined Analysis, showing the Signal Region (SR), High and Low ΔR control region (CR), and the top CR.

2.3.3 Top Control Region

The top control region (topCR) is used to constrain the rather significant top background that peaks at signal-like values of the discriminant variables. Indeed, when the candidate jets selected correspond to the b - and c -jet from a $t\bar{t}$ decay, the invariant mass of the pair peaks at 120 GeV, exactly the region of interest for a Higgs decay search. The topCR is defined by requiring at least one c -tagged jet in combination with at least one b -tagged jet using the *AllSignal* strategy, as previously described. This tagging requirement renders it orthogonal to the signal region of the analysis and targets the decay topology of the different top processes:

- Semi-leptonic $t\bar{t}$ decay: both t follow the usual decay chain $t \rightarrow b + W$, with one of the W decaying leptonically and the other one to a pair of quarks. Some events from this process can enter the signal region when some of the quarks are c -tagged or if the b -jets are mis-tagged or flew out of the detector acceptance. Requiring the combination of a b -tag and a c -tag effectively selects this process, the b coming from the direct t decay and the c from a subsequent W decay.

| Analysis regime | $VH(H \rightarrow b\bar{b})$ | $VH(H \rightarrow c\bar{c})$ |
|---|---|--|
| Common Selections | | |
| Jets | | ≥ 2 signal jets |
| Candidate jets tagging | 2 B-tags | ≥ 1 T-tag, no B-tag |
| Leading Higgs candidate jet p_T | | > 45 GeV |
| Sub-leading Higgs candidate jet p_T | | > 20 GeV |
| Non-H candidate jet p_T | | > 30 GeV |
| $\Delta R(\text{jet1}, \text{jet2})$ | | Cuts applied (see above) |
| 0 Lepton | | |
| Trigger | E_T^{miss} triggers | |
| Jets | ≤ 3 jets | |
| Non-H candidate jets tagging | no T-tag | - |
| Leptons | | 0 VH -loose leptons |
| E_T^{miss} | | > 150 GeV |
| $E_{T,\text{trk}}^{\text{miss}}$ | - | > 30 GeV |
| $\sum_i p_T^{\text{jet}_i}$ | | > 120 (2 jets), > 150 GeV (3 jets) |
| $ \min \Delta\phi(E_T^{\text{miss}}, \text{jet}) $ | | $> 20^\circ$ (2 jets), $> 30^\circ$ (3 jets) |
| $ \Delta\phi(E_T^{\text{miss}}, H) $ | | $> 120^\circ$ |
| $ \Delta\phi(\text{jet1}, \text{jet2}) $ | | $< 140^\circ$ |
| $ \Delta\phi(E_T^{\text{miss}}, E_{T,\text{trk}}^{\text{miss}}) $ | | $< 90^\circ$ |
| p_T^V regions | | [150, 250] GeV, [250, 400] GeV |
| 1 Lepton | | |
| Trigger | e channel: single electron trigger μ channel: single muon trigger ($p_T^V < 150$ GeV) and E_T^{miss} triggers (above) | |
| Jets | ≤ 3 jets | |
| Candidate jets tagging | no T-tag | - |
| hadronic τ -veto | | no hadronic τ |
| Leptons | | 1 WH -signal lepton |
| E_T^{miss} | | > 1 VH -loose lepton veto |
| p_T^V regions | | > 30 GeV (e channel) |
| | | [150, 250] GeV, [250, 400] GeV |
| 2 Lepton | | |
| Trigger | as for 1L but the p_T^V limit for μ is 250 GeV | |
| Leptons | 2 VH -loose leptons (≥ 1 ZH -signal lepton) Same flavour, Opposite-charge for $\mu\mu$ | |
| m_{ll} | $81 < m_{ll} < 101$ GeV | |
| p_T^V regions | | [75, 150], [150, 250], [250, 400] GeV |

Table 4: Summary of the signal event selection in the 0-, 1- and 2-lepton channels (adapted from the internal note). Variables not presented in the text: $E_{T,\text{trk}}^{\text{miss}}$ is the missing transverse momentum calculated from the negative vector sum of the transverse momenta of tracks reconstructed in the inner detector and identified as originating from the primary vertex and m_{ll} is the invariant mass of the di-lepton pair.

- Single top t -quark: predominantly the Wt process $W t \rightarrow W + b + W$, with one W decaying leptonically and the other hadronically. Some of these background events can enter the signal region if the b -jet is missed and if a jet is c -tagged, from the extra W or if the b -jet is mis-tagged. Events from the single-top t - and s -channel of the process $t \rightarrow b + W$ bring a smaller contribution, as the c -tagged jet must come from *Initial State Radiation* (ISR) or *Final State Radiation* (FSR) if the b -jet is not mis-tagged. Single-top is a minor background in 0L and 1L, with the main component being the production of Wt pairs. The t -channel and s -channel contribute less than 1% of the total background.

Of the two processes, the $t\bar{t}$ is therefore the most important one and a main background in the 0L and 1L channels. Due to their similarities, the $t\bar{t}$ and Wt processes are considered as a single *top* background in the analysis. In 2L, because this top background is small, no flavour-based topCRs are introduced and a different strategy is employed where the top is directly constrained in a pure top- $e\mu$ control region defined by requiring two charged leptons of different flavours. For the 0L and 1L channels, the expected top background normalisation and its kinematic distributions, as given by the MC simulation, are adjusted using data in the topCRs; this is extrapolated to the signal regions under consideration of extrapolation effects (and corresponding extrapolation uncertainties) that account for differences between the topCRs and SRs.

The combined top background is separated into different components, depending on the true flavour of the two candidate jets, that can be combined during the statistical analysis. These are:

- $\text{top}(bb)$: in this case, the two b -jets produced during the $t\bar{t}$ decay are selected. This is a small component in the signal regions of the $VH(H \rightarrow c\bar{c})$ analysis, due to the 70% efficiency WP for b -tagging and the low mis-tag rate for b -jets in c -tagging. Naturally, in $VH(H \rightarrow b\bar{b})$ it is the leading contribution. Due to the origin of the candidate jets, a large ΔR_{bb} is expected between the two b -jets so this component is most effectively constrained by the High ΔR CR.
- $\text{top}(bc)$: where the b is from a t decay and the c from a subsequent W hadronic decay (or from ISR/FSR though this is less likely). Given the definition of the topCR, this is the dominating component in that region and the most important to constrain in the signal regions of the $VH(H \rightarrow b\bar{b}/c\bar{c})$ analyses due to its signal-like kinematics.
- $\text{top}(bl)$: where l stands for anything not b nor c (light jets predominantly but also some mis-tagged hadronic τ). This component is similar to the $\text{top}(bc)$ as it also consists of a $b + l$ from the W and can end up in the SRs and topCRs due to mis-tags.
- $\text{top}(lq)$: where l is as above and q can be any sort of jet except a b . This is a small component that mostly accumulates in the background-like part of the BDT score distribution. It is not constrained in the high ΔR regions nor the topCRs.

The signal region distributions in the 1L channel in the p_T^V range [150, 250] GeV are displayed in Figure ???. While the top is not the dominant background, except in the tighter tagged TT 3-jet region, its relative contribution to the background composition increases at signal-like values of the discriminant, as shown in Figure ??.

The components contributing the most in the $VH(H \rightarrow c\bar{c})$ side of the analysis are the $\text{top}(bc)$ and $\text{top}(bl)$, due to the tagging requirement. There is very little $\text{top}(bb)$ thanks to the good performance of the tagger. $\text{Top}(lq)$ is mostly found in the looser tag regions (NT, LT) and not where the signal peaks. Figure ?? displays the new top control regions proposed in this work: as expected, the bulk of the distributions is made of top background, centred around the expected Higgs mass. The philosophy behind the proposed new design leverages the pseudo-continuous tagging to select the highest p_T b -tagged and c -tagged jets as Higgs candidates. Thus, BL and BT regions are defined depending on whether the highest p_T c -tagged jet is loose- or tight-tagged. The regions are further split in the number of jets and the same definition is used in the 0L channel. The full tag compositions of each region are as follows:

- 2-jet: BL: BL ; BT: BT
- 3-jet: BL: BLN, BLL ; BT: BTN, BTL, BTT , and BBT

In the *AllSignal* strategy, the Higgs candidates in the topCR are always the highest p_T b - and c -tagged jets. This selection was observed to make the top control region distributions more closely match the distributions in the signal regions.

For $VH(H \rightarrow c\bar{c})$, the bc and bl components are the most important to constrain. In $VH(H \rightarrow b\bar{b})$, while the bc component is also significant and can benefit from the topCRs, the most important contribution comes from the bb one and is well constrained by the High ΔR CR, since in a $t\bar{t}$ decay the two produced b -jets tend to be separated by a large ΔR due to the event topology. For the Combined Analysis, the SRs and CRs of both analyses will be considered simultaneously. To show the impact on the $VH(H \rightarrow c\bar{c})$ standalone analysis, the High ΔR CR from $VH(H \rightarrow b\bar{b})$ is included to study the effect on the bc and bl components. The aim is to demonstrate these and the bb component can be well constrained with these regions

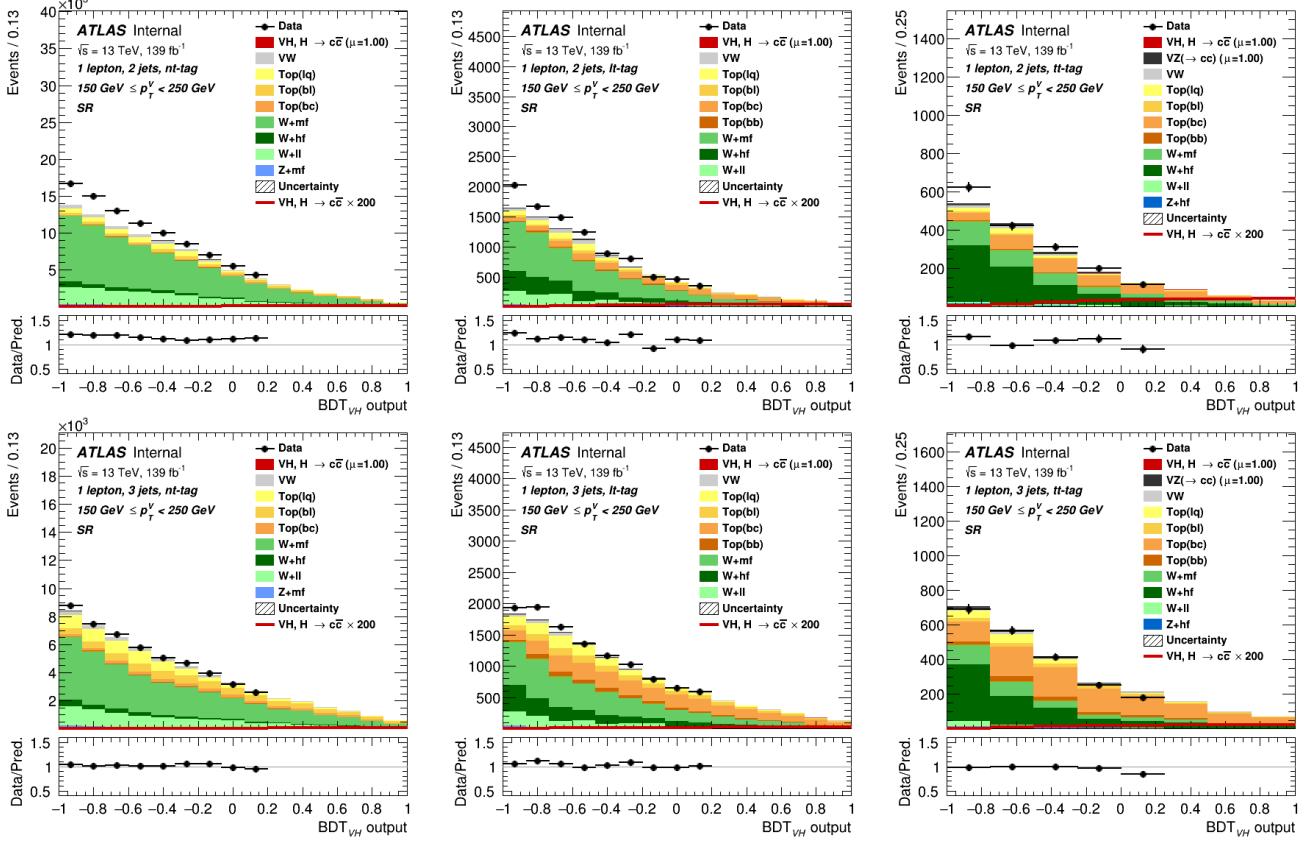


Figure 13: The 1L signal regions BDT distributions in the low [150-250] p_T^V range. Left: NT; centre: LT; right: TT. Top: 2 jets; bottom: 3 jets.

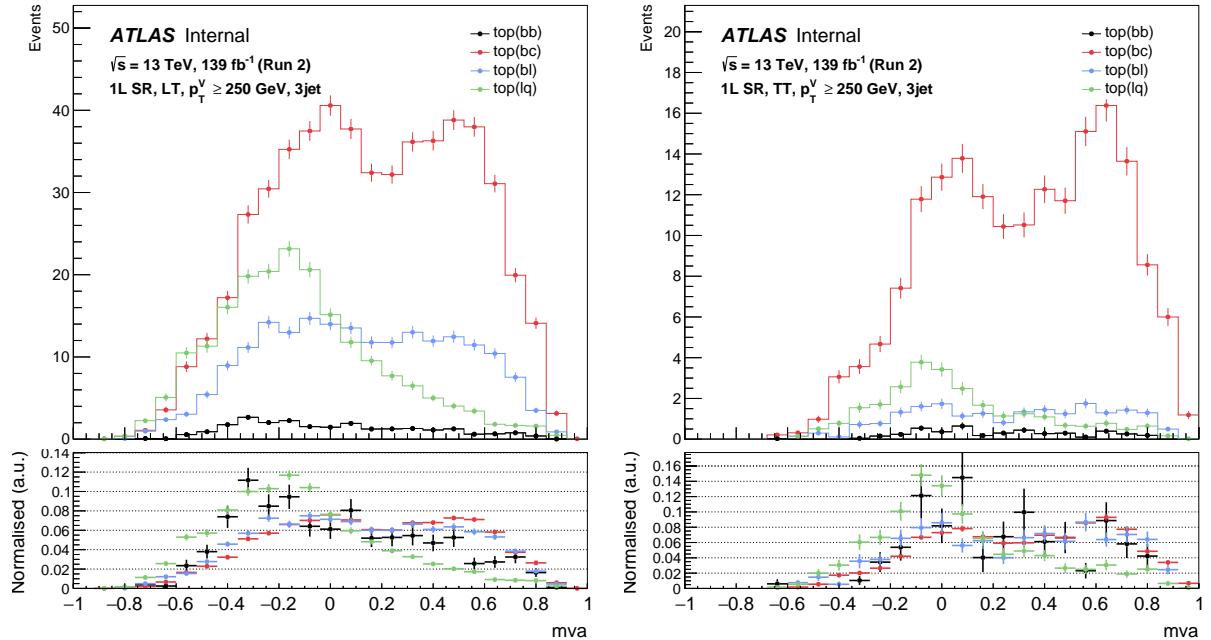


Figure 14: Top components in the 1L 3 jets signal regions BDT distributions in the ≥ 250 GeV p_T^V range. Left: LT; right: TT.

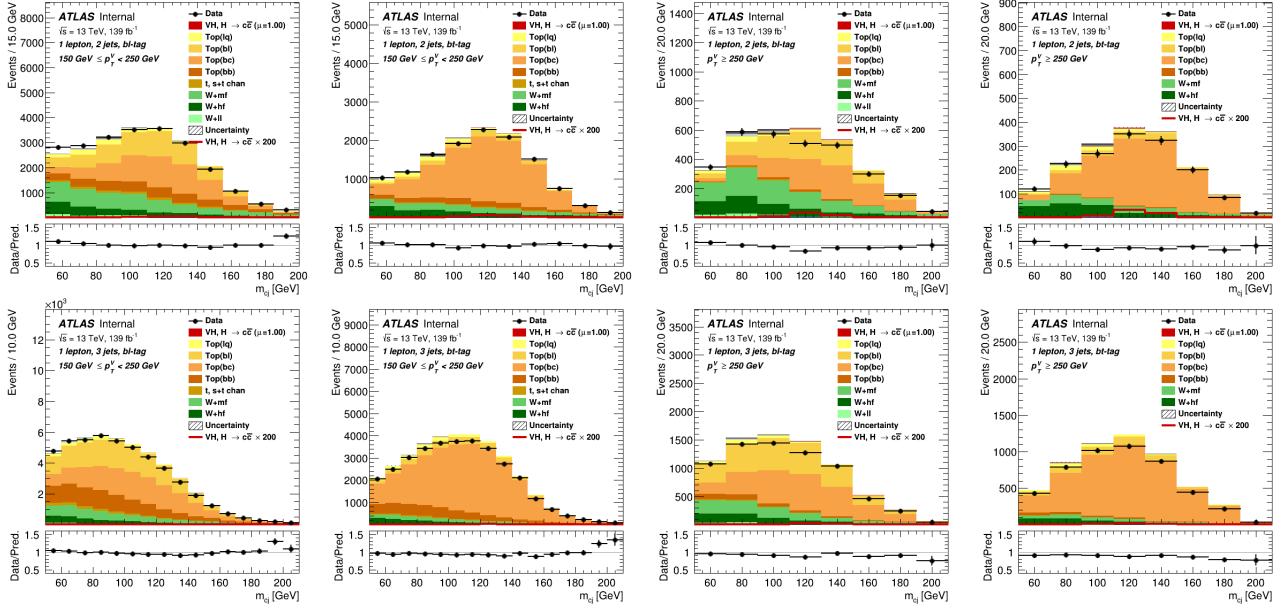


Figure 15: The 1L top control regions $m_{c\bar{c}}$ distributions in both p_T^V ranges (left two columns are $[150, 250]$ GeV, right two are > 250 GeV). Per group of two adjacent: left is BL, right is BT. Top: 2 jets; bottom: 3 jets.

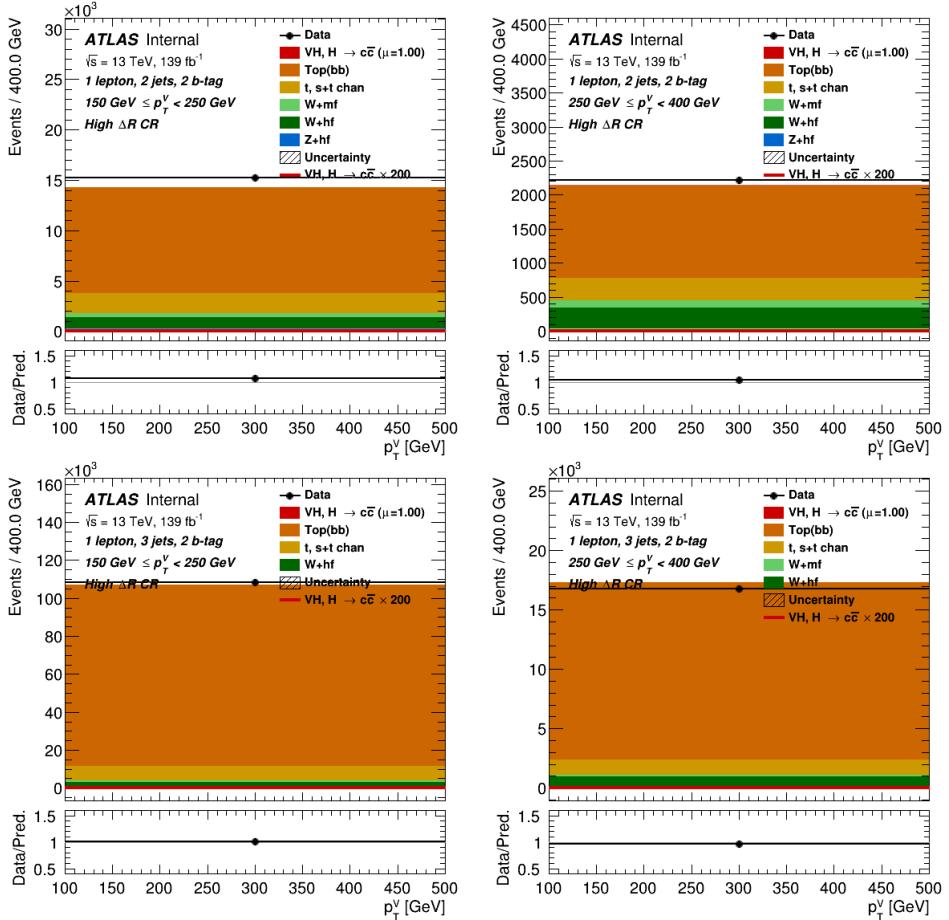


Figure 16: The 1L 1-bin $VH(H \rightarrow b\bar{b})$ High ΔR CR in both p_T^V ranges. Left: $[150, 250]$ GeV; right: $[250, 400]$ GeV. Top: 2 jets; bottom: 3 jets.

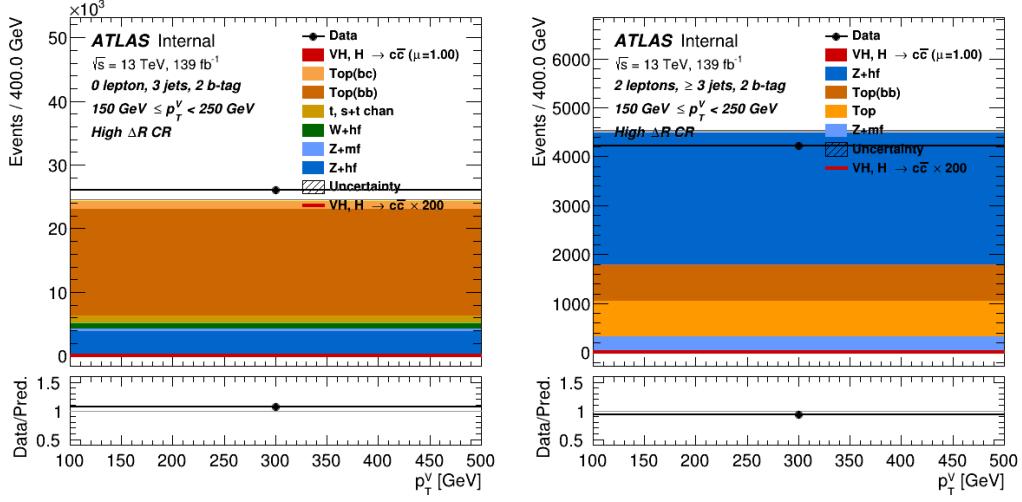


Figure 17: The $VH(H \rightarrow b\bar{b})$ High ΔR CR in the $[150, 250]$ GeV p_T^V range region with 3 jets, for 0L on the left, and 2L on the right.

alone, and in particular without the SRs of $VH(H \rightarrow b\bar{b})$. The $VH(H \rightarrow b\bar{b})$ High ΔR control regions are taken as a single bin of p_T^V , because the interest is solely to constrain the top(bb) normalisation. Figure ?? displays the 1L $VH(H \rightarrow b\bar{b})$ High ΔR CR, which is visibly dominated by the top(bb). Figure ?? shows the same region for the $[150, 250]$ GeV p_T^V range with 3 jets in 0L and 2L, showing a significant proportion of $Z+jets$ is also included in these regions.

2.3.4 Statistical analysis

After collecting the data and simulating the various samples, including the detector effects, reconstructing the physics objects, and applying the different cuts and the event categorisation, the last step in the analysis is to measure the signal yield normalised to the expected SM yield (from theory, i.e., $\sigma \times BR$). The analysis aims to define a 95% confidence level on the maximum signal enhancement factor μ . This is done by maximising the binned-likelihood distribution in all of the analysis regions simultaneously as a function of the signal strength and statistical and systematic uncertainties. The full binned-likelihood function can be decomposed into three terms [**Mironova:2837159**]:

1. A Poisson probability term based on the expected and observed event yields in all bins of the considered distributions tracks the signal strength: $\mathcal{L} = \prod_{i \in \text{bins}} \text{Pois}(N_i | \mu s_i + b_i)$, where N_i , s_i , and b_i are respectively the number of measured data events, the expected (simulated) signal yield, and the expected background yield in bin i . The signal strength parameter μ is the ratio of the measured $\sigma \times BR$ divided by the SM expectation for the signal process.
2. Systematics uncertainties enter the fit as Nuisance Parameters (NP) $\vec{\theta}$ which can modify the expected signal and background yields $s_i(\vec{\theta})$ and $b_i(\vec{\theta})$ in each bin. They are modelled as standardised Gaussian penalty terms: $\mathcal{L}_{\text{NP}} = \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}$. After the fit, the values of the NPs can be moved upwards or downwards, and this deviation from 0 (from 1 for the normalisation factors) is called a *pull*. The *constraint* indicates the certainty on the value of the NP after the fit. NPs can have a prior, based on pre-existing knowledge or empirical estimates (e.g. auxiliary measurements, MC simulation model differences), or be left *free-floating* (with a prefit value of 1), such as for the normalisation of the major backgrounds which is determined from data in control regions where these processes are enhanced.
3. Uncertainties tracking the limited available statistics of the simulations are introduced as γ_i -parameters, with one such parameter per bin. They give the fit the flexibility to adjust the expected background yield in a particular bin as $b_i(\vec{\theta}) \rightarrow \gamma_i b_i(\vec{\theta})$. They are introduced as: $\mathcal{L}_{\text{BkgStat}}(\vec{\gamma}) = \prod_{i \in \text{bins}} \text{Gauss}(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i})$, where $\beta_i = 1/\sigma_{\text{rel}}^2$ and σ_{rel} is the relative statistical uncertainty on the expected total background yield.

The full likelihood function is then described as the product of these three contributions as

$$\mathcal{L} = \prod_{i \in \text{bins}} \text{Pois}(N_i | \mu s_i(\vec{\theta}) + \gamma_i b_i(\vec{\theta})) \times \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} \times \prod_{i \in \text{bins}} \text{Gauss}(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}). \quad (5)$$

For the signal, as previously described, a BDT score is used as a discriminant variable. Also called a *Multivariate Analysis Discriminant* (MVA), one such BDT is trained per lepton channel, separately for $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$. For the topCRs and the ΔR control regions, the invariant mass of the candidate pair $m_{c\bar{c}}$ is used.

Systematic uncertainties, that act as priors and are constrained in the fit, on the simulated top background events (*nominal samples*) are assessed by comparing the output of different setups of the simulator as well as using alternative simulators. Variations studied in these *alternative samples* concern the matrix element generation (hard scattering), the renormalisation scale μ_R and factorisation scale μ_F for the ISR and FSR, the *Parton Shower* (PS), the underlying event simulation, and multiple parton interactions [[Mironova:2837159](#)]. Similar uncertainties are introduced for each of the other background processes. In addition, experimental uncertainties that affect all processes are introduced for the triggers, object reconstruction, flavour tagging, and the recorded data luminosity. They cover resolution effects, reconstruction efficiencies, and differences between data and simulations.

The different top components normalisations in 0L and 1L are determined from data in the profile likelihood fit by free-floating norm-factors (NF). A prior normalisation uncertainty is applied to account for potential differences between 0L and 1L. Acceptance ratios, in the number of jets and also in the extrapolation from the control region to the signal region, and shape uncertainties are also applied, after being derived for the top background in the 0L and 1L. When two top components are jointly floated, the dominated component is given an extra flavour ratio uncertainty to add flexibility to the fit. For example, when floating top(bl) with top(bc), the normalisation is mostly driven by top(bc), the dominating component, and a ratio bl/bc is added to let the bl components adjust to differences between the two.

The objective of this section is to study the impact of the modified top control regions on the combined 012L $VH(H \rightarrow c\bar{c})$ fit. The aim is to study for which top components the fit regions, and in particular the topCRs, provide enough information to determine their normalisation and constrain their shape from data. Furthermore, the data in the SRs should be sufficiently well described using the complete fit setup and the top background control regions should not significantly impact the behaviour (the NPs) of other backgrounds. While every bin of the analysis is used in the fit, bins with a large fraction of the signal (e.g., at high BDT score in the SRs) are not displayed as the analysis is still blinded.

Several iterations of the fit studies were performed to test the new topCRs proposed in this work and in particular which components can be constrained and which NPs can be correlated. The correctness of each fit is evaluated by verifying several diagnostics information, such as changes to the constraining of non-top-related NPs and to the predicted postfit yields. In the present report, three setups are presented to represent the evolution of the implementation of the topCR:

- Nominal: this is the first implementation to use all available topCRs (BT and BL) and the $VH(H \rightarrow b\bar{b})$ High ΔR CR with minimal changes to the other NPs. Other novelties introduced here are to free-float separately the top(bc) and top(bl) and to float the top(bb) normalisation. The 2-jet and 3-jet regions share the same NFs and an additional 3-to-2-jet extrapolation uncertainty is implemented and applied in 2-jet. TopCR-SR extrapolation uncertainties are applied in the SRs except for top(lq) where it is applied in the topCR.
- Baseline 2: uses only the topCR BT. Compared to the Nominal, the topCR BL is removed as it has a large $V+jets$ background, especially in 0L, and impacts the NPs associated with this process, which is not the aim of the top control region. The top(bl) is now floated jointly with top(bc), as they are similar physics-wise since they both select a b -quark from the top-quark decay and a jet from the subsequent W decay. An extra NP is introduced to model the bl to bc ratio. The main selection difference between the topCRs and the SRs is the tagging criteria. This difference should be covered by the flavour tagging uncertainties and the topCR-SR extrapolation uncertainties are therefore dropped. The flavour components represent different parts of the top components being selected as candidate jets. The difference induced in the 2-jet and 3-jet regions is not expected to be the same between the flavour components, hence the 3-to-2 jet extrapolation is decorrelated by flavour, separately for $t\bar{t}$ and Wt , marking the last modification with respect to the nominal setup.
- Baseline 2 + nJetDec: is similar to Baseline 2 but now the topCR NFs are decorrelated on the number of jets. This is the preferred scenario if the data and CRs are powerful enough for the fit to converge since it maximally exploits the knowledge from the data and altogether avoids the need for 3-to-2 jet extrapolation uncertainties. These latter uncertainties can be very large because they cover different levels of mis-reconstructed or out-of-acceptance objects involved in the distinction of the top background into the 2-jet and 3-jet regions.

Figures ?? and ?? compare the pulls obtained by the different baselines considered. The top of Figure ?? shows the floating normalisations of the top background. The [150, 250] GeV region is indicated by the *BMin150* suffix and the region above

250 GeV by *BMin250*, while the number of jets region is indicated by a suffix *J2* or *J3*. All of the top components NFs are well constrained and the values obtained are consistent between the different p_T^V regions and, for Baseline 2+nJetDec (in red), the number of jets regions. One small exception is the top(*lq*) in Baseline 2+nJetDec. This component is however less relevant to the analysis and the top(*lq*) can simply be kept inclusive in the number of jets in the final setup. Thanks to the inclusion of the CRHigh from $VH(H \rightarrow b\bar{b})$, even the top(*bb*) component is well constrained in $VH(H \rightarrow c\bar{c})$. When removing the topCR BL, a small difference is induced in the NF for the V+jets background (not shown in the Figure). For example, the postfit yield of the W+jets background in the 1L [150, 250] GeV TT SR changes by 0.69% between the Nominal and Baseline 2 + nJetDec (from 461000 to 457800 expected events).

The rest of Figure ?? shows the most important top-related NPs. The METTrigTop, covering the differences between data and simulation in the MET trigger efficiency for top processes, is strongly pulled for all fit setups, something that requires further investigation from the analysis team. The NPs with name of the form *NJetAcc_Top** list the nJet acceptances, a single one for the Nominal in blue and split in flavour for Baseline 2 in green. They are removed in Baseline 2 + nJetDec as in this case the NFs are floated separately for each number of jets region. This last setup is able to avoid incurring the rather large systematics induced by these 3-to-2 jets extrapolation uncertainties. Underneath the nJet acceptances, the pulls with names starting with *Top_mCC* represent shape systematics derived by considering the difference between the nominal samples and the alternative generators (such as Herwig and Powheg) or internal variations (like for the ISR). The *topCR_extrap* are the topCR-SR extrapolation uncertainties which are only applied in the Nominal setup. For the Baseline 2 variations, the flavour tagging NPs and associated pulls are able to cover this extrapolation with minimal changes to the Nominal, as shown in Figure ???. Through these different groups, the interesting observation is that the Baseline 2 + nJetDec performs similarly to the Baseline 2 and Nominal, despite having a much simpler structure without nJet acceptances nor topCR extrapolations. A closer look at other NPs, such as the b-tagging pulls, confirms there is a good agreement between the different alternatives and no over-constraining is observed.

The expected prefit 95% confidence limits on the signal strength of $VH(H \rightarrow c\bar{c})$ of these conditional data fits are:

- Nominal: $12.60^{+4.94}_{-3.52} \times \text{SM}$
- Baseline 2: $12.75^{+4.99}_{-3.56} \times \text{SM}$
- Baseline 2 + nJetDec: $12.71^{+4.98}_{-3.55} \times \text{SM}$

These values are consistent between the different fits. Comparing the fit setups, Baseline 2 + nJetDec is preferred to serve as a new nominal thanks to its simplified fit structure, its lower number of NPs and normalisation factors, the ability to constrain the components in several number of jet regions, and the freedom to use the topCR BL region as a validation region rather than as a control region. The obtained limits are significantly improved compared to the previous ATLAS result of $26 \times \text{SM}$ ($31 \times \text{SM}$) [[Collaboration:2721696](#)] and are competitive with the recent CMS results of an observed (expected) upper limit of $14 \times \text{SM}$ ($7.6 \times \text{SM}$) [[arXiv:2205.05550](#)]. The analysis is however still not concluded and improvements are still being pursued by fine-tuning the event selection to avoid large uncertainties as much as possible, including new MC samples with lower MC statistical uncertainties, as well as revisiting the modelling uncertainties all while monitoring the fit behaviour and carrying out similar studies to the present one for other regions and backgrounds.

2.4 Conclusion

The $VH(H \rightarrow b\bar{b}/c\bar{c})$ analysis of Run 2 provides an exciting avenue to improve the competitiveness of the ATLAS measurement of the charm Yukawa coupling. While the analysis is still ongoing, the harmonisation of the $VH(H \rightarrow b\bar{b})$ and $VH(H \rightarrow c\bar{c})$, the refinement of the top control region, the introduction of MVA as discriminating variables, and the many other modifications pursued in the Combined Analysis already indicate significant gains will be made on the previously published result. The analysis is planned to wrap up in the following months and will be described in detail in the final DPhil dissertation.



(a) Top pulls.

(b) Flavour tagging pulls.

Figure 18: Pulls for the 012L fit with the several baselines. Blue: nominal; green: baseline 2; red: baseline 2 + nJetDec.