University of Oxford

Lincoln College

Doctorate of Philosophy

Particle Physics

# Advanced Machine Learning Applications for the Higgs and Heavy Flavour Quarks at ATLAS

Candidate

Maxence DRAGUET

Supervisor

Daniela BORTOLETTO

2020-2024

# Contents

# List of Abbreviations

**AUC** Area Under the Curve
**BDT** Boosted Decision Trees
**BSM** Beyond the Standar Model
**CARL** Calibrated Likelihood Ratio Estimator
**CL** Confidence Level
**CR** Control Region
**DL1r** Deep Learner 1 Model with RNNIP
**DNN** Deep Neural Network
**EW** Electroweak
**FN** Floating Normalisation
**FSR** Final State Radiation
**GN2** Graph Network 2 Model
**GNN** Graph Neural Network
**ID** Inner Detector
**ISR** Initial State Radiation
**JVT** Jet Vertex Tagger
**LHC** Large Hadron Collider
**MC** Monte Carlo
**ME** Matrix Element
**MS** Muon Spectrometer
**MVA** Multivariate Analysis
**NN** Neural Network
**NP** Nuisance Parameter
**PCFT** Pseudo-Continuous Flavour Tagging
**PDF** Parton Distribution Function
**POI** Parameter Of Interest
**PS** Parton Shower
**PV** Primary Vertex
**QCD** Quantum Chromodynamics
**RNN** Recurrent Neural Network
**ROC** Receiver Operating Characteristic
**SF** Scale Factors
**SM** Standard Model
**SR** Signal Region
**STXS** Simplified Template Cross-Section
**UE** Underlying Event
**VR** Variable Radius
**WP** Working Point

# CHAPTER 1

# $VH(H \to b\bar{b}/c\bar{c})$ Combined Analysis

*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 [1, 2]. Theorised in 1964 by two independent papers introducing the mechanism of spontaneous symmetry breaking to give mass to the gauge bosons [3, 4], its discovery almost fifty years later marked one of the greatest achievements of the particle physics community. The Higgs boson is an essential part of the Standard Model (SM). It is tied to the mechanism through which particles acquire mass without breaking the electroweak gauge invariance, as described in Chapter* **??**. *While the gauge bosons W and Z gain mass through symmetry breaking, in the Standard Model (SM) the fermions - quarks and leptons - acquire theirs by interacting with the Higgs fields - a scalar field carried by the Higgs boson H, as described by the Yukawa mechanism [5].*

## 1.1 Introduction

The Higgs boson $H$ [3, 4, 6, 7] was discovered in 2012 by the ATlAS and CMS Collaborations using the data of the Large Hadron Collider (LHC) Run 1 [1, 2]. This triggered a race by both experiments to study the specific properties of the discovered particle, and in particular

to observe its different production processes and decay channels. The initial decay channels studied for the groundbreaking discovery where the bosonic decays of the Higgs to final states of photons and leptons: $H \to \gamma\gamma$, $H \to ZZ$, and $H \to WW$. These channels benefit from clean experimental conditions, reliable measurements, and limited backgrounds. The new particle is now being studied in ever finer details, confirming its coupling to many massive particles of the SM and showing remarkable agreement with the properties dictated by the theory. 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, providing the first measurement of the top Yukawa coupling [8, 9]. Additionally, the decay of Higgs bosons to a pair of $\tau$-lepton is now well established and different cross-sections measurements have been performed [10, 11]. Importantly, the decay channel of the Higgs boson to a $b\bar{b}$ pair was observed by both ATLAS and CMS [12, 13]. This last decay channel is of particular significance since it has the largest predicted branching ratio of 58% for $m_H = 125$ GeV in the SM.

Concerning the second generation of fermions, there now is evidence of the decay to a $\mu^-\mu^+$ pair by CMS [14] and a 2 $\sigma$ excess over the background-only hypothesis by ATLAS [15]. 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 $H \to c\bar{c}$ decay mode [16]. 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 2.9% [17], the large background rates, and the experimental difficulties in identifying $c$-jets. It is a fertile ground for new physics Beyond the Standar Model (BSM) due to the smallness of the predicted Yukawa coupling $y_c^{\mathrm{SM}} \approx 3.99 \times 10^{-3}$ [18] for the $c$-quark as well as an important test of the validity of the model [19, 20, 21, 22, 23, 24, 25]. The fermion Yukawa couplings in the SM are indeed largly added ad-hoc and do not explain the distinct mass hierarchy between the three generations of quarks. This problme can be probed by studying the quarks coupling strength to the Higgs boson. The $VH(H \to b\bar{b}/c\bar{c})$ analysis, to which this chapter is dedicated, probes the hierarchy of mass between the $b$- and $c$-quark, respectively a 3$^{\mathrm{rd}}$ and 2$^{\mathrm{nd}}$ generation quark.

## 1.2 The $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ ATLAS Analyses

While $H \to b\bar{b}$ 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 *associ-*

*ated production mode*, where the $H$ is produced in addition to an extra vector boson $V$ ($W$ or $Z$) decaying leptonically, to electrons ($e$), muons ($\mu$), neutrinos ($\nu$), or a combination $e\nu$ or $\mu\nu$. 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 and greatly reduce the contribution of the multi-jet background. Other analyses relying on full-hadronic final states in the associated or other production modes are also performed by the ATLAS Collaboration, but they are less sensitive to the Higgs coupling to heavy-flavour quarks. In the ATLAS Collaboration, the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analyses adopt very similar strategies. The main ingredient is the ability to reliably tag the flavour of jets produced in an event to reconstruct the heavy quark pair produced in the $H$ decay, using the tools described in Chapter **??**.

Using the full Run 2 dataset with a total integrated luminosity of 140 fb$^{-1}$, the published $VH(H \to c\bar{c})$ ATLAS analysis obtained the following upper limits on the signal strength of the $VH(H \to c\bar{c})$ as predicted by the SM: an observed (expected) upper limit of 26 $\times$ SM (31 $\times$ SM) [26]. The measurement also provided the first constraint on the Higgs-charm coupling modified $|\kappa_c| < 8.5$. For comparison, CMS reported an observed (expected) upper limit of 14.4 $\times$ SM (7.6 $\times$ SM) and a constraint of $1.1 < \kappa_c < 5.5$ [27].

For the $VH(H \to b\bar{b})$, thanks to a larger expected signal, the ATLAS analysis reaches a sensitivity of 6.7 standard deviations [28]. Following observation, the focus of this analysis has shifted towards a precision differential measurement of the fiducial cross-sections 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* [28] and the *boosted* [29] analyses, with the latter restricting to values of the transverse momentum of the associated vector boson $p_T^V$ above 250 GeV - a variables highly correlated with the $p_T$ of the Higgs $p_T^H$. The name of these analyses comes from the approach to reconstruct the Higgs boson candidate. At low $p_T^V$, the two $b$-jets from the $H$-boson decay can be independently resolved into two distinct small cone radius (small-$R$) jets. At high $p_T^V$, the $H$-boson is highly Lorentz-boosted requiring a change of strategy: the candidate $H$ are efficiently reconstructed as a single large-radius ($R = 1$) jet merging the two $b$-jets. 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 [28]. 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 $H \to b\bar{b}$ and $V \to$ leptons branchings fractions ($\sigma \times BR$) are reported in the reduced Simplified Template Cross-Section (STXS) scheme. Finally, limits are set on the coefficients of effective Lagrangian operators which can affect the $VH$ production and the $H \to b\bar{b}$ 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 [29].

Some preliminary studies aiming at combining the different analyses have already been performed, with the resolved $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analyses combined in Ref [26] and the resolved and boosted $VH(H \to b\bar{b})$ combined[1] in Ref [31]. These combinations require careful studies to remove the overlap between the analyses, such as by introducing a switch in $p_T^V$ at 400 GeV between the resolved and boosted strategies. However, they rely on the published analyses and are therefore not optimised. The objective of the Combined Analysis presented here is to define a common analysis strategy, correlating as much as possible the experimental and modelling uncertainties for both Higgs decay modes and $p_T^V$ regimes, thereby improving the measurements of $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ simultaneously. This new combined measurement has several additional benefits:

- The Higgs-charm and -beauty coupling modifiers, $\kappa_c$ and $\kappa_b$, can be measured directly, as well as their ratio $\kappa_c/\kappa_b$.

- 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 combined analysis benefits from improved signal selection thanks to upgraded physics objects and event reconstruction techniques. In particular, new machine learning-based techniques are integrated for both the event selection and flavour tagging.

This chapter details the current state of the $VH(H \to b\bar{b}/c\bar{c})$ Combined Analysis, an analysis which, at the time of writing, was not yet concluded and is therefore still blinded. The stage

[1] CMS published an analoguous combination in Ref [30].

described corresponds to that attained at the end of the third unblinding approval review. Some modifications to the analysis are expected in the soon-to-be-published final result, in particular to the modelling strategy and the categories entering the fit. The work presented here is largely based on the internal documentation of the experimental team and personal work carried out during the duration of the DPhil project.

## 1.3 Overview of the Combined $VH(H \to b\bar{b}/c\bar{c})$ Analysis

The Combined Analysis is performed with the full ATLAS Run 2 proton-proton collision data, collected from 2015 to 2018, for a total integrated luminosity of 140 fb$^{-1}$ at a centre of mass energy $\sqrt{s} = 13$ TeV. The regions and boundaries between the different regimes of the analysis are illustrated in Figure 1.1. The $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ parts are separated by the required presence of two $b$-tagged jets or a $c$-tagged jets respectively[2]. The $p_T^V$ cut marks the difference between the Higgs candidate reconstruction scheme of the resolved and boosted $VH(H \to b\bar{b})$: two small radius (R = 0.4) jets for $p_T^V < 400$ GeV and, above, one large radius (R = 1) jets with two $b$-tagged Variable Radius (VR) track-jets associated to the large-$R$ jet.

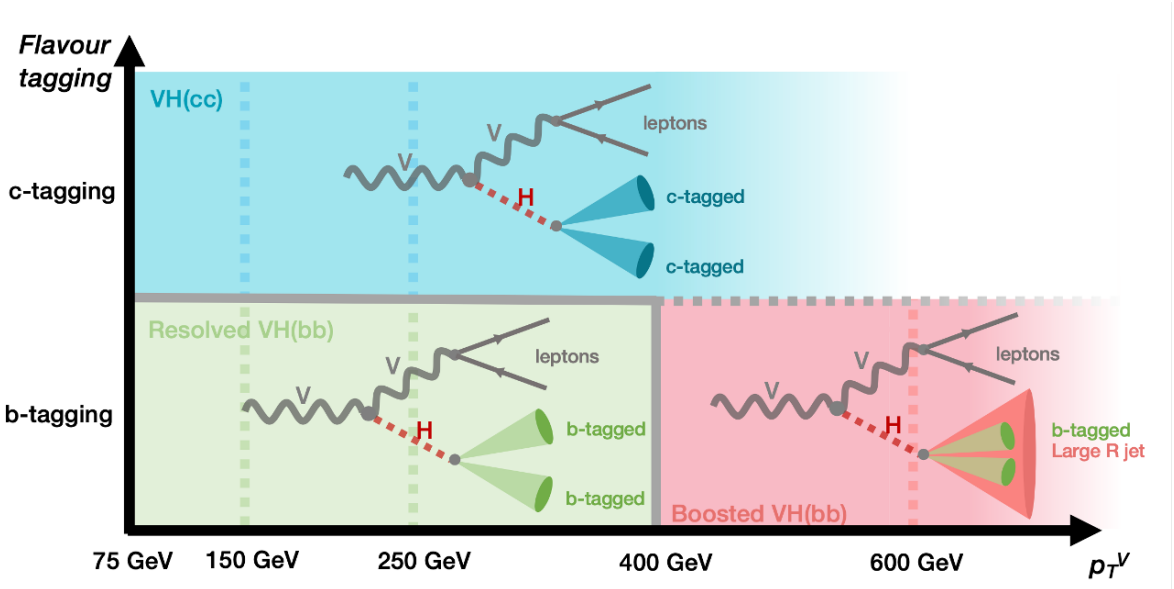



Figure 1.1: The analysis regimes considered in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

For each analysis region, three channels are defined based on the decay mode of the vector boson $V$: $Z \to \nu\nu$ defines the *0-lepton* (0L), $W \to \ell\nu$ the *1-lepton* (1L), and $Z \to \ell^+\ell^-$ the *2-lepton* (2L), where $l$ refers to an electron or a muon and $\nu$ to a neutrino. The signal considered are the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ processes, with the SM diboson processes $VZ(Z \to b\bar{b})$ and $VZ(Z \to c\bar{c})$ considered as signals in a cross-check analysis. Having a larger cross-section

[2]More precisely a tight $c$-tagged jet, as described in Section 1.5

and being kinematically similar to the signals, these processes can be measured with good statistical significance, thereby offering a suitable test to verify the validity of the strategy deployed. The main backgrounds are the production of a vector boson with additional jets($V$+jets, mostly $Z$+jets in 0L and 2L, $W$+jets in 1L) and the top-quark processes (*Top*, 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 with an extra $W$ boson, both in 0L and 1L). Minor backgrounds are the Quantum Chromodynamics (QCD) multi-jet[3], the single-top process (without an extra associated $W$ boson) and non-signal diboson pair productions ($VV$). The processes are further described in the Section 1.4.

Flavour tagging plays an essential role in the analysis, splitting the analysis phase space into different regimes of studies. The most important backgrounds are also split based on flavour components. The $V$+jets is split into three components: $V$+ heavy flavour jets ($V$+*hf*, including $V + bb$ and $V + cc$), $V$+ mixed flavour ($V$+*mf*, including the $V + bc$, $V + bl$, and $V + cl$), and the $V$+ light flavours ($V$+*lf*, including all other possible flavour selection including $\tau$'s). The top-quark background is also split by flavour: the Top($bb$), in which the two selected jets are $b$-tagged, is treated separately from the Top($bq/qq$) which groups all other flavours ($bc$, $bl$, and $qq$). The former is important in $VH(H \to b\bar{b})$ while the latter is the dominant flavour background in $VH(H \to c\bar{c})$.

All backgrounds are simulated using Monte Carlo (MC) simulation packages, except for the multi-jet and the top background in 2-leptonc channel which are estimated from a data-driven method. The multi-jet is only included in the 1-lepton channel, as it is negligible in the other channels. The chapter is separated into different section introducing the datasets and Monte Carlo (MC) simulations (Section 1.4), describing the objects reconstruction techniques (Section 1.5.1), the analysis selection and regimes (Section 1.5.2), the event categorisation (Section 1.5.3), the experimental and process modelling (Section 1.8 and 1.9), the fit framework (Section 1.10.1), and finally the main results (Section 1.10.2).

## 1.4 Data and Simulated Samples

The combined analysis is performed on data collected during the Run 2 of the LHC, with proton-proton collisions recorded between 2015 and 2018 at a $\sqrt{s} = 13$ TeV for an integrated luminosity

[3]Thanks to the required presence of leptons in the final state.

| Process | Matrix Element | PDF Set (ME) | Parton Shower | $\sigma$ order | $\sigma\times$ Br [pb] |
|---|---|---|---|---|---|
| $qq \to WH \to \ell\nu b\bar{b}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO(QCD)+ NLO(EW) | $2.69 \times 10^{-1}$ |
| $qq \to ZH \to \nu\nu b\bar{b}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO(QCD)+ NLO(EW) | $8.91 \times 10^{-2}$ |
| $qq \to ZH \to \ell\ell b\bar{b}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO (QCD)+NLO(EW) | $4.48 \times 10^{-2}$ |
| $gg \to ZH \to \nu\nu b\bar{b}$ | PowHeg-Box v2 | NNPDF3.0NLO | Pythia-8.307 | NLO+NLL | $1.43 \times 10^{-2}$ |
| $gg \to ZH \to \ell\ell b\bar{b}$ | PowHeg-Box v2 | NNPDF3.0NLO | Pythia-8.307 | NLO+NLL | $7.23 \times 10^{-3}$ |
| $qq \to WH \to \ell\nu c\bar{c}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO(QCD)+ NLO(EW) | $1.34 \times 10^{-2}$ |
| $qq \to ZH \to \nu\nu c\bar{c}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO(QCD)+ NLO(EW) | $4.42 \times 10^{-3}$ |
| $qq \to ZH \to \ell\ell c\bar{c}$ | PowHeg-Box v2 + GoSam + MiNLO | NNPDF3.0NLO | Pythia-8.245 | NNLO (QCD)+NLO(EW) | $2.23 \times 10^{-3}$ |
| $gg \to ZH \to \nu\nu c\bar{c}$ | PowHeg-Box v2 | NNPDF3.0NLO | Pythia-8.307 | NLO+NLL | $7.10 \times 10^{-4}$ |
| $gg \to ZH \to \ell\ell c\bar{c}$ | PowHeg-Box v2 | NNPDF3.0NLO | Pythia-8.307 | NLO+NLL | $3.59 \times 10^{-4}$ |
| $W \to \ell\nu$ + jets | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NNLO | 60242 |
| $Z \to \ell\ell$ + jets | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NNLO | 6201 |
| $Z \to \nu\nu$ + jets | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NNLO | 416.05 |
| $t\bar{t}$ | Powheg-Box v2 | NNPDF3.0NLO | Pythia-8.230 | NNLO+NNLL | 704 |
| single-top ($Wt$) | Powheg-Box v2 | NNPDF3.0NLO | Pythia-8.230 | Approx. NNLO | 80.03 |
| single-top ($t$) | Powheg-Box v2 | NNPDF3.0NLO | Pythia-8.230 | NLO | 70.7 |
| single-top ($s$) | Powheg-Box v2 | NNPDF3.0NLO | Pythia-8.230 | NLO | 3.35 |
| $qq \to WW$ | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NLO | 47.93 |
| $qq \to WZ$ | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NLO | 20.85 |
| $qq \to ZZ$ | Sherpa 2.2.11 | NNPDF3.0NNLO | Sherpa 2.2.11 | NLO | 6.33 |
| $gg \to VV$ | Sherpa 2.2.2 | NNPDF3.0NNLO | Sherpa 2.2.2 | NLO | 2.78 |

Table 1.1: The nominal Monte Carlo samples used in the $VH(H \to b\bar{b}/c\bar{c})$ analysis, and the corresponding process cross-sections at $\sqrt{s} = 13$ TeV. The PDF sets mentioned in the table are used for the matrix element.

of 140 fb$^{-1}$ [32]. Data events passing some quality requirement are selected, ensuring for example that all the sub-detectors were correctly operating. The analysis requires extensive and accurate Monte Carlo (MC)-based modelling of the signal and the background processes, except for the QCD multi-jet and the $t\bar{t}$ background in the 2-lepton channel which have data-driven estimations. All MC samples are simulated with ATLAS detector effects [33] using GEANT4 [34]. The nominal samples are produced using the prescriptions described in Table 1.1, detailing the Matrix Element (ME) generators, Parton Shower (PS), and Parton Distribution Function (PDF) releases used as well as the cross-sections. Samples are normalised either to the best theoretical cross-section predictions or the generator cross-sections.

Both simulated samples and data are reconstructed with the offline reconstruction software of ATLAS. The EVTGEN 1.6.0 program is used to simulate the properties of $b$- and $c$-hadrons decays[4] [35]. Pile-up is included in the simulation, both from multiple interactions in the same and adjacent bunch crossing. This is performed by overlaying events with minimum bias simulated using PYTHIA 8 with A3 tune and interfaced with the NNPDF 2.3 Parton Distribution Function (PDF)s [36]. The rest of this section gives more details behind the simulation of the different processes. When relevant, alternative samples generated from a different setup to the nominal samples are introduced. These alternative samples are used to assess process modelling uncertainties in Section 1.9.1, as summarised in Figure 1.14.

[4] EVTGEN 1.7.0 is used for the SHERPA generated samples.

### 1.4.1 Signal Processes

The analysis targets the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ processes, here called *signals*. The Leading Order (LO) Feynman diagrams contributing to the associated production $VH$ are the $qq$-initiated modes depicted in Figure 1.2. A gluon-initiated production of $ZH$ is also possible at Next-to-Leading Order (NLO) with a quark loop (mostly top-quark), as depicted in Figure 1.3. The ME calculations are based on the POWHEG-BOX V2 generator [37, 38]. The $qq$-initiated $VH$ samples are simulated with the POWHEG generator with the multiscale improved NLO (MiNLO) procedure [39], with one-loop amplitudes computed with the GoSam automated software [40]. The $qq$-initiated samples simulate Parton Shower (PS), Underlying Event (UE), and multiple parton interactions with PYTHIA 8.245, while the $gg$-initiated use PYTHIA 8.307 [36]. Both use the AZNLO tune [41] with PDFs based on the NNPDF3.0NLO for matrix elements [42].

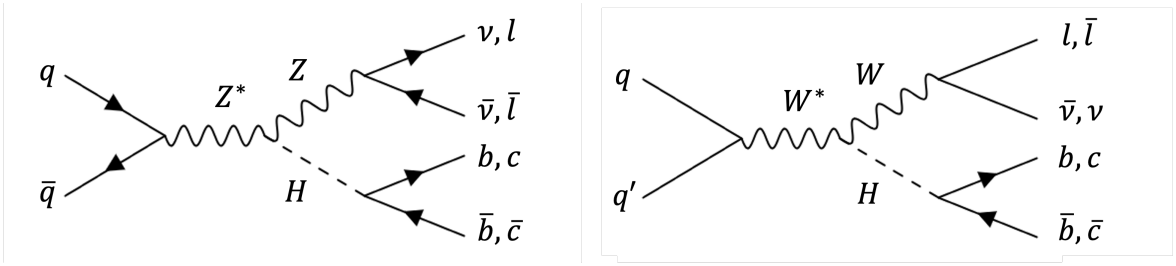



Figure 1.2: Leading order Feynman diagrams for $VH(H \to b\bar{b}/c\bar{c})$, $qq$-initiated with $V$ decaying into leptons.

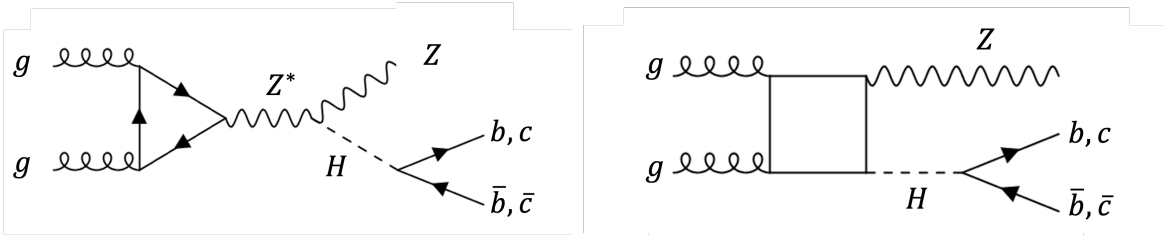



Figure 1.3: Feynman diagrams of the $gg$-initiated contributions to $ZH(H \to b\bar{b}/c\bar{c})$.

The inclusive cross-sections for $WH$ and $ZH$ are calculated at NNLO in QCD [43] and NLO in Electroweak (EW) [42]. The $gg$-initiaed $ZH$ contribution relies on the LO prediction from POWHEG instantiated with PYTHIA 8.

**Alternative samples:** are simulated with POWHEG+MiNLO+HERWIG 7.0, with the same simulation stack as the nominal but replacing PYTHIA 8 by HERWIG 7.0 [44] for the simulation of the PS, hadronisation, UE, and multiple parton interactions.

### 1.4.2 Background Processes

The most important backgrounds in the analysis are simulated with MC, except for the multi-jet and the top backgrounds in the 2-lepton channel. The major background processes and their simulations are detailed in this section.

#### $V$+jets

The production of a gauge vector boson $V$ in association with jets is the largest background in the analysis. Some leading contributing Feynman diagrams to this process are presented in Figure 1.4. Both the $Z$+jets and $W$+jets are simulated with SHERPA 2.2.11 [45], which delivers NLO precision on ME computation for up to 2 jets and LO accuracy for between 3 and 5 jets. PS and hadronisation are treated by the default SHERPA generator, with the NNLO PDFs based on NNPDF3.0nnlo [42]. Uncertainties from missing higher orders are evaluated by varying the QCD renormalisation and factorisation scales, $\mu_R$ and $\mu_F$, in the matrix elements by respective factors 0.5 and 2. Flavour filtering is applied to generate samples enriched with heavy flavour quarks.

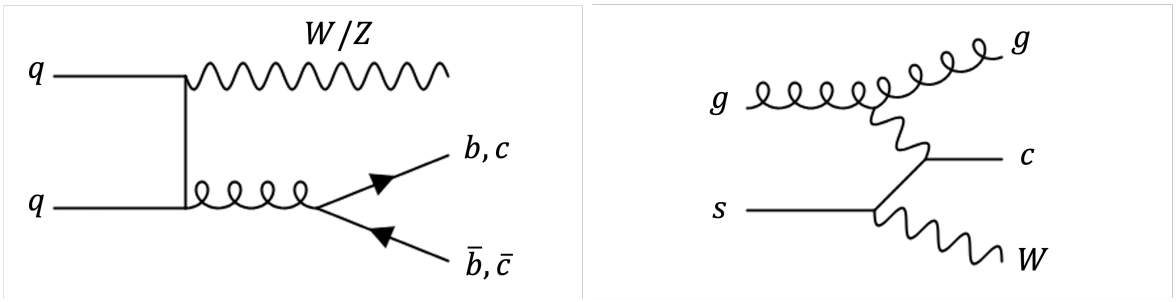



Figure 1.4: Examples of Feynman diagrams of the $V$+jets process. The left diagram gives a same flavour jet pair due to the gluon splitting, while the right one can give mixed flavours.

**Alternative samples** Two sets of alternative samples are available:

- MADGRAPH FXFX samples are produced for the modelling studies, using the MADGRAPH5_AMC@NLO 2.6.5 program [46]. This generates events with $V$-boson and up to three additional partons in the final state at NLO accuracy. The scales $\mu_R$ and $\mu_F$ are set to 1/2 the transverse mass of all final-state partons + leptons. PYTHIA 8.240 is interfaced for showering and hadronisation, with the A14 tune and the NNPDF2.3lo PDF set with $\alpha_s = 0.13$.

- SHERPA 2.2.1 [47] samples are used as alternative generator. They give different $p_T^V$ distributions to 2.2.11 [48], an important modification given the observed data-MC disagree-

ments in the $p_T^V$ distributions when using SHERPA 2.2.11. These samples are similar to those used in the standalone $VH(H \to c\bar{c})$ [26].

### Top-pair Production

The $t\bar{t}$ process is the second most important background in the analysis. The leading order Feynman diagram for this process is shown in Figure 1.5. The nominal samples are generated for the 0L and 1L channels with POWHEG at NLO calculation of the matrix element [49, 50]. It is interfaced with PYTHIA 8.230 with the NNPDF3.0NLO PDFs using the A14 tune for PS, hadronisation, and UE description. Filtering is applied while simulating to enhance statistics. Cross-sections are calculated to NNLO in QCD, with resummation of the Next-to-Next-to-Leading Logarithmic (NNLL) soft-gluon terms [51].

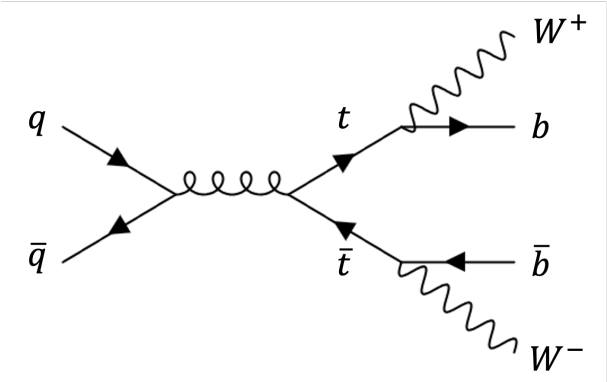



Figure 1.5: Feynman diagrams of the $t\bar{t}$ production and decay, where each $W$ can decay leptonically or hadronically.

**Alternative samples** Several alternatives are simulated for modelling studies:

- Replacing PYTHIA by HERWING 7.0 with H7UE tune [52] for parton shower, hadronisation, and UEs modelling while keeping the same nominal POWHEG setup. This sample is used to systematically assess variations to the parton shower model.

- Replacing POWHEG by MADGRAPH5_AMC@NLO [46] for NLO hard-scattering matrix element modelling with the nominal PYTHIA for PS, hadronisation, and the UE simulation. This sample is used to systematically assess variation the matrix element prediction.

- Weights variations tuning the initial and final state radiations (Initial State Radiation (ISR) and Final State Radiation (FSR)) contributions relative to the nominal setup. There are 4 such variations, based on the nominal POWHEG + PYTHIA 8.230:

  – A high- and low-variations of ISR, where the $\mu_R$ and $\mu_F$ scales are doubled and halved.

– An up- and down-variations of FSR, obtained by doubling (halving) the renormalisation scale $\mu_{R,FSR}$.

### Single-top Production

The so-called single-top process combines different minor channels, with the leading Feynman diagrams depicted in Figure 1.6. The dominant contribution is the associated top-production $Wt$ channel, with the $t \to Wb$. Two other contributions are the $t$- and $s$-channel, with the former having a significantly increased cross-section over the $s$-channel. These processes are simulated similarly to the $t\bar{t}$, with the cross-sections calculated for a top-quark mass of $m_t = 172.5$ GeV at NLO in QCD for the $s$- and $t$-channel [53, 54] and with approximate NNLO accuracy from NNLL soft-gluon resummation for the fiducial $Wt$ production cross-section [55, 56].

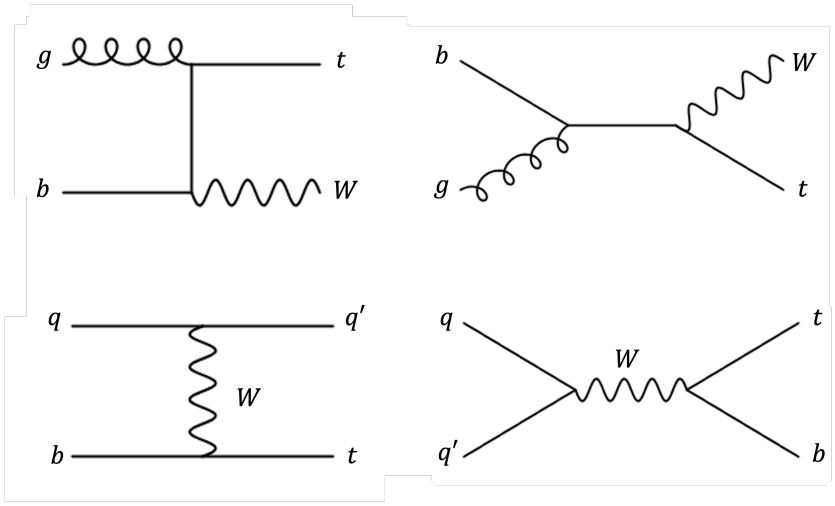



Figure 1.6: Feynman diagrams of the $Wt$-production (top) and the single top production (bottom) in the $t$-channel (left) and the $s$-channel (right).

The $Wt$ production channel has diagrams overlapping with the $t\bar{t}$ production at NLO in QCD. In the analysis, a diagram sustraction ($DS$) scheme is applied to remove the overlap with $t\bar{t}$ by locally cancelling the $t\bar{t}$ contribution in the NLO $Wt$ cross-section calculation [57].

**Alternative samples** Alternatives are used for the modelling of the single-top $Wt$- and $t$-channels[5]:

- The 2 alternative generators and the 4 changes to the ISR and FSR used for the alternatives of $t\bar{t}$ are also applied to the $Wt$- and $t$-channels.
- For $Wt$ only, a sample using an alternative overlap removal procedure is produced with the alternative diagram removal ($DR$) scheme [57] to systematically model the overlap with $t\bar{t}$.

[5]No alternatives are derived for the single-top $s$-channel due to its small contribution in the analysis.

This scheme removes the diagrams in the NLO $Wt$ amplitudes that are doubly-resonnant, when both $t$-quark are on-shell. $DR$ was the default scheme in prior iterations of this analysis, but the $DS$ samples showed better agreement with data in the boosted regime and were therefore chosen as nominal.

### Diboson Process

The diboson processes $WW$, $WZ$, and $ZZ$ enter the analysis both as a background, with a hadronically decaying $V$-boson mistaken for the Higgs, and as a cross-check signal when decaying into a $b\bar{b}$ or $c\bar{c}$ pair. Some leading $qq$-initated Feynman diagrams are depicted in Figure 1.7, with gluon-initiated diagrams also possible via quark-loops. The $qq$-initiated diboson are simulated similarly to the $V$+jets, using SHERPA 2.2.11 [45]. The $gg$-initiated are simulated with the older SHERPA 2.2.2 version. In both case, the cross-sections are computed at NLO precision, with the NNLO PDFs based on NNPDF3.0nnlo [42] for both the matrix element and parton shower.

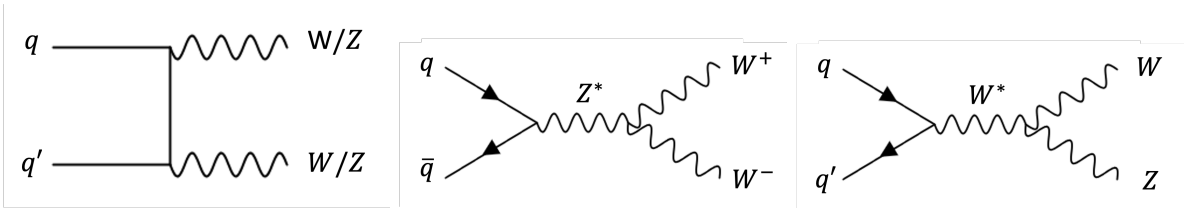



Figure 1.7: Feynman diagrams of the diboson production in the $t$- (left) and $s$-channel (centre & right). The $t$-channel can lead to any combination of $W$ and $Z$ dependingon the initial quark-pair.

### Alternative samples:

- POWHEG v2 interfaced with PYTHIA 8 samples are produced to systematically assess ME and PS variations.
- SHERPA 2.2.1 samples are produces to systematically model the impact of varying the fragmentation function.

### QCD Multi-jet

This process is estimated from data instead of simulations because of the difficulty in generating a sufficient statistics samples due to the low selection efficiency, despite having a much larger production cross-section than the Higgs. QCD multi-jet events can be selected when heavy flavour hadrons decay semi-leptonically or jets are mis-identified as leptons. Such leptons are normally not isolated, and only a small fraction passes the lepton requirements. The multi-jet

is negligible in the 0-lepton and 2-lepton channels thanks to the strict selections available. In the 1-lepton resolved channel, the remaining contribution is assessed from data-driven templates for $VH(H \to b\bar{b})$ or as a side control region for $VH(H \to c\bar{c})$. In both cases, a region enriched in multi-jet is defined by inverting the lepton isolation requirements. The residual multi-jet is mostly present at low $p_T^V$ values and is therefore ignored in the boosted regime.

## 1.5 Selection and Categorisation

Data collected by the ATLAS reconstruction consists of different types of low-level information measured in various sub-detectors. Different processing steps, collectively referred to as *reconstruction*, must be applied to unlock a higher-level physical interpretation of the information: constructing tracks from hit in the silicon detectors, identifying electrons and muons, etc. This section introduces the object selection, listing the recipes applied to identify the different relevant physics objects. The analysis event selection, requiring different reconstructed objects to be identified in data and simulations, is then presented as well as the final categorisation seperating events into defined analysis regions.

### 1.5.1 Object Selection

As introduced in Chapter **??**, the ATLAS software allows for different object reconstruction techniques. The reconstruction strategies of the relevant objects to the $VH(H \to b\bar{b}/c\bar{c})$ analysis are presented in this section.

**Primary Vertex:** all events considered in the analysis are required to have at least one primary vertex reconstructed from tracks in the Inner Detector (ID) [58].

**Electrons:** are reconstructed by matching a deposit in the electromagnetic calorimeter with a track in the ID [59, 60]. Electrons are required to have $p_T > 7$ GeV and $|\eta| < 2.47$. They are identified with a *loose* working point of a likelihood discriminant, matching the calorimeter shower shape to an associated track. The $e$ candidates must satisfy $p_T$-dependent isolation criteria in both the ID and calorimeter. In the 1L channel, the *tight* likelihood criterion is used with stricter calorimeter isolation requirements to better reject the multi-jet background. Additional requirements on the electron selection are lepton channel-dependent and summarised in Table 1.2. $VH$-Loose electrons require a loose likelihood identification is and are applied in all

channels. Additionally, the $WH$-Signal and $ZH$-Signal criteria are respectively applied in the 1L and 2L channels, 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.

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

Table 1.2: Electron Selection requirements.

**Muons:** are reconstructed by matching an energy deposit in the muon detector with information from the ID and Muon Spectrometer (MS) [61]. They are required to have $p_T > 7$ GeV, $|\eta| < 2.7$, to satisfy a *loose* identification criteria, and be isolated in the ID according to $p_T$-dependant criteria. These requirements are summarised in Table 1.3, and vary depending on the lepton channel simularly to the electron requirements. The $VH$-Loose requirements are applied to muons in all channels. The $WH$-Signal and $ZH$-Signal are additionally applied to the 1L and 2L channels respectively, with a stricter track-based isolation used in 1L to suppress the multi-jet background.

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

Table 1.3: Muon Selection requirements.

**Taus:** hadronically decaying $\tau$-leptons are identified and vetoed in 1L using an Recurrent Neural Network (RNN)-based tagger [62]. Taus are required to have a $p_T > 20$ GeV, $|\eta| < 2.5$, and to have 1 or 3 associated tracks. In 0L and 2L, if the jet passes a *loose* working point requirement for hadronically decaying $\tau$-leptons, it is no longer considered as a jet and cannot be considered as a candidate for reconstruction of the Higgs boson.

**Missing Transverse Energy:** neutrinos are not detectable in ATLAS and their presence is inferred from momentum imbalance in the transverse plane to the beam axis. $E_T^{\text{miss}}$, also called MET, is the negative vectorial sum of the transverse momentum of physics objects, namely electrons, muons, hadronic $\tau$, and jets. An additional track-based *soft term* is added, to include

a contribution from good quality tracks associated with the Primary Vertex (PV) but not to any reconstructed physics object [63].

**Jets** Three types of jets are used by the analysis, all reconstructed with the anti-$k_t$ algorithm [64]:

1. Small-$R$ jets: are reconstructed from topological clusters of energy deposit in the hadrnoic calorimeter based on the reconstructed PFlow objects with a radius $R = 0.4$. A jet is considered as *central* if $|\eta| < 2.5$ and $p_T > 20$ GeV, and as *forward* if $2.5 \leq |\eta| < 4.5$ and $p_T > 30$ GeV. Central (forward) jets with a $p_T < 60$ GeV ($p_T < 120$ GeV) are required to originate for the primary vertex as identified by the jet vertex tagger (Jet Vertex Tagger (JVT)) to limit the pile-up background [65]. *Tight* jet cleaning criteria are applied to suppress non-collision background. Central jets are used in the resolved regime to reconstruct the Higgs candidate, based on flavour tagging.

2. Large-$R$ jets: are similar to small-$R$ jets with a larger radius $R = 1.0$, and required to have $p_T > 250$ GeV and $|\eta| < 2$. They are used in the boosted regime to reconstruct the Higgs candidate.

3. Variable-$R$ (Variable Radius (VR)) track-jets: are reconstructed with a $p_T$-dependent radius, optimised for double $b$-tagging efficiency of the boosted $H \to b\bar{b}$ decays [66]. They are required to have $p_T > 10$ GeV and $|\eta| < 2.5$. These track-jets are used to reconstruct the $b$-tagged objects inside the large-$R$ jet and to define a boosted control regions rich in top background.

**Flavour Tagging:** Jet flavour tagging is perhaps the most important part of the object reconstruction. The latest available Deep Learner 1 Model with RNNIP (DL1r) tagger from Run 2 is used in the analysis for both $b$- and $c$-tagging in the resolved and boosted regime [67]. The methodology differs slightly between the two regimes of the analysis due to the different flavour tagging needs:

- In the resolved $VH(H \to b\bar{b}/c\bar{c})$, DL1r is used to tag both $b$- and $c$-jets. The so-called Pseudo-Continuous Flavour Tagging (PCFT) scheme, illustrated in Figure 1.8, is deployed to allow for a coherent joint definition and simultaneous calibration of $b$- and $c$-tagged jets, adopting the technique first introduced for 2D $c$-tagging in the $VH(H \to c\bar{c})$ analysis [26]. The DL1r tagger assigns a $b$-score[6] and a $c$-score[7] to every selected jets. To tag a jet,

[6] With an $f_c = 0.018$ value.
[7] With an $f_b = 0.3$ value.

the associated score must be higher than a specific cutoff value, defined to give a specific selection efficiency in simulated data, also known as a Working Point (WP). From this score, the jet is assigned one of 4 possible labels, base on 2 $b$-tagging Working Point (WP) and 2 $c$-tagging WP. These WP aretested in strict successive order, with first a 60% tight $b$-tagging working point (bin 4) followed by a looser 70% $b$-WP (bin 3). A jet passing these selections is labelled $B$[8]. Otherwise, it is considered for $c$-tagging with first a *tight* working point at 20% efficiency (bin 2), followed by a *loose* WP at an exclusive efficiency of 20% (bin 1) on the remaining jets - so that 40% of the $c$-jets are effectively selected in the combined tight and loose bins. A jet selected by the tight $c$-tagging WP is labelled $T$, and $L$ if it only passes the loose WP. A jet failing to pass all WPs is not tagged and labelled $N$ (bin 0). The $b$-tagging WPs correspond to official ATLAS ones for DL1r [67], while those for $c$-tagging are optimised for the purpose of the analysis. The tagging efficiency of each bin is displayed in Table 1.4, shown for the main flavours as well as $\tau$-leptons reconstructed as jets. The calibration of all five bins of Figure 1.8 is performed simultaneously for the

| PCFT bin | PCFT bin name | Jet tagging efficiency $\epsilon_{jet}$ | | | |
|---|---|---|---|---|---|
| | | $b$-jet | $c$-jet | light-jet | $\tau$-jet |
| 1 | $c$-loose | 11.5% | 20.5% | 6.5% | 18.5% |
| 2 | $c$-tight | 4.8% | 24.2% | 0.9% | 19.5% |
| 3 | $b$-70% | 11.2% | 5.2% | 0.13% | 1.7% |
| 4 | $b$-60% | 58% | 2.65% | 0.051% | 0.49% |

Table 1.4: Jet tagging efficiency for $b$-jet, $c$-jet, light-jet and $\tau$-jet in the Pseudo-Continuous Flavour Tagging (PCFT) bins, measured from a POWHEG+PYTHIA 8 simulated sample of semi-leptonic $t\bar{t}$ events.

analysis following the methodology described in Ref [67], with some results presented in Appendix A.1.

- The boosted regime only targets $b$-jets, with the single-jet DL1r tagger used. As such, the standard pseudo-continuous $b$-tagging method is used [67]. The track-jets associated to the leading large-$R$ jet are given a $b$-score based on the per-flavour probabilities outputted by DL1r. The 85% working point is adopted to maximise the signal yield, due to the important statistical limitations in the boosted regime. Track-jets passing this working point are $B$-tagged, otherwise they are untagged $N$. Studies showed that the very loose DL1r WP gives a better expected statistical significance than the then-available $X_{bb}$ tagger. The official calibration from Ref [67] is used and extended to higher $p_T$ with uncertainty

[8] The difference between these $b$-tagged is used in the discriminant Multivariate Analysis (MVA) of the analysis

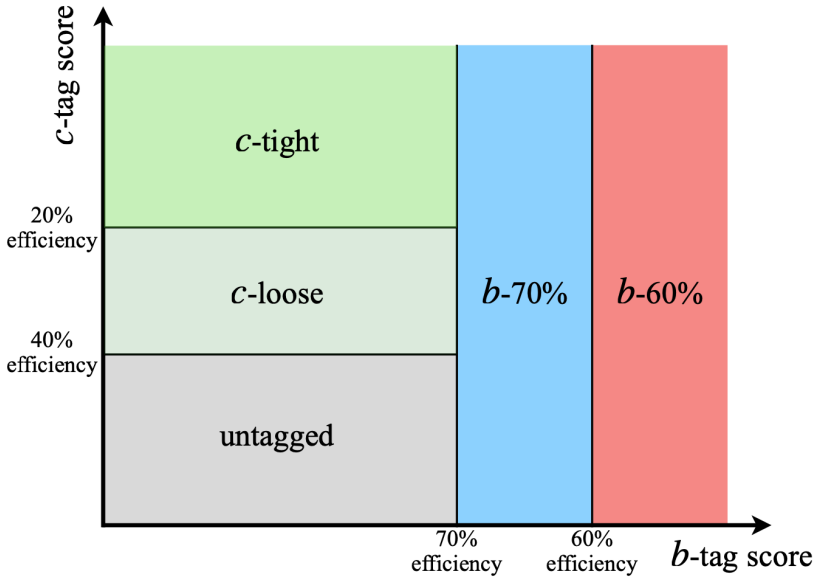



Figure 1.8: The Pseudo-Continuous Flavour Tagging (PCFT) scheme used to simultaneously define 2 $b$-tagged, a tight $c$-tagged, a loose $c$-tagged, and a non-tagged bins.

extrapolation due to the wide range of $p_T$ probed in the analysis, as presented in Appendix A.1.

While the analysis was underway, the superior single-jet Graph Neural Network (GNN) taggers and the boosted objects $GN2X$ tagger introduced in Chapter **??** were not yet available as their calibration was an ongoing effort. Furthermore, switching to these new taggers was 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, in both the resolved and boosted regimes.

**Object Overlap:** overlap removal is applied to avoid double-counting electrons, muons, small-$R$ and large-$R$ jets, and hadronically decaying $\tau$-leptons passing the object selection.

### 1.5.2 Event Selection

A subset of all the events recorded by ATLAS during Run 2 is selected for further selection based on specific triggers. The trigger selections of the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ have been harmonised for the combined analysis, and are specified per lepton channel in 0-lepton (0L), 1-lepton (1L), and 2-lepton (2L). In 0L, events are selected using the lowest un-prescaled $E_T^{\text{miss}}$ trigger, with an increasing lower threshold rising from 70 GeV for data recorded in 2015, 90 to 110 GeV for 2016, and to 110 GeV for 2017 and 2018 due to higher trigger rate later in Run 3. The 1L channel triggers cover both the $e$ and the $\mu$ sub-channels. For the $e$-channel, single electron events must be triggered by the lowest un-prescaled single electron trigger. For muons, the $E_T^{\text{miss}}$ trigger of 0L is used for events with $p_T^V > 150$ GeV, while the lowest un-prescaled single

muon trigger is used for events with a lower $p_T^V$. Finally, the triggers for 2L are equivalent to 1L except for the muon channel where the $p_T^V$ threshold for switching between triggers is raised to 250 GeV. The use of $E_T^{\text{miss}}$ trigger at high $p_T^V$ for muons was shown to increase the signal acceptance by $\sim 5\%$.

The different regimes of the analysis are defined by flavour tagging and the strategy to reconstruct the Higgs boson. In the resolved regime, an event must have at least two central jets. Two candidate jets are selected to reconstruct the Higgs using the so-called *All Signal Jets* strategy, and define an event-tag by combining their individual tags. A hierarchy of tags is introduced, following the ordering: $B > T > L > N$. The pair of candidates is selected from the two central jets having the highest tags, and the highest $p_T$ in case of ties. Events are labelled based on the tag of the selected jets, e.g., $TT$ is assigned to events with 2 tight $c$-tagged jet and no $b$-jet and $BL$ to events with a $b$-tagged and a loose $c$-tagged jets. In the boosted regime, at least 2 track-jets are required to be associated to the large-$R$ jet leading by $p_T$, and the tags of the 3 track-jets with the highest $p_T$ are considered for the event. This labelling and the reconstructed $p_T^V$ define the different regimes of the resolved and boosted $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ parts of the combined analysis. To fully reconstruct the Higgs boson from the two candidates jets or the large-$R$ jet, additional selections are applied.

**Higgs candidates in resolved regime:** for $VH(H \to b\bar{b})$, the two candidates must be $b$-tagged (bins 3 and 4) with no additional $B$- and tight $c$-tagged jets allowed[9], while in $VH(H \to c\bar{c})$ no $b$-tagged jet is allowed and at least one of the candidate must be tight $c$-tagged ($T$). As described in the next section, two control regions are defined by changing this flavour selection: a Top CR, combining at least 1 $B$-tag with at least 1 $T$-tag, and the $V + l$ requiring 1 loose $c$-tagged jet ($L$) with an untagged $N$ jet for $VH(H \to c\bar{c})$. The Higgs candidates are sorted by $p_T$ into a leading $j_1$ and sub-leading $j_2$ candidates. The leading candidate must have $p_T > 45$ GeV, while other jets are required to have $p_T > 20GeV$. The invariant mass of the Higgs candidate as measured by $m_{bb}$ ($m_{cc}$) must be above 50 GeV before applying energy corrections, to avoid the $V$+jets gluon splitting mis-modelling at low masses.

**Higgs candidates in boosted regime:** the selection requires exactly 2 $B$-tags among the 3 track-jets leading by $p_T$ associated to the leading large-$R$ jet. The reconstructed mass of the Higgs candidate based on the leading-$R$ jet mass $m_J$ must satisfy $m_J > 50$ GeV, with a leading

[9] In 2L, additional $T$-tagged jets are allowed due to the limited statistics and the different derivation of the top Control Region (CR) in 2L, requiring mixed leptons $e\mu$.

large-$R$ jet $p_T > 250$ GeV.

In all regimes, the number of reconstructed charged lepton in the final state defines three channels as the 0-lepton (0L), 1-lepton (1L), and 2-lepton (2L). The objective of this leptonic selection is to reconstruct the associated $V$-boson. The selection of events in the resolved regime is presented in Table 1.6 and in Table 1.5 for the boosted regime. Additional channel-specific requirements are also introduced to limit backgrounds contamination and reviewed in this section.

**Selection specific to the 0-lepton channel**

In the 0-lepton channel, 0 $VH$-loose leptons should be selected and $E_T^{\text{miss}}$ should be $> 150$ GeV ($> 250$ GeV) in the resolved (boosted) regime, to identify the decay $Z \to \nu\nu$.

In the resolved regime, the scalar sum $S_T$ of the jet $p_T$ in the events must be $> 120$ GeV ($> 150$ GeV) for 2-jets ($\geq 3$ jets) to avoid a mis-modelled region in simulation due to the triggers. In the case of a decaying $W \to \tau\nu$ followed by a hadronic decay of the $\tau$-lepton which is reconstructed as a jet, there are no electron nor muon in the final state. To limit this $\tau$-contamination in the 0L channel, an extra selection is applied when at least 1 hadronic $\tau$ is reconstructed: the reconstructed transverse $W$ mass

$$m_T^W = \sqrt{2p_T^l E_T^{\text{miss}}(1 - \cos(\Delta\phi(l, E_T^{\text{miss}})))}$$

is required to be $m_T^W \geq 10$ GeV, with the $W$-boson $p_T$ estimated from the vectorial sum of the leading hadronic $\tau$ momentum ($p_T^l$) and $E_T^{\text{miss}}$ instead of $p_T^V$.

To limit the multi-jet background, so-called *anti-QCD cuts* using the azimuthal angle $\phi$ are applied in all regimes:

- For resolved only, $|\Delta\phi(j_1, j_2)| < 140°$.
- $|\Delta\phi(E_t^{\text{miss}}, H)| > 120°$.
- $\min|\Delta\phi(E_T^{\text{miss}},\ \text{jet})|$ must be $> 20°$ ($> 30°$) for resolved 2-jet (3-jet) events and $> 30°$ for the boosted regime.

The cuts are tuned to limit the multi-jet contamination to a fraction of order 1% of the total background in 0L, making the mutli-jet negligible in the 0-lepton channel.

### Selection specific to the 1-lepton channel

In the 1L channel, the targeted decaying vector boson is a $W \to \ell\nu$, with $\ell = e, \mu$. Exactly 1 $WH$-signal lepton is required, with events having more than 1 $VH$-loose lepton vetoed[10]. The vector boson is reconstructed from the vectorial sum of the $E_T^{\text{miss}}$ and the lepton transverse momentum identified in the event, with a $p_T^V > 75$ GeV. To suppress the multi-jet background, events with one electron are required to have an $E_T^{\text{miss}} > 30$ GeV ($> 50$ GeV) in the resolved (boosted) regime, with a reconstructed $m_T^W > 20$ GeV for events with low $W$ transverse momentum $p_T^V < 150$ GeV. For the resolved $\mu$-channel, as the same $E_T^{\text{miss}}$ trigger is used as in the 0L, the scalar sum of $p_T$ is similarly restrained with $S_T > 120$ GeV ($> 150$ GeV) for 2-jets ($\geq 3$ jets). A significant background in the 1-lepton channel is the $t\bar{t}$, with both $t$-quarks decaying into a $W$ boson and a $b$-quark. Events where one of the $W$ boson decay follows $W \to \tau\nu$ with the $\tau$ decaying hadronically and the other $W$ decays into an $e$ or a $\mu$ have the same leptonic signature as the signal. A strict hadronic $\tau$-veto is applied in all regimes to suppress this background. Events passing the 0-lepton selection with $\geq 1$ hadronic taus are moved to the 1-lepton channel with the leading hadronic $\tau$ used to reconstruct variables requiring an $e$ or a $\mu$. This migration is performed to recover the estimated 10% ($\sim 20\%$) of $WH$ signal whre $W \to \tau\nu$ with a hadronically decaying $\tau$-lepton in the resolved (boosted) regime, and help decorrelate the $WH$ and $ZH$ measurements in the $VH(H \to b\bar{b})$ side.

### Selection specific to the 2-lepton channel

The 2L channel targets the associated bosonic decay where $Z \to \ell\ell$, with the $Z$ reconstructed from two $VH$-loose leptons required to have the same flavour, with at least one passing the $ZH$-signal lepton requirements. In the di-muon channel, the leptons are further required to be of opposite charges[11]. The invariant mass of the di-lepton system is required to be consistent with the $Z$ mass with $81 < m_{ll} < 101$ GeV in the resolved and $66 < m_{ll} < 111$ GeV in the boosted regime, to suppress non-resonant lepton-pair producing background such a $t\bar{t}$ and multi-jet. The leptons must satisfy $p_T > 25$ GeV, with a stricter $p_T > 27$ GeV required for the leading muon when the event is selected by the muon trigger.

[10] The $VH$-loose lepton is at best the $WH$-signal lepton.
[11] This is not applied to the di-electron channel due to a significantly higher charge mis-identification.

| Selection | 0-Lepton | 1-Lepton | | 2-Lepton | |
|---|---|---|---|---|---|
| | | $e$-channel | $\mu$-channel | $e$-channel | $\mu$-channel |
| Trigger | $E_T^{\text{miss}}$ | Single electron | $E_T^{\text{miss}}$ | Single electron | $E_T^{\text{miss}}$ |
| Leptons | 0 $VH$-loose lepton | 1 $WH$-signal lepton | | $\geq 1$ $ZH$-signal lepton | |
| | | No second $VH$-loose lepton | | 2 $VH$-loose leptons | |
| | | No hadronic $\tau$ | | Same flavour leptons | |
| | | | | Opposite charge for $\mu\mu$ | |
| $p_T^V$ | $> 400$ GeV | | | | |
| Large-$R$ jet | $\geq 1$ large-$R$ jet ($R = 1.0$), $p_T > 250$ GeV, $\lvert\eta\rvert < 2$ | | | | |
| Track-Jets | $\geq 2$ track-jets ($p_T > 10$ GeV, $\lvert\eta\rvert < 2.5$) matched to the leading large-$R$ jet | | | | |
| Tagging | Exactly 2 of the 3 leading track-jets matched to the large-$R$ jet must be $b$-tagged | | | | |
| $m_J$ | $> 50$ GeV | | | | |
| $E_T^{\text{miss}}$ | $> 200$ GeV | $> 50$ GeV | - | - | |
| $\lvert\Delta\phi(E_T^{\text{miss}}, H)\rvert$ | $> 120°$ | - | | - | |
| $\min\lvert\Delta\phi(E_T^{\text{miss}}, \text{jets})\rvert$ | $> 30°$ | - | | - | |
| $m_{\ell\ell}$ | - | - | | 66 GeV $< m_{\ell\ell} <$ 116 GeV | |

Table 1.5: Summary of the event selection in the 0-, 1- and 2-lepton channels of the boosted $VH(H \to b\bar{b})$ regime.

| Resolved Analysis Regime | $VH(H \to b\bar{b})$ | $VH(H \to c\bar{c})$ |
|---|---|---|
| | Common Selections | |
| Jets | $\geq 2$ signal jets | |
| Candidate jets tagging | 2 $B$-tags | $\geq 1$ $T$-tag, no $B$-tag |
| Leading Higgs ($H$) candidate jet $p_T$ | $> 45$ GeV | |
| Sub-leading $H$ candidate jet $p_T$ | $> 20$ GeV | |
| $m_{bb}$ or $m_{cc}$ | $> 50$ GeV (before correction) | |
| Non-$H$ candidate jet $p_T$ | $> 20$ GeV ($> 30$ GeV for nJet categorisation only) | |
| Candidate jets $\Delta R$ | Upper cut $\Delta R \leq \pi$ | |
| | 0-Lepton (0L) | |
| Trigger | $E_T^{\text{miss}}$ triggers | |
| Jets | $\leq 4$ jets | $\leq 3$ jets |
| Additional jets tagging | no $T$-tag | no $B$-tag |
| Top CR tagging | $\geq 1$ $B$-tag + 1 $T$-tag | |
| Leptons | 0 $VH$-loose lepton | |
| $E_T^{\text{miss}}$ | $> 150$ GeV | |
| $E_{T,trk}^{\text{miss}}$ | - | $> 30$ GeV |
| $S_T = \sum p_T^{\text{jets}}$ | $> 120$ GeV (2 jets), $>150$ GeV ($\geq 3$ jets) | |
| $m_T^W$ | $> 10$ GeV when $\geq 1$ hadronic $\tau$ | |
| $\lvert\Delta\phi(j_1, j_2)\rvert$ | $< 140°$ | |
| $\lvert\Delta\phi(E_T^{\text{miss}}, H)\rvert$ | $> 120°$ | |
| $\min\lvert\Delta\phi(E_T^{\text{miss}}, \text{jet})\rvert$ | $> 20°$ (2 jets), $> 30°$(3 jets) | |
| | 1-Lepton (1L) | |
| Trigger | $e$-channel: single electron trigger | |
| | $\mu$-channel: single muon trigger ($p_T^V < 150$ GeV) | |
| | and 0L $E_T^{\text{miss}}$ triggers | |
| Jets | $\leq 3$ jets | |
| Additional jets tagging | no $T$-tag | no $B$-tag |
| Top CR tagging | $\geq 1$ $B$-tag + 1 $T$-tag | |
| hadronic $\tau$-veto | no hadronic $\tau$ | |
| Leptons | 1 $WH$-signal lepton | |
| | veto if $> 1$ $VH$-loose lepton | |
| $E_T^{\text{miss}}$ | $> 30$ GeV ($e$-channel) | |
| $S_T$ | Same as 0L for $\mu$ with $E_T^{\text{miss}}$ trigger | |
| $m_T^W$ | $> 20$ GeV for $75 < p_T^V < 150$ GeV | |
| | 2-Lepton (2L) | |
| Trigger | Same as 1L, $p_T^V < 250$ GeV for single $\mu$ trigger | |
| Additional jets tagging | - | no $B$-tag |
| Leptons | 2 $VH$-loose leptons | |
| | ($\geq 1$ $ZH$-signal lepton) | |
| | Same flavour, opposite-charge for $\mu\mu$ | |
| Top CR | Mixed $e\mu$ flavour | |
| $m_{\ell\ell}$ | $81 < m_{\ell\ell} < 101$ GeV | |

Table 1.6: Summary of the event selection in the 0-, 1- and 2-lepton channels of the resolved $VH(H \to b\bar{b}/c\bar{c})$ regimes. The resolved 1L & 2L Top CR tagging definition ignores the candidate jet tagging requirements For $VH(H \to c\bar{c})$, an extra CR for $V$+$lf$ changes the candidates tagging to one $L$-tag + no-tag ($LN$).

### 1.5.3 Event Categorisation

Selected events are futher categorised following a successive decomposition into regions of defined tag, vector boson $V$ transverse momentum $p_T^V$, and number of jets $N_{\text{jet}}$. The full categorisation gives rise to signal and control regions that enter the statistical analysis defined in the fit framework of Section 1.10. The control regions are defined to help constrain the modelling of specific backgrounds, such as by correcting their yields. The definition of the regions depend on the analysis regime and the targeted Higgs decay, with Figure 1.17 providing a condensed global overview of the different regions defined.

#### Resolved Regime Categorisation

In the resolved regime, the number of central + forward jets in an event defines different $N_{\text{jet}}$ categories, separated to maximise the signal sensitivity. A $p_T > 30$ GeV cut is considered for non-Higgs candidate jets just to determine this number of jet categorisation. This was found to limit the signal migration across STXS bins in $VH(H \to b\bar{b})$ while having almost no impact on the $VH(H \to c\bar{c})$ sensitivity[12]. All distributions of the resolved regime regions with processes normalised to their posfit expectations from the fit decribed in Section 1.10 are presented in Appendix A.6.1. The plots presented in this section shown the prefit and blinded distributions in the different regions, and simulated samples processes are therefore not normalised to data. The variables displayed correspond to those chosen for the fit, as detailed in Section 1.7. The precise definitions of the analysis regions are reviewed in this section.

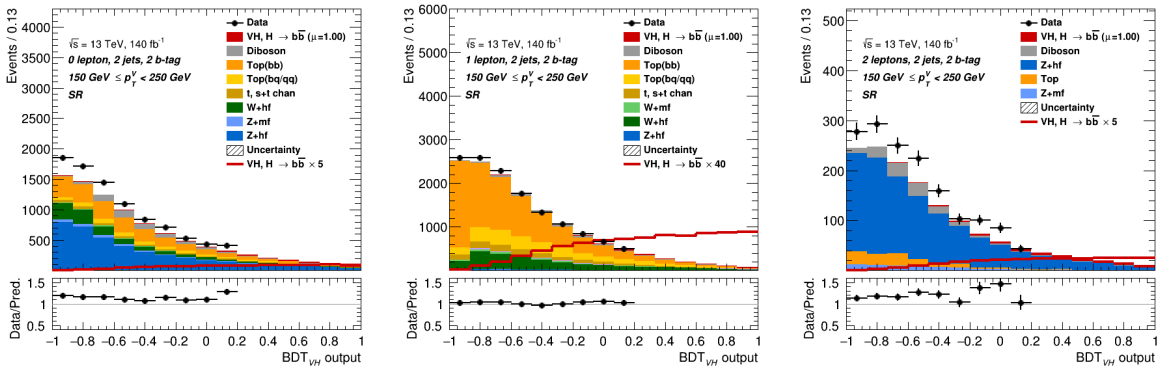
**Resolved $VH(H \to b\bar{b})$ SRs:** requires exactly 2 $b$-tagged jets ($BB$), with no extra $B$- nor any $T$-tags, and events are separated into different categories based on $N_{\text{jet}}$. All lepton channels have a 2-jet and a 3-jet categories. The 0L channel has an addittional 4-jet category, and the 2L an extra 4 or more jets (4p or $\geq$4) category[13]. They are included to improve the STXS measurements sensitivy in bins with at least one additional jets. All regions are further split into different bins of $p_T^V$ as [75, 150] GeV (except for 0L[14]), [150, 250] GeV, and [250, 400] GeV. Selected $VH(H \to b\bar{b})$ signal regions in the analysis are presented in Figure 1.9.

**Resolved $VH(H \to c\bar{c})$ SRs:** adopts a similar event categorisation to the resolved $VH(H \to b\bar{b})$, with now at least one candidate jet being tight c-tagged $T$. The categorisation of the signal

[12] It is nonetheless applied to harmonise the resolved regime.

[13] The 4/4p-jet category in 1L overlaps with the region used to calibrate the tagger and is rejected due to the $t\bar{t}$ limiting the sensitivity.

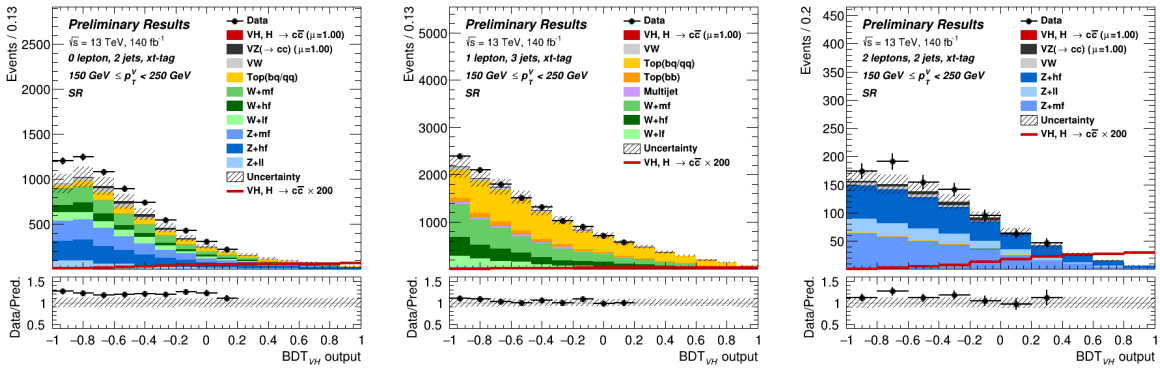
[14] It is not feasible in 0L due to the trigger threshold on $E_T^{\text{miss}}$.




(a) 0-lepton. (b) 1-lepton (c) 2-lepton.

Figure 1.9: $BB$-tagged 2-jet $150 < p_T^V < 250$ GeV signal regions in all lepton channels.

region is then split based on the remaining candidate tag into a 2 $c$-tags region and a 1 $c$-tag region: the former requiring an extra loose ($LT$) or tight $c$-tag ($TT$)[15], the latter an additional non-tagged jet $N$ ($NT$). The $p_T^V$ bins are similar to $VH(H \to b\bar{b})$, except for the higgest $p_T^V$ one that is relaxed to $\geq 250$ GeV given the limited impact of the overlap with the boosted $VH(H \to b\bar{b})$. Adding the $p_T^V$ region above 400 GeV was found to increase the total $VH(H \to c\bar{c})$ sensitivity by 10%. The jet multiplicity $N_{\text{jet}}$ defines a 2 and a 3 jets categories, with the latter being 3 or more jets (3p or $\geq$3) only in 2L thanks to a reduced $t\bar{t}$ background. A selection of 2 $c$-tagged signal regions is presented in Figure 1.10, with Figure 1.11 presenting some 1 $c$-tagged signal regions. The 1 $c$-tag Signal Region (SR)s in the 75 GeV $< p_T^V <$ 150 GeV range are not included in the fit to their significant background yield.

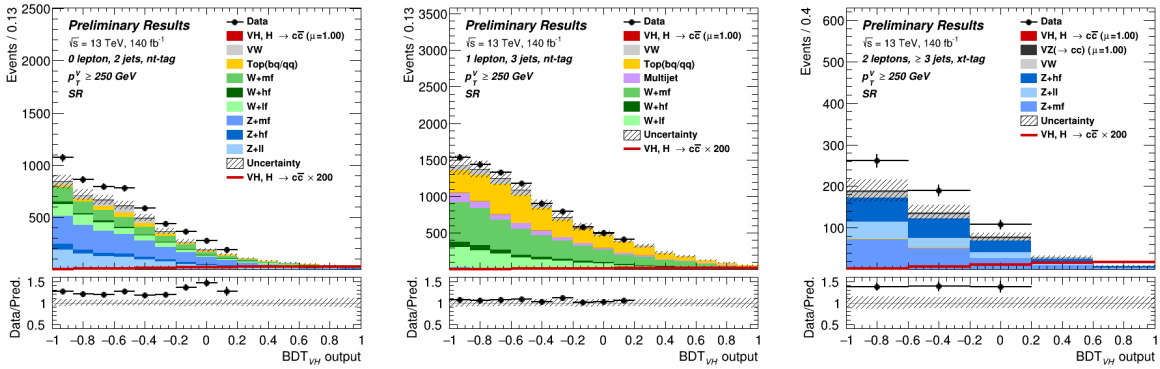



(a) 0-lepton, 2-jet. (b) 1-lepton, 3-jet. (c) 2-lepton, 2-jet.

Figure 1.10: Selection of 2 $c$-tagged ($TT$ + $LT$) $150 < p_T^V < 250$ GeV signal regions.

**The High $\Delta R$ Control Regions:** are designed for the resolved regime to constrain the normalisation and shape of the $V$+jets and the $t\bar{t}$ background when the 2 candidate jets are the $b$-quarks. They are defined by a further split from the SRs based on the angular separation

[15]The 2 $c$-tagged labelled $LT + TT$ is summarised as $XT$ in the plots.




(a) 0-lepton, 2-jet. (b) 1-lepton, 3-jet. (c) 2-lepton, $\geq 3$ (3p) jets.

Figure 1.11: Selection of 1 $c$-tagged $250 < p_T^V$ signal regions.

$\Delta R(j_1, j_2)$[16] (shortened as $\Delta R$ in this document) between the Higgs-candidate jets. This split is governed by a $p_T^V$-dependent cut on the $\Delta R$ that is derived to give a specific signal purity in the SR: keeping 95% (85%) of the signal yield in the 2-jet (3 or more jets) SRs. The mathematical expression of the cuts are presented in Table 1.7, and illustrated in Figure 1.12, with more details given in Appendix A.2.1. Events with a $\Delta R$ below the cutting line enter the signal region, while those above go in a High $\Delta R$ CR, also called *CRHigh*. To avoid some mis-modeling effect at high $\Delta R$ and keep the High $\Delta R$ CR kinematically close to the SR, an upper cut of $\Delta R \leq \pi$ is applied to all events. This effectively remove $\sim 40\%$ of events in the High $\Delta R$ CR, with a negligible impact on the signal region. For $VH(H \to c\bar{c})$, CRHighs are considered for every 1 and 2 $c$-tagged SRs[17], with the $TT$- and $LT$-tagged events separated in the CRHigh to respectively constrain the $V$+$hf$ and $V$+$mf$instead of being merged as in the SR. In $VH(H \to b\bar{b})$, the CRHighs are used to extract the normalisation of the backgrounds while in $VH(H \to c\bar{c})$ the shapes of the $m_{c\bar{c}}$ and $p_T^V$ spectrum are also used, as detailed in Section 1.7. Some High $\Delta R$ CRs are shown in Figure 1.13.

| Category | High $\Delta R$ Cut | Low $\Delta R$ Cut |
|---|---|---|
| 2-jet | $\Delta R > 0.787 + e^{1.387-0.0070\times p_T^V}$ | $\Delta R < 0.410 + e^{0.818-0.0106\times p_T^V}$ |
| 3-jet | $\Delta R > 0.684 + e^{1.204-0.0060\times p_T^V}$ | $\Delta R < 0.430 + e^{0.399-0.0093\times p_T^V}$ |
| 4-jet | $\Delta R > 0.863 + e^{0.984-0.0041\times p_T^V}$ | $\Delta R < 0.411 + e^{1.204-0.0060\times p_T^V}$ |
| $\geq$5-jet | $\Delta R > 1.667 + e^{0.519-0.0050\times p_T^V}$ | $\Delta R < 0.501 + e^{1.192-0.0075\times p_T^V}$ |

Table 1.7: Cuts defining the High $\Delta R$ (centre) and Low $\Delta R$ (right) control regions, CRHigh & CRLow. The inequalities are set to enter the control regions, with $p_T^V$ expressed in GeV.

[16] $\Delta R(j_1, j_2) = \sqrt{(\eta_{j_1} - \eta_{j_2})^2 + (\phi_{j_1} - \phi_{j_2})^2}$.

[17] Also for the in the 1L-chhanel in the 1 $c$-tagged 75 GeV $< p_T^V <$ 150 GeV, where the equivalent SR is not included.

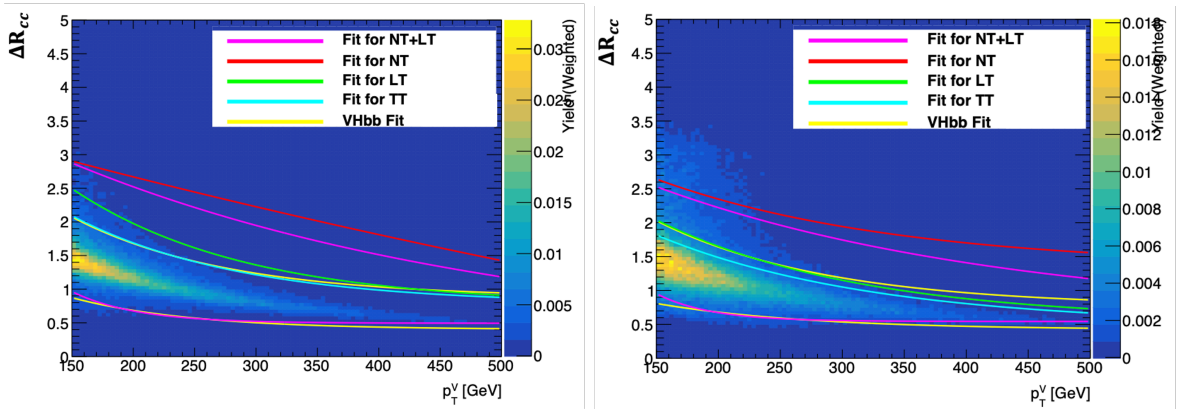
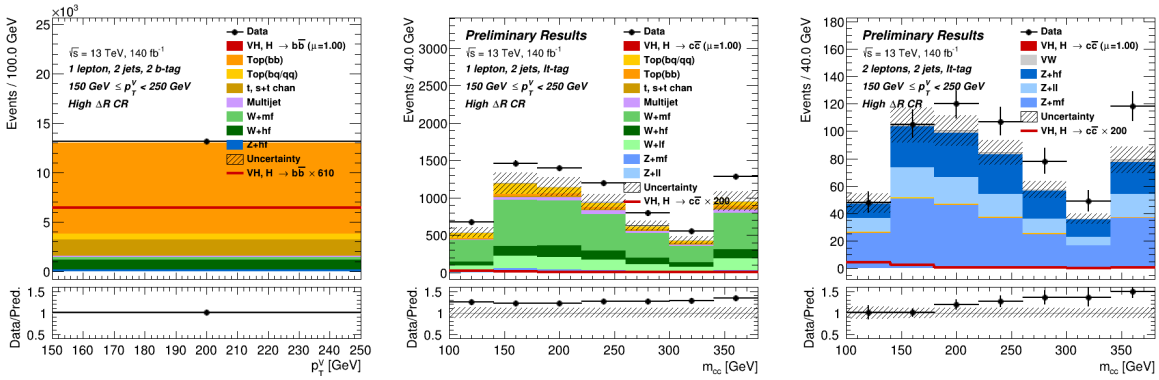



Figure 1.12: The $p_T^V$-$\Delta R_{c\bar{c}}$ 2D signal yield map of the 1-lepton $VH(H \to c\bar{c})$, for the 2-jet (left) and 3-jet (right) regions. The lines are the results of fitting the high and low $\Delta R_{c\bar{c}}(p_T^V)$ cuts for various signal tags, with the yellow curve showing the $VH(H \to b\bar{b})$ $\Delta R_{b\bar{b}}$ cut used in the analysis, with the CRHigh above the top yellow line, and the SR below. A Low $\Delta R$ CR can be defined by the bottom lines, splitting an extra region from the signal region for the $VH(H \to b\bar{b})$ only.




(a) 1-lepton, $BB$. (b) 1-lepton $LT$. (c) 2-lepton $LT$.

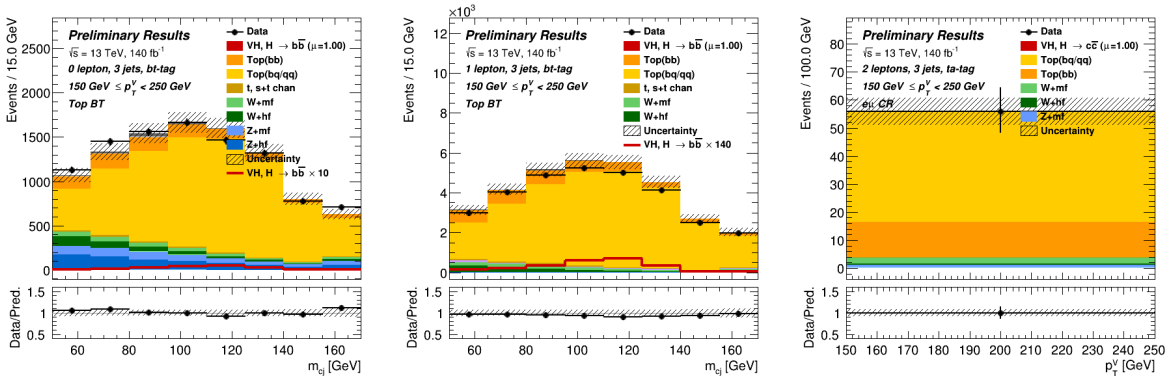
Figure 1.13: Some High $\Delta R$ CRs (CRHigh) with 2 jets and $150 < p_T^V < 250$ GeV.

**The Low $\Delta R$ Control Regions:** low-$\Delta R$ control regions ($CRLow$) are defined in $VH(H \to b\bar{b})$ 1L to better constrain the $W$+hf contribution. They are based on $p_T^V$-dependant cuts defined similarly to the High $\Delta R$ ones, separating 10% of the diboson events from the signal regions, as displayed in the bottom parts of the plots in Figure 1.12. The cuts used are defined in the right of Table 1.7, where events with a $\Delta R$ above the cutting line enter the signal region while those below go to the CRLow. In $VH(H \to c\bar{c})$ and the 0L and 1L $VH(H \to b\bar{b})$, the CRLow is not separated from the signal region as having it has little impact on the sensitivty of the fit. One of the CRLow region is presented in the left of Figure 1.15.

**Top Control Regions in 0L and 1L:** are defined to constrain the Top background Top($bc$) and Top($bl$) components[18]. The so-called *Top BT CRs* are shared by theresolved $VH(H \to b\bar{b})$

[18]The component in the parenthesis refers to the flavour of the Higgs-candidate jets. As explained later in this chapter, they are floated together in the fit as the Top($bq/qq$).

and $VH(H \to c\bar{c})$, with similar $p_T^V$ and jet multiplicity categorisation as the SRs. In 0L and 1L, they are defined by requiring events to have at least one $B$-tag and at least one tight $c$-tag $T$, making them orthogonal to the $VH(H \to b\bar{b})$ signal regions. The Higgs candidate is reconstructed from the leading $B$ jet among $B$-tagged jet and leading $T$ jet among $T$-tagged jet, for kinematic similarity to the SRs. The Top($b\bar{b}$) component, which is significant for $VH(H \to b\bar{b})$ due to the required tag, is controlled from the previously defined CRHighs in 0L and 1L, thanks to the large $\Delta R$ between the produced $b$ jets in a $t\bar{t}$, as shown in Figure 1.13a. Two Top $BT$ control regions are presented in the left of Figure 1.14.



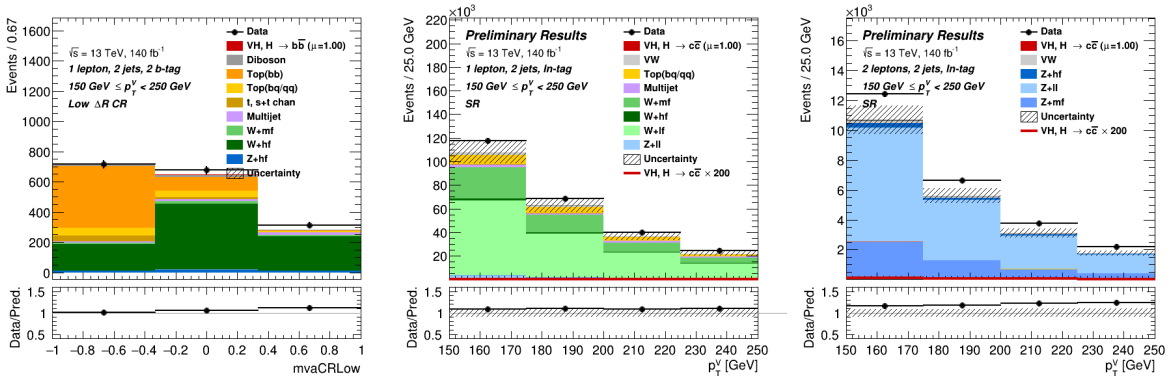

(a) 0-lepton, $BT$, 3-jet. (b) 1-lepton, $BT$, 3-jet. (c) 2-lepton, $e\mu$, $\geq 1$ $T$-tag.

Figure 1.14: Two Top CR $BT$-tagged (left & centre) and a Top $e\mu$ CR (right), all with 3 jets and $150 < p_T^V < 250$ GeV.

**Top Control Regions in 2L:** there, the Top background is mostly made of di-leptonic $t\bar{t}$ decays, with both subsequent $W$ decaying leptonically. High purity Top CRs are derived for the 2-lepton channels by requiring leptons of different flavours ($e\mu$ or $\mu e$) instead of the same flavour ($ee$ or $\mu\mu$). This mixed of flavour is possible as the leptons are produced in distinct $W$-boson decays. These so-called *Top $e\mu$ CRs* are used to derive a $t\bar{t}$ background template in a data-driven way for the 2-lepton SRs in $VH(H \to b\bar{b})$. For $VH(H \to c\bar{c})$, the $t\bar{t}$ is a less significant component due to the flavour tagging requirements, and the Top $e\mu$ CRs contribute to the fit as single-bin 2-lepton CRs per $p_T^V$ and jet multiplicity, with at least one $T$-tag jet. An example of such a CR is presented in Figure 1.14c.

**$V$ + light-jets Control Regions:** the $V$ + light-jets background is particularly significant for $VH(H \to c\bar{c})$, due to the difficulties in discriminating $c$-jets from light-jets. Dedicated CRs, labelled *$V + l$ CR*, in the 1L and 2L channels target, respectively, the $W$+*lf* and $Z$+*lf* backgrounds[19]. They are defined by requiring exactly one loose $L$-tag $c$-jet without any $T$- nor

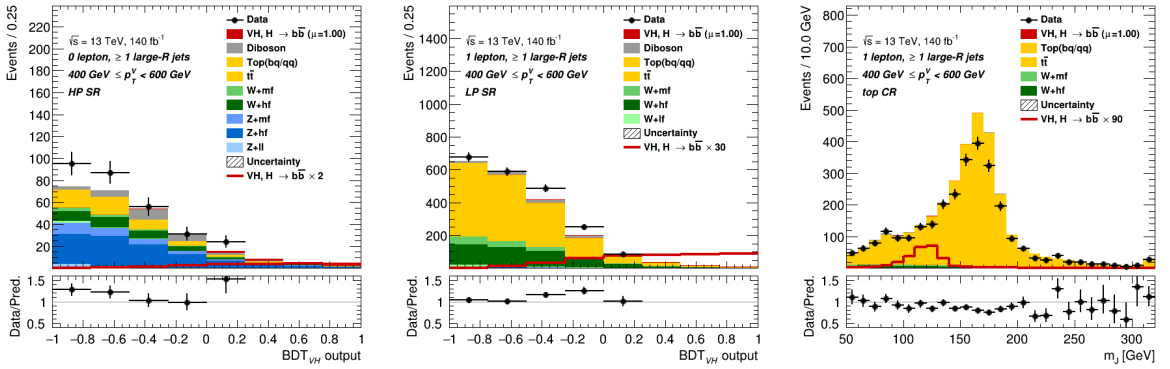
[19]The $V$+*lf* corresponds to a grouping of the $V$+jets components with light-flavour jets, as introduced in Section 1.9.3.




(a) 1-lepton Low $\Delta R$ CR, $BB$. (b) 1-lepton $V+l$ CR $LN$. (c) 2-lepton $V+l$ CR $LN$.

Figure 1.15: A $BB$-tagged Low $\Delta R$ CR (left) and two $LN$-tagged $V+l$ CRs (centre & right), both with 2 jets and $150 < p_T^V < 250$ GeV.

$B$-tagged jet in the event. The selection is otherwise similar to that of the 1 $c$-tagged signal regions[20], with the candidate pair now tagged as $LN$, where $N$ is the leading untagged central jet. The 1L $V+l$ CRs are 60% pure in $W$+$lf$, while the 2L $V+l$ CRs reach a 70% $Z$+$lf$ purity. An example of the former is shown in Figure 1.15b, while a 2L $V+l$ CR is shown in Figure 1.15c.




(a) 0-lepton high purity SR. (b) 1-lepton low purity SR. (c) 1-lepton top CR.

Figure 1.16: Some boosted $BB$-tagged with $400 < p_T^V < 600$ GeV regions signal regions (left & centre) and boosted Top CR (right).

### Boosted Regime Categorisation

In the boosted $VH(H \to b\bar{b})$, two $p_T^V$ bins are defined at [400, 600] GeV and $\geq 600$ GeV to avoid overlap with the resolved $VH(H \to b\bar{b})$. The SRs are defined by requiring exactly two of the three leading subjets (track-jets) associated to a single leading large-$R$ jet to be $b$-tagged, with no additional track-jet outside the large-$R$ jet being $B$-tagged to enhance the top background rejection. All boosted regions, with processes normalised to their postfit expectations, are presented in Appendix Section A.6.2. In the plots, the SRs are further separated into a high- (HP) and low-purity, *HP SR* and *LP SR*, when there are no or $\geq 1$ additional small-$R$ jet not

[20] Similarly to these SRs, there is no 1L $V+l$ CR for 75 GeV $< p_T^V <$ 150 GeV.

associated to the Higgs-candidate large-$R$ jet. These regions are however combined into single signal regions in the final fit.

**Boosted Top Control Regions in 0L and 1L:** events that have an additional $B$-tagged track-jet outside the large-$R$ jet as defined by an angular separation of

$$\Delta R(\text{VR-track jet, large-}R\text{ jet}) > 1$$

are moved toa the boosted Top control regions in the 0L and 1L channels. The $t\bar{t}$ process is the main background in these lepton channels, where a $t$-quark decay is captured as a single large-$R$ jet merging the produced $b$ and a hadronically decaying $W$. The boosted Top CRs effectively capture this signature by identifying the $b$-quark from the other decaying $t$-quark in the $t\bar{t}$ pair, with the same 85% $b$-tagging WP. An example of such a region is displayed in Figure 1.16.

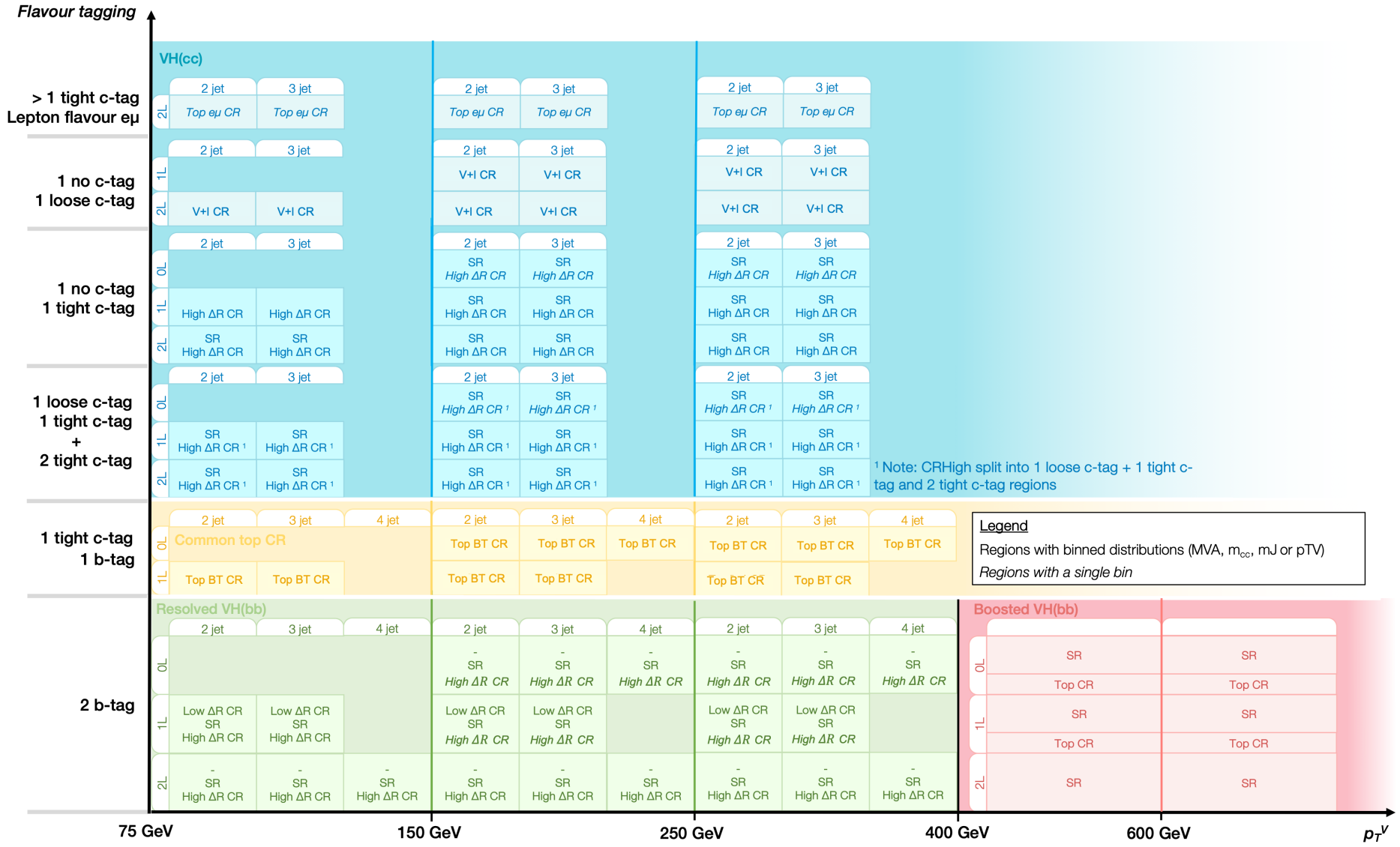



Figure 1.17: The $VH(H \to b\bar{b}/c\bar{c})$ combined analysis regions, showing the Signal Regions (SR), High and Low $\Delta R$ control regions (CRHigh & LowCR), the Top $BT$ CR, the Top $e\mu$ CR, the $V + l$ $LN$-tagged CR, and the boosted Top CR. $VH(H \to c\bar{c})$ is in blue, the shared Top $BT$ CR in the resolved regime in yellow, and $VH(H \to b\bar{b})$ in green and red for the resolved and boosted regimes respectively. Regions used in the fit as single-bin distributions to derive an absolute normalisation are indicated in italic.

## 1.6 Tagged-jets corrections

Several corrections to the energy are applied to tagged jet from the previously introduced selection. The objective is to improve the energy resolution of the pair of jets selected to form the Higgs candidate. All jets benefit from a default jet energy calibration called the *Global Sequential Calibration (GSC)* [68], as introduced in Section **??**. This global correction is not optimal for $b$- and $c$-jets that benefit from special features, motivating the use of additional dedicated corrections for such jets. Table 1.8 summarises the additional corrections presented in this section.

| Scheme | Lepton channel | Muon-in-jet | $p_T$-Reco | Kinematic fit | FSR Recovery |
|---|---|---|---|---|---|
| Resolved $VH(H \to b\bar{b})$ | 0L | ✓ | ✓ | | |
| | 1L | ✓ | ✓ | | |
| | 2L | ✓ | ✓($N_{\text{jet}} \geq 4$) | ✓($N_{\text{jet}} \leq 3$) | ✓($N_{\text{jet}} \leq 4$) |
| $VH(H \to c\bar{c})$ | 0L | ✓ | | | |
| | 1L | ✓ | | | |
| | 2L | ✓ | | ✓($N_{\text{jet}} \leq 3$) | ✓($N_{\text{jet}} \leq 4$) |
| boosted $VH(H \to b\bar{b})$ | 0L | ✓ | | | |
| | 1L | ✓ | | | |
| | 2L | ✓ | | ✓ | |

Table 1.8: The different $H$-candidate jet energy correction.

***Muon-in-jet* correction** is applied to all events to correct the energy of semi-leptonically decaying $b$- and $c$-jets with a muon in the jet cone. The energy of this $\mu$ is not measured in the calorimeter but deduced from the curvature of the muon track. For the resolved regime, the closest muon's 4-momentum $p_T^\mu$ is added to the jet if its angular separation from the jet axis satisfies

$$\Delta R(\text{jet}, \mu) \leq \min\left(0.4,\, 0.04 + \frac{10\ \text{GeV}}{p_T^\mu}\right).$$

In the boosted scheme, the angular separation is measured with respect to the track-jets but the muon 4-momentum $p_T^\mu$ is added to the large-$R$ jet in case of match.

***PtReco* correction** accounts for missing energy from neutrinos in the semi-leptonic decays or from the out-of-cone effect for $b$-jets. It is only applied to $b$-tagged jets in the resolved $VH(H \to b\bar{b})$ 0L and 1L channels, and the $\geq$ 4-jets 2L channel. The correction is derived from the signal samples of $VH(H \to b\bar{b})$ by comparing the truth jet $p_T$ and the reconstructed $p_T$ after the muon-in-jet correction. The correction is not applied to $VH(H \to c\bar{c})$ as it does not have a significant effect due to the lower likelihood of semi-leptonic decays and out-of-cone effects for

$c$-jets.

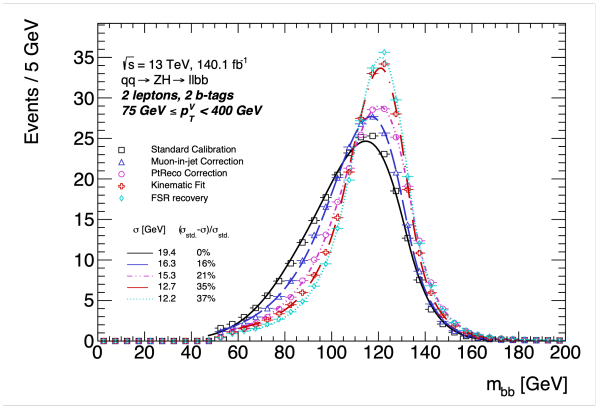
***Kinematic* fit correction** is applied in the 2L channel of the resolved regime, for events with 2 or 3 jets only. The $ZH \to \ell\ell b\bar{b}/\ell\ell c\bar{c}$ is fully reconstructed and a kinematif fit is applied to improve the $m_{jj}$ resolution after the muon-in-jet correction. The fit is performed using a likelihood function with terms covering the object resolution, the jet transfer function, a $Z$-mass constraint term, and system $p_T$ balance. The boosted 2L channel has a similar kinematic fit based on a Gaussian term instead. The procedure is not applied to events with more than 3 jets as the benefits are smeared out by the additional jets.

**FSR recovery** is deployed for events with 3 or 4 jets in the 2L resolved regime, to further improve the resolution of the $m_{bb}$ or $m_{cc}$ peak after the kinematic fit correction. Such events are likely to have jets emanating as Final State Radiation (FSR), whereby a quark or a gluon is emitted by a final state particle. A continuous cut on the sum $\Delta R_{j,j_1} + \Delta R_{j,j_2}$ of angular separations between a third or fourth jet ($j$) to the Higgs-candidate jets ($j_1$ and $j_2$) is applied as a function of $p_T^V$. Any additional jet below the cut is considered a radiation and is added to the closest candidate jet. This effectively corrects the reconstructed mass of Higgs bosons as well as the jet multiplicity, leading to an expected 7% improvement in $VH(H \to b\bar{b})$ STXS sensitivity by reducing the migration between measurement bins. Due to the possible increase acceptance of the $t\bar{t}$ background in the sensitive region from the reduction of additional jets, this correction is not applied to 0L nor 1L.

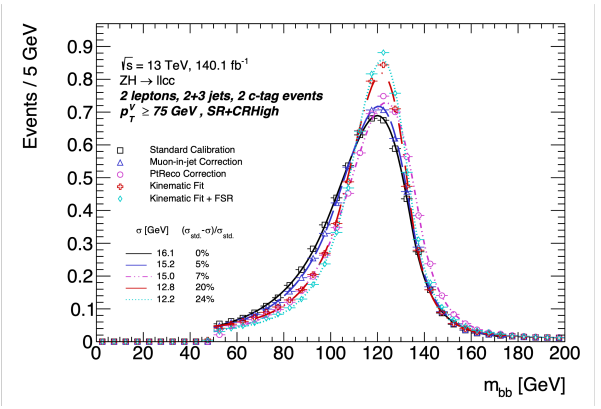
The effects of the different reconstruction techniques are illustrated in Figure 1.18 for some selected 2-lepton resolved and boosted distributions.

## 1.7 Discriminant Variables

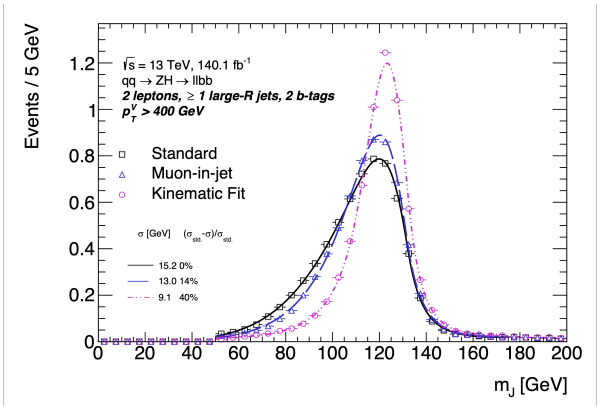
The analysis leverages a varied set of reconstructed physical variables in the fit to constrain the different processes and control mis-modelling effects. Figure 1.19 displays the variables used for each control region in the fit in the resolved regime. The reconstructed Higgs mass directly offer some separation power of the signals from their major backgrounds, hence some control regions such as the Top $BT$ CR and CRHigh are modelled with these relevant variables: $m_{bb}$, $m_{cc}$, and $m_J$, depending on the targeted decay and the regime. Some CRs passed to the fit are modelled with the $p_T^V$ distribution, such as the CRHigh in the 2L-channel $NT$ and $BB$ tagged-regions and the $V + l$ CR ($LN$-tag). Directly fitting this distribution helps constrain a Monte-Carlo mis-




(a) Resolved $VH(H \to b\bar{b})$.




(b) Resolved $VH(H \to c\bar{c})$ with 2 $c$-tags.




(c) Boosted $VH(H \to b\bar{b})$.

Figure 1.18: Performance of the energy corrections on simulated samples of different analysis schemes in the 2-lepton channels, inclusive in $p_T^V$ and number of jets.

modelling in the $p_T^V$ distributions of the SHERPA 2.2.11 $V$+jets samples, as detailed in Section 1.9. To optimise signal and background seperation in the statistical analysis, dedicated Boosted Decision Trees (BDT), also called MVA, are trained with the TMVA Root framework [69] in the signal regions of the combined analysis. Simple one-dimensional discriminants are built from the outputs of fine-tuned BDTs trained on specific sets of event-level input variables, as described in greated details in this section.

### 1.7.1 Multivariate Analysis

Three sets of discriminants are trained for the analysis: one set of called *MVA* discrimants for the signal region modelling, and a specific set called *mvaCRLow* for the CRLow distribution in the resolved $VH(H \to b\bar{b})$. An additional set of BDTs is trained for the cross-check analysis targetting the diboson process as signal, where one of the bosons is a $Z$ ($WZ$ or $ZZ$n summarised as $VZ$) with a $b\bar{b}$ or $c\bar{c}$ pair in the final state. For this purpose, the signal is set to the diboson process decayig into the expected pair of jets $VZ(\to b\bar{b})$ or $VZ(\to c\bar{c})$, and the non diboson pro-

cesses as well as the $VH$ processes are set as backgrounds. All multivariate discriminants predict a continous score in the range $[-1, 1]$, with higher values indicating a signal-like component and lower values background-like.

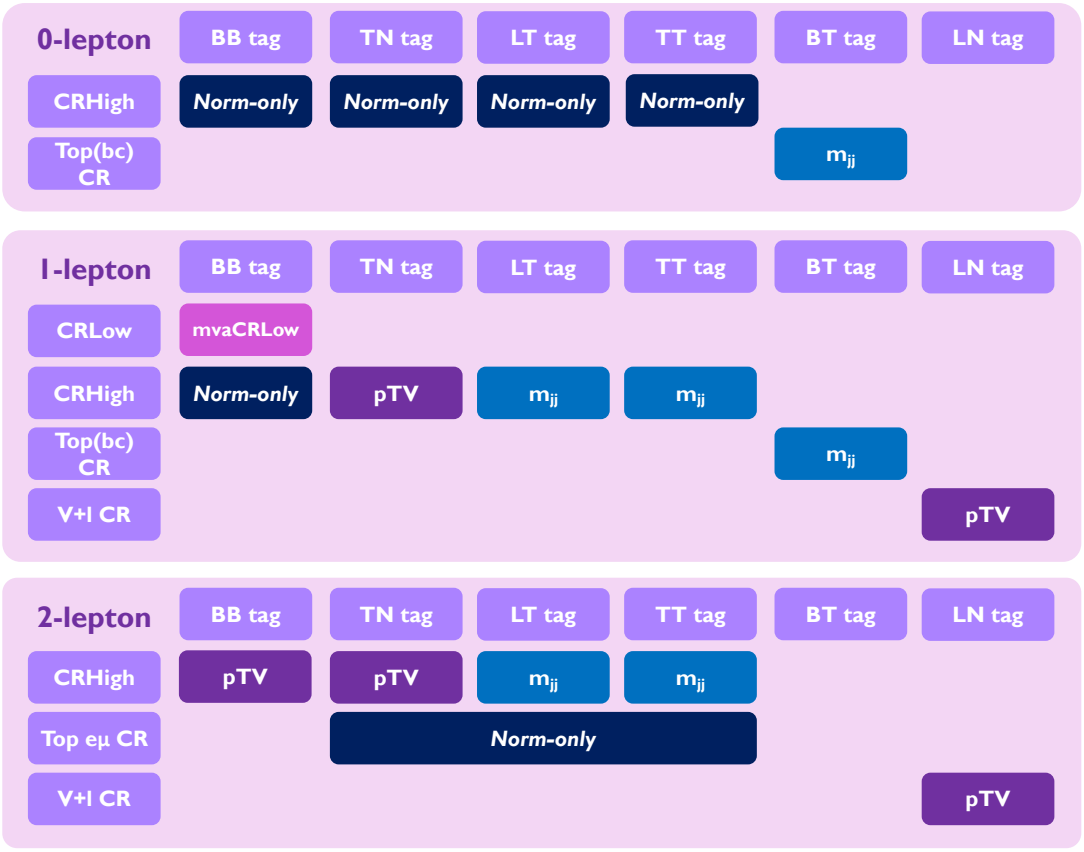



Figure 1.19: Illustration of the discriminant variables used per control regions of the resolved regime in the fit. Norm-only indicates a region to extract a global normalisation and not binned by a variable.

The wide-adoption of BDT-discriminants in all regime of the analysis marks a significant improvement over the standalone $VH(H \to c\bar{c})$ and boosted $VH(H \to b\bar{b})$ analyses, generalising the successful approach first introduced in the resolved $VH(H \to b\bar{b})$ [28]. Prior to training, the object and event selections of Section 1.5 and the jet energy corrections of Section 1.6 are applied. To limit the number of training runs and the risk of overtraining from the low statistics of some kinematic regions, the BDTs are trained on inclusive regions combining the SRs and the $\Delta R$-based CRs (CRHigh and CRLow). The BDTs are trained to discriminate the respective signal of the different targeted decays [21] from background samples, including $V$+jets, $t\bar{t}$, single-top, and diboson. BDTs were specifically trained in the following specific categories, covering the fine analysis categorisation with different categories limits to guarantee sufficient statistics and avoid overtraining:

[21] The $VH(H \to b\bar{b})$ samples for the $BB$-tagged events and $VH(H \to c\bar{c})$ samples for the 1 and 2 $c$-tagged events.

- **Resolved $VH(H \to b\bar{b}/c\bar{c})$:** BDTs are trained separately for the $BB$-, 2 $c$- ($TT + TL$ abbreviated $XT$), and 1 $c$-tags ($NT$). Seperate trainings are run for each lepton channel and for the following jet multiplicities[22] and $p_T^V$ bins:
  - **0L**: separate BDTs are trained for the 2-, 3-, and 4-jet categories, each in one inclusive $p_T^V \geq 150$ GeV region.
  - **1L**: separate BDTs are trained for the 2- and 3-jet categories, each in two $p_T^V$ bins: $p_T^V \in [75, 150]$ GeV and $p_T^V \geq 150$ GeV.
  - **2L**: separate BDTs are trained for the 2- and ≥3-jet categories in two $p_T^V$ bins: $p_T^V \in [75, 150]$ GeV and $p_T^V \geq 150$ GeV.

  The low $p_T^V$ bin is separated from the higher $p_T^V > 150$ GeV due to its large statistics and different background compositions.
- **Boosted $VH(H \to b\bar{b})$:** one BDT is trained per lepton channel in an inclusive bin of $p_T^V \geq 400$ GeV.

For training, the full MC samples statistics in all analysis regimes are used thanks to the so-called *GNN truth tagging*. Instead of filtering down the simulated samples by cutting away events failing to pass the flavour tagging requirements - the standard application of the selection called *direct tagging* -, this technique applies a weight to each event to represent its probability of passing the tagging selection. The result is a weighted distribution possessing the statistical precision of the full MC-samples but shaped as the directly tagged distributions. The weights in this truth tagging procedure are predicted by a GNN-based neural network passed event-level information, with more details given in Appendix A.2.3. The truth tagging procedure is applied to $BB$ for $VH(H \to b\bar{b})$, and to $TT$, $TL$, and $NT$ for $VH(H \to c\bar{c})$. The variables used for each lepton channel in both regimes are listed in Table 1.9, with more precise definitions of each variable given in Appendix A.3. Features with long tails are clipped to contain 99% of the centred distributions, and given a specifically chosen default value when they are not defined for an event. The sets of features used are the result of hyperparameter optimisation campaigns, with many other variables tested but eventually not included due to their negligible impact on the performance.

The architectures of the different BDTs are optimised, with the gradient boosting technique of Section **??** deployed in the resolved regime to improve performance and to capture effects outside the bulk of the distributions. In the boosted regime, due to the lower statistics available

[22] The jet multiplicity only accounts for jets with $p_T > 30$ GeV.

| | $VH(H \to b\bar{b}/c\bar{c})$ Resolved | | | $VH(H \to b\bar{b})$ Boosted | | | |
|---|---|---|---|---|---|---|---|
| Variable | 0L | 1L | 2L | 0L | 1L | 2L | |
| $m_{j_1 j_2}$ or $m_J$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Mass of Higgs candidate |
| $m_{j_1 j_2 j_3}$ | ✓ | ✓ | ✓ | | | | Mass of Higgs candidates and leading additional jet |
| $p_T^{j_1}$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Leading Higgs candidate $p_T$ |
| $p_T^{j_2}$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Sub-leading Higgs candidate $p_T$ |
| $p_T^{j_3}$ | | | | ✓ | ✓ | ✓ | Leading non-Higgs candidate $p_T$ |
| $\sum_{i\neq 1,2} p_T^{j_i}$ | ✓ | ✓ | ✓ | | | | Sum of non-Higgs jet $p_T$ |
| $\Delta R(\boldsymbol{j_1}, \boldsymbol{j_2})$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Angular separation of Higgs candidates |
| $\text{bin}_{\text{DL1r}}(j_1)$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Tag bin of $j_1$ |
| $\text{bin}_{\text{DL1r}}(j_2)$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Tag bin of $j_2$ |
| $p_T^V$ | $\equiv E_T^{\text{miss}}$ | ✓ | ✓ | $\equiv E_T^{\text{miss}}$ | ✓ | ✓ | Vector boson $p_T$ |
| $E_T^{\text{miss}}$ | ✓ | ✓ | | ✓ | ✓ | | Missing transverse energy |
| $E_T^{\text{miss}}/\sqrt{S_T}$ | | | ✓ | | | | Ratio of $E_T^{\text{miss}}$to sum of jets $p_T$ |
| $\lvert\Delta y(\boldsymbol{V}, \boldsymbol{H})\rvert$ | | ✓ | ✓ | | ✓ | ✓ | Rapidity difference between $V$ and $H$ |
| $\lvert\Delta\phi(\boldsymbol{V}, \boldsymbol{H})\rvert$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Azimuthal angle between $V$ and $H$ |
| $\lvert\Delta\eta(\boldsymbol{j_1}, \boldsymbol{j_2})\rvert$ | ✓ | | | | | | Pseudorapidity distance between Higgs candidates |
| $\min \Delta R(j_i, j)_{i=1,2}$ | ✓ | ✓ | | | | | Smallest angular distance between a Higgs and non-Higgs candidates |
| $\min[\Delta\phi(\boldsymbol{\ell}, \boldsymbol{j_1} \text{ or } \boldsymbol{j_2})]$ | | ✓ | | | | | Smallest $\phi$ between the lepton and a Higgs candidate |
| $m_{\text{eff}}$ | ✓ | | | | | | Scalar sum of $p_T$ of all small-$R$ jet and $E_T^{\text{miss}}$ |
| $m_T^W$ | | ✓ | | | | | Transverse mass of the $W$ |
| $m_{\text{top}}$ | | ✓ | | | | | Mass of reconstructed leptonically decaying top-quark |
| $m_{\ell\ell}$ | | | ✓ | | | | Mass of di-lepton system |
| $\cos\theta(\boldsymbol{\ell^-}, \boldsymbol{Z})$ | | | ✓ | | | ✓ | $Z$ boson polarisation sensitive angle |
| $(p_T^{\ell_1} - E_T^{\text{miss}})/p_T^W$ | | | | | ✓ | | |
| $p_T^{\ell}$ | | | | | ✓ | | $p_T$imabalance of the lepton and neutrino from $W$ |
| $N$(track-jets in $J$) | | | | ✓ | ✓ | ✓ | Number of track-jets associated to leading-$R$ jet |
| $N$(add. small R-jets) | | | | ✓ | ✓ | ✓ | Number of additional small-$R$ jets not matched |
| Colour | | | | ✓ | ✓ | ✓ | Variable modelling colour-flow from QCD |

Table 1.9: The variables used for the 0-, 1- and 2L channels MVA's in the resolved and boosted regimes for the $VH(H \to b\bar{b}/c\bar{c})$combined analysis. The variables are further described in Appendix A.3.

and large tails in the distributions, the AdaBoost method - introduced in Section **??** - is adopted to help stabilise the trainings [70]. Tables 1.10 and 1.11 list the architectures used for the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ BDTs respectively, with the main and diboson BDTs sharing the same hyperparameters. For $VH(H \to c\bar{c})$, the hyperparameters are further tuned to avoid overtraining from the smaller available statistics in the 2L channel and the diboson cross-check.

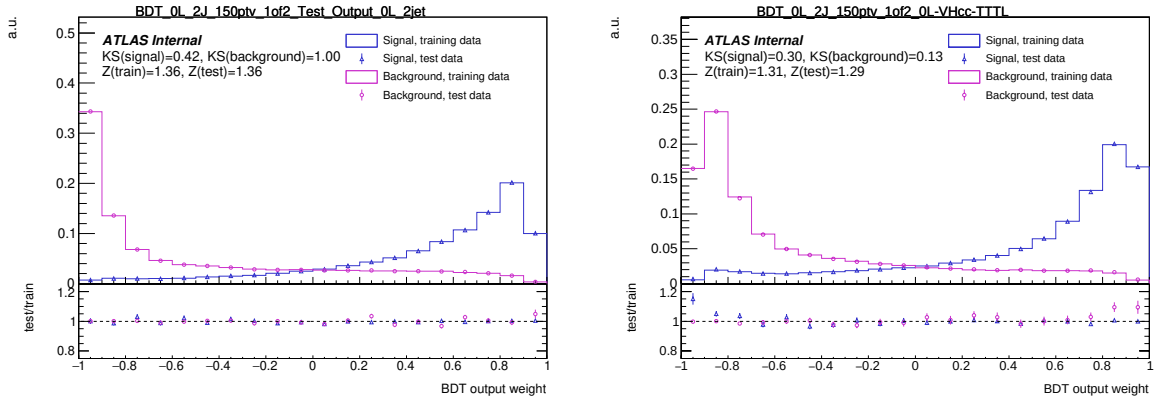
| | Resolved $VH(H \to b\bar{b})$ | | | Boosted $VH(H \to b\bar{b})$ | | |
|---|---|---|---|---|---|---|
| **Settings** | 0L | 1L | 2L | 0L | 1L | 2L |
| Boost type | Gradient boost | Gradient boost | Gradient boost | Adaboost | Adaboost | Adaboost |
| Number of trees | 200 | 600 | 200 | 800 | 800 | 400 |
| Maximum depth | 3 | 4 | 4 | 3 | 3 | 3 |
| Learning rate ($\beta$) | 0.5 | 0.5 | 0.5 | 0.5 | 0.35 | 0.3 |
| Number of cuts | 100 | 100 | 100 | 60 | 60 | 100 |
| Minimum node size | 5% | 5% | 5% | 2% | 2% | 7% |

Table 1.10: Hyperparameters of the BDTs per lepton channel of the $VH(H \to b\bar{b})$ resolved and boosted. All models used the Gini index as separation method, without pruning.

| | $VH(H \to c\bar{c})$ | | $VZ \to c\bar{c}$ Cross-check |
|---|---|---|---|
| **Settings** | 0L, 1L & most 2L regions | 2- & $\geq$3-jet, low $p_T^V$ | 0L, 1L, 2L |
| Boost type | Gradient boost | Adaboost | Adaboost |
| Number of trees | 600 | 200 | 200 |
| Maximum depth | 4 | 4 | 4 |
| Learning rate ($\beta$) | 0.5 | 0.15 | 0.15 |
| Number of cuts | 100 | 100 | 100 |
| Minimum node size | 5% | 5% | 5% |

Table 1.11: Hyperparameters of the BDTs per lepton channel of $VH(H \to c\bar{c})$. The 2L low $p_T^V$ region mentioned covers 75 GeV < pTV < 150 GeV. All models used the Gini index as separation method, without pruning.

Trainings are performed with the $k$-fold method, setting $k = 2$, to use the full statistics while assessing the overtraining risk. In other words, each BDT is doubly trained: once on odd events, and once on even events. The performance is assessed on the held-out fold and the final discriminant is the combination of the odd- and even-trained BDTs, thereby exposed to the whole training set. Additional overtraining checks are performed on each fold-training, comparing the trained distribution to a test distribution obtained by applying the BDT on the heldout set for the fold, as presented in Figure 1.20. The BDTs deliver a good discrimination performance, with a typical Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) of $\sim 0.9$ and a large increase on the expected statistical significance of the analysis compared to using the Higgs candidate mass as discriminating variable.




(a) $BB$-tagged model, test AUC = 0.9. (b) 2 $c$-tagged model, test AUC = 0.898.

Figure 1.20: Overtraining checks for the BDTs trained for the resolved $VH(H \to b\bar{b})$ (left) and $VH(H \to c\bar{c})$ (right) in the 0L 2-jet region with $p_T^V \geq 150$ GeV. The binned histrograms are the training data (blue) and background (purple) distributions, while the data points are the equivalent test distributions - the bottom plots show the ratio of test / train.

In addition to the signal and cross-checks MVAs, extra MVAs are trained for the $VH(H \to b\bar{b})$ resolved 1L channel in the Low $\Delta R$ CR (CRLow). This region is dedicated to the $W$+jets process, with a rich contribution of the important $W + bb$ background. At low $p_T^V$, there is unfortunately also a large contribution from $t\bar{t}$, reducing the purity of the $W+bb$ in the CR. To recover a higher sensitivity to this background, MVAs are specially trained to discriminate the $W + bb$ process from other backgrounds in the CRLow $BB$-tagged events. They are similarly trained with 2-fold on truth tagged samples, separately for the $p_T^V < 150$ GeV and $p_T^V > 150$ GeV and in a single inclusive jet multiplicity bin by combining the 2- and 3-jet categories. The typical AUC of these discriminants is $\sim 0.84$, with no overtraining observed.

### 1.7.2 Output Variable Transformation

The output of the BDTs introduced in the previous section delivers a fine-binned MVA variable maximising the separation of signal from backgrounds. To optimise the sensitivity of the statistical analysis, the MVA distributions are rebinned such that low BDT scores are still indicative of a background-like event while large values are signal-like. This rebinning is performed with attention given the statistical uncertainty in each bin and the final sensitivity of the discriminant score. The combined analysis applies the so-called *Transformation D* algorithm. The technique relies on a per bin score $Z$ defined as

$$Z = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b}, \tag{1.1}$$

where $N_s$ ($N_b$) is the total number of signal (background) events, $n_s$ ($n_b$) the number of signal (background) events in a specific bin, and $z_s$ and $z_b$ are tunable parameters indirectly controlling the number of signal- and background-enriched bins desired in the region. For a given choice of $z_s$ and $z_b$, the algorithm starts from the initial binning of the BDTs and successively recombines bins from the higher bin values (right) to the lower values (left). Successive bins of the original distribution are merged until the combined bin reaches a score $Z > 1$, thanks to increases in $n_s$ and $n_b$. Once a combined bin reaches the desired scores, it is removed from consideration and the algorithm re-starts from the highest bin not yet recombined (one bin to the left of the last rebinned one).

The $z_s$ and $z_b$ parameters are manually tuned for each analysis regime and lepton channel, giving signal regions with a final amount of BDT bins varying from 4 to 15, as displayed in the postfit plots of Appendix A.6. An additional protection is added to avoid bins with too few data or MC statistics, by requiring at least 3 signal + background events per bin after transformation. The specific tunes of the parameters for the different regime of the combined analysis are presented in Table 1.12.

| | $75 < p_T^V < 150$ GeV | $150 < p_T^V < 250$ GeV | $250 < p_T^V < 400$ GeV | $400 < p_T^V < 600$ GeV | $p_T^V > 600$ GeV |
|---|---|---|---|---|---|
| $VH(H \to b\bar{b})$ | $z_s = 10,\ z_b = 5$ | | $z_s = 5,\ z_b = 3$ | | $z_s = \begin{cases} 3 & \text{for 0L \& 1L} \\ 2 & \text{for 2L} \end{cases}, z_b = 2$ |
| $VH(H \to c\bar{c})$ | $\begin{cases} TT\text{: } z_s = 5, z_b = 3 \\ \text{Else: } z_s = 10,\ z_b = 5 \end{cases}$ | $\begin{cases} \text{0L/1L} & \begin{cases} TT\text{: } z_s = 5, z_b = 3 \\ \text{Else: } z_s = 10,\ z_b = 5 \end{cases} \\ \text{2L} & \begin{cases} TT\text{: } z_s = 2,\ z_b = 2 \\ LT/XT\text{: } z_s = 5,\ z_b = 5 \\ \text{Else: } z_s = 10,\ z_b = 5 \end{cases} \end{cases}$ | $\begin{cases} TT\text{: } z_s = 2, z_b = 2 \\ LT/XT\text{: } z_s = 5,\ z_b = 3 \\ \text{Else: } z_s = 10,\ z_b = 5 \end{cases}$ | | |

Table 1.12: The optimised tune of the $z_s$ and $z_b$ parameter to rebin the MVAs with the *Transformation D* algorithm in different phase spaces of the combined analysis. $XT$ is the 2 $c$-tagged region.

## 1.8 Experimental Uncertainties

While efforts are made to correctly similate the collection and reconstruction of information with the ATLAS detector, inaccuracies permeate this procedure and must be taken into account in the statistical analysis of Section 1.10. Several types of experimental uncertainties are considered in the combined analysis, to cover systematics effect due to the detector performance, the reconstruction of objects such as leptons and jets, and the effects of flavour tagging. Table 1.13 summarises the various contributions that are sources of uncertainty, which are detailed in this section.

| Systematic uncertainty name | Description | Regime |
|---|---|---|
| | Luminosity & Pile-up | |
| LUMI_2015_2018 | Uncertainty on total integrated luminosity | All |
| PRW_DATASF | Uncertainty on pile-up modelling | All |
| | $E_T^{\text{miss}}$and $\mathbf{E}_{\text{T,trk}}^{\text{miss}}$ | |
| MET_SoftTrk_ResoPara(Perp) | Soft term longitudinal (transverse) resolution uncertainty | All |
| MET_SoftTrk_Scale | Soft term scale uncertainty | All |
| MET_JetTrk_Scale | $\mathbf{E}_{\text{T,trk}}^{\text{miss}}$ scale uncertainty | All |
| METTrig{Stat,Top,Z,Sumpt} | Trigger efficiency uncertainty | Resolved |
| | Electrons | |
| EL_EFF_Trigger_TOTAL | Trigger efficiency uncertainty | All |
| EL_EFF_Reco_TOTAL | Reconstruction efficiency uncertainty | All |
| EL_EFF_ID_TOTAL | Identification (ID) efficiency uncertainty | All |
| EL_EFF_Iso_TOTAL | Isolation efficiency uncertainty | All |
| EG_SCALE_ALL | Energy scale uncertainty | all |
| EG_RESOLUTION_ALL | Energy resolution uncertainty | All |
| | Muons | |
| MUON_EFF_RECO_{STAT,SYS} | Reconstruction and ID efficiency uncertainty for muons with $p_T > 15$ GeV | All |
| MUON_EFF_RECO_{STAT,SYS}_LOWPT | Reconstruction and ID efficiency uncertainty for muons with $p_T \leq 15$ GeV | All |
| MUON_EFF_ISO_{STAT,SYS} | Isolation efficiency uncertainty | All |
| MUON_EFF_TTVA_{STAT,SYS} | Track-to-vertex association efficiency uncertainty | All |
| MUON_SCALE | Momentum scale uncertainty | All |
| MUON_SAGITTA_RHO(RESBIAS) | Momentum scale uncertainty to cover charge-dependent local misalignment effects | All |
| MUON_ID(MS) | Momentum resolution uncertainty of the inner detector (muon spectrometer) | All |
| MUON_EFF_Trig{Stat,Sys}Uncertainty | Trigger efficiency uncertainty | All |
| | Taus | |
| TAUS_TRUEHADTAU_EFF_RECO_TOTAL | Reconstruction efficiency | All |
| TAUS_TRUEHADTAU_EFF_RNNID_* | RNN ID efficiency | All |
| TAUS_TRUEHADTAU_SME_TES_* | In-Situ tau energy scale correction | All |
| TAUS_TRUEELECTRON_EFF_ELEBDT_* | Electron Veto efficiency SF | All |
| | Small-R jets | |
| JET_CR_BJES_Response | Energy scale uncertainties for $b$-jets | All |
| JET_CR_EffectiveNP_Detector{1-2} | Energy scale uncertainties due to in-situ calibration | All |
| JET_CR_EffectiveNP_Mixed{1-3} | Energy scale uncertainties due to in-situ calibration | All |
| JET_CR_EffectiveNP_Modelling{1-4} | Energy scale uncertainties due to in-situ calibration | All |
| JET_CR_EffectiveNP_Statistical{1-6} | Energy scale uncertainties due to in-situ calibration | All |
| JET_CR_EtaIntercal_Modelling | Energy scale uncertainties to cover $\eta$-intercalibration non-closure | All |
| JET_CR_EtaIntercal_NonClosure_highE | Energy scale uncertainties to cover $\eta$-intercalibration non-closure | All |
| JET_CR_EtaIntercal_NonClosure_negEta | Energy scale uncertainties to cover $\eta$-intercalibration non-closure | All |
| JET_CR_EtaIntercal_NonClosure_posEta | Energy scale uncertainties to cover $\eta$-intercalibration non-closure | All |
| JET_CR_EtaIntercal_TotalStat | Energy scale uncertainties to cover $\eta$-intercalibration non-closure | All |
| JET_CR_Flav_Comp(Flavor_Response) | Energy scale uncertainty related to flavour composition (response) | All |
| JET_CR_PunchTroughMC16 | Energy scale uncertainty for 'punch-through' | All |
| JET_CR_SingleParticle_HighPt | Energy scale uncertainty for the behavior of high-$p_T$ single hadrons | All |
| JET_CR_JER_DataVsMC | Energy resolution total uncertainty | All |
| JET_CR_JER_EffectiveNP_{1-6,7restTerm} | Energy resolution total uncertainties | All |
| JET_JvtEfficiency | JVT efficiency uncertainty | All |
| JET_PU_{OffsetMu(NPV),PtTerm,RhoTopology} | Energy scale uncertainties due to pile-up effects | All |
| | Large-R jets | |
| FJ_JMSJES_Baseline_Kin | Energy and mass scale uncertainty due to basic data-simulation differences | Boosted |
| FJ_JMSJES_Modelling_Kin | Energy and mass scale uncertainty due to simulation differences | Boosted |
| FJ_JMSJES_Tracking_Kin | Energy and mass scale uncertainty on reference tracks | Boosted |
| FJ_JMSJES_TotalStat_Kin | Energy and mass scale uncertainty from stat. unc. on the measurement | Boosted |
| FJ_JER | Energy resolution uncertainty | Boosted |
| FJ_JMR | Mass resolution uncertainty | Boosted |
| | Flavour tagging: PFlow jets | |
| FT_EFF_PFlow_Eigen_B_{0-44} | Tagging efficiency uncertainties for $b$-jets | Resolved |
| FT_EFF_PFlow_Eigen_C_{0-19} | Tagging efficiency uncertainties for $c$-jets | Resolved |
| FT_EFF_PFlow_Eigen_Light_{0-19} | Tagging efficiency uncertainties for light-jets | Resolved |
| FT_EFF_PFlow_extrapolation | Tagging efficiency uncertainty for high-$p_T$ jets | Resolved |
| | $b$-tagging: VR track jets | |
| FT_EFF_VR_Eigen_B_{0-4} | $b$-tagging efficiency uncertainties for $b$-jets | Boosted |
| FT_EFF_VR_Eigen_C_{0-3} | $b$-tagging efficiency uncertainties for $c$-jets | Boosted |
| FT_EFF_VR_Eigen_Light_{0-3} | $b$-tagging efficiency uncertainties for light-jets | Boosted |
| FT_EFF_VR_extrapolation | $b$-tagging efficiency uncertainty for high-$p_T$ jets | Boosted |

Table 1.13: Summary of all experimental systematic uncertainties.

**Luminosity & Pile-up:** The measured Run 2 luminosity for ATLAS is 140 fb$^{-1}$ with an uncertainty of 0.83% [32]. The measurement is performed by $x - y$ beam separation scans and is combined with information from dedicated lumonisity-sensitive detectors. The pile-up uncertainty for simulated events is obtained by varying the data rescaling factor of the nominal average pile-up $\langle\mu\rangle$. These factor is introduced to the observation that MC-simulated samples match data at a higher $\mu$ than used in their simulation. This rescaling factor is threfore used to reweight the data, matching a simulated-$\mu = 1.0$ to a data-$\mu = 1.09$, a rescaling summarised as 1.0/1.09. A $1\sigma$ uncertainty on the average pile-up is measured by varying the factor from 1.0/1.0 to 1.0/1.18.

**Triggers** Uncertainties on the trigger efficiencies are derived for the electron, muon, and $E_T^{\text{miss}}$ triggers. Statistical and systematics effects are combined for the electron trigger uncertainty, while they are considered separately for the muon triggers. Scale factors for the $E_T^{\text{miss}}$ trigger efficiency are derived on $W$+jets events, taking into acount the statistics of the dataset, assessing systematics effects by deriving Scale Factors (SF)s with alternative top and $Z$+jets samples, and a final uncertainty modelling the efficiency dependency on the scalar sum of all final state jets.

**Leptons & $E_T^{\text{miss}}$ Reconstruction** Leptons and $E_T^{\text{miss}}$ reconstructions are calibrated in dedicated analyses, with a reduced set of uncertainties propagated to the combined $VH(H \to b\bar{b}/c\bar{c})$. These consists of:

- $E_T^{\text{miss}}$: SFs factors are included to account for the direction of the $E_T^{\text{miss}}$ as well as the soft term contributions.
- Electrons: uncertainties on the reconstructed values, the identification efficiency, isolation efficiency, and the energy scale and resolution are included. These are derived by comparing in data and simulations kinematic distributions in $Z \to e^+e^-$, $W \to e\nu$, and $J/\psi \to e^+e^-$ events [59].
- Muons: uncertainties on the reconstruction and identification efficiencies for muons with $p_T > 15$ GeV and $p_T < 15$ are included separately, using respectively samples of $Z \to \mu^+\mu^-$ and $J/\psi \to \mu^+\mu^-$ [61]. Additional, uncertainties on the isolation efficiency, track-to-vertex association efficiency, momentum scale and resolution as well as charge-dependent misalignement effects are considered.
- Taus: hadronically decaying $\tau$-leptons[23] uncertainties on the reconstruction and RNN-

[23] About 65% of $\tau$ decays are hadronic.

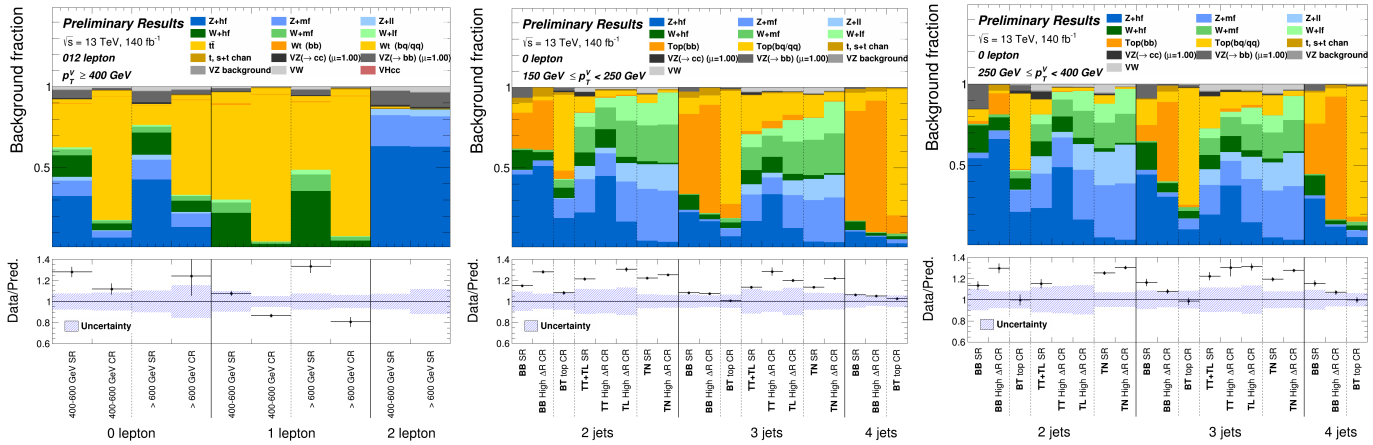
based identification efficiencies as well as the electron veto efficiencies are derived from samples of $Z \to \tau^+\tau^-$ and top-quark decays to taus [62, 71, 72].

**Jets** Jets are calibrated in dedicated analyses, of which two reduced sets of uncertainties are propagated to the combined $VH(H \to b\bar{b}/c\bar{c})$ for small- and large-$R$ jets. For the small-$R$ jets, these uncertainties cover *in-situ* analyses, $\eta$-intercalibration, flavour composition, punch-through jets, high-$p_T$ hadrons, and pile-up effects as well as the jet energy scale and resolution measured in data [73, 74]. The reduced set is derived by a principal component analysis to preserve the largest correlations in certain regions of jet kinematics. Large-$R$ jets uncertainties for the energy scale and resolution are also estimated from data [75]. An uncertainty covering the calibration discrepancy between data and MC-simulations is also included.

**Flavour Tagging** A dedicated calibration is performed to derive flavour tagging scale factors in the resolved regime, as described in Section 1.5, while the common ATLAS uncertainties are used for the boosted regime, as described in **??**. These flavour tagging calibration SFs are derived by combining data-MC efficiency modelling SFs and MC-MC SFs to account for variations to parton showering and hadronisation. These scale factors are smoothed using a local polynomial kernel estimator to avoid distortions in the kinematic variables [76]. For each jet flavour, there is one uncertainty per $p_T$ bin in the calibration. A $\tau$-jet uncertainty is derived by copying the $c$-jet values and decorrelating them. A principal component analysis is deployed to reduce the large set of systematics uncertainties to 45 (5) for $b$-jet, 20 (4)for $c$-jet, and 20 (4) for light-jets in the resolved (boosted) regime. Additional uncertainties are added to model to extrapolate the performance to high-$p_T$ jets. Truth tagging uncertainties are expected to be covered by these flavour tagging uncertainties, so no dedicated uncertainties are considered.

## 1.9 Signals & Backgrounds Modelling

Similarly to the experimental process, the simulations of the signals and backgrounds cannot entirely be accurate and mis-modelling are to be expected in the derived samples. These inaccuracies must be taken into account in the fit to avoid introducing bias. The modelling of the signals and backgrounds in the $VH(H \to b\bar{b}/c\bar{c})$ combined analysis is discussed in this section. The composition of the different processes changes depending on the lepton channel and the analysis category, as shown in Figure 1.21. The $V$+jets backgrounds are the dominant ones in the signal regions of the 0-lepton and 2-lepton channels, while the top processes contribute more



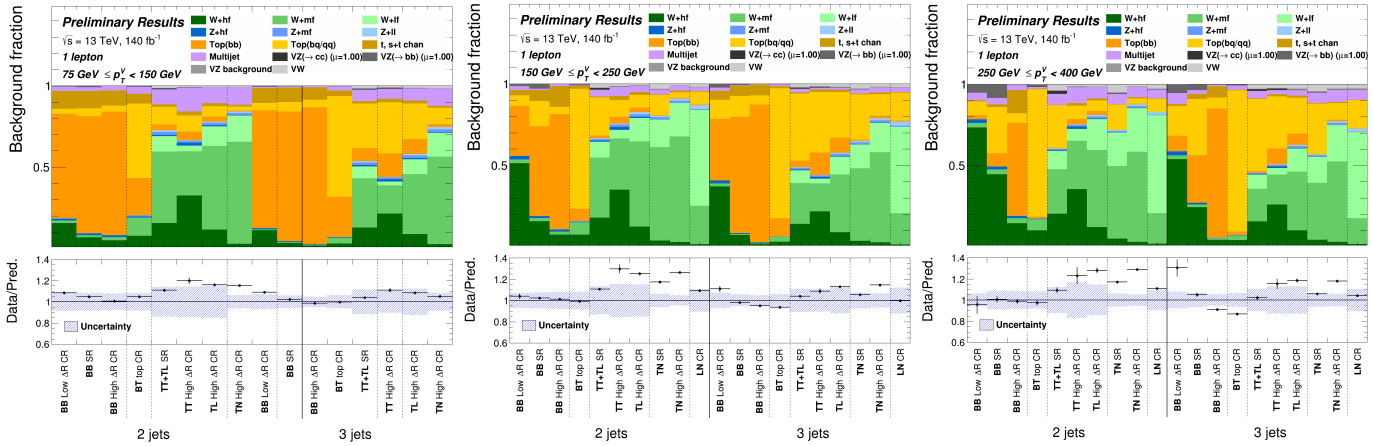

(a) Boosted regime ($\geq 400$ GeV).




(b) 0L, $p_T^V \in [150, 250]$ GeV.




(c) 0L, $p_T^V \in [250, 400]$ GeV.



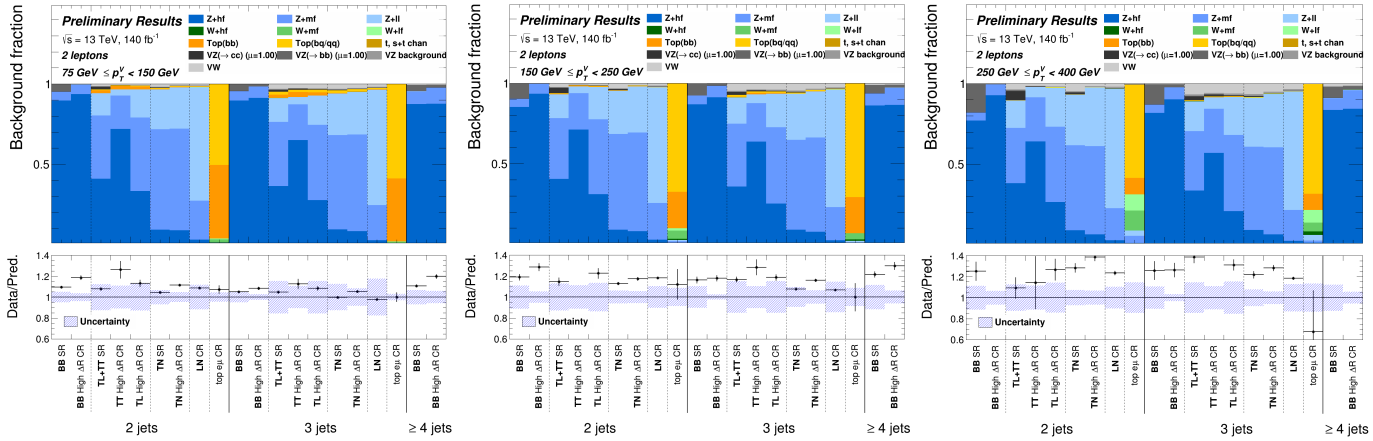

(d) 1L, $p_T^V \in [75, 150]$ GeV.




(e) 1L, $p_T^V \in [150, 250]$ GeV.




(f) 1L, $p_T^V \in [250, 400]$ GeV.




(g) 2L, $p_T^V \in [75, 150]$ GeV.




(h) 2L, $p_T^V \in [150, 250]$ GeV.




(i) 2L, $p_T^V \in [250, 400]$ GeV.

Figure 1.21: The background composition of the different analysis regimes and lepton channels, with the data - Monte Carlo prefit agreement displayed in the bottom panels.

in the 1-lepton channel, and globally at larger jet multiplicities and lower $p_T^V$. Due to flavour tagging, $VH(H \to b\bar{b})$ primarily selects the $bb$-component of the background while $VH(H \to c\bar{c})$ has more diverse flavour compositions with the 2 $c$-tag as an intermediate step between the $BB$ and 1 $c$-tag. This translates into increase $V{+}hf$ fractions in $VH(H \to b\bar{b})$ and $V{+}mf$ and $V{+}lf$ in $VH(H \to c\bar{c})$. In summary:

- **0-lepton**: the dominant background is the $Z$+jets with a sizeable $W$+jets component, particularly in $VH(H \to c\bar{c})$ due to large $E_T^{\text{miss}}$ or a miss-identified hadronic $\tau$ and some top backgrounds for the $VH(H \to b\bar{b})$ side particularly. In $VH(H \to b\bar{b})$ there is a significiant top-background contribution, with this process dominating in 3- and 4-jets. In addition, this lepton channel has some diboson contribution.

- **1-lepton**: the top process is dominant for $VH(H \to b\bar{b})$, while for $VH(H \to c\bar{c})$ a sizeable $W$+jets leads followed by the top. There is also a visible multi-jet contribution.

- **2-lepton**: most of the background is made of $Z$+jets, followed by the diboson and some residual top process at low $p_T^V$ for $VH(H \to b\bar{b})$.

The different background contributions to each analysis region require an adequate strategy to constrain their modelling in the fit, as detailed in this section.

### 1.9.1 General Modelling Strategy

The combined analysis adopts some common strategy to model the backgrounds and signals that are described in this section before reviewing specifities adopted for each process. A guideline for the modelling is to treat backgrounds coherently across analysis regimes and correlate uncertainties between the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ sides when possible. The normalisations of the major backgrounds, the $V$+jets and Top, are free to float in the fit, with Floating Normalisation (FN) split by $p_T^V$ and jet multiplicity when the statitics allows. Minor backgrounds are fixed at MC predictions with a normalisation uncertainty. To account for MC-generator modelling uncertainty, comparisons of the nominal samples to alternative samples introduced in section 1.4 and summarised in Table 1.14 are performed. For each process, the uncertainties are split into normalisation, relative acceptance, and shape uncertainties.

| **Sample** | **Nominal Generator** | **Alternative Generators** | **Systematics Effects** |
|---|---|---|---|
| $VH(H \to b\bar{b})$ | POWHEG + PYTHIA 8 | POWHEG + HERWIG 7 | $\mu_R$, $\mu_F$, ISR, FSR, PDF |
| $VH(H \to c\bar{c})$ | POWHEG + PYTHIA 8 | POWHEG + HERWIG 7 | $\mu_R$, $\mu_F$, ISR, FSR, PDF |
| $V$+jets | SHERPA 2.2.11 | MADGRAPH5 FXFX, SHERPA 2.2.1 | $\mu_R$, $\mu_F$, PDF, EW corrections |
| $t\bar{t}$ & single-top | POWHEG+PYTHIA 8 | POWHEG+HERWIG 7, MADGRAPH5+PYTHIA 8 | ISR, FSR, DS/DR (single-top $Wt$) |
| Diboson | SHERPA 2.2.11 | POWHEG+PYTHIA 8, SHERPA 2.2.1 | $\mu_R$, $\mu_F$, $PDF$, EW corrections |

Table 1.14: Summary of nominal and alternative samples in the analysis. Alternative samples include different generator and systematics effects from modification to the nominal setup.

**Normalisation uncertainties** are an overall uncertainty on the yield of a process, computed in and applied to all regions. These uncertainties are applied when the yield of a background to derive its normalisation from data, e.g., for the diboson and single-top $s$ processes.

**Acceptance uncertainties:** relative acceptance uncertainties for each process cover possible changes in the distribution of events of the specific process across the different regions of the analysis phase space. They account for migration of events between these regions and are assessed by measuring the change of ratio of events between regions when switching to differently generated samples (indexed by $i$ here). The priors on these uncertainties are calculated with the double ratio of Equation 1.2:

$$\text{Acceptance Unc}_i = \frac{\text{Acceptance}[\text{Cat.}^B(\text{alternative}_i\ \text{MC})]}{\text{Acceptance}[\text{Cat.}^A(\text{alternative}_i\ \text{MC})]} \Bigg/ \frac{\text{Acceptance}[\text{Cat.}^B(\text{nominal MC})]}{\text{Acceptance}[\text{Cat.}^A(\text{nominal MC})]}, \tag{1.2}$$

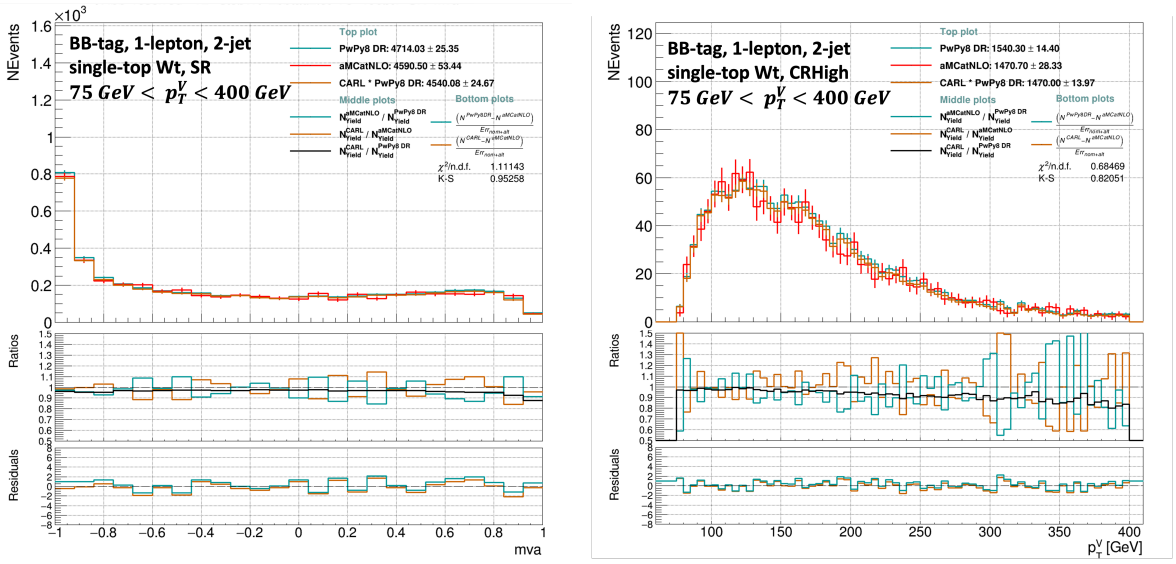
where category $A$ ($\text{Cat.}^A$) is the region with the highest purity in the studied process, and $B$ ($\text{Cat.}^B$) is the region extrapolated to. If several alternative generators are used ($i > 1$), their respective double ratios are summed in quadrature:

$$\text{Total Acceptance Unc} = \sqrt{\sum_i (\text{Acceptance Unc}_i)^2}.$$

If the extrapolation is across several regions $A$, $B$, $C$ ordered by decreasing purity, the acceptance ratio is decomposed into two extrapolations: a first one from $A \to B + C$ with an additional $B \to C$ uncertainty. Due to their similar kinematic definition, for the purpose of acceptance uncertainties between distinct analysis regions in the resolved regime, the signal and Top $BT$ control regions are considered jointly. The acceptance uncertainties between these two regions are modelled by the flavour tagging uncertianties.

**Shape uncertainties:** the BDTs, $m_{bb}$, $m_{cc}$, and $p_T^V$ shapes of the processes in the different regions are given some flexibility in the fit by introducing shape uncertainties derived from a comparison of the nominal to the alternative samples. The combined analysis introduces the novel Calibrated Likelihood Ratio Estimator (CARL) technique to derive a reweighted shape uncertainty using a neural network [77]. A Deep Neural Network (DNN) is trained to discriminate nominal events from alternative ones, with the process repeated for each alternative sample. The output of the CARL network is a score representing the probability for an event to belong to the alternative sample. This is used to reweight the nominal distribution into the alternative

distribution, similarly to the process of truth tagging. The advantage of this technique is that the reweighted nominal distributions benefit from a much larger statistics than the alternative ones, thus smoothing out bin fluctuations and reducing the MC statistics uncertainties. Examples of such derived CARL shape uncertainties modelling the PS with MADGRAPH5_AMC@NLO for the single-top $Wt$ process in 1-lepton are presented in Figure 1.22. Additional shape uncertainties from Electroweak (EW) corrections, QCD scales, $V$+jets and diboson $p_T^V$ modelling with SHERPA 2.2.1, parton shower alternative for the signal samples, and uncertainties for the single-top $Wt$ DS / DR shapes are directly derived by comparing samples.




(a) SR BDT distribution.

(b) High $\Delta R$ CR $m_{bb}$ distribution.

Figure 1.22: CARL closure plots, between the nominal POWHEGPYTHIA8 (*PwPy8*, with the DR scheme) and the alternative MADGRAPH5_AMC@NLO (*aMCatNLO*), for the single-top $Wt$ production in $VH(H \to b\bar{b})$, 1-lepton, 75 GeV $< p_T^V <$ 400 GeV, and 2 jets. The CARL interpolation (orange) of the nominal (blue) into the alternative (red) is smoother and with lower MC-stats. uncertainty. The top plots show the distributions, middle plots the ratios, and bottom plots the residuals.

An overview of the signals and backgrounds modelling systematics considered is presented in Figure 1.15, with the full details listed in Appendix A.5. All uncertainties presented here are further processed before entering the fit. To remove large statistical fluctuations potentially present in shape systematics, these shapes are smoothed by iteratively rebinning the distribution until the statistical uncertainty in each merged bin of the nominal distribution is smaller than 5%. When a systematics has a negligible impact on the distributions entering the fit, it is pruned away to ease convergence and reduce the fit complexity. This is applied to systematics causing a normalisation effect smaller than 0.5% or when both the up- and down-variations have the same sign. Shape uncertainties are pruned if no bin in the distribution has a deviation above 0.5% after the overall normalisation, or if only one of the up- or down-variation is non-zero. For very

| Uncertainties | Resolved | Boosted |
|---|---|---|
| **Signal** | | |
| $qqWH$ / $qqZH$ / $ggZH$ normalisations / acceptance | Values from previous analyses [28, 29, 26] | |
| $H \to bb$ Branching Ratio | 1.61% | |
| $H \to cc$ Branching Ratio | From +5.53% to -1.99% | |
| **$Z$+jets** | | |
| $Z$+$hf$ normalisations | Floating | |
| $Z$+$mf$ normalisation | Floating | 35% |
| $Z$+$lf$ normalisation | Floating | 35% |
| $Z$+$hf$ flavour composition ratios | 8% - 12% | 6% - 9% |
| $Z$+$mf$ flavour composition ratios | 4% - 10% | 6% - 9% |
| High-$\Delta R$ CR-SR ratios | 5% - 30% | - |
| Top CR-SR extrapolation ratios | - | 15% - 25% |
| 2L to 0L acceptance ratios | 2% - 10% | 3% |
| $p_T^V$ extrapolation ratios | - | 15% |
| **$W$+jets** | | |
| $W$+$hf$ normalisations | Floating | |
| $W$+$mf$ normalisation | Floating | 36% |
| $W$+$lf$ normalisation | Floating | 38% |
| $W$+$hf$ flavour composition ratios | 4% - 25% | 11% |
| $W$+$mf$ flavour composition ratios | 14% - 29% | 9% - 15% |
| $W$+$lf$ flavour composition ratios | 9% | - |
| High / Low $\Delta R$ CR-SR extrapolation ratios | 2% - 63% | - |
| Top CR-SR extrapolation ratios | - | 16% - 27% |
| 1L to 0L acceptance ratios | 3% - 30% | 20% |
| $p_T^V$ extrapolation ratios | - | 3% |
| $N_{\text{jet}}$ extrapolation ratios | 12% - 20% | - |
| **Top ($t\bar{t}$ + single-top $Wt$) 0L & 1L resolved** | | |
| Top($bb$) normalisations | Floating | - |
| Top($bq/qq$) normalisations | Floating | - |
| Flavour acceptance ratios | 5% - 10% | - |
| 1L to 0L acceptance ratios | 2% - 8% | - |
| High / Low $\Delta R$ CR-SR extrapolation ratios | 2% - 10% | - |
| $Wt$ / $t\bar{t}$ ratios | 12% - 48% | - |
| **Top ($t\bar{t}$ + single-top $Wt$) 2L resolved** | | |
| Normalisations in $VH(H \to c\bar{c})$ | Floating | - |
| Normalisation in $VH(H \to b\bar{b})$ | 0.08% | - |
| **Single-top ($t$-channel) 0L & 1L resolved** | | |
| Normalisations $s$ - $t$ | 4.6% - 17% | - |
| High / Low $\Delta R$ CR-SR extrapolation ratios | 3% - 17% | - |
| $p_T^V$ extrapolation ratios | 7% - 15% | - |
| $N_{\text{jet}}$ acceptance ratios | 15% | - |
| 1L to 0L acceptance ratio | 6% | - |
| **$t\bar{t}$ and single-top boosted** | | |
| $t\bar{t}$ normalisations | - | Floating |
| single-top $s$, $t$, $Wt$, normalisations | - | 4.6% - 10% - 25% |
| $t\bar{t}$ 1L to 0L acceptance ratios | - | 6% - 20% |
| $t\bar{t}$ Top CR-SR acceptance ratios | - | 10% |
| $Wt$ $p_T^V$ extrapolation ratio | - | 20% |
| $Wt$ 1L to 0L acceptance ratios | - | 20% - 40% |
| **Diboson** | | |
| $WW$ / $ZZ$ / $WZ$ normalisations | 16% / 17% / 19% | 16% / 17% / 27% |
| $ggVV$ normalisation | 30% | |
| Lepton channel acceptance | 2% - 23% | 7% |
| $N_{\text{jet}}$ acceptance | 10% - 30% | - |
| $p_T^V$ acceptance | 3% - 16% | 8% - 40% |
| SR / CR acceptance | 6% - 16% | - |
| STXS binning acceptance | - | 1.2% - 42.2% |
| **Multi-jet (1L)** | | |
| Normalisations | 20% - 100% | - |

Table 1.15: Summary of the modelling systematic uncertainties in the analysis. The values given refer to the size of the uncertainty affecting the yield of each background. Uncertainties in the shapes of the distributions are not shown but taken into account for all backgrounds.

small background processes, both shape and normalisation uncertainties are pruned: if this is a signal-sensitive region - if the signal yield is $> 2\%$ of the total in the region -, the uncertainties are pruned if the process is $\leq 2\%$ of the signal, while in non-signal sensitive region the process must be $\leq 0.5\%$ of the total background. The rest of this section goes into the details of the modelling, highlighting some specificities and subtelties related to each process.

### 1.9.2 Signal Modelling

The three main signal productions $qq \to WH$, $q\bar{q} \to ZH$, and $gg \to ZH$ are modelled separately, with uncertainties addressing the production and the decay mode of the Higgs into $b\bar{b}$ or $c\bar{c}$. The goal of the analysis is to measure the fiducial cross-sections of the $VH(H \to b\bar{b})$ and the signal strength of the $VH(H \to c\bar{c})$. This first objective is approached with the adoption of the Simplified Template Cross-Section (STXS) in the reduced scheme of stage 1.2 [78, 79], depicted in Figure 1.23. The bins are defined in successive regions of transverse momentum of the vector boson $p_T^V$, from truth information in the simulated samples, and the number of additional jets $N_{\text{jet}}$ in the event, at 0 or more than 1 additional jet.

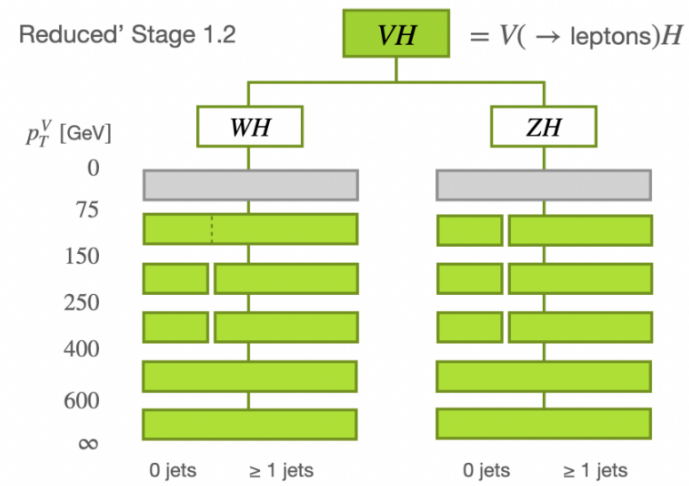



Figure 1.23: The Standard Template Cross-Section scheme in the reduced stage 1.2 used for the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

The signal samples are finely binned following the STXS prescription, with 5 $p_T^V$ bins for the $ZH$ covering $[75, 150[$ GeV, $[150, 250[$ GeV, $[250, 400[$ GeV, $[400, 600[$ GeV, and $\geq 600$ GeV. The first three bins, corresponding to the resolved regime, are further split in $N_{\text{jet}}$ with 0 additional jet or $\geq 1$ jet, for a total of 8 different Parameter Of Interest (POI)s measured in $ZH$. For $WH$, the binning is similar to $ZH$ but there is no $N_{\text{jet}}$ splitting of the $[75, 150[$ GeV bin, giving a total number of 7 POIs for $WH$. The full STXS categorisation is used for $VH(H \to b\bar{b})$ and also for the $VH(H \to c\bar{c})$, to enable correlation of the $VH$ uncertainties. For the $VH(H \to c\bar{c})$ however,

the templates are merged and only one POI is extracted: the global signal strength.

The signal is coherently modelled across the resolved and boosted regimes and targeted final state $VH(H \to b\bar{b})$ or $VH(H \to c\bar{c})$. Several uncertainties are implemented to model the $VH$ production of the $H \to b\bar{b}/c\bar{c}$ decay. These uncertainties include:

- *QCD scale uncertainties:* obtained by varying the renormalisation and factorisation scales $\mu_R$ and $\mu_F$. These variations are the most impactful in the theoretical prediction of the $VH$ production cross-sections. They are considered as shape uncertainties, implemented to cover modifications to the inclusive cross-sections and to parametise possible migrations across $p_T^V$ and additional jet multiplicity bins, following Ref [80]. The quark- and gluon-initiated signal processes have cross-sections modifications parametrised separately.

- *PDF + $\alpha_s$ uncertainties*: alternative parton distributions from the PDF4LHC15_30 modifying the $VH$ cross-sections in STXS bins are considered [81]. The $VH$ cross-sections in each STXS bin is systematicaly modified by comparing the nominal PDF to 30 alternatives. Furthermore, the $\alpha_s$ estimated at the $Z$-mass is varied for the nominal setup within its uncertainties. These uncertainties are separately calculated for $qq$-initiaded $WH$ and $ZH$, and for $gg$-initiated $ZH$. Shape effects on the resolved regime $p_T^V$ distributiond are considered, while variations to the boosted large-$R$ mass $m_J$ and the invariant mass $m_{bb}$ or $m_{cc}$ are negligible.

- *EW corrections*: NLO electroweak corrections from NNLO EW effects are considered with uncertainties modifying the $p_T^V$ distributions.

- *Branching ratio*: a theoretical uncertainty of 1.61% on the $H \to b\bar{b}$ branching ratio and an uncertainty covering the range from -1.99% to +5.53% for the $H \to c\bar{c}$ branching ratio are used [82]. The $ZH$ ($WH$) cross-sections cover 96.52% to 104.11% (97.95% to 101.98%) of their values thanks to additional uncertainties.

- *Parton shower and underlying event uncertainties*: variations to the PS and UE can affect the properties of the $H \to b\bar{b}/c\bar{c}$ decays. Uncertainties are introduced to model these effects on the signal acceptance. In the resolved regime, the effects of an alternative PS model on the signal acceptance are evaluated on truth information in a similar phase space to the analysis selection. Acceptance uncertainties are derived by comparing the signal acceptance in the analysis categories between the nominal PYTHIA 8 and the alternative HERWIG 7. Additional sub-leading acceptance uncertainties are evaluated by

modifying the PYTHIA AZNLO tune. Differences in $p_T^V$ and $m_{bb}$ ($m_{cc}$) between PYTHIA and HERWIG are also considered, and the shape difference in the MVA distribution when adopting POWHEG+HERWIG 7 is used in the final stage of the analysis. In the boosted regime, the same strategy with the same PS models is employed but the full detector response and event reconstruction are simulated, with uncertainties covering modifications to the $m_J$ distributions.

### 1.9.3 $V$+jets Modelling

The $V$+jets processes are modelled separately for $Z$+jets and $W$+jets, depending on the flavour of the reconstructed vector boson. Their modelling nonetheless share many simularities.

#### $Z$+jets

This background is dominant in the 0L and 2L channels, and limited in 1L. The background is split into different components from the flavour compositions of jets selected to form the Higgs candidate, grouping compositions with similar kinematic performance as:

- *$Z$+ heavy flavours ($Z$+hf)*: $Z + bb$ and $Z + cc$.
- *$Z$+ mixed flavours ($Z$+mf)*: $Z + bc$, $Z + bl$, and $Z + cl$.
- *$Z$+ light flavours ($Z$+lf)*: $Z + l$.

Each grouping has its own free-floating normalisations in 0L and 2L, with $Z$+$hf$ dominant in $VH(H \to b\bar{b})$ and the other two components significant in $VH(H \to c\bar{c})$. These FNs are decorrelated in $p_T^V$ and jet multiplicities $N_{\text{jet}}$[24]. The modelling of $Z$+jets includes several types of acceptance uncertainties that are applied only in 0L and 2L:

- *Channel extrapolation 2L → 0L uncertainties*: for the $Z$+$hf$, $Z$+$mf$, and $Z$+$lf$ respectively.
- *Flavour composition uncertainties*: accounting for the variation on the yields of different flavour in the combinations with the double ratio of Equation 1.2. These include a ratio of $cc$ to $bb$ for $Z$+$hf$, and of $bc$ and $bl$ to $cl$ for $Z$+$mf$. They are decorrelated in $p_T^V$ and jet multiplicity $N_{\text{jet}}$ bins and cover the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ sides of the resolved regime.
- *Region extrapolation uncertainties*: are included to model the acceptance of different regions, and derived with the double ratio Equation 1.2 from a high purity region to a lower purity as:

[24] Except for the 2L with 75 GeV $< p_T^V <$ 150, where the $VH(H \to b\bar{b})$ 4-jet is merged with 3-jet: to solve this, $VH(H \to b\bar{b})$ has an extra $Z$+$hf$ FN for 3p-jet.

– $Z$+*hf* and $Z$+*mf*: constrained mostly in the CRHigh and applied to the SR.

– $Z$+*lf*: constrained mostly in 1 *LN*-tagged $V+l$ CR and the SR, thus applied in CRHigh.

The values of the acceptance uncertainties are presented in Table A.3 of Appendix A.5, with a summary mentioned in Table 1.14. In addition, 4 different types of shape uncertainty are considered:

- CARL shape: modelling the difference between SHERPA 2.2.11 and MADGRAPH FXFX, derived for all components and applied in all analysis regions.
- SHERPA 2.2.1 $p_T^V$ shape uncertainties to model the data-MC mis-modelling of $p_T^V$ in SHERPA 2.2.11. The growing data-MC disagreement with higher $p_T^V$ is visible in the plots of Appendix Figure **??**.
- QCD scale shape uncertainties by varying $\mu_R$ and $\mu_F$.
- EW shape variations that are typically quite small.

**Boosted regime:** the modelling strategy adopted is roughly the same as in the resolved regime, with the uncertainties fully detailed in the Appendix Table A.4. The $Z$+*hf* component is left free-floating in 0L and 2L, while the $Z$+*mf* and $Z$+*lf* components both have overall acceptance uncertainties of 35%. The $Z$+*lf* has no other acceptance uncertainty since it is negligible in the boosted regime. Flavour acceptance uncertainties for $Z$+*hf* and $Z$+*mf* are applied in 0L and 2L. They also have channel acceptance uncertainties and SR → Top CR acceptance ratios, both applied in the 0L. Additional $p_T^V$ extrapolation uncertainties from [400, 600] GeV to > 600 GeV is considered in 0L and 2L. Shape uncertainties are derived similarly to the resolved regime.

### *W*+jets

This background is dominant in the 1-lepton channel, with a residual contribution in 0-lepton mostly due to hadronically decaying $\tau$-lepton. It is split equivalenty to the $Z$+jets background as:

- $W$+ *heavy flavours (W+hf)*: $W + bb$ and $W + cc$
- $W$+ *mixed flavours (W+mf)*: $W + bc$, $W + bl$, $W + b\tau$, $W + cl$, and $W + c\tau$.
- $W$+ *light flavours (W+lf)*: $W + l$, $W + l\tau$, $W + \tau\tau$.

Each grouping has its own floating normalisation, with $W$+$hf$ signficant in $VH(H \to b\bar{b})$, while $W$+$mf$ and $W$+$lf$ are more important in $VH(H \to c\bar{c})$. These FNs are decorrelated in $p_T^V$ and jet multiplicities $N_{\text{jet}}$[25]. Acceptance uncertainties, listed in the Appendix Table A.5, are applied in 0L and 1L. They include:

- *Channel extrapolation 1L → 0L uncertainties*: applied in 0L for all components separately.
- *Flavour composition uncertainties*: include a comparison of $cc$ to $bb$ for $W$+$hf$, of [$bc$, $bl$, $c\tau$, $b\tau$] to $cl$ for $W$+$mf$, and of [$l\tau$, $\tau\tau$] to $Wl$ for $W$+$lf$. They are decorrelated in $p_T^V$ and $N_{\text{jet}}$, and cover the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ sides.
- *Region extrapolation uncertainties* are defined differently for the combinations:
  - $W$+$hf$: constrained mostly in the SR and the $BB$-tagged CRLow[26], applied to CRHigh in different $p_T^V$ regions. For $VH(H \to b\bar{b})$ 1L, an extra CRLow → SR is applied.
  - $W$+$mf$: constrained mostly in 2 $c$-tagged CRHigh and applied to SR and CRLow[26].
  - $W$+$lf$: constrained mostly in the SR and the 1 $LN$-tagged $V + l$ CR and applied in CRHigh.
- *Jet multiplicity $N_{jet}$ acceptance*: FNs are left free-floating in $N_{\text{jet}}$ (2-jet and 3-jet). For $VH(H \to b\bar{b})$, the 4-jet category has no dedicated CR and a 3-jet → 4-jet acceptance is applied to $W$+$hf$ (other components are negligible).

In addition, 4 different types of shape uncertainties are considered similarly to the $Z$+jets.

**Boosted regime:** roughly the same modelling strategy is applied, with the uncertainties fully detailed in Appendix Table A.6. The $W$+$hf$ component is left free-floating, while the $W$+$mf$ and $W$+$lf$ components have overall acceptance uncertainties of 36% and 38% respectively. Flavour acceptance uncertainties are considered for $W$+$hf$ from $bb$, and for $W$+$mf$ from $bc$ (components with $\tau$ are negligible). The different components also have channel acceptance uncertainties applied in 0L channel and SR → Top CR acceptance ratios applied in the 0L and 1L channels. Additional $p_T^V$ extrapolation uncertainties from [400, 600] GeV to > 600 GeV are considered in 0L and 1L. Shape uncertainties are derived similarly to the resolved regime.

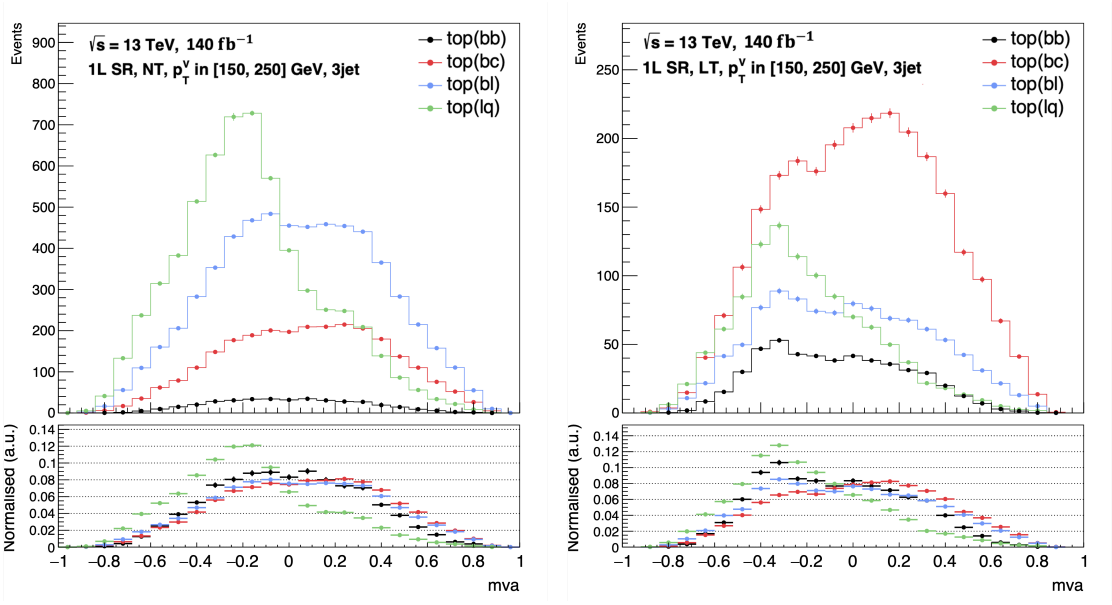
[25]The only exception is the 1L $W$+$lf$ in 75 GeV < $p_T^V$ < 150 GeV, where a normalisation uncertainty of 25% is considered.

[26]The CRLow is considered only in $VH(H \to b\bar{b})$ 1L.

### 1.9.4 Top Modelling

The backgrounds including the decay of a top-quark $t$ are considered here, distinguishing between the $t\bar{t}$ pair-production and the single-top $Wt$ production as well as the single-top $t$- and $s$-channels, by decreasing order of relative importance. The $t\bar{t}$ and single-top $Wt$ are combined into a unified *Top* component[27] in the resolved regime, and the single-top $t$- and $s$-channels are considered seperately. The Top backgrounds in 0L and 1L are estimated from MC and dedicated Top $BT$ control region, with the 2L case described later in this section. In the resolved regime, the Top is grouped into different components based on three truth flavour categories:

- Top($bb$): which is mostly found in the $VH(H \to b\bar{b})$ phase space of the signal regions, and the High $\Delta R$ CRs thanks to the large initial angle between the emitted top-quark, passed over to the two $b$-quarks.
- Top($bq$): combining top($bc$) and top($bl$). It is mostly in the $VH(H \to c\bar{c})$ phase space and is well-selected by the $BT$-tagged Top CR.
- Top($qq$): combines top($cc$), top($cl$) and top($ll$), where $l$ is a light-jet ($u$, $d$, $s$, or a gluon). Mostly in the $TN$ and $TL$ regions of the $VH(H \to c\bar{c})$.




(a) (b)

Figure 1.24: The non-rebinned MVA distributions of the top background components (direct tagged) in the $VH(H \to c\bar{c})$ signal regions ($TN$-tagged on the left, $TL$-tagged on the right) with 150 GeV $< p_T^V <$ 250 GeV and 3 jets. Top($bb$) in black, top($bc$) in red, top($bl$) in blue, and top($qq$) in green. The bottom panels show the normalised distributions.

[27] Throughout this chapter, Top will refer to the combination of the $t\bar{t}$ & $Wt$ processes.

These grouping are based on the shared kinematics of the components, where the selected jets are either both $b$-jets and thus likely to directly come from the top-decays ($bb$), 1 $b$-jet likely from a top decay and 1 non $b$-jet from a subsequent hadronic $W$-decay or a radiated jet ($bc$ and $bl$, summarised $bq$), or neither directly from the top-decay ($cc$, $cl$, and $ll$, summarised $qq$). The $bc$ and $bl$ are combined into a single top($bq$) component as they indeed share the same kinematics, as illustrated in Figures 1.24 in the signal regions of $VH(H \to c\bar{c})$.This top($bq$) background is particularly significant in the $VH(H \to c\bar{c})$ analysis as it peaks at the signal mass (having a mass $\sim (m_{\text{top}} + m_W)/2 \approx m_H$) and therefore exhibits signal-like properties such as reaching high MVA scores, as shown in Figure 1.24. Due to the small contribution of the top($qq$) component, it is merged with the top($bq$) into a single top($bq/qq$) component, with the different sub-components shapes modelled by flavour composition uncertainties. The rest of this section details the modelling of the top backgrounds in the analysis regimes for 0L and 1L, followed by the single-top $t$- and $s$-channels in resolved, and finally the modelling adopted for the boosted regime.

### The $t\bar{t}$ and $Wt$ Resolved 0L & 1L Modelling

There are three main elements in the top background modelling scheme in the 0L and 1L resolved regime: floating normalisation, acceptance uncertainties, and shape uncertainties. On the first point, free-floating normalisations are applied for the top($bb$) and the top($bq/qq$) components, constrained respectively by the $BB$-tagged High $\Delta R$ CR and the $BT$-tagged Top CR. These FNs are separated in jet multiplicity $N_{\text{jet}}$ (2-jet, 3-jet, and only for the 0L channel 4-jet) as well as $p_T^V$, for a total of 16 FNs. Concerning the second point, several types of acceptance uncertainties are applied, as summarised in Table 1.14 and detailed in the Appendix Table A.7:

- *Channel extrapolation 1L → 0L uncertainties*: the Top is dominant in 1L, hence the FNs derivation is driven by the 1-lepton channel and applied to the 0-lepton one. This uncertainty is split in $p_T^V$: 2% in [150, 250] GeV and 8% in [250, 400] GeV.
- *Flavour composition uncertainties*: the top($bq/qq$) includes differently shaped sub-components. Uncertainties are derived from the alternative samples with the double ratio Equation 1.2, comparing $bl$ and $qq$ to $bc$ (of 5% and 10% respectively).
- *Region extrapolation uncertainties*: the top($bb$) is dominant in the CRHigh while the top($bq/qq$) leads in the Top $BT$ CR, hence the extrapolations differ for the components. They are all derived from the double ratios with alternative samples.

  - Top($bb$): extrapolation uncertainties are derived from the CRHigh and applied in the SR, the Top CR and the CRLow[26]. Additional uncertainties are applied from the SR to the Top CR and CRLow[26]. All uncertainties are split per $p_T^V$.
  - Top($bq/qq$): the uncertainties are derived from the SR + Top CR + CRLow[26], due to their shared kinematic, and applied to the CRHigh. Additional uncertainties are applied from the SR and Top CR to the CRLow[26]. All uncertainties are split per $p_T^V$.

- *Process acceptance ratios*: in the $t\bar{t}$ and $Wt$ combination, the $t\bar{t}$ dominates and drives the normalisation. Additional acceptance uncertainties are included and applied to the $Wt$ to model differences in the relative contributions of the two processes. These are calculated with a double ratio in the different $p_T^V$ regions, lepton channels, and flavour components. They range from 12% to 48%.

In addition, several different shape uncertainties are considered for the Top backgrounds:

- CARL shapes: modelling the difference between the nominal samples (POWHEG+PYTHIA 8) and the alternative modelling of the parton shower (POWHEG+HERWIG 7) and matrix element (MADGRAPH5_AMC@NLO+PYTHIA 8). These CARL models are trained separately for $t\bar{t}$ and $Wt$ and per lepton channel, inclusively in flavour compositions and $N_{\text{jet}}$. The DR scheme is used as nominal for these training of $Wt$, because the alternative samples use the same $t\bar{t}$ overlap removal scheme.

- A DS-DR shape uncertainty is derived uniquely for $Wt$ to account for possible shape effects from modifications to the overlap removal procedure with $t\bar{t}$. The POWHEG+PYTHIA 8 samples with DS scheme are directly used in the fit as template, thanks to their sufficient statistiscs. This shape uncertainty is unique in the combined analysis: it simultanesouly applies a normalisation uncertainty, to account for the different yields of the DS- and DR-schemes.

- ISR and FSR shape uncertainties are derived by varying the scales $\mu_R$ and $\mu_F$ from the nominal setup. For both, an up- and a down-variation are considered, with the variations being symmetric for the ISR while the down -variation of FSR is smaller than the up-variation.

#### The Single-Top $t$- & $s$-channels in Resolved 0L & 1L Modelling

The single-top $t$- and $s$-channels are almost negligible in the analysis, except in the $VH(H \to b\bar{b})$ resolved at low $p_T^V$, where the $t$-channel reaches a total backgrounds fraction of ∼8% in the 1L 75

GeV $< p_T^V <$ 150 GeV. The importance of single-top $t$ quickly reduces with increasing energy[28]. In 0L and 1L, the single-top $t$- and $s$-channels are only applied cross-sections uncertainties of 17% and 4.6%, respectively. The single-top $t$-channel has several additional acceptance uncertainties derived by double ratio computations with alternative samples to model:

- *channel extrapolation uncertainty*: of 6% from 1L to 0L.
- *Region extrapolations uncertainties*: depends on the $p_T^V$. For $p_T^V <$ 150 GeV, the uncertainty is applied from SR → CRLow+CRHigh, with an additional CRHigh → CRLow uncertainty in 1L. For the higher $p_T^V$ regions, the extrapolations are instead from CRHigh → SR+CRLow[26], with an additional SR → CRLow[26] uncertainty, due to the higher purity of the CRHigh.
- *Jet multiplicity extrapolations*: are considered from the 3-jet to the 2-jet, and from the 2+3-jet to the 4-jet in 0L.
- $p_T^V$ *extrapolation uncertainties*: since the single-top $t$-channel is mostly present in the lowest $p_T^V$ regions, $p_T^V$ extrapolation uncertainties are included from [75, 150] GeV to [150, 400] GeV, with an additional [150, 250] GeV to [250, 400] GeV uncertainty.

In addition, CARL and ISR/FSR shape uncertainties are considered for the single-top $t$-channel in 1L only, as is done for the Top background. Table A.9 of the Appendix details the various single-top uncertainties considered.

### Resolved Regime Top Backgrounds in 2L

For the 2L $VH(H \to b\bar{b})$ resolved, data-driven estimates are used, deriving a template in the Top $e\mu$ region for the Top background with an 0.8% extrapolation uncertainty to the signal region. For $VH(H \to c\bar{c})$, the Top $e\mu$ region is used as a control region with the Top background left free-floating.

### Boosted Regime Top Backgrounds

In the boosted regime, the $t\bar{t}$ benefits from a good Top CR and is not combined with the $Wt$ in the presented results[29]. The modelling in the boosted regime, detailed in the Appendix Tables A.8 and A.10, covers:

[28] Except in the CRHigh region where the ratio stays in the 7%-9% range.
[29] Studies were, at the time of writing, ongoing to also merge these two processes in the boosted regime.

- $t\bar{t}$: 1 FN per $p_T^V$ region for 0L and 1L. In 2L, a 20% normalisation uncertainty is applied. *Channel extrapolation uncertainties* are derived from 1L $\to$ 0L, split per $p_T^V$. Finally, *region extrapolation uncertainties* of 10% are applied in 0L an 1L from the boosted Top CR to the SR.

- Single-top $Wt$-, $t$-, and $s$-channels are not free floated but insead have respectively a 25%, 10%, and 4.6% normalisation uncertainty. The $Wt$ has additional acceptance uncertainties, to cover the lepton channel extrapolation and a $p_T^V$ extrapolation from [400, 600] GeV to $>$ 600 GeV.

In addition, boosted shape uncertainties are considered simularly to what is done in the resolved regime 0L and 1L.

### 1.9.5 Diboson Modelling

The diboson production background consists of the $WW$, $WZ$, and $ZZ$ processes. In $VH(H \to b\bar{b})$, the $ZZ$ primarily contributes to the 2L channel, while $WZ$ with $W$ leptonically and $Z$ hadronically decaying contributes to the 1L. Both equally contribute to 0L. In $VH(H \to c\bar{c})$, the main contributor to 2L is the $WZ$ with $W$ hadronically decaying for a $Z$ leptonically decaying, while in 1L it is the $WW$ process that contributes the most. Again, both contribute almost similarly to 0L. The resolved and boosted acceptance uncertainties are detailed in Table A.11 and table A.12.

In the resolved regime, the diboson processes are a small background in the analysis, so only normalisations uncertainties are used for $ZZ$ (17%), $WW$ (16%), and $WZ$ (19%) for the $qq$-initiated, and $ggVV$ (30%) for the $gg$-initiated. All uncertainties are correlated between $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$. The $VZ(\to b\bar{b})$ and $VZ(\to c\bar{c})$ are considered as signals of the cross-check analysis, and denoted as $VZbb$ and $VZcc$ in the diboson modelling. The rest of the $WW$, $WZ$, and $VZ$ are classified as background components, denoted as $VV$bkg. Acceptance uncertainties, summarised in Table 1.14 and detailed in the Appendix Table A.11. For the signal components, the uncertainties are split between the $ZZ$ and $WZ$ components and include:

- *Channel extrapolation uncertainties*: two sets are considered due to the differences between the two components. One covers the 1L $\to$ 0L for $WZ$bb and $WZcc$, and the other the 2L $\to$ 0L for $ZZbb$ and $ZZcc$. They are split by $N_{\text{jet}}$.

- *Acceptance in jet multiplicity*: are considered from low (2-jet) to high jet-multiplicity. First to

3-jet, with a different value for the low $p_T^V$ region, then from 3-jet to 4-jet inclusively in $p_T^V$ for 0L and 2L. They are decorrelated between the different lepton channels.

- *Region extrapolation uncertainties*: from the SR to the CRHigh and CRLow[26], due to the higher diboson purity of the SR, with an additional SR to CRLow in $VH(H \to b\bar{b})$ 1L. These uncertainties are separated for the different lepton channels.

- $p_T^V$ *extrapolation uncertainties*: the $150 < p_T^V < 250$ GeV region is the purest in signal diboson and is therefore used to extrapolate to the other $p_T^V$ regions, separately for the different lepton channels and $N_{\text{jet}}$.

- *STXS binning acceptance uncertainties*: are included between $N_{\text{jet}}$ and $p_T^V$ regions for all $VZ$ signal diboson processes. They are modelled by QCD scale variations.

For the background components[30] of $WW$, $W_{\text{had}}Z_{\text{lep}}$, $W_{\text{lep}}Z_{\text{had}}$, and $ZZ$, with the "had" or "lep" index specifying the type of decay of each boson, the acceptances uncertainties are similar to those of the signal components and include:

- *Channel extrapolation uncertainties*: two sets covering 1L $\to$ 0L (for $WW$ and $W_{\text{lep}}Z_{\text{had}}$) and 2L $\to$ 0L (for $ZZ$bkg and $W_{\text{had}}Z_{\text{lep}}$) are included due to difference in purity.

- *Acceptance in jet multiplicity*: go from low (2-jet) to high jet-multiplicity. First to 3-jet, with a different value for the low $p_T^V < 150$ GeV region. Then from 3-jet to 4-jet inclusively in $p_T^V$ for 0L and 2L. They are derived separately for the different lepton channels.

- *Region extrapolation uncertainties*: go from the SR to the CRHigh, due to the higher diboson purity of the SR, separately for the different channels.

- $p_T^V$ *extrapolation uncertainties*: all extrapolation go from the $150 < p_T^V < 250$ GeV region to the other $p_T^V$ regions, due to the higher purity in diboson of the medium $p_T^V$ range, separately for the different channels.

In addition, the diboson processes are modelled with different shape uncertainties:

- CARL shape uncertainties comparing the nominal SHERPA 2.2.11 samples to the two alternative samples POWHEG+PYTHIA8 and SHERPA 2.2.1. The former accounts for differences to the matrix-element and parton shower, while the latter accounts for the mis-modelled $p_T^V$ shape. These uncertainties are applied to all regions.

[30] Thus excluding the signal-like $VZbb$ and $VZcc$ described above.

- QCD scale shape uncertainties are included to model changes to the scales $\mu_R$ and $\mu_F$, similarly to the $V$+jets.
- PDF shape uncertainties modelling variation to $\alpha_s$ are considered.
- EW shape uncertainties are considered, similarly to $V$+jets.

**Boosted regime:** is modelled similarly to the resolved regime, with the uncertainties fully detailed in Table A.12 of the Appendix. Small contributions from mis-identified $W$ decays as jets or mis-reconstructed leptons are taken into account. The $ZZ$ and $WZ$ have normalisation uncertainties of 17% and 27% respectively. Acceptances uncertainties considered cover the lepton channel acceptance, $p_T^V$ acceptance, and STXS-like uncertainty convering the $p_T^V$ and $N_{\text{jet}}$ bins, as is done in the resolved regime.

### 1.9.6 Multi-jet Modelling

The multi-jet background is negligible in 0L and 2L and in the boosted regime. In 1L, a data-driven estimate is used from a high purity multi-jet control region obtained by inverting the lepton isolation requirements. Shapes are derived by a template fit on the $m_T^W$ distributions in the multi-jet CRs. The shapes of the multi-jet are extracted to the SRs of the resolved regime, primarily in $VH(H \to c\bar{c})$, with extrapolation and normalisation uncertainties applied. Top and $W$+jets scale factors are applied to the template to account for the non-insignificant contributions of these processes in the mutli-jet CRs.

## 1.10 Statistical Analysis

After collecting the data and various simulated samples, including detector effects, reconstructing the physics objects, and applying the complex events selection and categorisation, the final step in the analysis is to measure the different *Parameters of interest* (POI) with the modelling strategy defined in the previous section. The combined analysis targets several deliverables to be measured, namely:

- $VH(H \to b\bar{b})$:
  - Inclusive signal strength $\mu_{VH_{bb}}$ and significance: 1 POI.
  - Signal strengths for $WH(\to b\bar{b})$ and $ZH(\to b\bar{b})$: 2 POIs.
  - Fiducial STXS measurements in the reduced stage 1.2, described in Section 1.9.2 and Figure 1.23: 15 POIs (8 for $ZH$, 7 for $WH$).

  - Constraints on Yukawa Higgs-bottom coupling modifier $\kappa_b$.

- $VH(H \to c\bar{c})$:
  - The signal strength $\mu_{VH_{cc}}$ and upper limits at the 95% Confidence Level (CL) : 1 POI.
  - Constraints on Yukawa Higgs-charm coupling modifier $\kappa_c$.

- *Combined* $VH(H \to b\bar{b}/c\bar{c})$:
  - Effective field theory interpretation.
  - Limits on the ratio of Yukawa coupling modifiers $\kappa_c/\kappa_c$.

The *signal strength or enhancement factor* $\mu$ is the ratio of the measurement signal yield to the expected yield in the SM, from the process $\sigma_{VH}\times$ branching ratio of the decay targeted.

### 1.10.1 Likelihoof Function Definition

All parameters of interest are estimated by comparing theory-based expactations baked into MC-simulated samples to real collected data in a fit. This fit is performed by maximising the binned-likelihood function in all the analysis regions simultaneously, as a function of the signals strengths and statistical and systematic uncertainties. The full binned-likelihood function is composed of three terms, accounting, respectively, for the number of events per bin $\mathcal{L}_{\text{Events}}$, the impact of systematics $\mathcal{L}_{\text{Systematics}}$, and finally the impact of the limited statistics of the simulated sample $\mathcal{L}_{\text{MC-stats}}$. These are combined into the likelihood function of Equation 1.3:

$$\mathcal{L} = \mathcal{L}_{\text{Events}} \times \mathcal{L}_{\text{Systematics}} \times \mathcal{L}_{\text{MC-stats}}. \tag{1.3}$$

The first part $\mathcal{L}_{\text{Events}}$ is statistically modelled with Poisson distributed ($\mathcal{P}$) probabilities for every bin $i$ in the analaysis, comparing the number of measured data events $N_i$ to the expectations of the signal $s_i$ and backgrounds $b_i$ in simulations. The $\mu$ signals strengths POIs enter this term as parameter modifying the expected signal contributions:

$$\mathcal{L}_{\text{Events}} = \prod_{i\in\text{bins}} \mathcal{P}(N_i \,|\, \mu s_i + b_i) = \prod_{i\in\text{bins}} \frac{(\mu s_i + b_i)^{N_i}}{N_i} e^{-(\mu s_i + b_i)}.$$

For $VH(H \to b\bar{b})$, the several POIs in the STXS measurement sets the signal strengths as a vector $\boldsymbol{\mu}$ with one entry per STXS bin.

The systematics uncertainties are introduced in the fit by the $\mathcal{L}_{\text{Systematics}}$ term as Nuisance Parameter (NP) $\boldsymbol{\theta}$, accounting for possible perturbations to the expected signal and background yields $\{s_i, b_i\} \to \{s_i(\boldsymbol{\theta}), b_i(\boldsymbol{\theta})\}$ in each bin. The NPs are statistically modelled as standard Gaussian $\mathcal{N}(0,1)$ penalties of mean 0 and unit-variance:

$$\mathcal{L}_{\text{Systematics}}(\boldsymbol{\theta}) = \prod_{\theta \in \boldsymbol{\theta}} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}.$$

The nominal value is by convention set at $\theta_0 = 0$, with $\theta = \pm 1$ representing a $\pm 1$ $\sigma$ variation. The effect of each NP is determined in auxiliary measurements, following the prescriptions introduced in the modelling Sections 1.8 and 1.9. For example, if an NP tracking the normalisation of a background with a 10% prior is moved upwards by 1 standard deviation in the fit, the yield of the background is increased by 10%. After the fit, the central values of the NPs can be moved upwards or downwards, with a deviation from the initial central value called a *pull* and defined as

$$\text{pull}_\theta = \frac{\hat{\theta} - \theta_0}{\sigma_{\theta_0}},$$

where the prefit value $\theta_0 = 0$ and $\sigma_{\theta_0} = 1$ for all NPs. The *constraint* indicates the change in certainty on the NP after the fit, estimated by the variance $\sigma_{\hat{\theta}}$ measured in the inverse Hesse matrix at the maximal likelihood point. For the normalisation of the major backgrounds, special unconstrained NPs are included with no likelihood penalty and said to be *free-floating* (FNs). They are free to vary, and determined from data in control regions with an enhanced purity of the processes they normalised. These special NPs and have prefit values $\theta_0$ set at 1.

The final part covers the uncertainties linked to the limited statistics of the MC-samples, statistically modelling $\mathcal{L}_{\text{MC-stats}}$ with $\gamma$-parameters. One such $\gamma_i$ is introduced per bin, with the freedom to modify the expected background yield as $b_i(\boldsymbol{\theta}) \to \gamma_i b_i(\boldsymbol{\theta})$. The $\boldsymbol{\gamma}$ factors are Gaussian distributed with a likelihood function:

$$\mathcal{L}_{\text{MC-stats}}(\boldsymbol{\gamma}) = \prod_{i \in \text{bins}} \mathcal{N}\left(\beta_i \,|\, \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}\right),$$

with $\beta_i = 1/\sigma_{\text{rel}}^2$ introducing the relative statistical uncertainty ($\sigma_{\text{rel}}$) on the expected yield of the sum $b_i$ of backgrounds in bin $i$.

Bringing every components together, the full likelihood function of Equation 1.3 is thus

defined as

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i \in \text{bins}} \mathcal{P}(N_i \,|\, \boldsymbol{\mu} s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}, \boldsymbol{\gamma})) \times \prod_{\theta \in \boldsymbol{\theta}} \mathcal{N}(\theta \,|\, 0, 1) \times \prod_{i \in \text{bins}} \mathcal{N}(\beta_i \,|\, \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}). \quad (1.4)$$

The parameters $\{\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\gamma}\}$ jointly maximising the likelihood are written as $\{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\}$ while those maximising the likelihood conditionned on a fixed value of $\boldsymbol{\mu}$ are written as $\{\hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\boldsymbol{\gamma}}}\}$. These two sets are used to define a profile likelihood ratio $\lambda(\boldsymbol{\mu})$ to test a certain hypothesis about the values of $\boldsymbol{\mu}$ with

$$\lambda(\boldsymbol{\mu}) = \frac{\mathcal{L}\left(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\boldsymbol{\gamma}}}\right)}{\mathcal{L}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\right)}. \quad (1.5)$$

The $\lambda$ ratio is bound to the range [0, 1], with higher values implying a good agreement between the data and the chosen $\boldsymbol{\mu}$ while lower values are signs of disagreements. This pattern permits the construction of a likelihood ratio test statistics $t_{\boldsymbol{\mu}}$, defined as [83]

$$t_{\boldsymbol{\mu}} = \begin{cases} -2 \ln \lambda(\boldsymbol{\mu}) & \hat{\boldsymbol{\mu}} \geq \boldsymbol{\mu} \\ 0 & \hat{\boldsymbol{\mu}} < \boldsymbol{\mu} \end{cases}, \quad (1.6)$$

since in this case the signal can only have a positive contribution to the yield. This statistics is leveraged to perform two types of test: the no signal hypothesis $\boldsymbol{\mu} = \mathbf{0}$ and the nominal signal hypothesis $\boldsymbol{\mu} = \mathbf{1}$. In the no signal hypothesis test, also called null hypothesis, the $p$-value quantifies the compatibility of the observed data with a background-only hypothesis ($\mu = 0$). It is defined as:

$$p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}} \,|\, \mathbf{0}) dt_{\boldsymbol{\mu}}, \quad (1.7)$$

where $t_{\boldsymbol{\mu},\text{obs}}$ is the observed test statistics (for the observed $\hat{\boldsymbol{\mu}}$) and $f(t_{\boldsymbol{\mu}} \,|\, 0)$ is the test statiscs $t_{\boldsymbol{\mu}}$ probability density function assuming $\boldsymbol{\mu} = \mathbf{0}$. The $p$-value is the probability of finding data that is equally or more incompatible with the null hypothesis. Therefore, a low $p$-value gives confidence to reject the null hypthosis. In particle physics, the $p$-value is often translated into the significance $Z$, measuring the number of Gaussian standard deviations ($\sigma$) above the background as

$$Z = \Phi^{-1}(1 - p_{\boldsymbol{\mu}}), \quad (1.8)$$

where $\Phi^{-1}$ is the inverse Gaussian cumulative distribution function. The standard for *observation* of a process is abritrarily set by the community at 5 $\sigma$ (correspond to a $p$-value $\approx 3 \times 10^{-7}$), with a 3 $\sigma$ signal strength significance ($p$-value $\approx 10^{-3}$) taken as *evidence* of a process.

To determine a 95% upper limit CL on a signal strength, a modified frequentist $CL_s$ method is deployed [83, 84], based on the test statistics $\tilde{t}$ defined as:

$$\tilde{t} = -2\ln\frac{\mathcal{L}_{s+b}}{\mathcal{L}_b} = -2\ln\frac{\mathcal{L}\left(\mu=1,\hat{\hat{\boldsymbol{\theta}}}(\mu=1),\hat{\hat{\boldsymbol{\gamma}}}(\mu=1)\right)}{\mathcal{L}\left(\mu=0,\hat{\hat{\boldsymbol{\theta}}}(\mu=0),\hat{\hat{\boldsymbol{\gamma}}}(\mu=0)\right)}, \tag{1.9}$$

where $\mathcal{L}_{s+b}$ is the nominal signal hypothesis ($\mu = 1$) and $\mathcal{L}_b$ the null hypothesis ($\mu = 0$), with the conditional likelihood optimisation of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ distinct between the two hypotheses for $\mu$. The upper 95% $CL_s$ limit on the signal strength $\mu$ is the $\mu$ value such that the $p$-value of the test statistics $\tilde{t}$ is equal to 0.05.

In addition to the fits performed between real and simulated datasets, so-called *Asimov* fits are perfomed. These leverage the *Asimov* datasets, corresponding to the sum all simulated processes (signal + backgrounds). Two such sets are used: a *prefit* Asimov dataset where the nuisance parameters are set to their initial values[31], and a *postfit* Asimov dataset where the NPs take their best-fit values from the fit to the real dataset. The postfit Asimov dataset can be used to define expected results, which quantify the sensitivity of an analysis to any similarly collected real data. Fits can be perfomed either conditionally or unconditionally, by setting the POIs to their SM expectations or letting them free-floating.

### 1.10.2 The $VH(H \to b\bar{b}/c\bar{c})$ fit

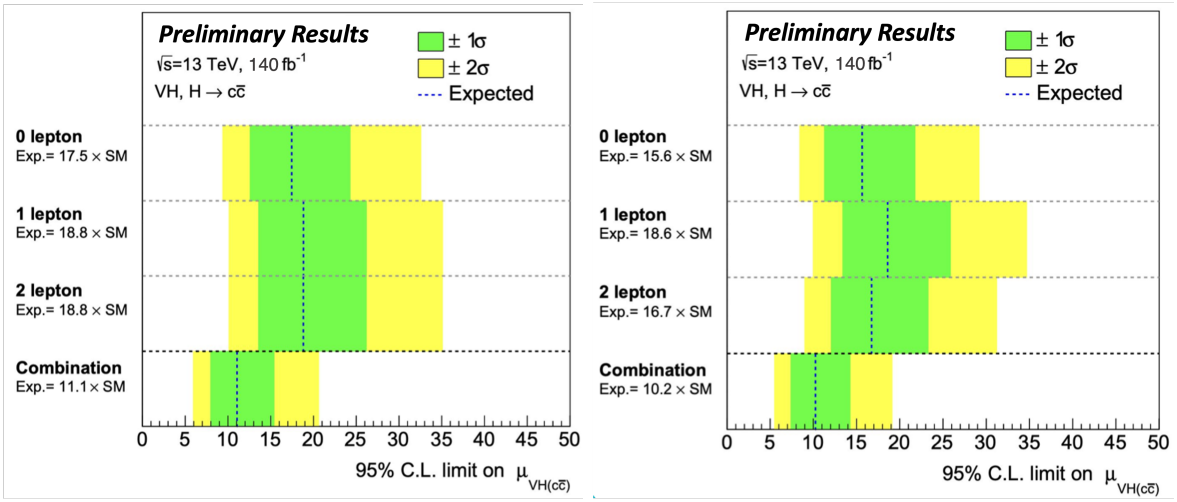
There are 15 POIs for the $VH(H \to b\bar{b})$ side and 1 POI for $VH(H \to c\bar{c})$. The binning used and regions included as well as the variables defining the underlying distributions entering the fits are detailed in Sections 1.5 and 1.7. A dense summary of the full categorisation is presented in Figure 1.17, underscoring the complexity of an analysis spanning 164 different regions, 84 of which are in $VH(H \to c\bar{c})$ (30 SRs, 6 Top $e\mu$ CRs, 10 $V+l$ CRs, 48 CRHigh), 48 in the resolved $VH(H \to b\bar{b})$ (21 SRs, 6 CRLow, 21 CRHigh), 12 $BT$-tagged Top CRs shared in the resolved regime, and 10 in the boosted regime (6 SRs, 4 boosted Top CRs). Experimental and modelling uncertainties are introduced to account for any mis-modelling and avoid biasing the fit, as described in Sections 1.8 and 1.9. The analysis described in this thesis is not yet concluded, with modifications to the modelling under active investigation at the time of writing. Consequentely, the fit is still blinded, with the data in signal regions bins most sensitive to the signal hidden. For $m_{bb}$ or $m_{cc}$

[31] 0 for all NPs but the FNs, which are set at 1.

distributions, the $H$-mass peak is blinded from 70 GeV to 140 GeV. For the MVA distributions, right-most (thus most signal-like) bins are iteratively blinded until at least 60% of the signal yield in the region is contained in blinded bins. For conditional fits, these blinded bins are used but the data is still not displayed in the plots. This thesis does not describe any unconditional fit to data, with any unconditional fits included performed with the Asimov dataset instead of the real data.

### $VH(H \to c\bar{c})$

Concerning the $VH(H \to c\bar{c})$ signal strength measurement, the 95% $CL_s$ expected upper limits are shown for the different lepton channels and combined in Figure 1.25a for the postfit Asimov and Figure 1.25b for the prefit Asimov.




(a) Combined postfit Asimov. (b) Combined prefit Asimov.

Figure 1.25: The 95% $CL_s$ upper limit on the $VH(H \to c\bar{c})$ signal strength from the combined analysis posfit (left) and prefit (right).

Significant improvements are expected for all lepton channels. The combination of all lepton channels leads to a remarkable improvement on the 95% $CL_s$ upper limit on $\mu_{VHcc}$ from 31 × SM in the latest published ATLAS $VH(H \to c\bar{c})$ analysis [26] to 11.1 × SM (10.2 × SM) in the postfit (prefit) Asimov combined one, a factor 2.8 improvements in sensitivity. Gains are expected to be made in all lepton channels, which now have very similar sensitivity thanks to modifications to the analysis strategy. Compared to the published analysis, the 0-lepton channel upper limits is reduced fom 40 × SM → 17.5 × SM, the lepton from 60 × SM → 18.8 × SM, and the 2-lepton from 51 × SM → 18.8 × SM [26]. These correspond to relative sensitivity improvement factors of 2.3, 3.2, and 2.7. Most of the gains are made in the 1- and 2-lepton channels, although the 0-lepton channel remains the most sensitive one.

### $VH(H \to b\bar{b})$

On the $VH(H \to b\bar{b})$ side, combining the resolved and boosted regime, the postfit expected significance on the $VH(H \to b\bar{b})$ signal strength is of 7.9 $\sigma$ over the background-only prediction, corresponding to a 23% improvement over the published expected significance at 6.3 $\sigma$ [31]. This is achieved thanks to a postfit expected significance of 4.7 $\sigma$ in the 0-lepton channel (15% improvement to published result), 5.3 $\sigma$ in 1-lepton (30% improvement), and 4.4 $\sigma$ in 2-lepton (3% improvement). The most sensitive channel is distinctively the 1-lepton channel in this case.

Separating the $VH(H \to b\bar{b})$ signal strength into two POIs for $WH(H \to b\bar{b})$ and $ZH(H \to b\bar{b})$, the prefit expected significances are 5.5 $\sigma$ for $WH$ and 6.2 $\sigma$ for $ZH$. This marks the first time a $H \to b\bar{b}$ analysis is expected to reach observation-level in $WH$, thanks to the large improvement in the 1-lepton channel sensitivity.

Finally, adopting the fine splitting of the STXS stage 1.2 with 15 bins defined by $p_T^V$ and additional jet multiplicity $N_{\text{jet}}$, with 8 bins in $ZH$ and 7 bins in $WH$, the $VH(H \to b\bar{b})$ analysis reaches the per bin sensitivities listed in Table 1.16, with evidence-level only attained for the 150 GeV $< p_T^V <$ 250 GeV with no additional jet $ZH$ bin. The systematics and statistical uncertainties impact on the signal strengths of the different bins are shown in Figure 1.26.

| $VH$ | Truth $p_T^V$ | 0 additional $N_{\text{jet}}$ | $\geq$1 additional $N_{\text{jet}}$ |
|---|---|---|---|
| WH | [75, 150[ GeV | 0.69 $\sigma$ | |
| WH | [150, 250[ GeV | 2.29 $\sigma$ | 0.55 $\sigma$ |
| WH | [250, 400[ GeV | 2.78 $\sigma$ | 0.94 $\sigma$ |
| WH | [400, 600[ GeV | 1.87 $\sigma$ | |
| WH | $\geq$ 600 GeV | 1.43 $\sigma$ | |
| ZH | [75, 150[ GeV | 1.48 $\sigma$ | 0.90 $\sigma$ |
| ZH | [150, 250[ GeV | 3.37 $\sigma$ | 1.64 $\sigma$ |
| ZH | [250, 400[ GeV | 2.85 $\sigma$ | 1.49 $\sigma$ |
| ZH | [400, 600[ GeV | 1.91 $\sigma$ | |
| ZH | $\geq$ 600 GeV | 1.07 $\sigma$ | |

Table 1.16: The expected prefit significance in the different STXS bins of the combined analysis.

#### The Diboson Cross-Check

The diboson cross-check analysis is performed with the $VZ(\to b\bar{b})$ and $VZ(\to c\bar{c})$ as signals in a similar fashion to the $VH(H \to b\bar{b}/c\bar{c})$ fit, with the intention to validate the strategy adopted.

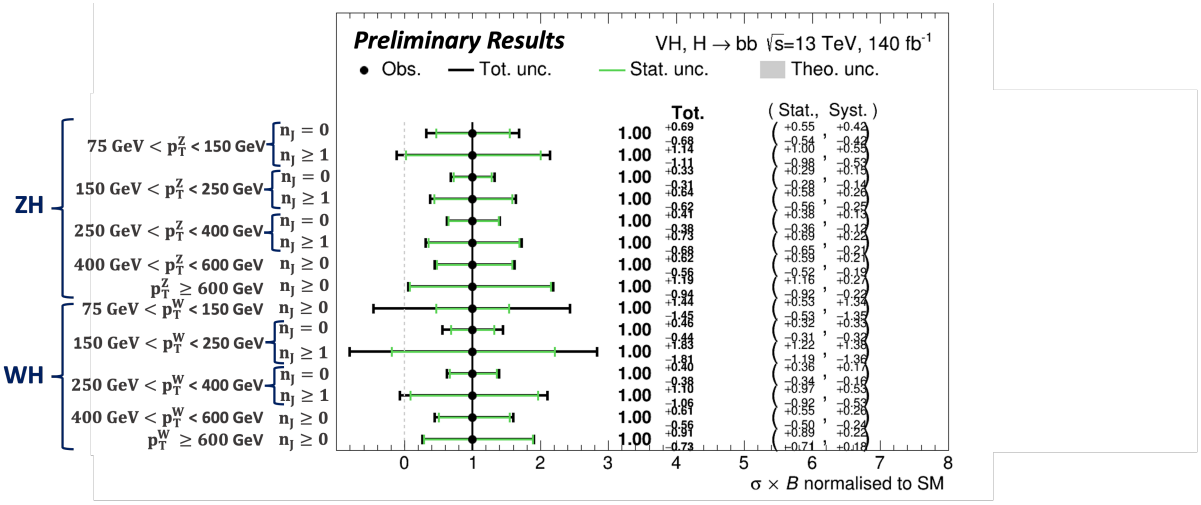


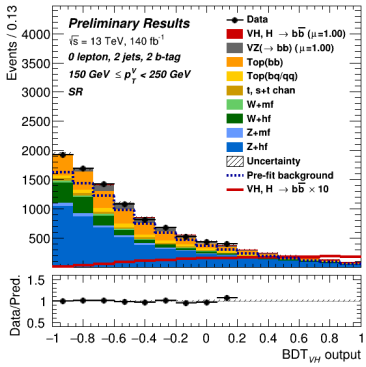

Figure 1.26: The constraints on the prefit STXS signal strength.

For the $VZ(\to b\bar{b})$ part, the postfit expected significance reaches a large value of 15.1 $\sigma$ when combining lepton channels. The 0-lepton channel reaches 11.2 $\sigma$, 1-lepton 6.2 $\sigma$, and 2-lepton 9 $\sigma$. On the $VZ(\to c\bar{c})$ side, the combined analysis expects to reach observation level for the first time, with a combined postfit expected significance of 5.1 $\sigma$. This represents a significant improvement of a factor 2.3 from the 2.2 $\sigma$ published expected result [26]. The combined analysis reaches a postfit expected significance of 3.9 $\sigma$ in 0-lepton, 2.6 $\sigma$ in 1-lepton, and 3.1 $\sigma$ in 2-lepton.

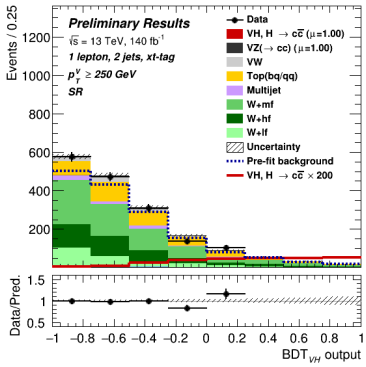
### Additional Fit Results

In addition to the main results highlighted above, some further insights into the output of the fits are given before concluding this chapter. To verify the MC-samples correctly reproduce the data after the fit, some posfit plots are presented in Figure 1.27 for selected signal regions and control regions, with all of them listed in Appendix A.6. Interestingly, good agreement between the data and posfit MC samples is also observed for validation distributions not directly constrained in the fit. Figure 1.28 displays posfit distributions for a $BL$-tagged region analogu-ous to the used $BT$-tagged Top CR, an $LL$-tagged region similar to a $c$-tagged signal region, and the $p_T^V$ spectrum of the whole 2L $BB$-tagged 2-jet signal region. The good agreement ob-served between data and MC-samples in all regions is a sign of the correct constraining by the fit.

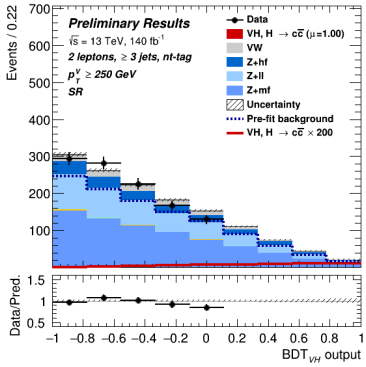
The breakdown of the uncertainties, presented in Table 1.17, is a measure of the different contributions of the uncertainties to the $VH(H \to b\bar{b}/c\bar{c})$ analysis. The NPs are grouped based on their origin, and their impact on the signal strengths of $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ is assessed by iteratively re-running fits with successive groups of NPs fixed at their postfit values. The notation adopted is to label the signal strengths (one for $VH(H \to b\bar{b})$ and one for



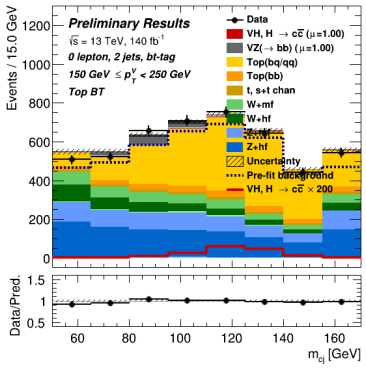

(a) 0L, 2-jet 150 GeV $< p_T^V <$ 250 GeV $BB$-tagged.



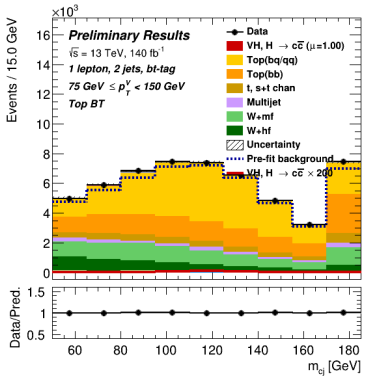

(b) 1L, 2-jet $p_T^V >$ 250 GeV 2 $c$-tagged.



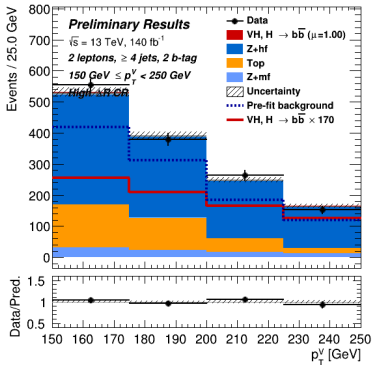

(c) 2L, ≥3-jet $p_T^V >$ 250 GeV 1 $c$-tagged.




(d) 0L, 2-jet 150 GeV $< p_T^V <$ 250 GeV Top $BT$ CR.

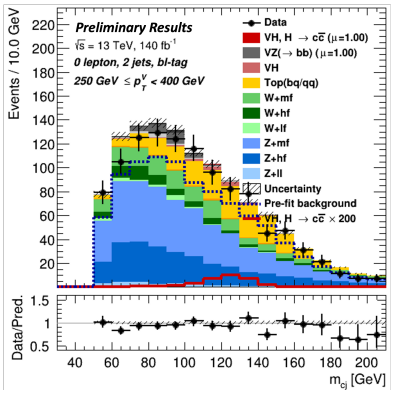



(e) 1L, 2-jet 75 GeV $< p_T^V <$ 150 GeV Top $BT$ CR.

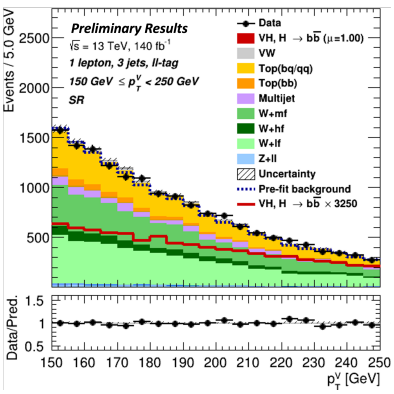



(f) 2L, ≥ 4-jet 250 GeV $< p_T^V <$ 400 GeV $BB$-tagged CRHigh.

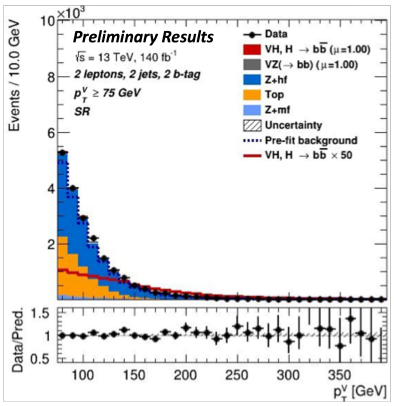
Figure 1.27: Selected posfit signal regions (top row) and control regions (bottom row), for the 0L (left), 1L (centre), and 2L (right).




(a) 0L $BL$-tagged, Top $BT$ CR-like.




(b) 1L $LL$, $c$-tagged SR-like.




(c) 2L $p_T^V$ in 2-jets $BB$ SR.

Figure 1.28: Posfit distributions in a $BL$-tagged Top CR-like (left) and $LL$-tagged SR-like (centre) validations regions and the 2L $p_T^V$ spectrum in the 2-jet $BB$-tagged SR.

$VH(H \to c\bar{c})$) of the nominal maximal likelihood fit as $\hat{\mu}$ with uncertainty $\sigma_{\hat{\mu}}$, and of a re-ran fit with a group of NPs fixed as $\hat{\mu}'$ with uncertainty $\sigma_{\hat{\mu}'}$. The impact of the fixed group of NPs is defined as the change in uncertainty measured by

$$\text{Impact} = \sqrt{\sigma_{\hat{\mu}}^2 - \sigma_{\hat{\mu}'}^2}. \tag{1.10}$$

| Source of Uncertainty | $\mu_{VH(H\to b\bar{b})}$ | $\mu_{VH(H\to c\bar{c})}$ |
|---|---|---|
| **Total** | 0.127 | 5.089 |
| **Statistics** | 0.095 | 3.791 |
| **Systematics** | 0.085 | 3.395 |
| **Statistical Uncertainties** | 0.095 | 3.791 |
| Data sample size | 0.088 | 3.538 |
| Floating normalisations | 0.029 | 1.247 |
| Top $e\mu$ CR statistics | 0.011 | 0.130 |
| **Systematics Uncertainties** | 0.085 | 3.395 |
| $VH(H \to b\bar{b}/c\bar{c})$ Modelling | 0.021 | 0.237 |
| **Backgrounds Modelling** | 0.069 | 2.739 |
| $Z$+jets | 0.036 | 1.587 |
| $W$+jets | 0.036 | 1.088 |
| Diboson | 0.020 | 0.546 |
| $t\bar{t}$ | 0.011 | 0.613 |
| single-top | 0.008 | 0.116 |
| Multi-jet | 0.007 | 0.691 |
| **Experimental Uncertainties** | 0.035 | 1.278 |
| Jet | 0.026 | 0.737 |
| Large-$R$ jet | 0.009 | 0.206 |
| $E_T^{\text{miss}}$ | 0.007 | 0.150 |
| Lepton | 0.004 | 0.115 |
| FTAG PFlow ($b$-jet) | 0.015 | 0.258 |
| FTAG PFlow ($c$-jet) | 0.008 | 0.769 |
| FTAG PFlow (light-jet) | 0.003 | 0.751 |
| FTAG PFlow (extrap) | 0.000 | 0.000 |
| FTAG VR ($b$-jet) | 0.004 | 0.049 |
| FTAG VR ($c$-jet) | 0.001 | 0.018 |
| FTAG VR (light-jet) | 0.001 | 0.009 |
| FTAG VR (extrap) | 0.001 | 0.037 |
| Pile-up | 0.005 | 0.052 |
| Luminosity | 0.007 | 0.035 |
| **MC-samples Size** | 0.020 | 1.410 |

Table 1.17: Breakdown of the different systematics and statistical uncertainties.

To define the impact of the statistical uncertainties, a fit is run with all NPs fixed except for the floating normalisations. The total systematics effect is set to the difference in quadrature between the total and the statistical uncertainties. For both the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ measurement, the statistical and systematics uncertainties are of similar size, with the statistical

uncertainties being slightly larger. The uncertainties are far smaller for the $VH(H \to b\bar{b})$ side, as expected from the larger statistics and better performance of both the experimental reconstruction and modelling. For $VH(H \to b\bar{b})$, the largest contributions to the systematics uncertainties come from the $V$+jets and diboson background modelling, and the signal modelling. The importance of the $V$+jets is expected, since the $W$+jets and $Z$+jets play a significant role in 1-lepton and in 0- and 2-lepton channels respectively. On the experimental side, the jet and flavour tagging uncertainties are leading. For the latter, the $b$-jets uncertainties contributes the most followed by the $c$-jets, as expected from the resemblance between heavy flavour jet species.

For $VH(H \to c\bar{c})$, similar observations can be made with several nuances. On the modelling side, the signal modelling is less paramount, with the top processes and multi-jet not contributing far more significantly. Additional, the $Z$+jets uncertainties are now clearly the leading one, with the $W$+jets proportionally less important. This latter observation is connected with the larger importance of the top processes, as $VH(H \to c\bar{c})$ has a much larger top contribution in the 1-lepton channel, competing with $W$+jets as the leading source of uncertainty there. On the experimental side, the flavour tagging uncertainties of the $c$- and light-jet are now dominant, with the jet uncertainties. This is expected from the challenges of tagging and reconstructing $c$-jets. The statistics of the MC-samples is far more important on the $VH(H \to c\bar{c})$ side, mostly due to the low $c$-tagging efficiency of the DL1r tagger used.

A second technique to assess the importance of different nuisance parameters on the signal strengths is to change their NP values upwards and downwards by their postfit uncertainties $\sigma_\theta$ and re-run the fit with the modified NP fixed. For each NP, this requires to run two fits in addition to the nominal fit from whitch $\hat{\theta}$ and $\sigma_{\hat{\theta}}$ are measured: one with the NP fixed at $\hat{\theta} + \sigma_{\hat{\theta}}$ and one with $\hat{\theta} - \sigma_{\hat{\theta}}$. NPs are ranked by the difference of the signal strengths between these new fits and the nominal one, as shown in Figure 1.29. In these plots, the central values of NPs are set at 0 (1 for FNs and $\gamma$-factor) as the dataset is the postfit Asimov set. For $VH(H \to b\bar{b})$, the $W$+$hf$ extrapolations have a significant impact on the predicted signal strength, with several of these systematics highly ranked. Shape uncertainties associated to the diboson process and Higgs modelling uncertainties as well as the $Wt$ DS-DR shape uncertainty and a $b$-jet tagging uncertainty also contribute meaningfully. The floating normalisation of $W$+$hf$ in the boosted region is the only FN to make the ranking, due to its significant pull as is shown below in Figure 1.30. For $VH(H \to c\bar{c})$, the Top process CARL shapes are the leading nuisance parameters,

with the $Z+cc$ shape and the $W$+jets $cc/bb$ acceptance ratio. The $Z$+$lf$ and, to a lesser extent, $W$+$lf$ floating normalisations have a large impact on the predicted signal strength, despite the constraining offered by the $V+l$ CR. The light- and $c$-jets uncertainties from flavour tagging are the biggest contributors in this category. Finally, the multi-jet process normalisation enters the ranking as this process contributes more in $VH(H \to c\bar{c})$. The $\gamma$-factor listed corresponds to the last unblinded bin in the 1L high $p_T^V$ 2-jet SR, shown in Figure 1.27b, where a large amount of signal is expected, and the effect of this NP should be reduced once the signal is no longer constrained to its SM expectations in a conditional data fit.

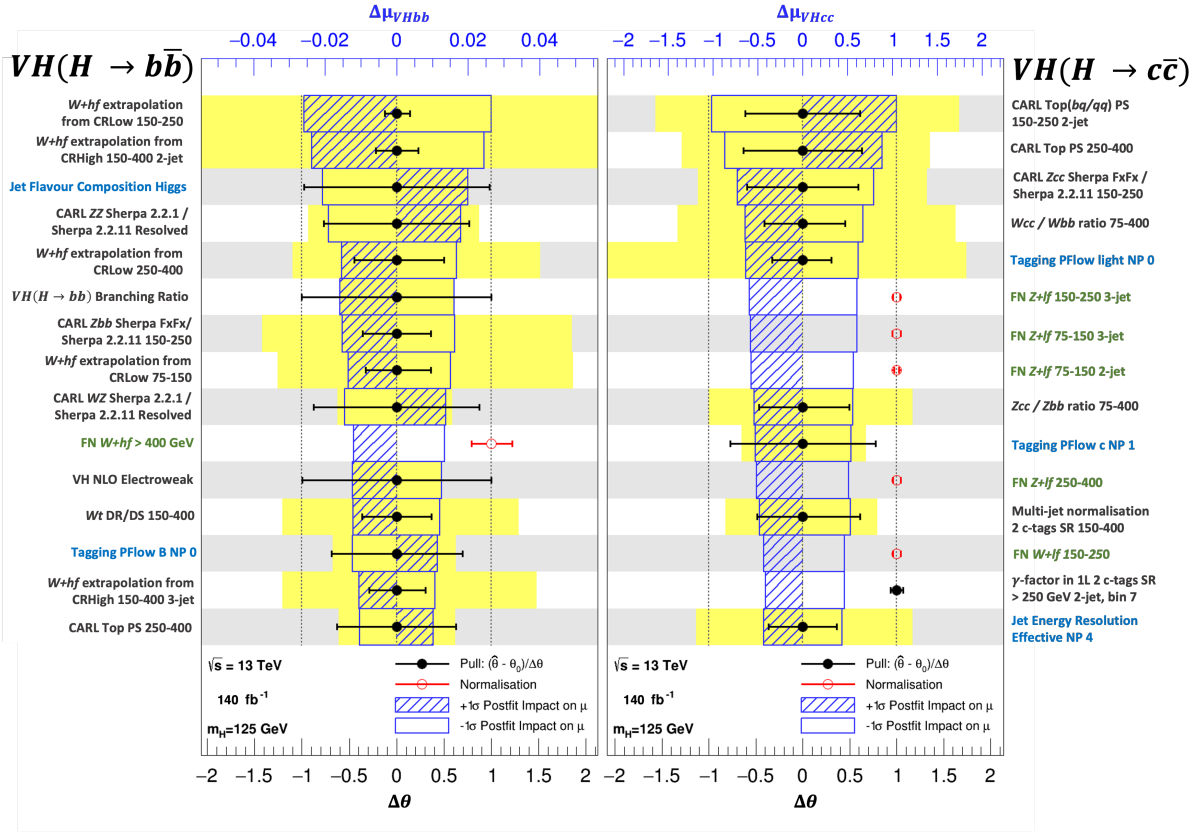



Figure 1.29: The 15 most highly ranked Asimov postfit nuisance parameters for the $VH(H \to b\bar{b})$ (left) and $VH(H \to c\bar{c})$ (right) signal strengths. Modelling NPs are written in black, experimental NPs in blue, and floating normalisation (and $\gamma$-factor) in green, with values indicated by the bottom axis showing $\Delta\theta = \hat{\theta} - \theta_0$. Black points are nuisance parameters with their central value at 0 showing the pull ($\gamma$-factor with central value at 1), red points are floating normalisation with central values at 1. The error bars on the point show the 1 $\sigma$ uncertainty of the NPs. The effect of changing the NP by +1 $\sigma$ (-1 $\sigma$) induces the change in signal strength $\Delta\mu$ shown by the hashes (empty) blue rectangle, with respect to the top axis.

In the combined analysis, the major backgrounds have free-floating normalisations decorrelated across the different $p_T^V$ and jet multiplicity bins. The values set by a conditional likelihood fit to data, where the $VH(H \to b\bar{b}/c\bar{c})$ signal strengths are set to the SM expectations, are presented in Figure 1.30. They are compared to the same FNs obtained in the cross-check analysis with $VZ(\to b\bar{b}/c\bar{c})$ as signals. Good agreement is observed between the two sets of floating

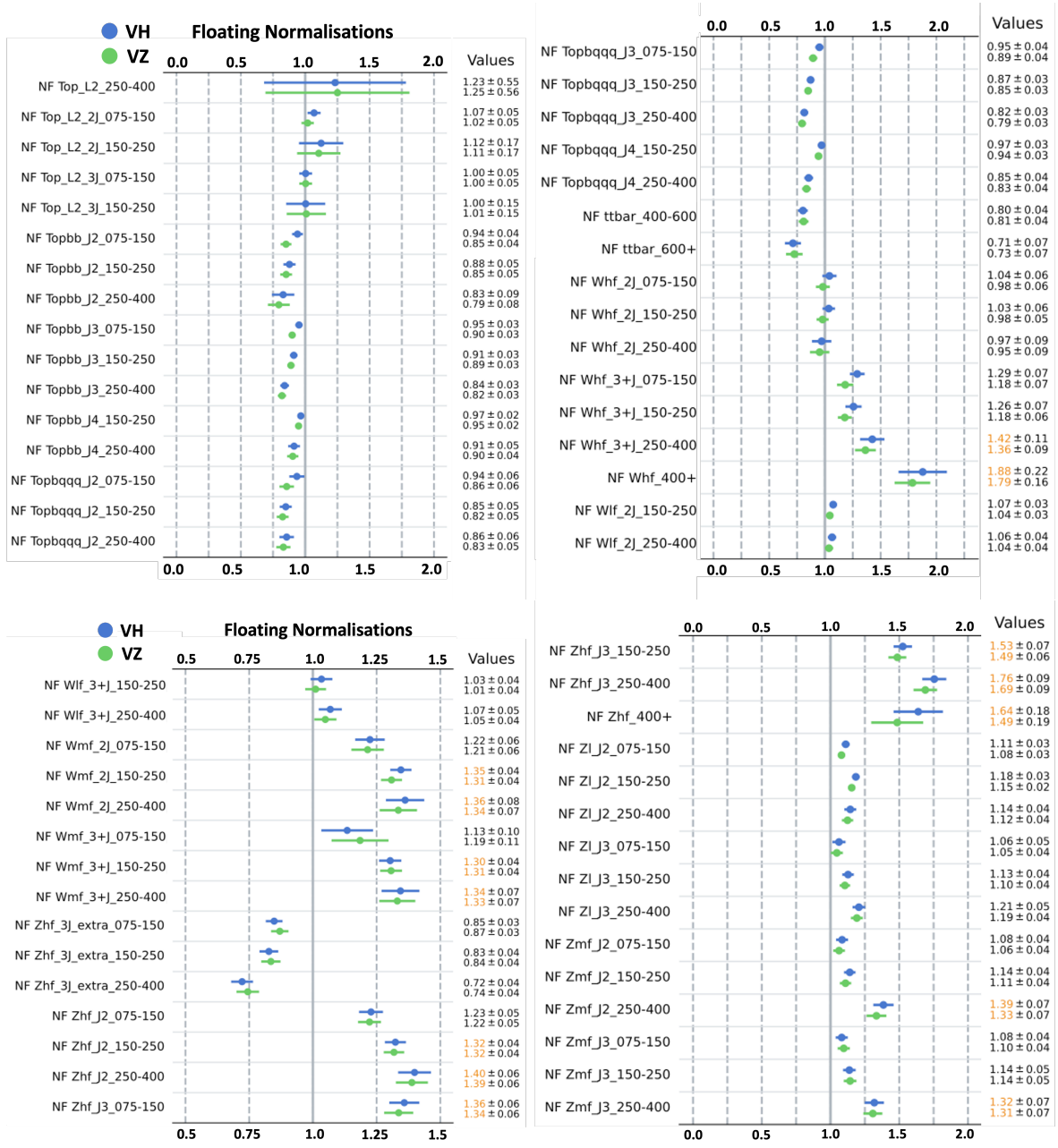



Figure 1.30: The floating normalisations of the major background in the combined analysis targeting the $VH(H \to b\bar{b}/c\bar{c})$ in blue, versus the cross-check analysis $VZ(\to b\bar{b}/c\bar{c})$ in green.

normalisations. Some common trends per process are present. Concerning the Top backgrounds, in 0L and 1L it seems mostly overestimated in the MC simulation, with the overestimation increasing with $p_T^V$. In 2L, the Top process seems well estimated in the Top $e\mu$ CR, but the lower statistics available at higher $p_T^V$ leads to a poor constraining of the floating normalisation. The FNs for the Top process are generally better constrained in the 3-jet then the 2-jet category, as expected from the larger yield available for this background at higher jet multiplicities. The Top($bq/qq$) and Top($bb$) have generally similar FN values. Concerning the $W$+jets, the $W$+$hf$ is well modelled in 2-jet across $p_T^V$ but less so in the ≥3-jet category, where the underestimation of the simulations grows with $p_T^V$. The boosted $W$+$hf$ normalisation in the ≥ 400 GeV range is

significantly distant from 1. The same observations hold for $W$+*lf*, which is also well modelled in 2-jet but gets higher FNs in 3-jet. The $W$+*mf* component also requires large corrections from the fit, with FN values $\sim 1.3$ in all the $N_{\text{jet}}$ and $p_T^V$ bins. The final background modelled with floating normalisations is $Z$+jets, which also requires significant yield modifications from the fit in all components, jet multiplicity, and $p_T^V$ bins. All $Z$+jets yield are corrected up, with larger FN values required at higher $p_T^V$. A special case for the $Z$+*hf* are the 3-jet and 3-jet-extra categories, adopted to account for the fact the $VH(H \to c\bar{c})$ side does not use 4-jet or separates 3- and ≥4-jet in 0L and 2L while the $VH(H \to b\bar{b})$ combines 3-jet with 4-jet into ≥3-jet in 2L. The 3-jet FNs in the figure, labelled "J3", cover the ≥ 3-jet (for $VH(H \to b\bar{b})$), while the 3-jet-extra, labelled "3J_extra" acount only for the 3-jet category[32]. There is therefore some overlap, with the latter set of FNs used to correct down the large normalisation of the ≥ 3-jet. Similarly to the $W$+*hf*, the boosted $Z$+*hf*FN value is significantly pulled away from 1.

The correlations between the different floating normalisations are displayed as a heat map in Figure 1.31. A rich structure of dependencies emerge from such a plot. As expected, FNs related to each individual process are highly correlated with the other FN of the same process, in different analysis $p_T^V$ and $N_{\text{jet}}$ categories. Some stricking exceptions are visible: the boosted $t\bar{t}$ displays a very small uncorrelation with the resolved top($bb$) and top($bq/qq$). Concerning correlations across processes, the Top($bb$) and Top($bq/qq$) are respectively seen to have large correlations with the $Z$+*hf* (and the $W$+*hf* in lesser extend) and the $W$+*mf* and $Z$+*mf*, as expected from the presence of the $Z$+jets and $W$+jets in the CRHigh and the 0L and 1L $BT$-tagged Top CR. The $W$+*hf* normalisations are slightly anti-correlated to the $V$+*lf*and slightly correlated to the $Z$+*hf* in the low $N_{\text{jet}}$Ṡimilarly, the $V$+*lf* are strongly correlated between the $W$ and $Z$.

## 1.11 Conclusion

This chapter introduces the $VH(H \to b\bar{b}/c\bar{c})$ combined ATLAS analysis using the 140 fb$^{-1}$ of data collected during Run 2 (2015-2018), in its state after its third unblinding approval meeting. Events are separated based on their $p_T^V$ and two highest tags of their jet or sub-jets into a resolved or boosted regime $VH(H \to b\bar{b})$ or $VH(H \to c\bar{c})$. Tagging is performed with the machine learning-based DL1r, and the ranking establishes a hierarchy of $b$-tagged > tight $c$-tagged > loose $c$-tagged > untagged. This is followed by a split into lepton channel based on the number of charged lepton ($e$, $\mu$) in the final state, to separate the $Z(\to \nu\nu)H$, $W(\to \ell\nu)H$,

[32]The contribution of the $Z$+*mf* and $Z$+*lf* are ignored in 4-jet.

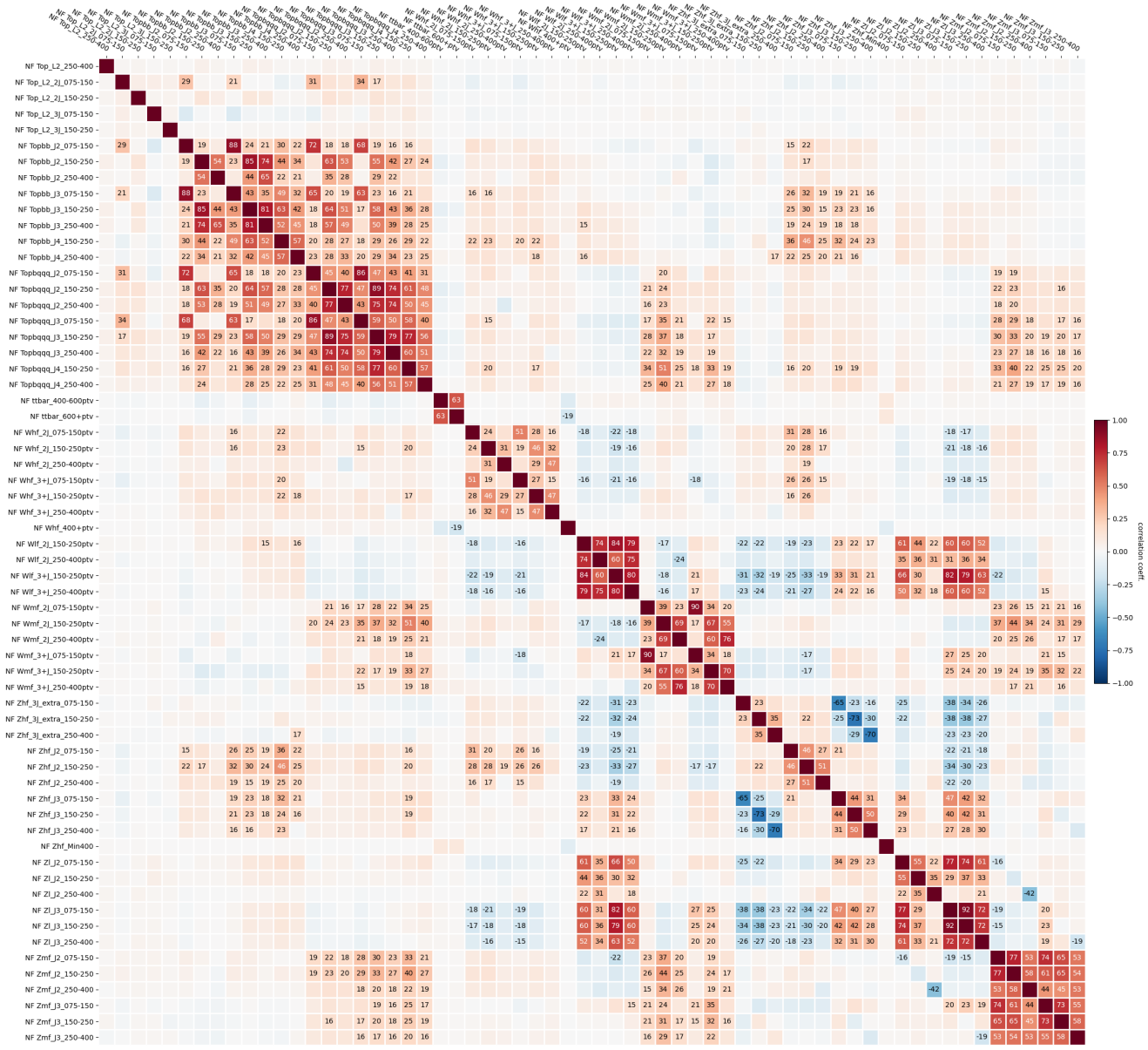


Figure 1.31: The correlations between the floating normalisations of the major background in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

and $Z(\to \ell^+\ell^-)H$, with the $H \to b\bar{b}$ or $H \to c\bar{c}$. To boost the sensitivity, a final categorisation further splits the analysis phase space into regions of defined $p_T^V$ and jet multiplicity. The major backgrounds of the analysis are the $V$+jets and the top processes, such as the $t\bar{t}$ pair production and the single-top $Wt$ production. These are respectively constrained from data in dedicated control regions defined by a cut on the angular separation of the Higgs-candidate jets and by an alternative event-tagging selection. To validate the adopted strategy, a cross-check analysis targeting the $VZ(\to b\bar{b}/c\bar{c})$ is performed.

The analysis promises to deliver exciting improvements in the sensitivitiy of the search for

the $H \to c\bar{c}$ process, and the finest measurements to date of the differential cross-section of the $VH(H \to b\bar{b})$. The analysis has already converged on the complex and wide-ranging categorisation strategy presented in this thesis. MVA discriminants are introduced throughout the different regimes to improve the sensitivity to the sought signals. The adoption of upgraded flavour tagging offered the opportunity to adopt a pseudo-continuous joint-tagging approach, paving the way for a more coherent joint measurements of the $VH$ to heavy flavour quarks decay. New Monte-Carlo simulated samples of higher statistics contributed to reduce the importance of uncertainties plaguing the final fit performance. The final step, still under study at the time of writing by the analysis team, is the modelling approach adopts to constrain the background within the limited knowledge of the detector and simulations uncertainties. Fine adjustements are required to understand how the complex fit introduced in this thesis constrains the different backgrounds, and whether this constraining is relying on physically motivated effects modelled by their respective nuisance parameters.

This analysis serves as the jointly combined legacy $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analyses of ATLAS for the full Run 2, using the entire data statistics available of 140 fb$^{-1}$. Excitingly, progress in the analysis sensitivity to the $VH(H \to c\bar{c})$ signal strength has greatly accelerated, with reductions in the upper limit fast approaching the realm of direct measurement of the central value. At current pace of improvement, the signal strength might be measurable in the next phase of the LHC: the High-Luminosity-LHC (HL-LHC). This will require continued improvement to the experimental tools and the analysis strategy. The former will primarily rely on improved flavour tagging abilities which, excitingly, is right around the corner for the $VH$ analysis: from the single tagger Graph Network 2 Model (GN2) to the boosted $X \to b\bar{b}/c\bar{c}$ decay tagger *GN2X*. The adoption of transformer-based neural networks is promising a significant increase in tagging performance. This will reverbates into an improved signal acceptance and a better separation from the backgrounds. The larger volume of data to be collected in the Run 4 of the LHC and future data taking campaigns will significantly enrich the prospects of this severely statistically-limited analysis. This comes with an additional challenge, to operate at higher pile-up, which will require additional upgrades to the experimental techniques, particularly in pile-up jets rejection.

# Bibliography


[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[2] CMS Collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[3] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[4] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[5] Hideki Yukawa. "On the Interaction of Elementary Particles. I". In: *Progress of Theoretical Physics Supplement* 1 (Jan. 1955), pp. 1–10. ISSN: 0375-9687. DOI: 10.1143/PTPS.1.1. eprint: https://academic.oup.com/ptps/article-pdf/doi/10.1143/PTPS.1.1/5310694/1-1.pdf. URL: https://doi.org/10.1143/PTPS.1.1.

[6] Peter W. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Phys. Lett.* 12 (1964), pp. 132–133. DOI: 10.1016/0031-9163(64)91136-9.

[7] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[8] ATLAS Collaboration. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". In: *Phys. Lett. B* 784 (2018), pp. 173–191. DOI: 10.1016/j.physletb.2018.07.035. arXiv: 1806.00425 [hep-ex].

[9] CMS Collaboration. "Observation of $t\bar{t}H$ production". In: *Phys. Rev. Lett.* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex].

[10] ATLAS Collaboration. "Measurements of Higgs boson production cross-sections in the $H \to \tau^+\tau^-$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Journal of High Energy Physics* 2022.8 (2022), p. 175. DOI: 10.1007/JHEP08(2022)175. URL: https://doi.org/10.1007/JHEP08(2022)175.

[11] CMS Collaboration. "Measurement of the inclusive and differential Higgs boson production cross-sections in the decay mode to a pair of $\tau$ leptons in pp collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. Lett.* 128.8 (2022), p. 081805. DOI: 10.1103/PhysRevLett.128.081805. arXiv: 2107.11486 [hep-ex].

[12] ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and $VH$ production with the ATLAS detector". In: *Phys. Lett. B* 786 (2018), pp. 59–86. DOI: 10.1016/j.physletb.2018.09.013. arXiv: 1808.08238 [hep-ex].

[13] CMS Collaboration. "Observation of Higgs boson decay to bottom quarks". In: *Phys. Rev. Lett.* 121.12 (2018), p. 121801. DOI: 10.1103/PhysRevLett.121.121801. arXiv: 1808.08242 [hep-ex].

[14] CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons". In: *JHEP* 01 (2021), p. 148. DOI: 10.1007/JHEP01(2021)148. arXiv: 2009.04363 [hep-ex].

[15] ATLAS Collaboration. "A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector". In: *Phys. Lett. B* 812 (2021), p. 135980. DOI: 10.1016/j.physletb.2020.135980. arXiv: 2007.07830 [hep-ex].

[16] ATLAS Collaboration. "Search for the Decay of the Higgs Boson to Charm Quarks with the ATLAS Experiment". In: *Phys. Rev. Lett.* 120.21 (2018), p. 211802. DOI: 10.1103/PhysRevLett.120.211802. arXiv: 1802.04329.

[17] A. Djouadi, J. Kalinowski, and M. Spira. "HDECAY: a program for Higgs boson decays in the Standard Model and its supersymmetric extension". In: *Computer Physics Communications* 108.1 (1998), pp. 56–74. ISSN: 0010-4655. DOI: https://doi.org/10.1016/S0010-4655(97)00123-9. URL: https://www.sciencedirect.com/science/article/pii/S0010465597001239.

[18] Tao Han et al. "Higgs boson decay to charmonia via c-quark fragmentation". In: *Journal of High Energy Physics* 2022.8 (2022), p. 73. DOI: 10.1007/JHEP08(2022)073. URL: https://doi.org/10.1007/JHEP08(2022)073.

[19] Cédric Delaunay et al. "Enhanced Higgs boson coupling to charm pairs". In: *Phys. Rev. D* 89 (3 Feb. 2014), p. 033014. DOI: 10.1103/PhysRevD.89.033014. URL: https://link.aps.org/doi/10.1103/PhysRevD.89.033014.

[20] Gilad Perez et al. "Constraining the charm Yukawa and Higgs-quark coupling universality". In: *Phys. Rev. D* 92 (3 Aug. 2015), p. 033016. DOI: 10.1103/PhysRevD.92.033016. URL: https://link.aps.org/doi/10.1103/PhysRevD.92.033016.

[21] F. J. Botella et al. "What if the masses of the first two quark families are not generated by the standard model Higgs boson?" In: *Phys. Rev. D* 94.11 (2016), p. 115031. DOI: 10.1103/PhysRevD.94.115031. arXiv: 1602.08011 [hep-ph].

[22] Shaouly Bar-Shalom and Amarjit Soni. "Universally enhanced light-quarks Yukawa couplings paradigm". In: *Phys. Rev. D* 98 (5 Sept. 2018), p. 055001. DOI: 10.1103/PhysRevD.98.055001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.055001.

[23] Diptimoy Ghosh, Rick Sandeepan Gupta, and Gilad Perez. "Is the Higgs mechanism of fermion mass generation a fact? A Yukawa-less first-two-generation model". In: *Physics Letters B* 755 (2016), pp. 504–508. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2016.02.059. URL: https://www.sciencedirect.com/science/article/pii/S0370269316001556.

[24] Daniel Egana-Ugrinovic, Samuel Homiller, and Patrick Meade. "Aligned and Spontaneous Flavor Violation". In: *Phys. Rev. Lett.* 123 (3 July 2019), p. 031802. DOI: 10.1103/PhysRevLett.123.031802. URL: https://link.aps.org/doi/10.1103/PhysRevLett.123.031802.

[25] Daniel Egana-Ugrinovic, Samuel Homiller, and Patrick Meade. "Higgs bosons with large couplings to light quarks". In: *Phys. Rev. D* 100 (11 Dec. 2019), p. 115041. DOI: 10.1103/PhysRevD.100.115041. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.115041.

[26] ATLAS Collaboration. *Direct constraint on the Higgs-charm coupling using Higgs boson decays to charm quarks with the ATLAS detector*. Tech. rep. Geneva: CERN, 2020. URL: %5Curl%7Bhttps://cds.cern.ch/record/2721696%7D.

[27] CMS Collaboration. *Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. arXiv: 2205.05550. URL: %5Curl%7Bhttps://cds.cern.ch/record/2809290%7D.

[28] ATLAS Collaboration. "Measurements of $WH$ and $ZH$ production in the $H \to b\bar{b}$ decay channel in $pp$ collisions at 13 TeV with the ATLAS detector". In: *Eur. Phys. J. C* 81.2 (2021), p. 178. DOI: 10.1140/epjc/s10052-020-08677-2. arXiv: 2007.02873 [hep-ex].

[29] ATLAS Collaboration. "Measurement of the associated production of a Higgs boson decaying into $b$-quarks with a vector boson at high transverse momentum in $pp$ collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Lett. B* 816 (2021), p. 136204. DOI: 10.1016/j.physletb.2021.136204. arXiv: 2008.02508 [hep-ex].

[30] CMS Collaboration. *Simplified template cross-section measurements of Higgs boson produced in association with vector bosons in the $H \to b\bar{b}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2827421.

[31] ATLAS Collaboration. *Combination of measurements of Higgs boson production in association with a $W$ or $Z$ boson in the $b\bar{b}$ decay channel with the ATLAS experiment at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2782535.

[32] ATLAS Collaboration. "Luminosity determination in $pp$ collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC". In: *Eur. Phys. J. C* 83.10 (2023), p. 982. DOI: 10.1140/epjc/s10052-023-11747-w. arXiv: 2212.09379 [hep-ex].

[33] ATLAS Collaboration. "The ATLAS Simulation Infrastructure". In: *The European Physical Journal C* 70.3 (2010), pp. 823–874. DOI: 10.1140/epjc/s10052-010-1429-9. URL: https://doi.org/10.1140/epjc/s10052-010-1429-9.

[34] GEANT4 Collaboration. "GEANT4, A Simulation toolkit". In: *Nucl. Instrum. Methods Phys. Res., A* 506.CERN-IT-2002-003 (2002), 250–303. 54 p. URL: %5Curl%7Bhttps://cds.cern.ch/record/602040%7D.

[35] David J. Lange. "The EvtGen particle decay simulation package". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 462.1 (2001). BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, pp. 152–155. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00089-4. URL: https://www.sciencedirect.com/science/article/pii/S0168900201000894.

[36] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159–177. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.01.024. URL: https://www.sciencedirect.com/science/article/pii/S0010465515000442.

[37] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *Journal of High Energy Physics* 2007.11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. URL: https://dx.doi.org/10.1088/1126-6708/2007/11/070.

[38] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *Journal of High Energy Physics* 2010.6 (2010), p. 43. DOI: 10.1007/JHEP06(2010)043. URL: https://doi.org/10.1007/JHEP06(2010)043.

[39] Gionata Luisoni et al. "HW±/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". In: *Journal of High Energy Physics* 2013.10 (2013), p. 83. DOI: 10.1007/JHEP10(2013)083. URL: https://doi.org/10.1007/JHEP10(2013)083.

[40] Gavin Cullen et al. "Automated one-loop calculations with GoSam". In: *The European Physical Journal C* 72.3 (2012), p. 1889. DOI: 10.1140/epjc/s10052-012-1889-1. URL: https://doi.org/10.1140/epjc/s10052-012-1889-1.

[41] ATLAS collaboration. "Measurement of the $Z/\gamma^*$ boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector". In: *Journal of High Energy Physics* 2014.9 (2014), p. 145. DOI: 10.1007/JHEP09(2014)145. URL: https://doi.org/10.1007/JHEP09(2014)145.

[42] Richard D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (2015), p. 40. DOI: 10.1007/JHEP04(2015)040. URL: https://doi.org/10.1007/JHEP04(2015)040.

[43] Oliver Brein, Abdelhak Djouadi, and Robert Harlander. "NNLO QCD corrections to the Higgs-strahlung processes at hadron colliders". In: *Physics Letters B* 579.1 (2004), pp. 149–156. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2003.10.112. URL: https://www.sciencedirect.com/science/article/pii/S0370269303017234.

[44] Johannes Bellm et al. *Herwig 7.1 Release Note*. 2017. arXiv: 1705.06919 [hep-ph].

[45] Enrico Bothmann et al. "Event generation with Sherpa 2.2". In: *SciPost Phys.* 7 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034. URL: https://scipost.org/10.21468/SciPostPhys.7.3.034.

[46] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross-sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 2014.7 (2014), p. 79. DOI: 10.1007/JHEP07(2014)079. URL: https://doi.org/10.1007/JHEP07(2014)079.

[47] Enrico Bothmann et al. "Event generation with Sherpa 2.2". In: *SciPost Phys.* 7 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034. URL: https://scipost.org/10.21468/SciPostPhys.7.3.034.

[48] ATLAS collaboration. "Modelling and computational improvements to the simulation of single vector-boson plus jet processes for the ATLAS experiment". In: *Journal of High Energy Physics* 2022.8 (2022), p. 89. DOI: 10.1007/JHEP08(2022)089. URL: https://doi.org/10.1007/JHEP08(2022)089.

[49] Stefano Frixione, Giovanni Ridolfi, and Paolo Nason. "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction". In: *Journal of High Energy Physics* 2007.09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. URL: https://dx.doi.org/10.1088/1126-6708/2007/09/126.

[50] Paolo Nason. "A new method for combining NLO QCD with shower Monte Carlo algorithms". In: *Journal of High Energy Physics* 2004.11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. URL: https://dx.doi.org/10.1088/1126-6708/2004/11/040.

[51] Michał Czakon and Alexander Mitov. "Top++: A program for the calculation of the top-pair cross-section at hadron colliders". In: *Computer Physics Communications* 185.11 (2014), pp. 2930–2938. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2014.06.021. URL: https://www.sciencedirect.com/science/article/pii/S0010465514002264.

[52] Johannes Bellm et al. "Herwig 7.0/Herwig++ 3.0 release note". In: *The European Physical Journal C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. URL: https://doi.org/10.1140/epjc/s10052-016-4018-8.

[53] M. Aliev et al. "HATHOR – Hadronic top and Heavy quarks cross-section calculator". In: *Computer Physics Communications* 182.4 (2011), pp. 1034–1046. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2010.12.040. URL: https://www.sciencedirect.com/science/article/pii/S0010465510005333.

[54] P. Kant et al. "HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions". In: *Computer Physics Communications* 191 (2015), pp. 74–89. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0010465515000454.

[55] Nikolaos Kidonakis. "Two-loop soft anomalous dimensions for single top quark associated production with a $W^-$ or $H^-$". In: *Phys. Rev. D* 82 (5 Sept. 2010), p. 054018. DOI: 10.1103/PhysRevD.82.054018. URL: https://link.aps.org/doi/10.1103/PhysRevD.82.054018.

[56] Nikolaos Kidonakis. *Top Quark Production*. 2013. arXiv: 1311.0283 [hep-ph].

[57] Stefano Frixione et al. "Single-top hadroproduction in association with a W boson". In: *Journal of High Energy Physics* 2008.07 (July 2008), p. 029. DOI: 10.1088/1126-6708/2008/07/029. URL: https://dx.doi.org/10.1088/1126-6708/2008/07/029.

[58] ATLAS Collaboration. *Vertex Reconstruction Performance of the ATLAS Detector at* $\sqrt{s} =$ 13 *TeV*. Tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2015-026. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2037717.

[59] ATLAS Collaboration. "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 79.8 (2019), p. 639. arXiv: 1902.04655. URL: %5Curl%7Bhttps://cds.cern.ch/record/2657964%7D.

[60] ATLAS Collaboration. "Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data". In: *Journal of Instrumentation* 14.12 (2019), P12006. DOI: 10.1088/1748-0221/14/12/P12006. URL: https://dx.doi.org/10.1088/1748-0221/14/12/P12006.

[61] "Muon reconstruction and identification efficiency in ATLAS using the full Run 2 *pp* collision data set at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J., C* 81 (2021), p. 578. arXiv: 2012.00578. URL: %5Curl%7Bhttps://cds.cern.ch/record/2746302%7D.

[62] ATLAS Collaboration. *Identification of hadronic tau lepton decays using neural networks in the ATLAS experiment*. Tech. rep. Geneva: CERN, 2019. URL: %5Curl%7Bhttps://cds.cern.ch/record/2688062%7D.

[63] ATLAS Collaboration. "Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV". In: *The European Physical Journal C* 78.11 (2018), p. 903. DOI: 10.1140/epjc/s10052-018-6288-9. URL: https://doi.org/10.1140/epjc/s10052-018-6288-9.

[64] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-$k_t$ jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[65] ATLAS Collaboration. "Performance of pile-up mitigation techniques for jets in $\sqrt{s} = 8$ TeV using the ATLAS detector". In: *The European Physical Journal C* 76.11 (2016), p. 581. DOI: 10.1140/epjc/s10052-016-4395-z. URL: https://doi.org/10.1140/epjc/s10052-016-4395-z.

[66] ATLAS Collaboration. *Variable Radius, Exclusive-$k_T$, and Center-of-Mass Subjet Reconstruction for Higgs($\to b\bar{b}$) Tagging in ATLAS*. Tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-010. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2268678.

[67] ATLAS Collaboration. "ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset". In: *The European Physical Journal C* 83.7 (2023), p. 681. DOI: 10.1140/epjc/s10052-023-11699-1. URL: https://doi.org/10.1140/epjc/s10052-023-11699-1.

[68] ATLAS Collaboration. "Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 96 (7 Oct. 2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. URL: https://link.aps.org/doi/10.1103/PhysRevD.96.072002.

[69] Jan Therhaag. "TMVA - Toolkit for Multivariate Data Analysis in ROOT". In: *PoS* ICHEP 2010 (2011), p. 510. DOI: 10.22323/1.120.0510.

[70] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[71] ATLAS Ccollaboration. *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC*. Tech. rep. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2064383.

[72] ATLAS Collaboration. *Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2261772.

[73] ATLAS Collaboration. "Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *The European Physical Journal C* 81.8 (2021), p. 689. URL: https://doi.org/10.1140/epjc/s10052-021-09402-3.

[74] ATLAS Collaboration. "New techniques for jet calibration with the ATLAS detector". In: *Eur. Phys. J. C* 83 (2023), p. 761. DOI: 10.1140/epjc/s10052-023-11837-9. arXiv: 2303.17312. URL: https://cds.cern.ch/record/2854733.

[75] ATLAS Collaboration. "In situ calibration of large-radius jet energy and mass in 13 TeV proton-proton collisions with the ATLAS detector". In: *Eur. Phys. J. C* 79.2 (2019), p. 135. DOI: 10.1140/epjc/s10052-019-6632-8. arXiv: 1807.09477 [hep-ex].

[76] ATLAS Collaboration. *Optimisation of the smoothing of b-jet identification efficiency and mistag rate simulation-to-data scale factors in ATLAS*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2710598.

[77] Gilles Louppe, Kyle Cranmer, and Juan Pavez. *carl: a likelihood-free inference toolbox*. Mar. 2016. DOI: 10.5281/zenodo.47798. URL: http://dx.doi.org/10.5281/zenodo.47798.

[78] S. Badger et al. *Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report*. 2016. arXiv: 1605.04692 [hep-ph].

[79] Nicolas Berger et al. *Simplified Template Cross-Sections - Stage 1.1*. 2019. arXiv: 1906.02754 [hep-ph].

[80] *Evaluation of theoretical uncertainties for simplified template cross-section measurements of $V$-associated production of the Higgs boson*. Tech. rep. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2649241.

[81] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph].

[82] LHC Higgs Cross Section Working Group. "Handbook of LHC Higgs Cross-Sections: 4. Deciphering the Nature of the Higgs Sector". In: 2/2017 (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[83] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011), p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. URL: https://doi.org/10.1140/epjc/s10052-011-1554-0.

[84] A L Read. "Presentation of search results: the CLs technique". In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313. URL: https://dx.doi.org/10.1088/0954-3899/28/10/313.

[85] ATLAS Collaboration. *Flavor Tagging Efficiency Parametrisations with Graph Neural Networks*. Tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2022-041. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2825433.

[86] Peter Battaglia et al. "Relational inductive biases, deep learning, and graph networks". In: *arXiv* (2018). URL: https://arxiv.org/pdf/1806.01261.pdf.

# Appendices

APPENDIX A

# Combined $VH(H \to b\bar{b}/c\bar{c})$ Analysis Appendix

*This Appendix lists some additional results in support of Chapter 1.*

## A.1 Flavour Tagging Calibrations

This section presents some results on the calibration of the flavour tagger used in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis. Figure A.1 shows some efficiency scale factors in the resolved regime for the different main flavours, while Figure A.2 displays the same information for the boosted tagger.

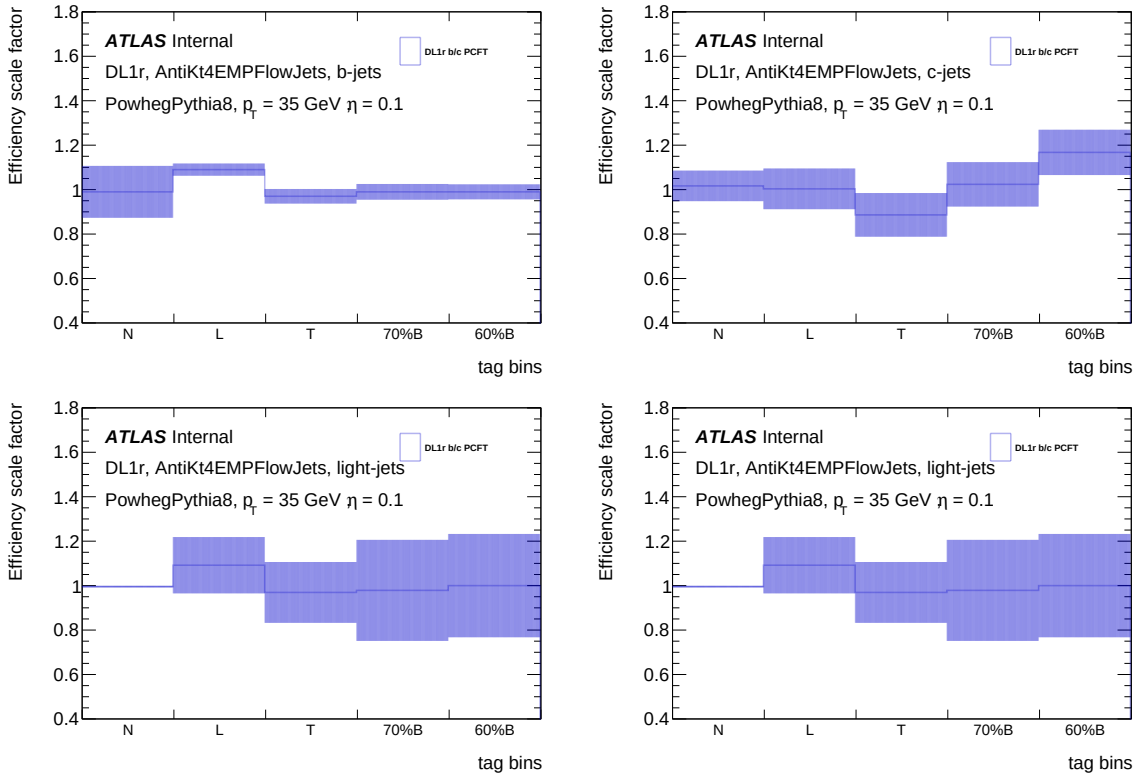



Figure A.1: Efficiency scale factors calibration results of the DL1r Pseudo-Continuous Flavour Tagging (PCFT) tagger on PowhegPythia8 $t\bar{t}$ samples for the resolved regime of the $VH(H \to b\bar{b}/c\bar{c})$. The scale factors of $\tau$-jets are from $c$-jet calibration. From the internal documentation.

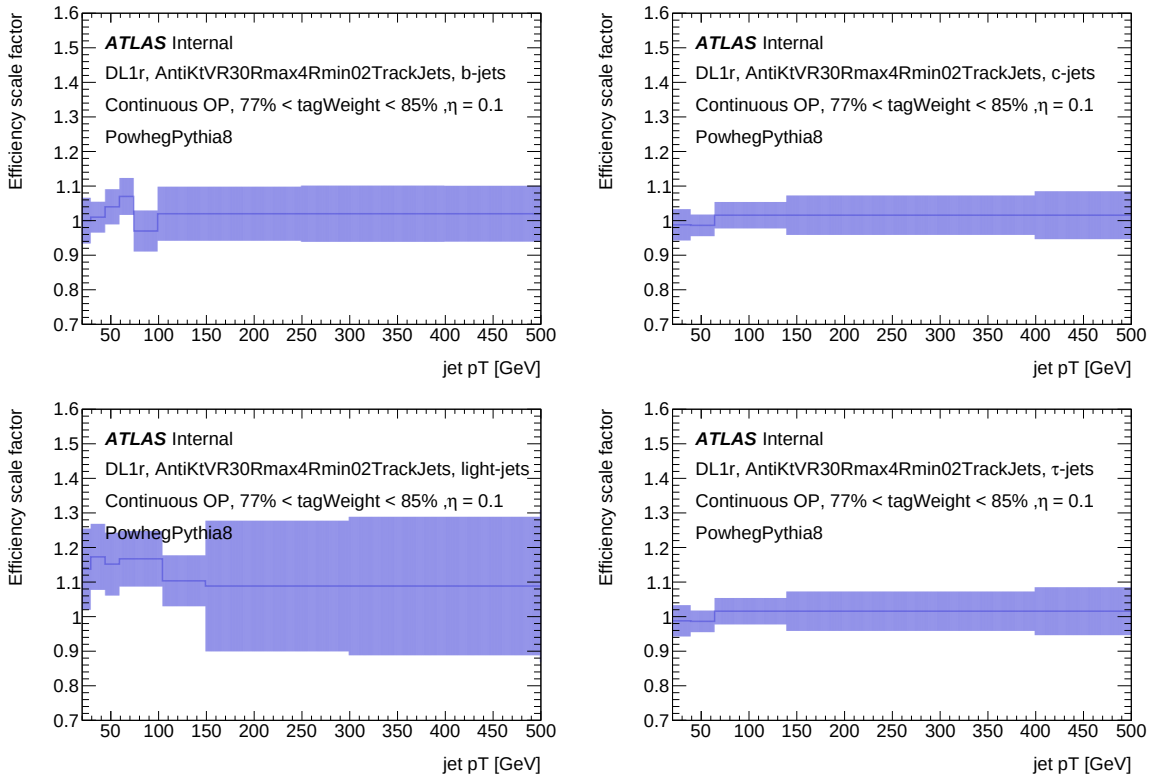



Figure A.2: Example of the boosted $VH(H \to b\bar{b})$ efficiency scale factor calibration results of the DL1r tagger on PowhegPythia8 $t\bar{t}$ samples for the $77\% < \epsilon_j < 85\%$ bin. The scale factors of $\tau$-jets are from $c$-jet calibration. From the internal documentation.

## A.2 Analysis Categorisation

This section offers more details in two elements of the categorisation in the resolved regime: the $\Delta R$ cut and the resolved top CR for the 0- and 1-lepton channels.

### A.2.1 The $\Delta R$ Cut Between Higgs Candidate 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 $\Delta R$. A *high* $\Delta R$ control region (High $\Delta R$ CR) is defined using parametrised cuts on $\Delta R$ between the Higgs candidate jets as a function of $p_T^V$. An additional *low* $\Delta R$ control region (Low $\Delta R$ CR) for the 1L channel in the resolved $VH(H \to b\bar{b})$ is also introduced (for $VH(H \to c\bar{c})$, it is merged with the signal region). 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. The cuts are defined by fitting a template function $c_1 \times e^{c_2 + c_3 \times p_T^V}$ to the $VH(H \to b\bar{b})$ selected events, so that:

- 95% (85%) of the $VH(H \to b\bar{b})$ 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 1.12, showing the signal yield in a 2-dimensional histogram ($p_T^V$ vs $\Delta R_{c\bar{c}}$) for different tags applied. Cuts derived on the $VH(H \to c\bar{c})$ selected events showed good agreement with the $VH(H \to b\bar{b})$ derived cuts. The $VH(H \to b\bar{b})$ cuts is chosen so that the kinematic selection of the two analyses is harmonised.

### A.2.2 Resolved Top Control Region in 0L and 1L

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 $t \to b + W$, 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 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.
- Single top $t$-quark: predominantly the $Wt$ process $W\ t \to W + b + W$, with one $W$ decaying leptonically and the other hadronically. Some of these background events can enter the signal region if the $b$- 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 $t \to b + W$ bring a smaller contribution, as the $c$-tagged jet must come from *Initial State Radiation* (ISR) or *Final State Radiation* (FSR) if the $b$ 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:

- 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 $VH(H \to c\bar{c})$ analysis, due to the 70% efficiency WP for $b$-tagging and the low mis-tag rate for $b$-jets in $c$-tagging. Naturally, in $VH(H \to b\bar{b})$ it is the leading contribution. Due to the origin of the candidate jets, a large $\Delta R_{b\bar{b}}$ is expected between the two $b$-jets so this component is most effectively constrained by the High $\Delta R$ CR.

- 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 $VH(H \to b\bar{b}/c\bar{c})$ analyses due to its signal-like kinematics.

- top($bl$): where $l$ stands for anything not $b$ nor $c$ (light jets predominantly but also some mis-tagged hadronic $\tau$). This component is similar to the top($bc$) as it also consists of a $b$ + a jet from the $W$ and can end up in the SRs and topCRs due to mis-tags.

- 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 $\Delta R$ regions nor the topCRs.

The signal region distributions in the 1L channel in the $p_T^V$ range $[150, 250]$ GeV are displayed in Section A.6 of the Appendix. 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.

The components contributing the most in the $VH(H \to c\bar{c})$ side of the analysis are the top($bc$) and top($bl$), due to the tagging requirement. There is very little top($bb$) thanks to the good performance of the tagger. Top($lq$) is mostly found in the looser tag regions (NT, LT) and not where the signal peaks. The philosophy behind the design of the top CR 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 the fit, only the $BT$ region is used, as it provides sufficient control on the important top background components. The $BL$ region is only used to validate the, assessing the data-MC agreement after correcting the yields of the major backgrounds, as is shown in Figure 1.28a.

For $VH(H \to c\bar{c})$, the $bc$ and $bl$ components are the most important to constrain. In $VH(H \to b\bar{b})$, 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 $\Delta R$ CR, since in a $t\bar{t}$ decay the two produced $b$-jets tend to be separated by a large $\Delta R$ due to the event topology. For the Combined Analysis, the SRs and CRs of both analyses are considered simultaneously. The High $\Delta R$ CR from $VH(H \to b\bar{b})$ are used to constrained the residual top($bb$) component in $VH(H \to c\bar{c})$.

### A.2.3 Truth Tagging

The tagging method described in Section 1.5, referred to as *direct tagging*, is a cut-based method where a jet passes or fails a threshold cut, as defined by dedicated working points in the PCFT or PCBT schemes. These WP have a large rejection for $b$-tagging due to the good performance of the method. For $c$-jets, the tagging efficiency is low and many $c$-jets end up rejected by the selection. This problem is compounded by the event selection criteria, requiring two $b$-tags or at least one tight $c$-tag to enter the analysis' regions. Only a part of the events in the simulated samples satisfy these requirements, and most are discarded from the analysis. Having sufficient MC statistics in all regions is essential to effectively model the backgrounds and reduce the MC statistical uncertainty. An alternative approach to direct tagging used in the analysis to retain the large MC statistics is *truth tagging*. Rather than applying a pass-fail decision, truth tagging reweighs events by their probability of being selected at a specific working point, based on truth information only available in the simulated samples. The tagging scale factors are applied in the analysis after truth tagging.

Mathematically, truth tagging derives a per event weight $w$ from the tagging efficiency $\epsilon_j(\mathbf{x}, \theta)$ for a given flavour jet $j$ to be tagged at a given working point of a classifier trained on a set of input variables $\mathbf{x}$, with the assumption that the efficiency is parametrisable as a function of several variables $\theta$, such as the jet $p_T$, $\eta$, ... For a set of $m$ jets with a tagged subset $T_i$ of cardinality $|T_i| = n$, and defining the efficiency at tagging the tagged jets as

$$\epsilon(T_i, \mathbf{x}, \theta) = \prod_{j \in T_i} \epsilon_j(\mathbf{x}, \theta),$$

and the efficiency at not tagging the set of untagged jets $\tilde{T}$, with $|\tilde{T}_i| = m - n$,

$$\epsilon_{in}(\tilde{T}_i, \mathbf{x}, \theta) = \prod_{j \in \tilde{T}_i} (1 - \epsilon_j(\mathbf{x}, \theta)),$$

the expression for $w$ can be factorised as [85]:

$$w = \sum_{i}^{C} \epsilon(T_i, \mathbf{x}, \theta) \cdot \epsilon_{in}(\tilde{T}_i, \mathbf{x}, \theta), \tag{A.1}$$

where the sum is over all possible permutations of tags $C$. The probability of a specific configuration $i$ is given by

$$P_i = \frac{\epsilon(T_i, \mathbf{x}, \theta) \cdot \epsilon_{in}(\tilde{T}_i, \mathbf{x}, \theta)}{w}.$$

When deploying the technique, one possible permutation is randomly sampled to keep distinct bins uncorrelated in the fit and the whole weight $w$ is applied to it.

Technically, truth tagging was deployed with map-based 2D histograms $p_T - \eta$ parametrising the tagging efficiency of the jets in the latest standalone $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analysis [28, 26]. Such histograms are called *efficiency map*, leading to the implementation being referred to *map-based truth tagging*. These maps were derived individually for each $b$-, $c$-, light-, and $\tau$-jets flavour and each working point. A further possibility is to combine direct tagging with truth tagging into the so-called *hybrid tagging* strategy, in which a portion of the events are direct tagged and the rest is truth tagged. This last approach limits the mis-modelling incurred by truth tagging and remove the need to correct for non-closure effects.

A new approach considered for the combined analysis relies on a GNN to perform the so-called *GNN truth tagging* [85]. This removes the statistical dispersion limitation of high-dimension efficiency maps. Interestingly, it also becomes possible to include more variables to parametrise the efficiency, leading to better agreement with the direct tagging distribution comparing to map-

| Jet features | Type of variable |
|---|---|
| Jet $p_T$ | Jet level feature |
| Jet $\eta$ | |
| Jet $\phi$ | |
| Jet flavour label | |
| Mass of $p_T$ leading b or c hadron in the jet $\phi$ | |
| $p_T$ of $p_T$ leading b or c hadron in the jet $\phi$ | |
| $\eta$ of $p_T$ leading b or c hadron in the jet $\phi$ | |
| $\phi$ of $p_T$ leading b or c hadron in the jet $\phi$ | |
| Average number of interactions per event $\langle\mu\rangle$ | Event level variable |
| Angular separation between two jets $\Delta R$ | Jet-pair variable |

Table A.1: The input features to parametrise the efficiency in GNN truth tagging.

based truth tagging. The network builds a fully-connected graph with several layers message-passing updates [86], where each node represents a jet in the event[1]. The features per node are the jet-level and event-level variables listed in Table A.1, with the angular separation between the jets set as edge between the nodes. Finally, a fully-connected Neural Network (NN) receives the last update graph and outputs all track-jets or jets flavour-tagging efficiencies.

In the combined analysis, truth tagging is deployed in all regimes and trained independently for samples with different MC generators[2], inclusively in all lepton channels. In the resolved regime, the training is further separated for each background samples. The GNN truth tagging is seen to improve the parametrisation of the efficiency, showing better closure with the direct-tagged distributions than the map-based approach. However, some unclosure remain for particular flavours. To limit this effect, hybrid tagging is also deployed in the combined analysis with GNN truth tagging. In this hybrid tagging, $b$-jets are direct tagged and other jets are GNN truth tagged in the resolved regime. In the boosted regime, all jets are truth tagged due to the limited MC statistics. The strategies deployed in the different regimes of the analysis are summarised in Table A.2.

| | $VH(H \to b\bar{b})$ Resolved | $VH(H \to c\bar{c})$ | $VH(H \to b\bar{b})$ Boosted |
|---|---|---|---|
| Hybrid tagging | Yes ($b$-jets are DT) | No (fully TT) | No (fully TT) |
| Truth tag WP | 70% $b$ & 70% $b$ | $c$-tight & $c$-tight | 85% $b$ & 85% $b$ |
| MC stat. % for TT regions | 100% | 8% | 100% |
| $V$+jets | HT | TT | TT |
| single-top $s/t$ | HT | TT | TT |
| single-top $Wt$ | DT | DT | TT |
| $t\bar{t}$ | DT | DT | TT |
| diboson | DT | DT | TT |
| signal | DT | DT | DT |

Table A.2: The tagging strategies to be used in the different regimes of the analysis, with truth tagging (TT), direct tagging (DT), and hybrid tagging (HT).

The tagging strategy is optimised to maximise the MC statistics of the different regions and boost the sensitivity. Truth and hybrid tagging are only deployed when they deliver a meaningful improvement to the analysis. The full tagging strategy of the analysis is:

- Resolved $VH(H \to b\bar{b})$: direct tagging is used except for the $V$+jets and single-top $s/t$

[1] Only central jets in the resolved regime and track-jets associated with the large-$R$ jet in the boosted regime.
[2] Since the Scale Factors (SF) are derived per generator.

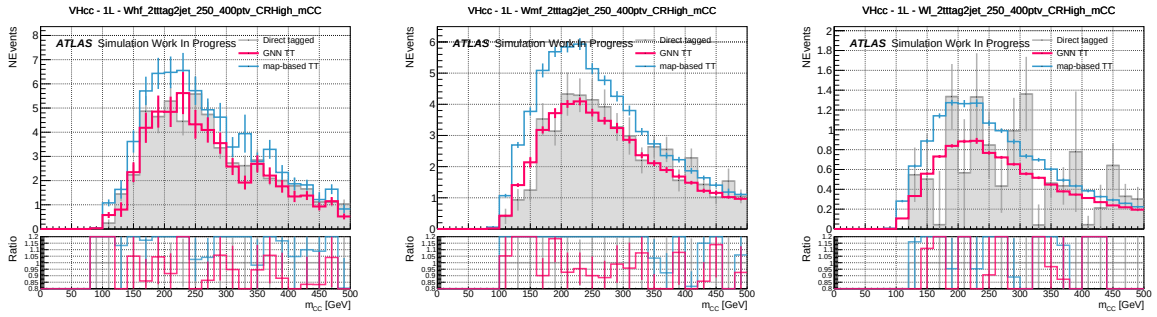
process where hybrid tagging is deployed, with both $b$-jets being direct tagged at the 70% WP.

- $VH(H \to c\bar{c})$: similar to the resolved $VH(H \to b\bar{b})$, with the $V$+jets and single-top $s/t$ now fully GNN truth tagged. For $VH(H \to c\bar{c})$, the samples are split based on the tag region to avoid reusing an event twice. For example, an initially $LN$ direct-tagged event could enter the $TT$ region with a low truth tag weight, thereby removing the statistical independence assumed between MC events. To correct this, only 8% of the MC statistics is randomly sampled and truth tagged to the $TT$-tag region, and the rest is passed to direct tagging (for the $TL$, $NT$, $LN$, and $BT$ tags).

- Boosted $VH(H \to b\bar{b})$: GNN truth tagging is applied for all background except the signal samples that are direct tagged.

Unfortunately, at the time of writing this thesis the analysis samples were not yet fully updated to the tagging scheme described here. Instead, the resolved $VH(H \to b\bar{b}/c\bar{c})$ all use direct tagging everywhere and the boosted regime uses full GNN truth tagging. Moving to the full tagging scheme outlined above should have a small positive effect on MC statistics uncertainty and bring smoother MC templates, reducing the noise in the fit.

To showcase the effectiveness of the method, the direct tagged, GNN truth tagged, and map-based truth tagged $m_{cc}$ distributions of the SHERPA 2.2.11 simulated $W$+jets in the 1-lepton 2-jet CRHigh $p_T^V \in [250, 400]$ GeV region of the $VH(H \to c\bar{c})$ is displayed in Figure A.3. The GNN truth tagging is found to be in better agreement with the direct tagged distributions in the regions of sufficient statistics. In the $W + l$ region, direct tagging leads to statistically depleted regions with large uncertainties: this is effectively corrected by the GNN-based truth tagging approach. No significant non-closures are observed for from GNN truth tagging with the outlined strategy.




(a) $W$+hf in CRHigh. (b) $W$+mf in CRHigh. (c) $W + l$ in CRHigh.

Figure A.3: Comparing the tagged $m_{cc}$ distribution for the $VH(H \to c\bar{c})$ of the SHERPA 2.2.11 simulated $W$+jets in 1L CRHigh 2-jet region, in the 250 GeV $< p_T^V <$ 400 GeV region. TT stands for truth tagging.

## A.3 MVA Variables

This section is dedicated to the set of variables used to train the various MVAs used in the analysis. Notice that the $H$ candidate is reconstructed by the selected jets sorted by $p_T$ and labelled $j_1$ and $j_2$.

**Input variables for the resolved regime:**

- $p_T^V$: transverse energy of the vector boson. In 0-lepton channel it is equivalent to the missing transverse energy ($E_T^{\text{miss}}$); in 1-lepton channel it is the vector sum of $E_T^{\text{miss}}$ and the lepton $p_T$; in 2-lepton channel, it is the vector sum of the 2 charged lepton $p_T$.
- $p_{\text{T}}^{j_1}$ and $p_{\text{T}}^{j_2}$: transverse momenta of the Higgs candidate jets. $j_1$ refers to the jet with higher $p_T$.
- $m_{j_1 j_2}$ or $m_J$: invariant mass of the reconstructed $H$ system, depending on the analysis regime.
- $\Delta R(j_1, j_2)$: angular distance between the two Higgs-candidate jets, defined as $\Delta R(i,j) = \sqrt{(\Delta\phi(i,j))^2 + (\Delta\eta(i,j))^2}$ with $\Delta\phi(i,j) = \phi_i - \phi_j$ the azimuthal and $\Delta\eta(i,j) = \eta_i - \eta_j$ the pseudorapidity distances.
- $m_{j_1 j_2 j_3}$: invariant mass of two Higgs-candidate jets and the remaining jet with highest $p_T$. When there are only 2 jets in an event, $m_{j_1 j_2 j_3} = m_{j_1 j_2}$.
- $\Delta\phi(V, H)$: azimuthal distance between the reconstructed vector boson $V$ and Higgs boson candidates $H$.
- $\text{bin}_{\text{DL1r}(j1)}$, $\text{bin}_{\text{DL1r}(j2)}$: variable showing the tagged-bin the jet or track-jet $j_1$ belongs to (5 possible bins, as defined in Section 1.5) - the untagged $N$, the loose (70% WP) and the tight (60% WP) $b$-tagged, and the loose and the tight $c$-tagged bins. In the MVA, the value of the two Higgs-candidate jets or track-jets are used.
- $\sum_{i \neq 1,2} p_T^{j_i}$:$p_T$ sum of non $H$ candidate jets that have $p_T > 20$ GeV.
- **0-lepton channel variables**:
  - $|\Delta\eta(j_1, j_2)|$: absolute value of the pseudorapidity distances between the two Higgs-candidate jets or track-jets.
  - $\min\{\Delta R(j_i, j)\}_{i=1,2}$: the distance in $R$ between the closest $b$- or $c$-tagged Higgs candidate jet and an additional jet with $p_T^V > 20$ GeV.
  - $m_{\text{eff}}$: the scalar sum of the $p_T$of all small-$R$ jets and $E_T^{\text{miss}}$ in the event.
- **1-lepton channel variables**:
  - $m_T^W$: transverse mass of the $W$ boson candidate reconstructed from the lepton and $E_T^{\text{miss}}$, as presented in the 1L-specific selection of Section 1.5.2.
  - $E_T^{\text{miss}}$: missing transverse energy.
  - $\Delta y(V, H)$: rapidity difference between the $V$ and $H$.
  - $\min\left[\Delta\phi(l, j_i)\right]_{i=1,2}$: distance in $\phi$ between the lepton and the closest $b$-tagged ($c$-tagged) $H$ candidate jet.
  - $m_{\text{top}}$: reconstructed mass of the leptonically decaying top quark. The longitudinal momentum of the neutrino ($p_z^\nu$) is first reconstructed the mass of the $W$ boson, and selected to minimise the reconstructed $m_{\text{top}}$ with the 2 Higgs candidates.
- **2-lepton channel variables**:

  - $m_{ll}$: invariant mass of the di-leptons system.
  - $\cos\theta(l^-, Z)$: $Z$ boson polarisation sensitive angle.
  - $E_T^{\text{miss}}/\sqrt{S_{\text{T}}}$: the quasi-significance of $E_T^{\text{miss}}$ with $S_{\text{T}}$ being the scalar sum of the $p_T$ of the leptons and jets in the event.
  - $\Delta y(V, H)$: rapidity difference between the vector boson and Higgs boson candidates.

**Input variables for the boosted regime:**

- $m_J$: leading-$R$ jet mass, the Higgs candidate.
- $p_T^V$: same as in the resolved regime.
- $p_T^{j_1}$, $p_T^{j_2}$ and $p_T^{j_3}$: transverse momenta of the track-jets inside the $H$ candidate large-$R$ jet, where $j_1$ and $j_2$ are the $b$-tagged sub-jets, and $j_3$ refers to the leading additional jet.
- $\Delta R(j_1, j_2)$: angular distance between the two $b$-tagged track-jets.
- $N$(track-jets in $J$): the number of track-jets that are associated to the leading large-$R$ jet.
- $N$(add. small $R$-jets): the number of additional small-$R$ jets that are not associated to the leading large-$R$ jet, such that $\Delta R$(small-$R$ jet, large-$R$ jet) $> 1.0$.
- $\Delta\phi(V, H)$: same as in the resolved regime.
- Colour: variable exploitting the difference in colour-flow between gluon splittings and decay from glsqcd singlets states. Colour is defined here as

$$\text{Colour} = \frac{\theta^2_{j_1 j_3} + \theta^2_{j_2 j_3}}{\theta^2_{j_1 j_2}},$$

  where $\theta$ is the angle between the indexed jets, $j_3$ is the leading additional jet, and $j_2$ are the $H$ candidate jets.

- $\text{bin}_{\text{DL1r}(j1,\text{trk})}$, $\text{bin}_{\text{DL1r}(j2,\text{trk})}$: corresponds to the tagged-bin the track-jet belongs to (4 possible bins): the 85%, the 77%, the 70% and 60% $b$-tagging efficiency bins.
- **0-lepton channel specific variables**
  - $E_T^{\text{miss}}$: missing transverse energy, same as $p_T^V$.
- **1-lepton channel specific variables**
  - $\Delta y(\boldsymbol{V}, \boldsymbol{H})$: same as in the resolved regime.
  - $p_T^l$: transverse momentum of the lepton.
  - $(p_T^l - E_T^{\text{miss}})/p_T^W$: proxy for the $p_T$ imbalance of the charged lepton and the neutrino of the $W$-boson.
- **2-lepton channel specific variables**
  - $\Delta y(V, H)$: same as in the resolved regime.
  - $\cos\theta(l^-, \mathbf{Z})$: same as in the resolved regime.

## A.4 Top Modelling Uncertainties in the Fit

There are many processes of relevance in a complex analysis such as the $VH(H \to b\bar{b}/c\bar{c})$. These must individually be modelled, with studies of the pulls of the different systematics required to verify the fit correctly accounts for the background's contribution to the analysis. The risk with such a complex fit structure with large numbers of NPs necessary to model a large variety of effects is to give the fit too much freedom and, in a sense, overfit to the data distributions. To highlight the process, some top-related pulls are shown in Figure **??**, with pulls displayed for both top acceptance uncertainties and CARL shape systematics. Again, there is good agreement for most pulls between the $VH$ and $VZ$ analyses. In the acceptance systematics part, a very large significant pull is observed for the so-called "*MetTrigTop*", an $E_T^{\text{miss}}$ trigger related experimental uncertainty derived from the top-process, as described in 1.8.

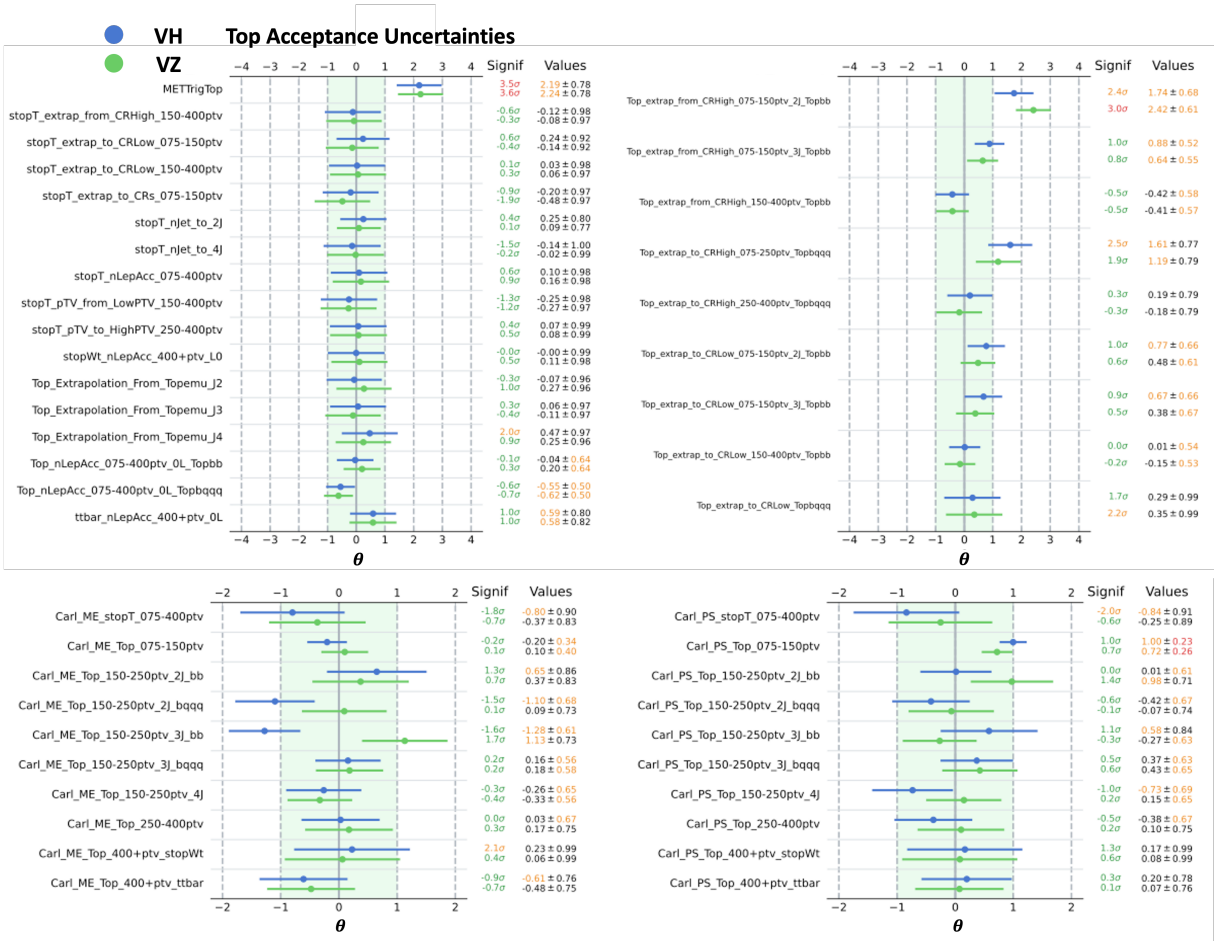



Figure A.4: Some Top nuisance parameters related to acceptance uncertainties (top) and CARL shapes (bottom) in the combined analysis targeting the $VH(H \to b\bar{b}/c\bar{c})$ in blue, versus the cross-check analysis $VZ(\to b\bar{b}/c\bar{c})$ in green.

Other uncertainties presented cover the region extrapolation for the single-top $t$ (left) and combined Top process (right), as well as $N_{\text{jet}}$ and $p_T^V$ (for single-top $t$), lepton channel extrapolations, and the extrapolation from the Top $e/\mu$ CR. Most of the NPs are not signficantly pulled, with little constraining. This indicates that the fit is not very sensitive nor requires the effect they implement. One exception is the Top($bb$) extrapolation from the CRHigh in 75 GeV $< p_T^V <$ 150 GeV with 2-jet: the NP is largely pulled, and even more so in the $VZ$ cross-check analysis. This feature can be understood from the large presence of $V$+jets and Top in the 0L and 1L CRHigh, leading to an interplay between the two processes when shifting the focus towards the signal part of $V$+jets. This interplay is visible in the correlation of the boosted $t\bar{t}$ and $W$+$hf$ in Figure 1.31.

| Acceptance Ratio Name | Applied | Value |
|---|---|---|
| $Z$+$hf$ normalistion | $Z$+$hf$ | floating |
| $Z$+$mf$ normalistion | $Z$+$mf$ | floating |
| $Z$+$lf$ normalistion | $Z$+$lf$ | floating |
| $Zcc/Zbb$ ratio | $Zcc$ | 12% |
| $Zcc/Zbb$ ratio | $Zcc$, $VH(H \to b\bar{b})$, 2-jet | 8% |
| $Zbl/Zbc$ ratio | $Zbl$ | 4% |
| $Zbc/Zcl$ ratio | $Zbc$ | 10% |
| $Z$+$hf$ SR/CR ratio | $Z$+$hf$, 2L, SR, $p_T^V$ 75-150 | 7% |
| $Z$+$hf$ SR/CR ratio | $Z$+$hf$, 2L, SR, $p_T^V$ >150 | 15% |
| $Z$+$hf$ SR/CR ratio | $Z$+$hf$, 0L, SR, TopCR, $p_T^V$ >150 | 10% |
| $Z$+$hf$ SR/CR ratio | $Z$+$hf$, 02L, SR, TopCR, $p_T^V$ >250, 2-jet | 30% |
| $Z$+$mf$ SR/CR ratio | $Z$+$mf$, 2L, SR, $p_T^V$ 75-150 | 7% |
| $Z$+$mf$ SR/CR ratio | $Z$+$mf$, 0L, SR, $p_T^V$ >150 | 5% |
| $Z$+$lf$ CR/SR ratio | $Z$+$lf$, 2L, CRLow, $p_T^V$ 75-150 | 7% |
| $Z$+$lf$ CR/SR ratio | $Z$+$lf$, 0L, CRHigh, $p_T^V$ >150 | 5% |
| $Z$+$hf$ 0L/2L ratio | $Z$+$hf$, 0L, 2-jet | 2% |
| $Z$+$hf$ 0L/2L ratio | $Z$+$hf$, 0L, 3-jet | 4% |
| $Z$+$hf$ 0L/2L ratio | $Z$+$hf$, $VH(H \to b\bar{b})$ 0L, 4-jet | 8% |
| $Z$+$mf$ 0L/2L ratio | $Z$+$hf$, 0L, 2-jet | 3% |
| $Z$+$mf$ 0L/2L ratio | $Z$+$mf$, 0L, 3-jet | 8% |
| $Z$+$lf$ 0L/2L ratio | $Z$+$lf$, 0L, 2-jet | 4% |
| $Z$+$lf$ 0L/2L ratio | $Z$+$lf$, 0L, 3-jet | 10% |

Table A.3: $Z$+jets acceptance uncertainties in the resolved regime.

## A.5 Signal and Background Modelling

Additional information on the signal and background modelling are given in this section. Tables A.3 and A.5 list the different acceptance uncertainties for the $Z$+jets and $W$+jets respectively in the resovled regime. Tables A.4 and A.6 present $V$+jets uncertainties in the boosted regime. The top-related uncertainties are detailed in Table A.7 for the resolved regime and Table A.8 for the boosted regime, while the single-top $t$ is described in Tables A.9 and A.10. The diboson uncertainties are described in Tables A.11 and A.12.

| Acceptance Ratio Name | Applied | Value |
|---|---|---|
| $Z$+$hf$ normalistion | $Z$+$hf$ | floating |
| $Z$+$mf$ normalistion | $Z$+$hf$ | 35% |
| $Z$+$lf$ normalistion | $Z$+$hf$ | 35% |
| $Zcc/Zbb$ ratio | $Zcc$ in 02L | 6% |
| $Zbl/Zbc$ ratio | $Zbl$ in 02L | 6% |
| $Zcl/Zbc$ ratio | $Zcl$ in 02L | 6% |
| $Z$+$hf$ TopCR/SR ratio | $Z$+$hf$, 0L, TopCR | 15% |
| $Z$+$mf$ TopCR/SR ratio | $Z$+$mf$, 0L, TopCR | 25% |
| 0L / 2L ratio | $Z$+$hf$& $Z$+$mf$, 0L | 3% |
| $p_T^V$ 600 / 400-600 ratio | $Z$+$hf$& $Z$+$mf$, 0L & 2L | 15% |

Table A.4: $Z$+jets acceptance uncertainties in the boosted regime.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| $W$+$hf$ normalistion | $W$+$hf$ | floating |
| $W$+$mf$ normalistion | $W$+$mf$ | floating |
| $W$+$lf$ normalistion | $W$+$lf$ | floating |
| $W$+$lf$ normalistion | $W$+$lf$1L $p_T^V$ 150-250 | 25% |
| $Wcc/Wbb$ ratio | $Wcc$, 1L $p_T^V$ 75-150 | 20% |
| $Wcc/Wbb$ ratio | $Wcc$, 1L $p_T^V$ >150, 2-jet | 4% |
| $Wcc/Wbb$ ratio | $Wcc$, 1L $p_T^V$ >150, 3-jet | 15% |
| $Wcc/Wbb$ ratio | $Wcc$, $VH(H \to b\bar{b})$, 0L, 2-jet | 4% |
| $Wcc/Wbb$ ratio | $Wcc$, $VH(H \to b\bar{b})$, 0L, 3-jet | 10% |
| $Wcc/Wbb$ ratio | $Wcc$, $VH(H \to b\bar{b})$, 0L, 4-jet | 10% |
| $Wcc/Wbb$ ratio | $Wcc$, $VH(H \to c\bar{c})$, 0L | 25% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ 75-150 | 24% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ 150-250, 2-jet | 24% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ 150-250, 3-jet | 14% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ >250 | 14% |
| $Wbl/Wcl$ ratio | $Wbl$, $p_T^V$ 75-150 | 29% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ 150-250, 2-jet | 29% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ 150-250, 3-jet | 22% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ >250, 2-jet | 19% |
| $Wbc/Wcl$ ratio | $Wbc$, $p_T^V$ >250, 3-jet | 12% |
| $Wbc/Wcl$ ratio | $Wbc$, 0L, 4-jet | 8% |
| $Wq\tau/Wcl$ ratio | $Wc\tau$, $Wb\tau$ | 20% |
| $Wl\tau/Wcl$ ratio | $Wl\tau$, $W\tau\tau$ | 9% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ 75-150, 2-jet | 3% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ 75-150, 3-jet | 7% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ 150-250, 2-jet | 30% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ 150-250, 3-jet | 10% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ >250, 2-jet | 50% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 1L, CRHigh, $p_T^V$ >250, 3-jet | 20% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 0L, CRHigh, 2-jet | 30% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 0L, CRHigh, 3-jet | 20% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 0L, CRHigh, $p_T^V$ 150-250, 4-jet | 10% |
| $W$+$hf$ CRHigh / SR+CRLow ratio | $W$+$hf$, 0L, CRHigh, $p_T^V$ >250, 4-jet | 15% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ 75-150, 2-jet | 33% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ 75-150, 3-jet | 3% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ 150-250, 2-jet | 65% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ 150-250, 3-jet | 7% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ >250, 2-jet | 20% |
| $W$+$hf$ SR / CRLow ratio | $W$+$hf$, 1L, SR, topCR, $p_T^V$ >250, 3-jet | 13% |
| $W$+$mf$ CRHigh / SR ratio | $W$+$mf$, 1L, CRHigh, CRLow, $p_T^V$ 75-150, 2-jet | 2% |
| $W$+$mf$ CRHigh / SR ratio | $W$+$mf$, 1L, CRHigh, CRLow, $p_T^V$ 75-150, 3-jet | 5% |
| $W$+$mf$ SR / CRHigh ratio | $W$+$mf$, 01L, SR, topCR, CRLow $p_T^V$ 150-250 | 7% |
| $W$+$mf$ SR / CRHigh ratio | $W$+$mf$, 01L, SR, topCR, CRLow $p_T^V$ >250 | 16% |
| $W$+$lf$ CRHigh / SR ratio | $W$+$lf$, 1L, CRHigh, $p_T^V$ 75-150, 2-jet | 5% |
| $W$+$lf$ CRHigh / SR ratio | $W$+$lf$, 1L, CRHigh, $p_T^V$ 75-150, 3-jet | 10% |
| $W$+$lf$ CRHigh / SR ratio | $W$+$lf$, 01L, CRHigh, $p_T^V$ 150-250, 2-jet | 5% |
| $W$+$lf$ CRHigh / SR ratio | $W$+$lf$, 01L, CRHigh, $p_T^V$ 150-250, 3-jet | 10% |
| $W$+$lf$ CRHigh / SR ratio | $W$+$lf$, 01L, CRHigh, $p_T^V$ >250, 2-jet, 3-jet | 17% |
| $W$+$hf$ 4-jet / 3-jet ratio | $W$+$hf$, 0L, $p_T^V$ 150-250, 4-jet | 12% |
| $W$+$hf$ 4-jet / 3-jet ratio | $W$+$hf$, 0L, $p_T^V$ >250, 4-jet | 20% |
| $W$+$hf$ 0L / 1L ratio | $W$+$hf$, 0L, $p_T^V$ 150-250, 2-jet | 30% |
| $W$+$hf$ 0L / 1L ratio | $W$+$hf$, 0L, $p_T^V$ 150-250, 3(+)-jet | 20% |
| $W$+$hf$ 0L / 1L ratio | $W$+$hf$, 0L, $p_T^V$ >250, 2-jet | 20% |
| $W$+$hf$ 0L / 1L ratio | $W$+$hf$, 0L, $p_T^V$ >250, 3(+)-jet | 13% |
| $W$+$mf$ 0L / 1L ratio | $W$+$mf$, 0L, $p_T^V$ 150-250, 2-jet | 3% |
| $W$+$mf$ 0L / 1L ratio | $W$+$mf$, 0L, $p_T^V$ 150-250, 3-jet | 8% |
| $W$+$mf$ 0L / 1L ratio | $W$+$mf$, 0L, $p_T^V$ >250 | 10% |
| $W$+$lf$ 0L / 1L ratio | $W$+$lf$, 0L | 4% |

Table A.5: The $W$+jets acceptance uncertainties in the resolved regime.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| $W$+$hf$ normalistion | $W$+$hf$ | floating |
| $W$+$mf$ normalistion | $W$+$hf$ | 36% |
| $W$+$lf$ normalistion | $W$+$hf$ | 38% |
| $Wcc/Wbb$ ratio | $Wcc$ | 11% |
| $Wcl/Wbc$ ratio | $Wcl$ | 15% |
| $Wbl/Wbc$ ratio | $Wbl$ | 9% |
| $W$+$hf$ TopCR / SR ratio | $W$+$hf$, 0L & 1L, TopCR | 27% |
| $W$+$mf$ TopCR / SR ratio | $W$+$mf$, 0L & 1L, TopCR | 20% |
| $W$+$lf$ TopCR / SR ratio | $W$+$lf$, 0L & 1L, TopCR | 16% |
| 0L / 1L ratio | All, 0L | 20% |
| $p_T^V$ >600 / 400-600 GeV ratio | $W$+$mf$& $W$+$lf$, 0L & 1L | 3% |

Table A.6: The $W$+jets acceptance uncertainties in the boosted regime.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| Top($bb$) normalisation | 0L & 1L, decorr in $N_{\text{jet}}$& $p_T^V$ | floating |
| Top($bb$) normalisation | $VH(H \to c\bar{c})$ $e\mu$ CR 2L | floating |
| Top($bq/qq$) normalisation | 0L & 1L, decorr in $N_{\text{jet}}$ & $p_T^V$ | floating |
| Top $bl$ / $bc$ Ratio | 01L, Top($bl$) | 5 % |
| Top $qq$ / $bc + bl$ ratio | 01L, Top($qq$) | 10 % |
| Top($bb$) CRLow+SR / CRHigh ratio | 01L, CRLow, SR, TopCR, Top($bb$) | 2 % (75-250 GeV) 8 % (250-400 GeV) |
| Top($bb$) CRLow / SR ratio | $VH(H \to b\bar{b})$ 1L, CRLow, Top($bb$) | 2.5 % (75-150 GeV) 9 % (150-400 GeV) |
| Top($bq/qq$) CRHigh / CRLow+SR ratio | 01L, CRHigh, Top($bq/qq$) | 4 % (75-250 GeV) 10 % (250-400 GeV) |
| Top($bq/qq$) CRLow / SR ratio | $VH(H \to b\bar{b})$ 1L, CRLow, Top($bq/qq$) | 2.5 % (75-250 GeV) 4 % (250-400 GeV) |
| Top SR / Top $e\mu$ CR | $VH(H \to b\bar{b})$ 2L | 0.8% |
| $Wt$ / $t\bar{t}$ ratio | 0L, $Wt$($bb$) | 22 % (150-250 GeV) 48 % (250-400 GeV) |
| $Wt$ / $t\bar{t}$ ratio | 1L, $Wt$($bb$) | 15 % (75-150 GeV) 13 % (150-400 GeV) |
| $Wt$ / $t\bar{t}$ ratio | 01L, $Wt$($bq/qq$) | 12 % (75-250 GeV) 18 % (250-400 GeV) |
| Top 0L / 1L ratio | 0L | 2 % (150-250 GeV) 8 % (250-400 GeV) |
| CARL ME Top shape | 01L | – |
| CARL PS Top shape | 01L | – |
| $Wt$ DS/DR shape + normalisation | $Wt$, 01L | – |
| ISR Top shape | 01L | – |
| FSR Top shape | 01L | – |

Table A.7: Resolved regime top ($t\bar{t}$ + $Wt$) uncertainties.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| $t\bar{t}$ normalisation | $t\bar{t}$, 01L, decorr. in $p_T^V$ | floating |
| $t\bar{t}$ normalisation | $t\bar{t}$, 2L | 20% |
| $Wt$ normalisation | $t\bar{t}$, 012L | 25% |
| $t\bar{t}$ SR / TopCR ratio | $t\bar{t}$, 01L, SR | 10% |
| $t\bar{t}$ 0L / 1L ratio | $t\bar{t}$, 0L | 6% (400-600 GeV) 20% (600+ GeV) |
| $Wt$ 0L / 1L ratio | $t\bar{t}$, 0L | 20% (400-600 GeV) 40% (600+ GeV) |
| $Wt$ $p_T^V$ >600 / 400-600 GeV ratio | $Wt$, 01L 400-600 GeV | 20% |
| CARL ME $t\bar{t}$ shape | $t\bar{t}$, 01L | – |
| CARL PS $t\bar{t}$ shape | $t\bar{t}$, 01L | – |
| CARL ME $wt$ shape | $t\bar{t}$, 01L | – |
| CARL PS $wt$ shape | $t\bar{t}$, 01L | – |
| ISR $t\bar{t}$ shape | $t\bar{t}$, 01L | – |
| FSR $t\bar{t}$ shape | $t\bar{t}$, 01L | – |
| ISR $wt$ shape | $t\bar{t}$, 01L | – |
| FSR $wt$ shape | $t\bar{t}$, 01L | – |

Table A.8: Boosted regime $t\bar{t}$ and $Wt$ uncertainties.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| stop-$t$ normalisation | 01L, all regions | 17 % |
| stop-$t$ CRLow+CRHigh / SR ratio | 1L, 75-150 GeV, CRHigh and CRLow | 3 % |
| stop-$t$ CRLow / CRHigh ratio | 1L, 75-150 GeV, CRLow | 6 % |
| stop-$t$ CRLow+SR / CRHigh ratio | 01L, SR, TopCR, CRLow, decorr. 150-250 and 250-400 GeV | 6 % |
| stop-$t$ CRLow / SRratio | 01L, CRLow, decorr. 150-250 and 250-400 GeV | 17 % |
| stop-$t$ 2-jet / 3-jet ratio | 01L, 2-jet region | 15 % |
| stop-$t$ 4-jet / 2+3-jet ratio | 0L, 4-jet region | 15 % |
| stop-$t$ $p_T^V$150-400 / 75-150 ratio | 01L, decorr. 150-250 and 250-400 GeV | 7 % |
| stop-$t$ $p_T^V$250-400 / 150-250 ratio | 01L, 250-400 GeV | 15 % |
| stop-$t$ 0L / 1L | 0L | 6 % |
| CARL ME stop-$t$ shape | 01L | – |
| CARL PS stop-$t$ shape | 01L | – |
| ISR stop-$t$ shape | 01L | – |
| FSR stop-$t$ shape | 01L | – |

Table A.9: Resolved regime single-top $t$ (stop-$t$) uncertainties. The single-top $s$ is applied a global 4.6% normalisation.

| **Acceptance Ratio Name** | **Applied** | **Value** |
|---|---|---|
| stop-$t$ normalisation | 01L, all regions | 10 % |
| ISR stop-$t$ shape | 01L | – |
| FSR stop-$t$ shape | 01L | – |

Table A.10: Boosted regime single-top $t$ (stop-$t$) uncertainties.

| Acceptance Ratio Name | Production mode | Decay component | Value & Application |
|---|---|---|---|
| $ZZ$ normalisation | $qqZZ$ | All | 17% |
| $WZ$ normalisation | $qqWZ$ | All | 19% |
| $WW$ normalisation | $qqWW$ | All | 16% |
| $ggVV$ normalisation | $ggVV$ | All | 30% |
| $ZZ$ CRHigh / SR ratio | $qqZZ$ | $VZbb$, $VZcc$ | 20% (0L) & 12%-20% (2L) |
| $WZ$ CRHigh / SR ratio | $qqWZ$ | $VZbb$, $VZcc$ | 12% (0L) & 13%-20% (1L) |
| $WZ$ CRLow / SR+CRHigh ratio | $qqWZ$ | $VZbb$, $VZcc$ | 50%-18% in 1L |
| $WW$ CRHigh / SR ratio | $qqWW$ | $VW$bkg | 10% (0L) & 16% (1L) |
| $W_{\text{had}}Z_{\text{lep}}$ CRHigh / SR ratio | $qqWZ$ | $VW$bkg | 14%-12%-17% in 0-1-2L |
| $W_{\text{lep}}Z_{\text{had}}$ CRHigh / SR ratio | $qqZZ$ | $VZ$bkg | 10% (0L) & 11 % (1L) |
| $ZZ$bkg CRHigh / SR ratio | $qqWZ$ | $VZbb$, $VZ$bkg | 6% (0L) & 7% (2L) |
| $ZZ$ 3-jet / 2-jet ratio | $qqZZ$ | $VZbb$, $VZcc$ | 10% in 02L |
| $WZ$ 3-jet / 2-jet ratio | $qqWZ$ | $VZbb$, $VZcc$ | 22% in 01L |
| $ZZ$ 4-jet / 3-jet ratio | $qqZZ$ | $VZbb$, $VZcc$ | 16% (0L) & 30% (2L) |
| $WZ$ 4-jet / 3-jet ratio | $qqWZ$ | $VZbb$, $VZcc$ | 16% in 0L |
| $WW$ 3p-jet / 2-jet ratio | $qqWW$ | $VW$bkg | 12% in 01L |
| $W_{\text{had}}Z_{\text{lep}}$ 3p-jet / 2-jet ratio | $qqWZ$ | $VW$bkg | 13%-10%-24% in 0L-1L-2L |
| $W_{\text{lep}}Z_{\text{had}}$ 3p-jet / 2-jet ratio | $qqWZ$ | $VZ$bkg | 14% in 0L & 11% in 1L |
| $ZZ$bkg 3-jet / 2-jet ratio | $qqZZ$ | $VZ$bkg | 10% (0L) & 13% (2L) |
| $W_{\text{lep}}Z_{\text{had}}$ 4p-jet / 3-jet ratio | $qqWZ$ | $VZ$bkg | 14% (0L) |
| $W_{\text{had}}Z_{\text{lep}}$ 4p-jet / 3-jet ratio | $qqWZ$ | $VW$bkg | 13% (0L) & 37% (2L) |
| $ZZ$bkg 4p-jet / 3-jet ratio | $qqZZ$ | $VZ$bkg | 10% (0L) & 42% (2L) |
| $ZZ$ 0L / 2L ratio | $qqZZ$ | $VZbb$, $VZcc$ | 2%-3.5%-23% in 2-, 3-, 4-jet 0L |
| $W_{\text{had}}Z_{\text{lep}}$ 0L / 2L ratio | $qqWZ$ | $VW$bkg | 10% in 0L |
| $ZZ$bkg 0L / 2L ratio | $qqZZ$ | $VZ$bkg | 13% in 0L |
| $WZ$ 0L / 1L ratio | $qqWZ$ | $VZbb$, $VZcc$ | 4%-10% in 2-, 3-jet 0L |
| $WW$ 0L / 1L ratio | $qqWW$ | $VW$bkg | 6% in 0L |
| $W_{\text{lep}}Z_{\text{had}}$ 0L / 1L ratio | $qqWZ$ | $VZ$bkg | 4% in 0L |
| $ZZ$ $p_T^V$ 250-400 / 150-250 ratio | $qqZZ$ | $VZbb$, $VZcc$ | 3%-9% in 02L |
| $ZZ$ $p_T^V$ 75-150 / 150-250 ratio | $qqZZ$ | $VZbb$, $VZcc$ | 6% in 2L |
| $WZ$ $p_T^V$ 250-400 / 150-250 ratio | $qqWZ$ | $VZbb$, $VZcc$ | 4%-16% (0L) & 4% (1L) |
| $WZ$ $p_T^V$ 75-150 / 150-250 ratio | $qqWZ$ | $VZbb$, $VZcc$ | 2%-5% in 1L |
| $WW$ $p_T^V$ 250-400 / 150-250 ratio | $qqWW$ | $VW$bkg | 7% in 1L |
| $WW$ $p_T^V$ 75-150 / 150-250 ratio | $qqWW$ | $VW$bkg | 7% in 1L |
| $W_{\text{had}}Z_{\text{lep}}$ $p_T^V$ 75-150 / 150-250 ratio | $qqWZ$ | $VW$bkg | 4% in 12L |
| $W_{\text{lep}}Z_{\text{had}}$ $p_T^V$ 75-150 / 150-250 ratio | $qqWZ$ | $VZ$bkg | 5% (1L) |
| $ZZ$bkg $p_T^V$ 150-250 / 75-150 ratio | $qqZZ$ | $VZ$bkg | 4% 2L |
| $WW$ $p_T^V$ 250-400 / 150-250 ratio | $qqWW$ | $VW$bkg | 7% in 01L |
| $W_{\text{had}}Z_{\text{lep}}$ $p_T^V$ 250-400 / 150-250 ratio | $qqWZ$ | $VW$bkg | 10% in 012L |
| $W_{\text{lep}}Z_{\text{had}}$ $p_T^V$ 250-400 / 150-250 ratio | $qqWZ$ | $VZ$bkg | 9% in 012L |
| $ZZ$bkg $p_T^V$ 250-400 / 150-250 ratio | $qqZZ$ | $VZ$bkg | 8% in 02L |
| QCD scale ZZ $p_T^V$ 150-400 / 75-150 | $qqZZ$ | $VZbb$, $VZcc$ | -3.2% to 7.8% in 12L |
| QCD scale WZ $p_T^V$ 150-400 / 75-150 | $qqWZ$ | $VZbb$, $VZcc$ | -3.1% to 5.8% in 12L |
| QCD scale ZZ $p_T^V$ 250-400 / 150-250 | $qqZZ$ | $VZbb$, $VZcc$ | -2.4% to 8.4% |
| QCD scale WZ $p_T^V$ 250-400 / 150-250 | $qqWZ$ | $VZbb$, $VZcc$ | -1.6% to 7.9% |
| QCD scale ZZ 3(p)-jet / 2-jet | $qqZZ$ | $VZbb$, $VZcc$ | -35.6% to 19.9% |
| QCD scale WZ 3(p)-jet / 2-jet | $qqWZ$ | $VZbb$, $VZcc$ | -37.4% to 16.2% |
| QCD scale ZZ 4(p)-jet / 3-jet | $qqZZ$ | $VZbb$, $VZcc$ | -30% to 32% in 02L |
| QCD scale WZ 4(p)-jet / 3-jet | $qqWZ$ | $VZbb$, $VZcc$ | -14.7% to 23.2% in 0L |
| Carl $ZZ$ PwPy8 / Sh2211 | $qqZZ$ | All | 02L |
| Carl $ZZ$ Sh221 / Sh2211 | $qqZZ$ | All | 02L |
| Carl $WZ$ PwPy8 / Sh2211 | $qqZZ$ | All | 01L (2L in $VH(H \to c\bar{c})$ only) |
| Carl $WZ$ Sh221 / Sh2211 | $qqZZ$ | All | 01L (2L in $VH(H \to c\bar{c})$ only) |
| Carl $WW$ PwPy8 / Sh2211 | $qqZZ$ | All | 01L in $VH(H \to c\bar{c})$ |
| Carl $WW$ Sh221 / Sh2211 | $qqZZ$ | All | 01L in $VH(H \to c\bar{c})$ |
| QCD scale largest shape | qqVV | All | Inclusive region in 12L |
| EW largest shape | qqVV | All | Inclusive region in 12L |

Table A.11: Diboson uncertainties in the resolved regime.

| **Acceptance Ratio Name** | **Production mode** | **Decay component** | **Value & Application** |
|---|---|---|---|
| $ZZ$ normalisation | $qqZZ$ | All | 17% |
| $WZ$ normalisation | $qqWZ$ | All | 27% |
| $WW$ normalisation | $qqWW$ | All | 16% |
| $ggVV$ normalisation | $ggVV$ | All | 30% |
| $ZZ$ LP / HP ratio | $qqZZ$ | $VZbb$, $VZcc$ | 10% in 0L LP |
| $WZ$ LP / HP ratio | $qqWZ$ | $VZbb$, $VZcc$ | 15% in 01L LP |
| $ZZ$ 0L / 2L ratio | $qqZZ$ | $VZbb$ | 7% |
| $ZZ$bkg 0L / 2L ratio | $qqZZ$ | $VZ$bkg | 10% |
| $W_{\text{had}}Z_{\text{lep}}$bkg 0L / 2L ratio | $qqWZ$ | $VW$bkg | 10% |
| $ZZ$ 0L / 2L ratio | $qqZZ$ | $VZbb$ | 7% |
| $WZ$ 0L / 1L ratio | $qqWZ$ | $VZbb$ | 7% |
| $WW$ 0L / 1L ratio | $qqWZ$ | $qqWW$ | 10% |
| $W_{\text{lep}}Z_{\text{had}}$bkg 0L / 1L ratio | $qqWZ$ | $VZ$bkg | 10% |
| $ZZ$ $p_T^V$ >600 / 400-600 ratio | $qqZZ$ | $ZZbb$, $ZZcc$ | 8% in 02L L |
| $WZ$ $p_T^V$ >600 / 400-600 ratio | $qqWZ$ | $VZbb$, $VZcc$ | 40% (0L) - 7% (1L) |
| $WW$ $p_T^V$ >600 / 400-600 ratio | $qqWW$ | $qqWW$ | 10% in 01L |
| $W_{\text{lep}}Z_{\text{had}}$bkg $p_T^V$ >600 / 400-600 ratio | $qqWZ$ | $VZ$bkg | 30% in 01L |
| $W_{\text{had}}Z_{\text{lep}}$bkg $p_T^V$ >600 / 400-600 ratio | $qqWZ$ | $VW$bkg | 30% in 02L |
| $ZZ$bkg $p_T^V$ >600 / 400-600 ratio | $qqZZ$ | $VZ$bkg | 10% in 02L |
| QCD scale $ZZ$ $p_T^V$ >600 / 400-600 | $qqZZ$ | $VZbb$, $VZcc$ | -1.6% to 7.6% in 02L |
| QCD scale $WZ$ $p_T^V$ >600 / 400-600 | $qqWZ$ | $VZbb$, $VZcc$ | -2.2% to 10.6% in 01L |
| QCD scale $ZZ$ LP / HP | $qqZZ$ | $VZbb$, $VZcc$ | -17.8% to 16.3% 0L |
| QCD scale $WZ$ LP / HP | $qqWZ$ | $VZbb$, $VZcc$ | -42.2% to 19.2% 01L |
| Carl $ZZ$ PwPy8 / Sh2211 | $qqZZ$ | All | 02L |
| Carl $ZZ$ Sh221 / Sh2211 | $qqZZ$ | All | 02L |
| Carl $WZ$ PwPy8 / Sh2211 | $qqZZ$ | All | 01L |
| Carl $WZ$ Sh221 / Sh2211 | $qqZZ$ | All | 01L |
| QCD scale largest shape | qqVV | All | 12L |
| EW largest shape | qqVV | All | 12L |

Table A.12: Diboson uncertainties in the boosted regime.

## A.6 Analysis Posfit Regions

### A.6.1 Resolved Posfit Regions

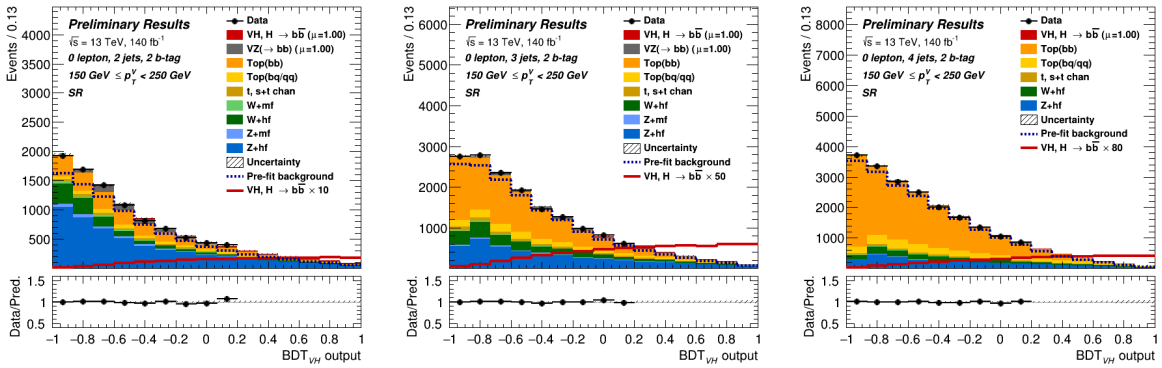
All regions in the resolved regime of the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis after the conditional fit to data of Section 1.10 are presented here, organised by increasing number of charged lepton channel (0L, 1L, 2L). The distributions indicate the pre-fit expactations of the sum of processes in dashed blue lines and highlight multiples of either the $VH(H \to b\bar{b})$ or $VH(H \to c\bar{c})$ signal distributions in red lines. Figures A.5, A.7, and A.10 are the $BB$-tagged signal regions. The 2 $c$-tagged SRs are displayed in Figures A.12, A.17, and A.23. The 1 $c$-tagged SRs are displayed in Figures A.13, A.18, and A.24.

The control regions are presented in:

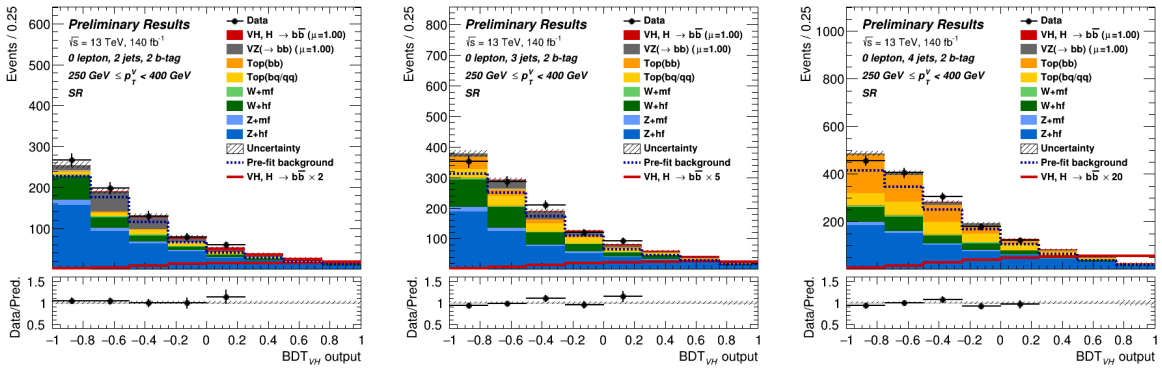
- The $BB$-tagged High $\Delta R$ CR in Figures A.6, A.8, and A.11.
- The $c$-tagged ($TN$, $TL$, and $TT$) High $\Delta R$ CR in Figures A.14, A.15, A.19, A.20, A.25, and A.26.
- The 1L $BB$-tagged Low $\Delta R$ CR in Figure A.9.
- The 1L and 2L $V + l$ CR ($LN$-tagged) in Figures A.22 and A.27.
- The 0L and 1L top CR $BT$-tagged in Figures A.16 and A.21.
- The 2L top $e\mu$ CR with $\geq 1$ $T$-tag in Figure A.28.

### A.6.2 Boosted Posfit Regions

This section presents the boosted regime regions after the conditional fit, with Figure A.29 presenting the 0L regions, Figure A.30 the 1L regions, and Figure A.31 the 2L regions.

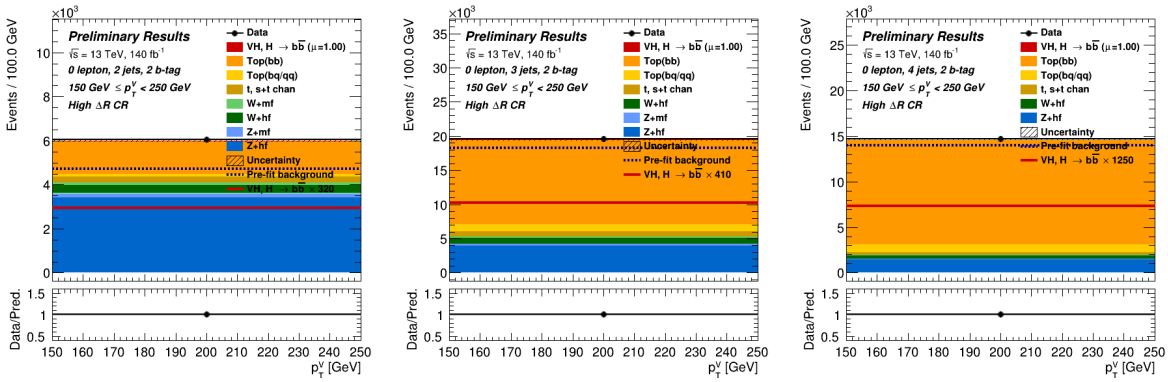



(a) $150 < p_T^V < 250$ GeV.

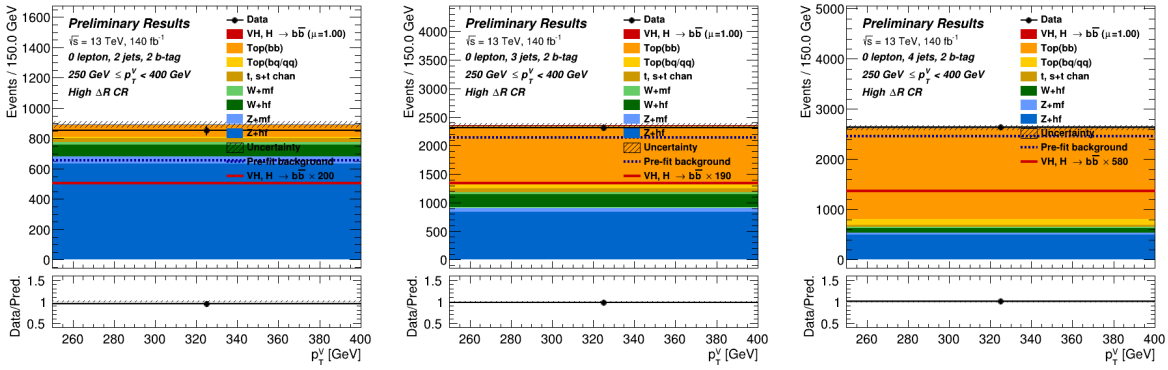



(b) $250 < p_T^V < 400$ GeV.

Figure A.5: The 0L signal regions in the $BB$-tagged 2-jet (left), 3-jet (centre), and 4-jet (right).

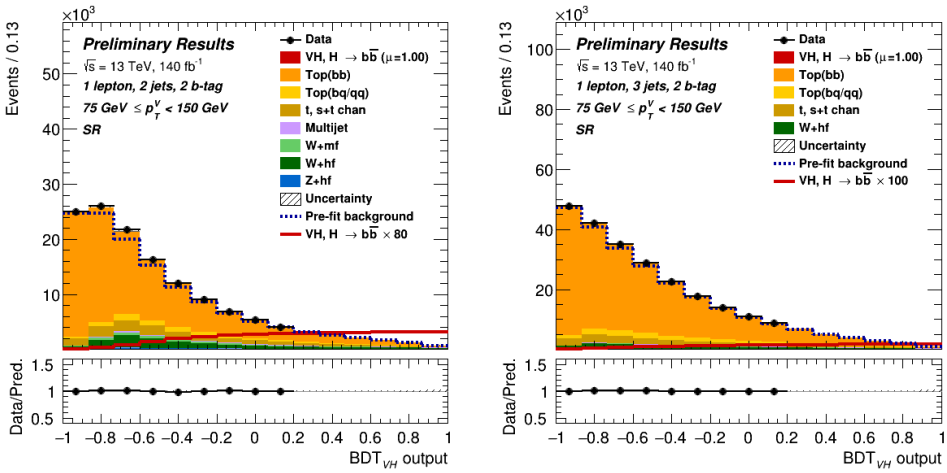



(a) $150 < p_T^V < 250$ GeV.

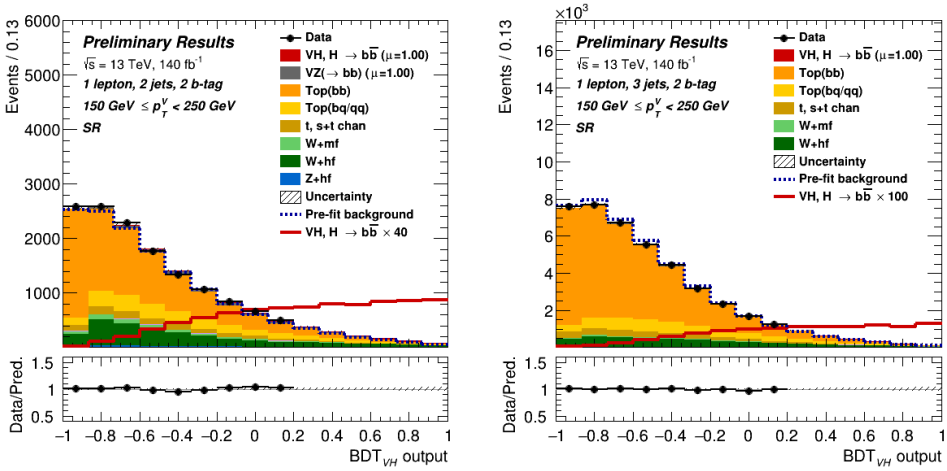



(b) $250 < p_T^V < 400$ GeV.

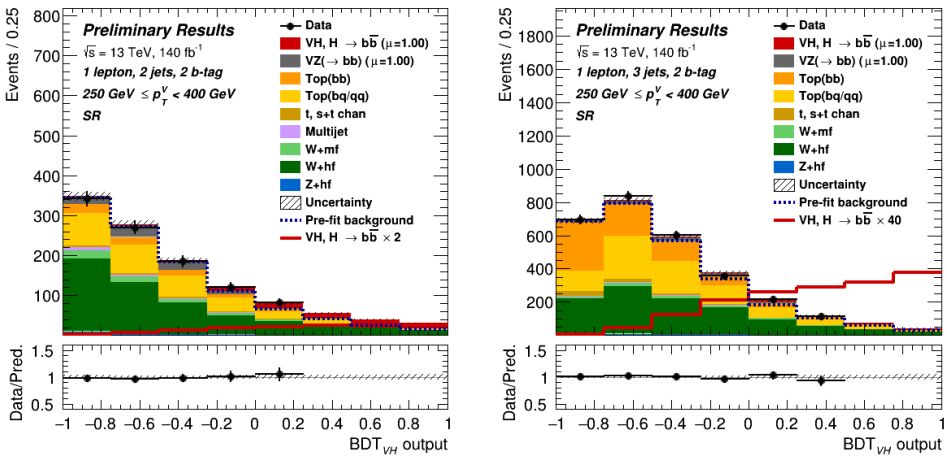
Figure A.6: The 0L High $\Delta R$ CR in the $BB$-tagged 2-jet (left), 3-jet (centre), and 4-jet (right).




(a) $75 < p_T^V < 150$ GeV.

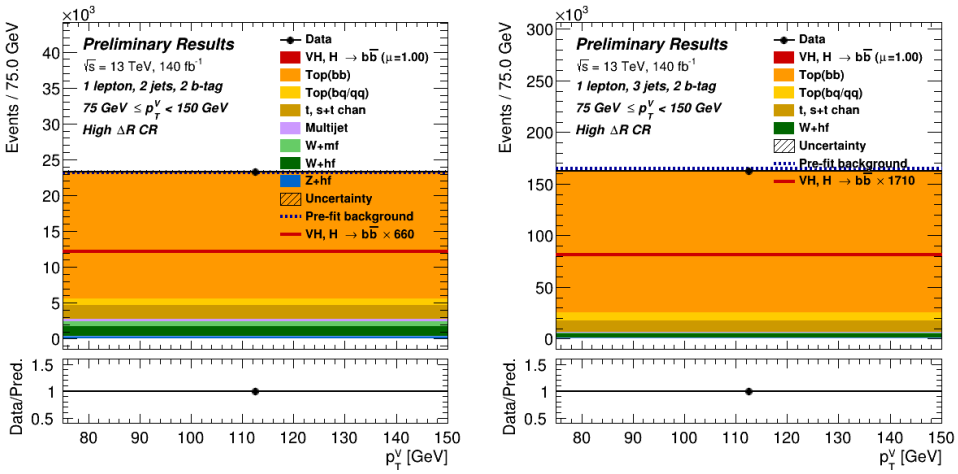



(b) $150 < p_T^V < 250$ GeV.

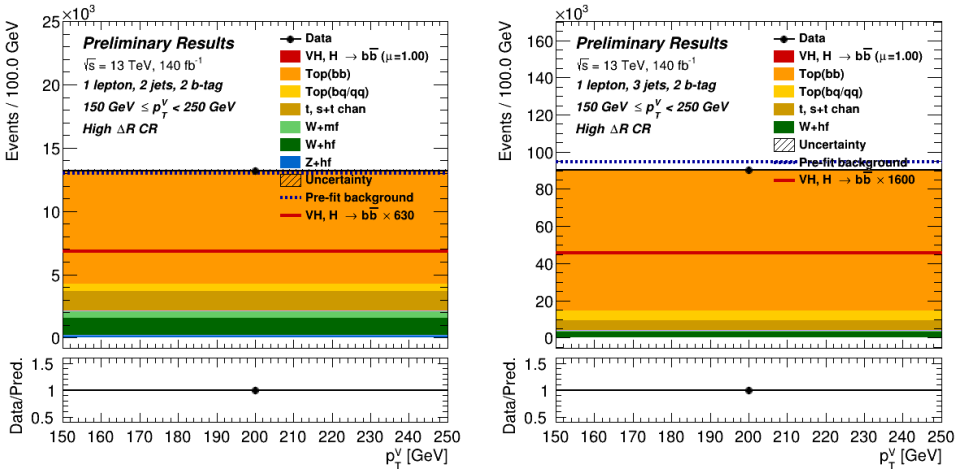



(c) $250 < p_T^V < 400$ GeV.

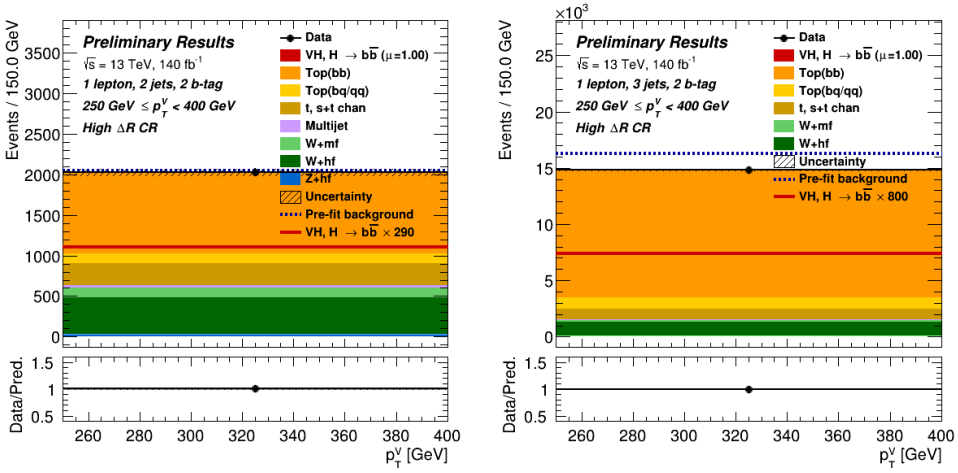
Figure A.7: The 1L signal regions in the $BB$-tagged 2-jet (left) and 3-jet (right).




(a) $75 < p_T^V < 150$ GeV.

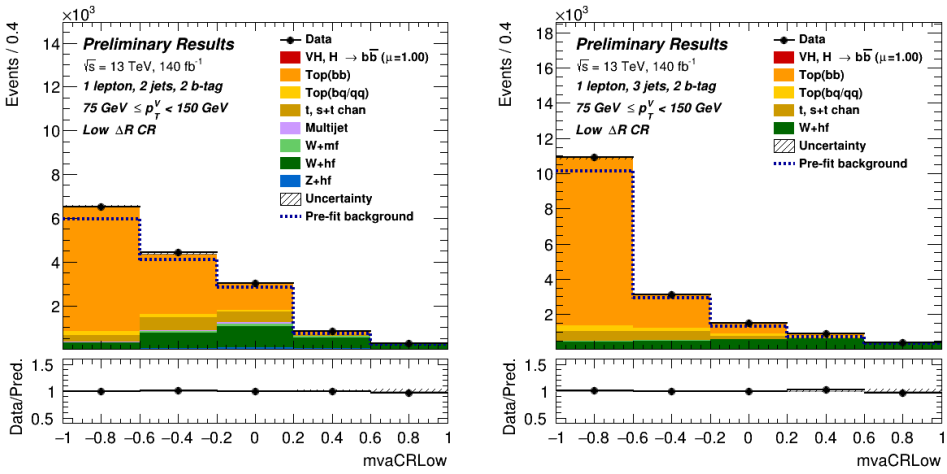



(b) $150 < p_T^V < 250$ GeV.

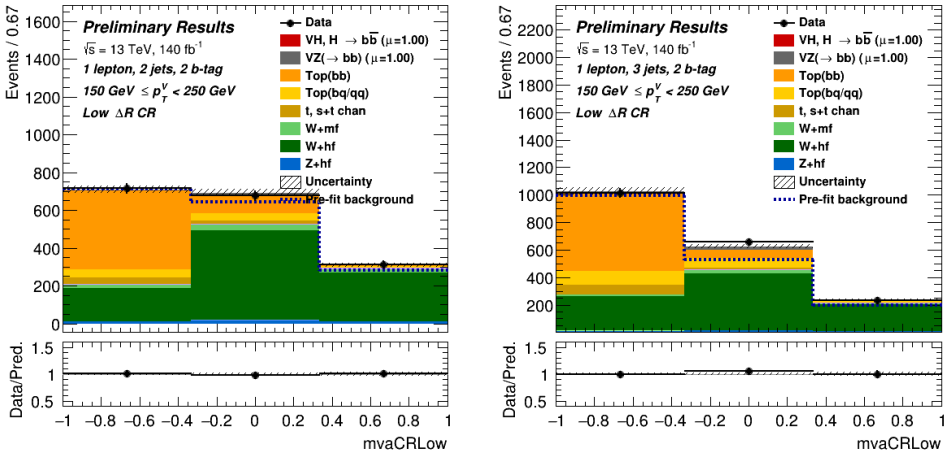



(c) $250 < p_T^V < 400$ GeV.

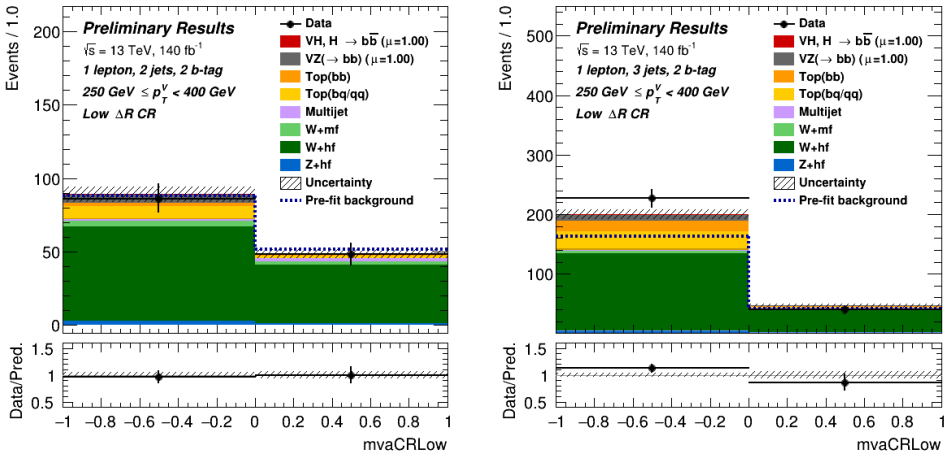
Figure A.8: The 1L High $\Delta R$ CR in the $BB$-tagged 2-jet (left) and 3-jet (right).




(a) $75 < p_T^V < 150$ GeV.

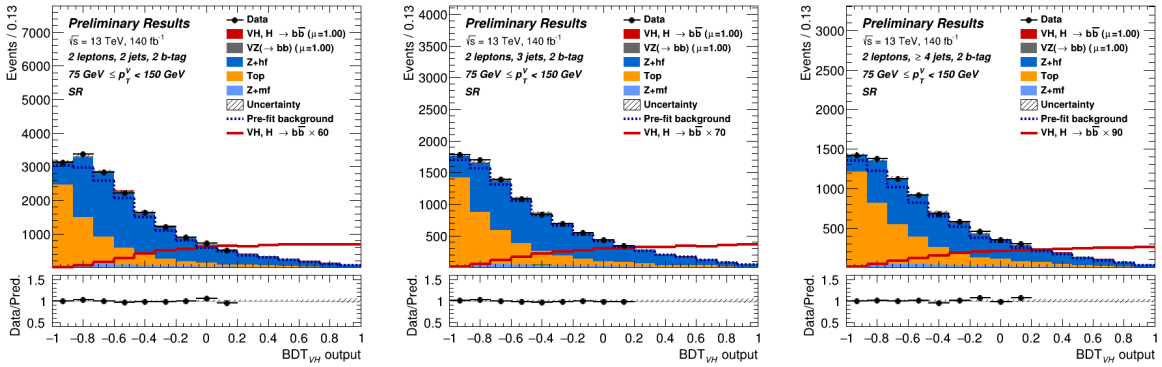



(b) $150 < p_T^V < 250$ GeV.

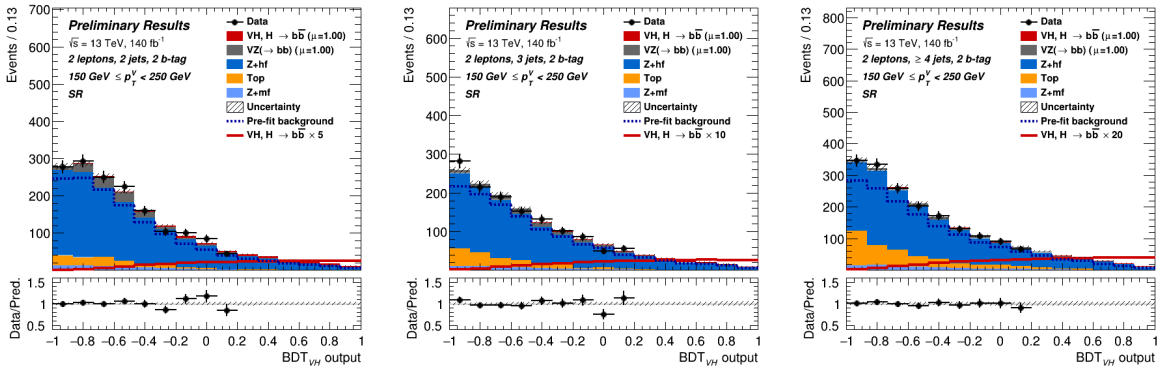



(c) $250 < p_T^V < 400$ GeV.

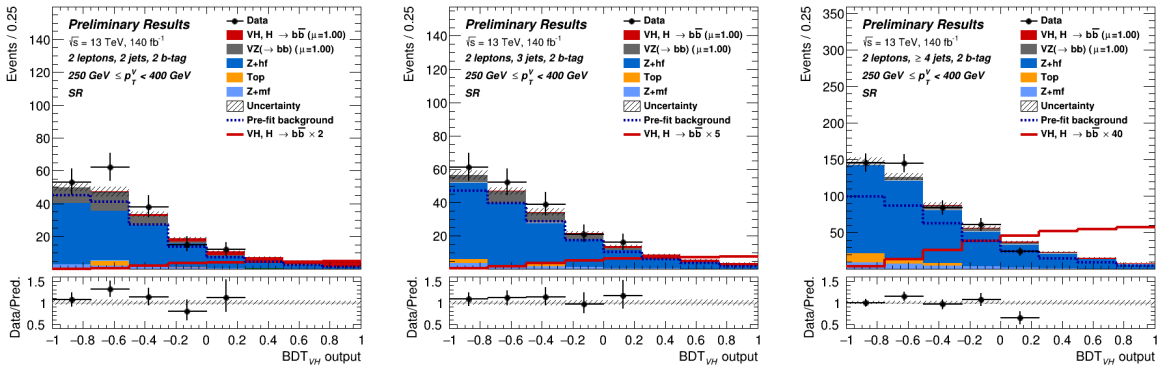
Figure A.9: The 1L Low $\Delta R$ CR in the $BB$-tagged.




(a) $75 < p_T^V < 150$ GeV.

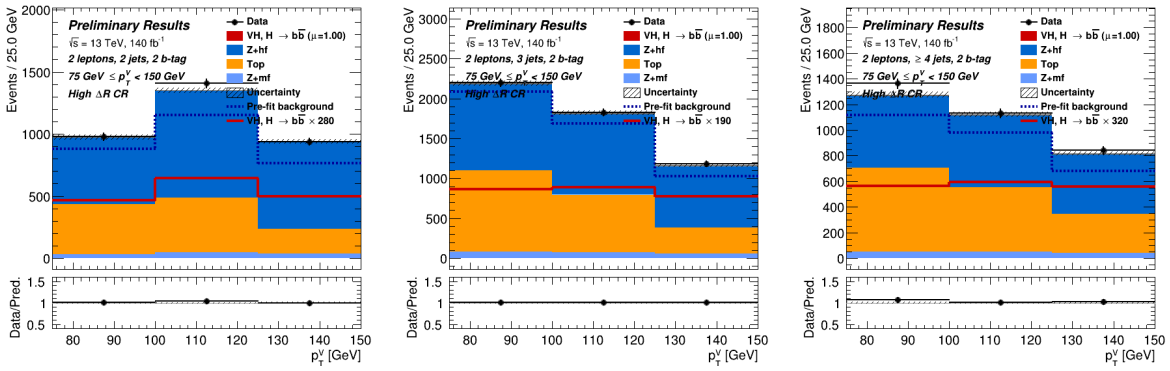



(b) $150 < p_T^V < 250$ GeV.

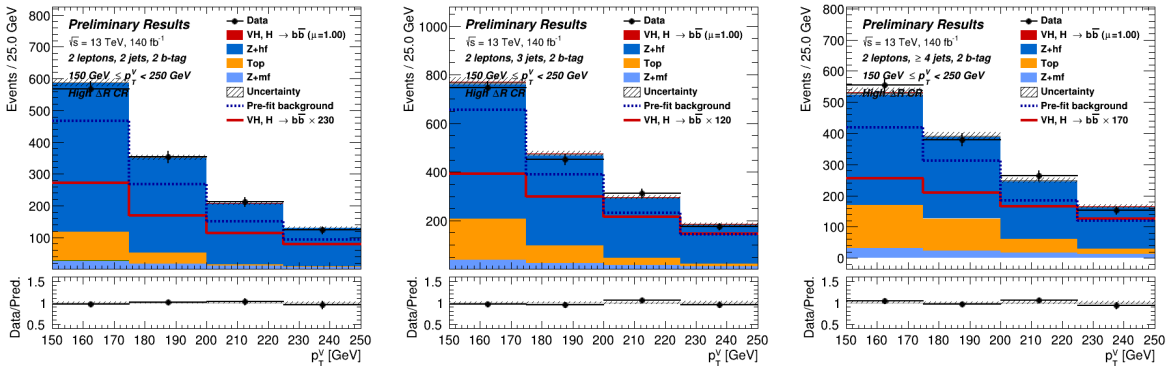



(c) $250 < p_T^V < 400$ GeV.

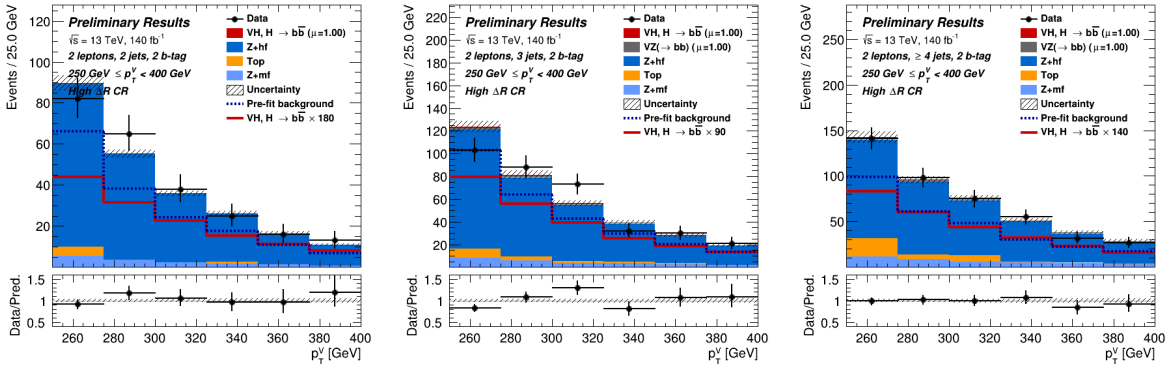
Figure A.10: The 2L signal regions in the $BB$-tagged 2-jet (left), 3-jet (centre), and ≥4-jet (right).




(a) $75 < p_T^V < 150$ GeV.

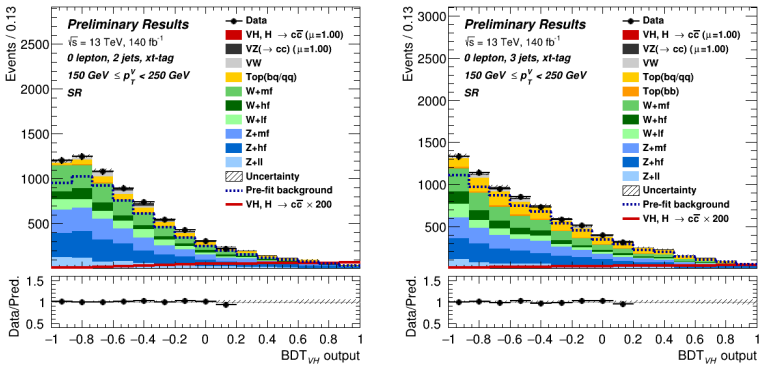



(b) $150 < p_T^V < 250$ GeV.

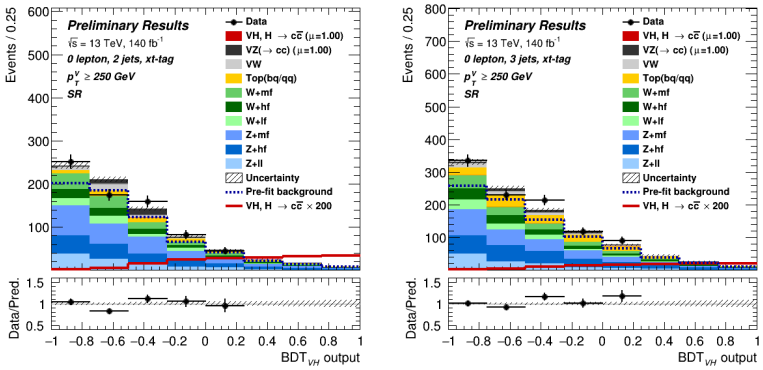



(c) $250 < p_T^V < 400$ GeV.

Figure A.11: The 2L High $\Delta R$ CR in the $BB$-tagged 2-jet (left), 3-jet (centre), and ≥4-jet (right).

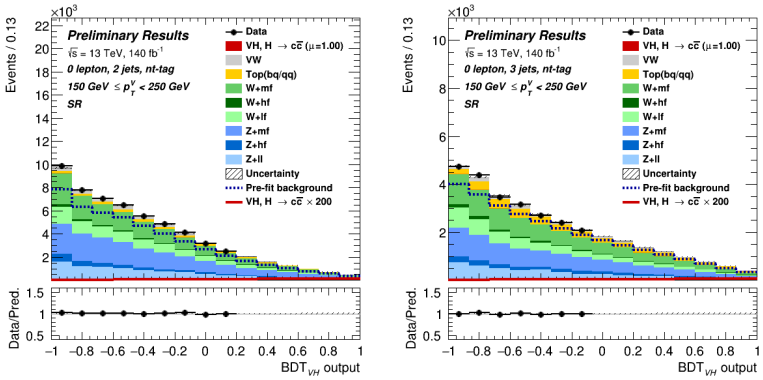



(a) $150 < p_T^V < 250$ GeV.

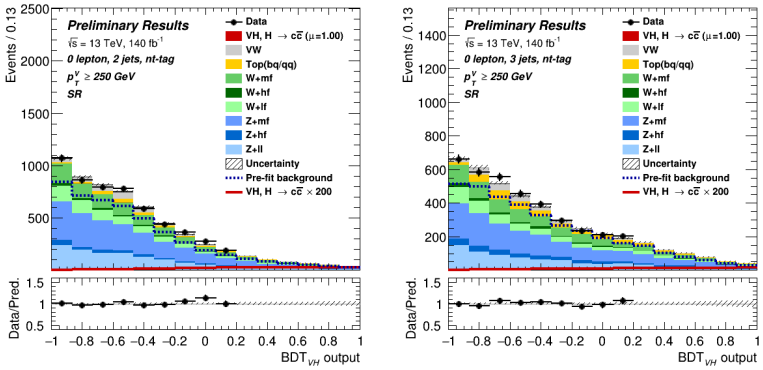



(b) 250 GeV $< p_T^V$.

Figure A.12: The 0L signal regions in the 2 $c$-tagged 2-jet (left) and 3-jet (right).

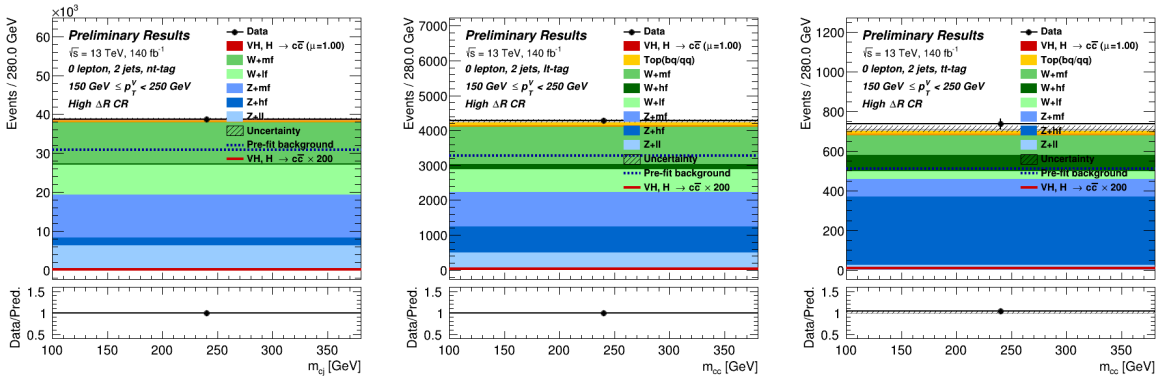



(a) $150 < p_T^V < 250$ GeV.

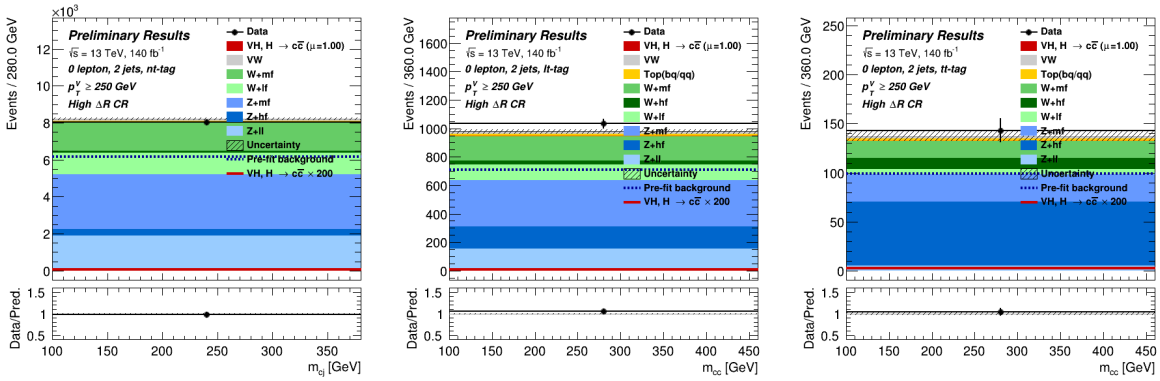



(b) 250 GeV $< p_T^V$.

Figure A.13: The 0L signal regions in the 1 $c$-tagged 2-jet (left) and 3-jet (right).

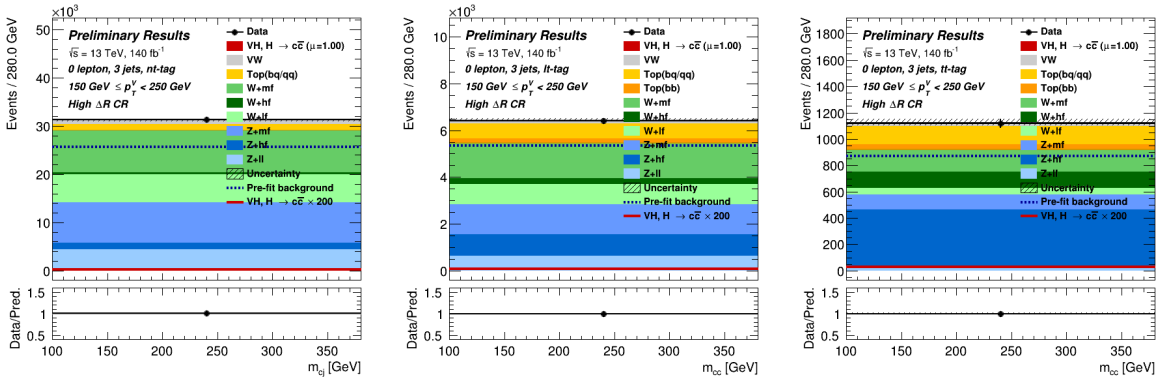



(a) $150 < p_T^V < 250$ GeV.

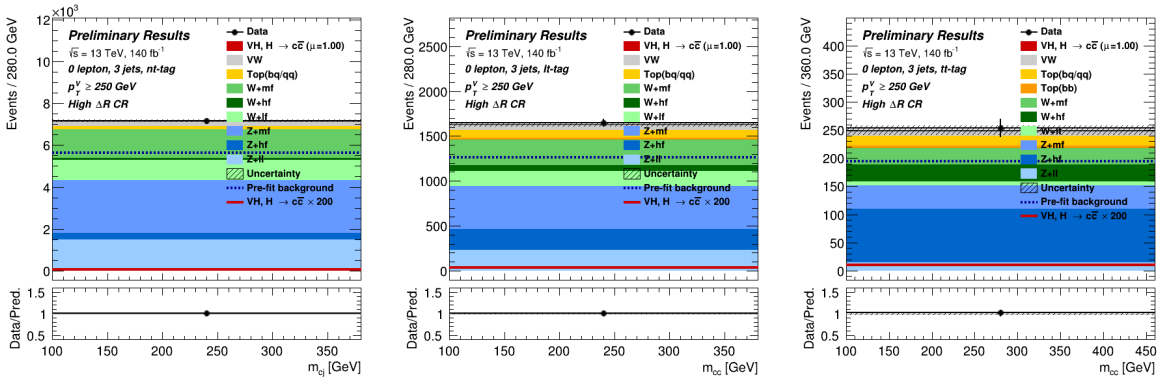



(b) 250 GeV $< p_T^V$.

Figure A.14: The 0L 2-jet High $\Delta R$ CR in the $TN$- (left), $LT$ -(centre), and $TT$-tagged (right).

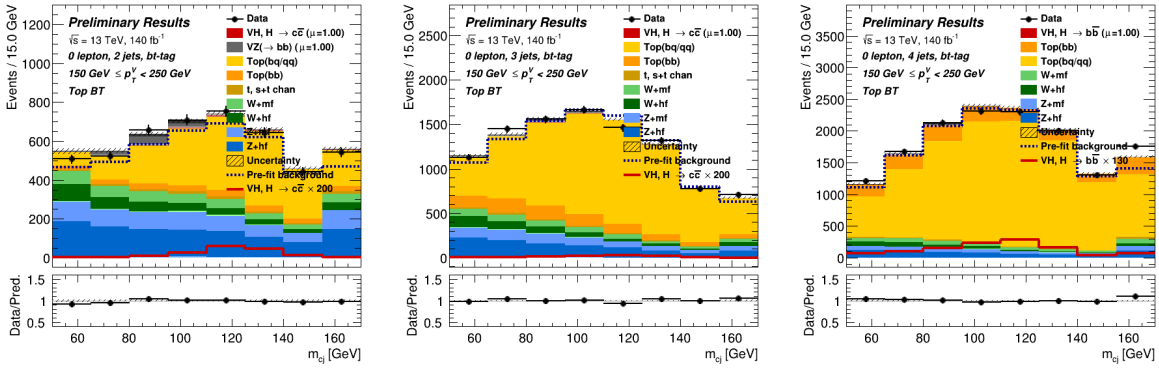



(a) 150 GeV $< p_T^V < 250$ GeV.

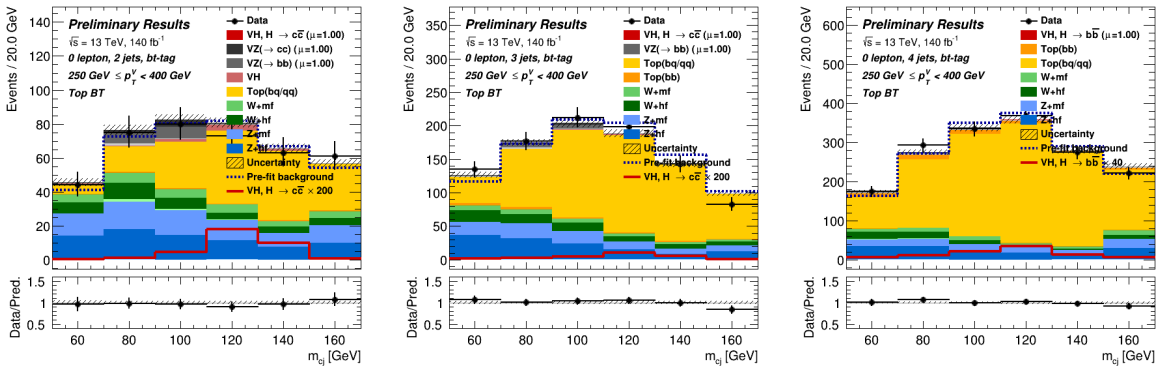



(b) 250 GeV $< p_T^V$.

Figure A.15: The 0L 3-jet High $\Delta R$ CR in the $TN$- (left), $LT$- (centre), and $TT$-tagged (right).

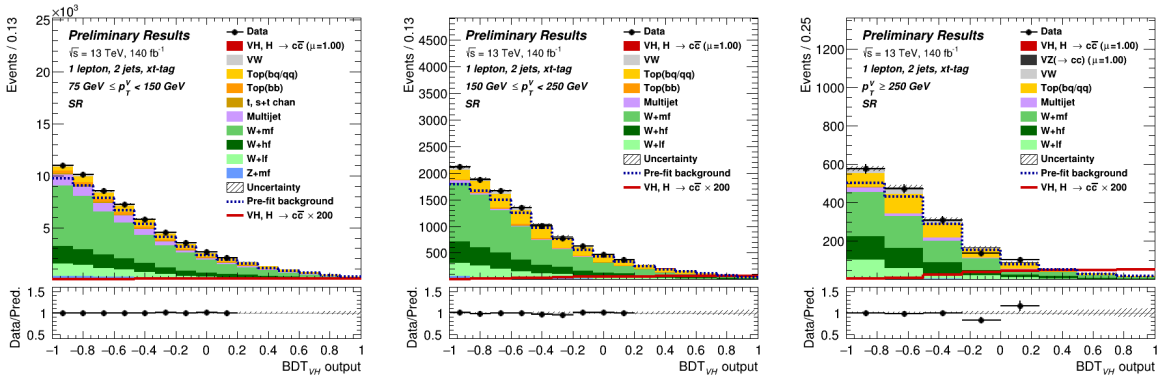



(a) $150 < p_T^V < 250$ GeV.

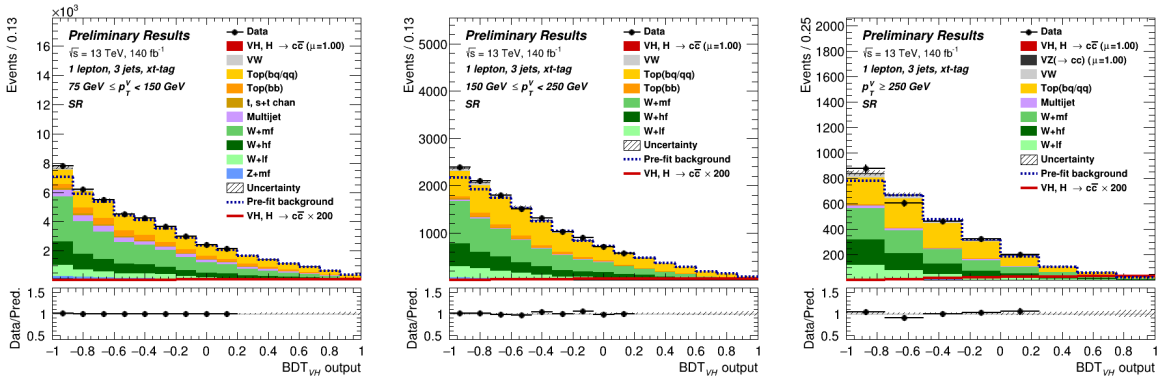



(b) 250 GeV $< p_T^V <$ 400 GeV.

Figure A.16: The 0L Top CR in the $BT$-tagged 2-jet (left), 3-jet (centre), and 4-jet (right).

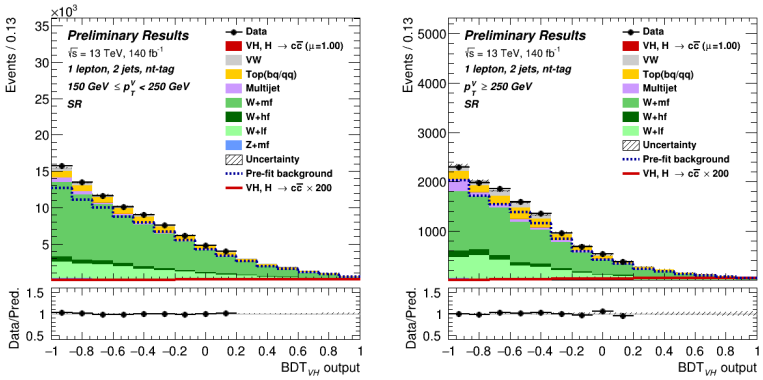



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

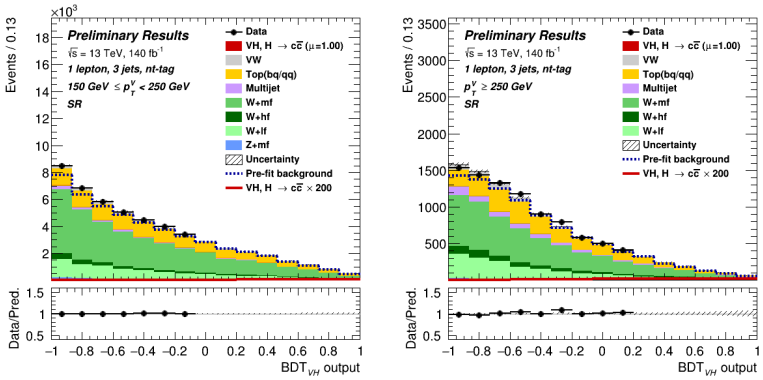



(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

Figure A.17: The 1L signal regions in the 2 $c$-tagged regionss.

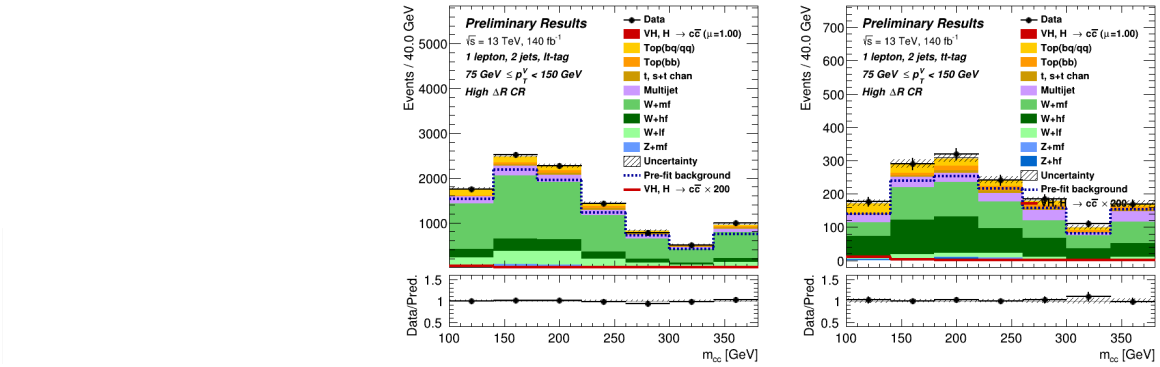



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

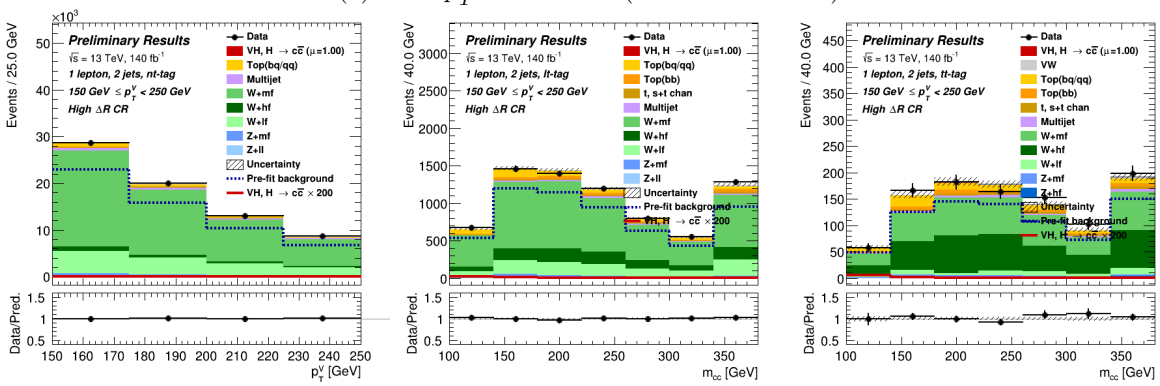



(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

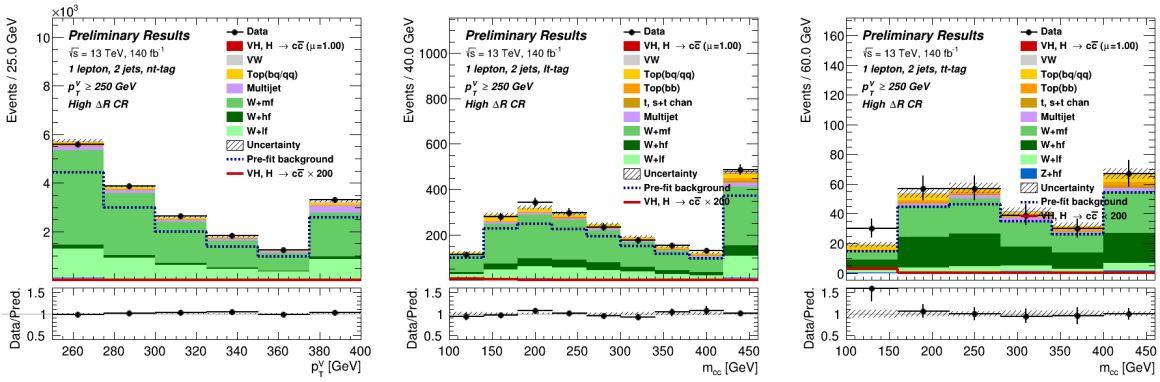
Figure A.18: The 1L signal regions in the 1 $c$-taggedged regions.




(a) $75 < p_T^V < 150$ GeV ($TN$ not included).

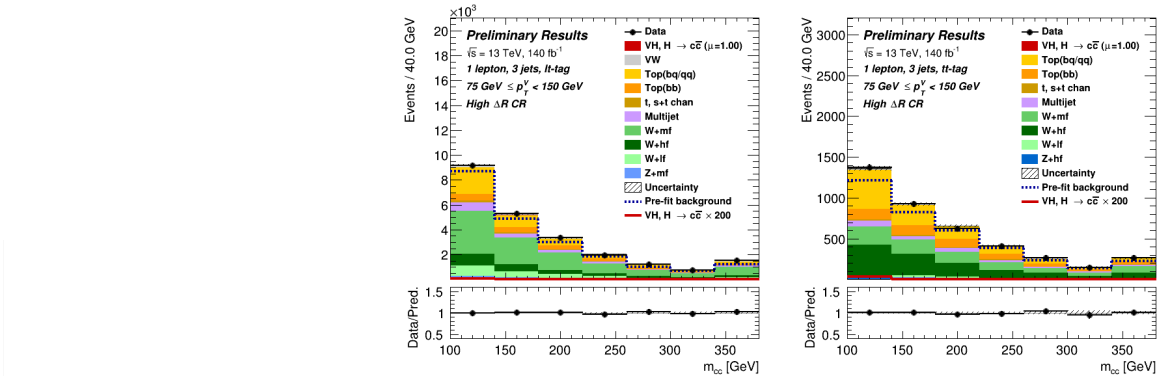



(b) $150 < p_T^V < 250$ GeV.

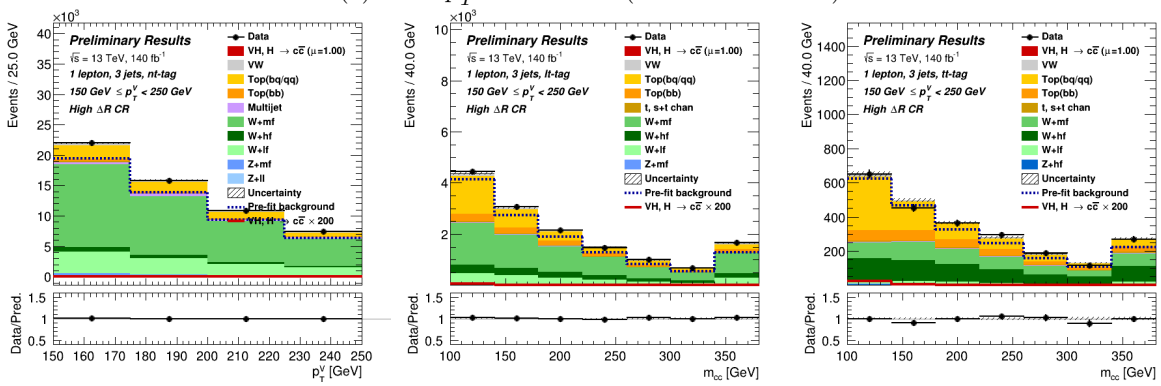



(c) 250 GeV $< p_T^V$.

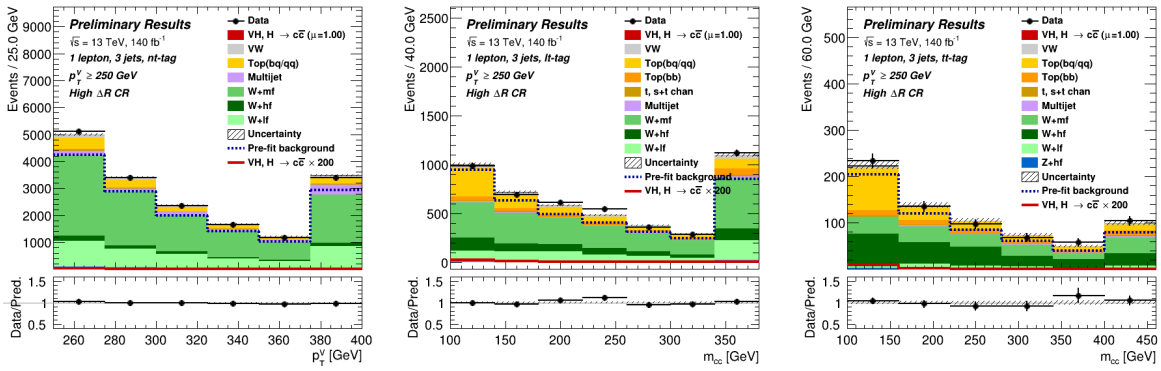
Figure A.19: The 1L High $\Delta R$ CR in the 2-jet $TN$- (left), $LT$- (centre), and $TT$-tagged (right) regions.




(a) $75 < p_T^V < 150$ GeV ($TN$ not included).

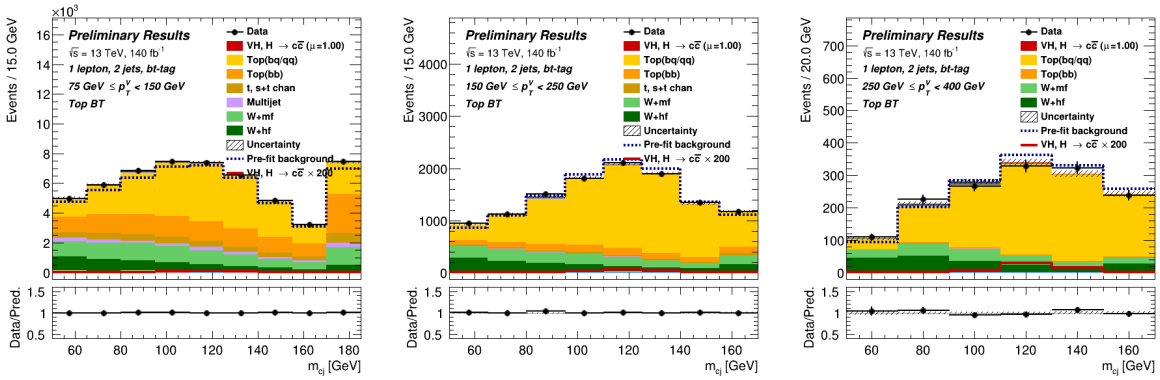



(b) $150 < p_T^V < 250$ GeV.

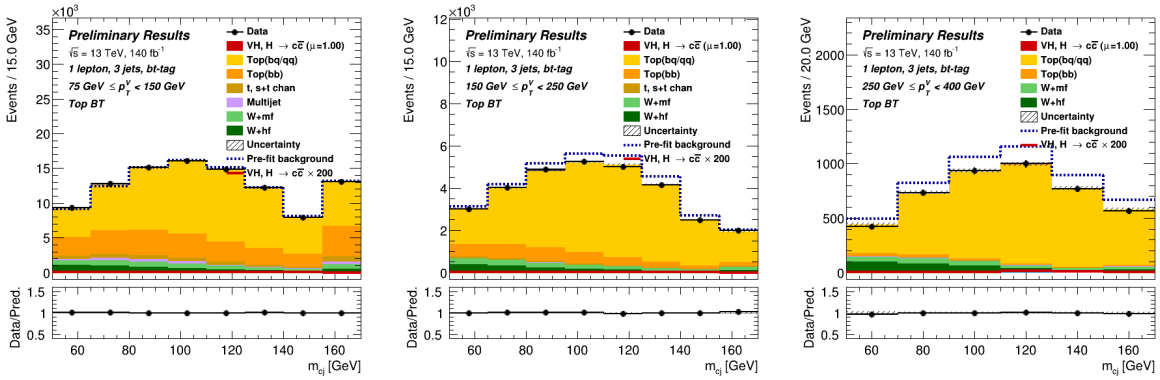



(c) 250 GeV $< p_T^V$.

Figure A.20: The 1L High $\Delta R$ CR in the 3-jet $TN$- (left), $LT$- (centre), and $TT$-tagged (right) regions.

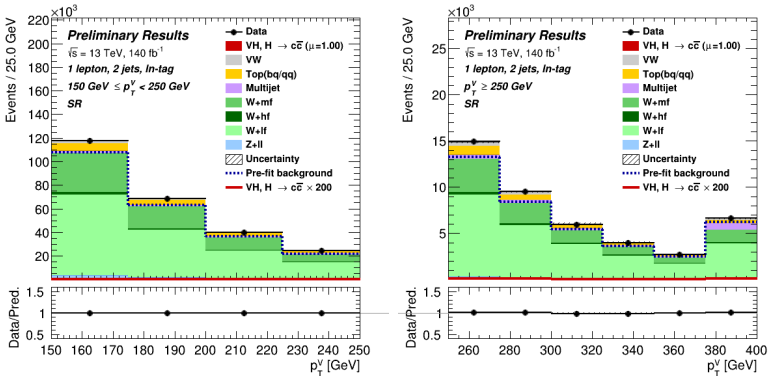



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) $p_T^V$ regions.

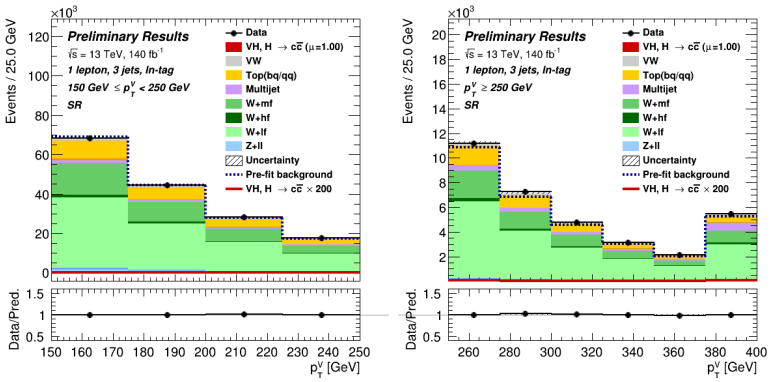



(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) $p_T^V$ regions.

Figure A.21: The 1L Top CR $BT$-tagged regions.

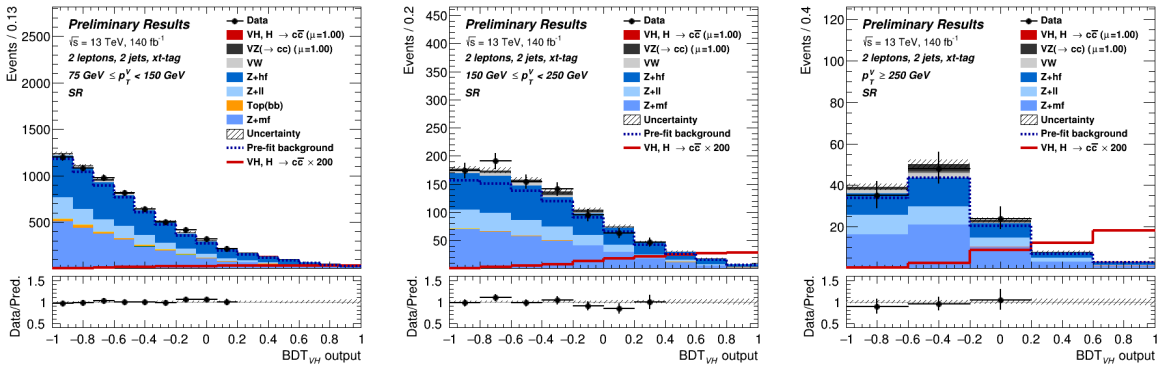



(a) 2-jet.

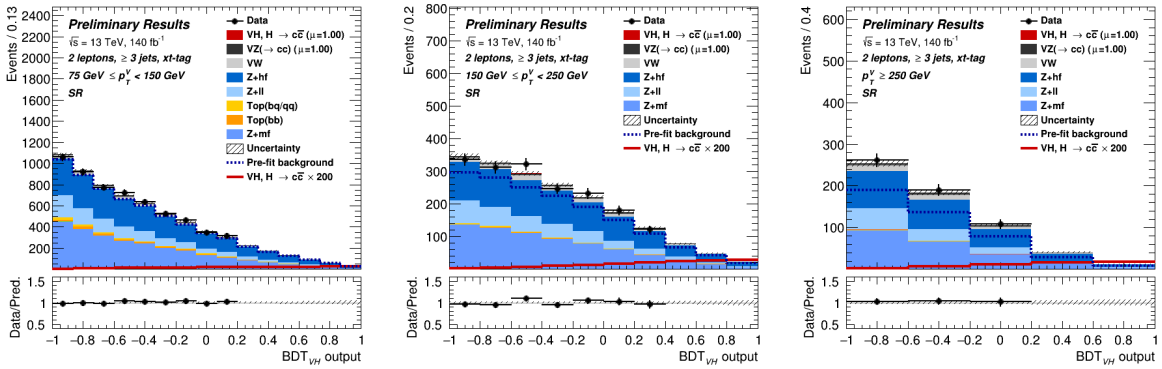



(b) 3-jet.

Figure A.22: The 1L $V + l$ CR in the $LN$-tagged, [150, 250] GeV (left) and 250 GeV ≤ (right) $p_T^V$ regions.

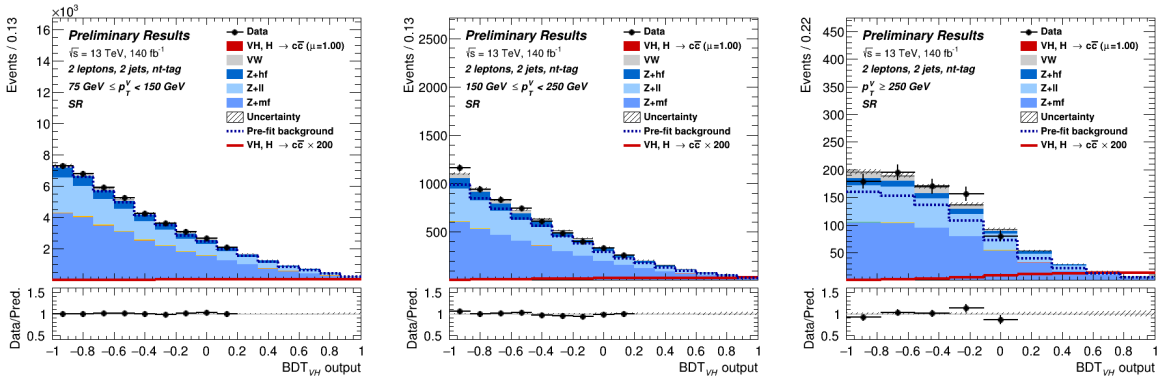



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

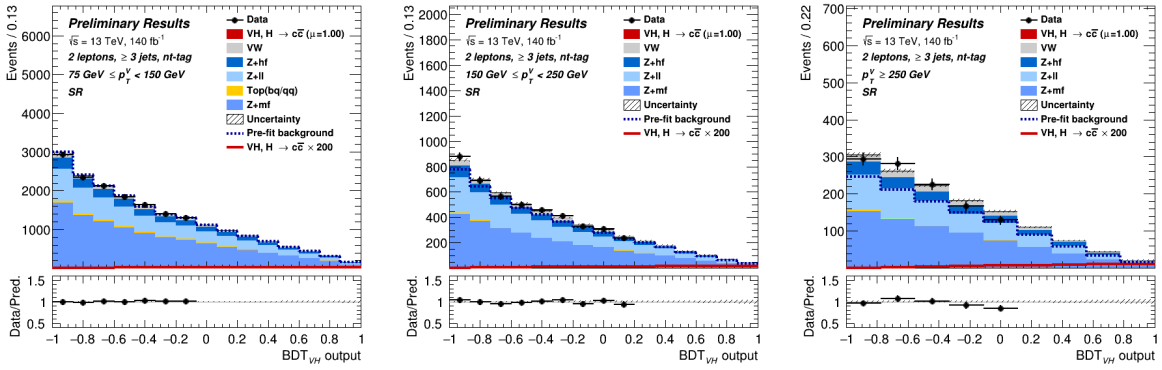



(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

Figure A.23: The 2L signal regions in the 2 $c$-tagged regionss.

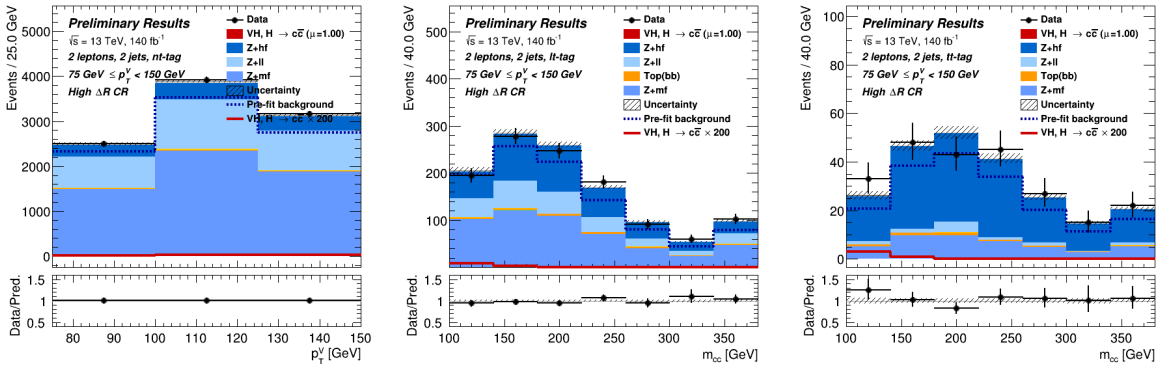



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

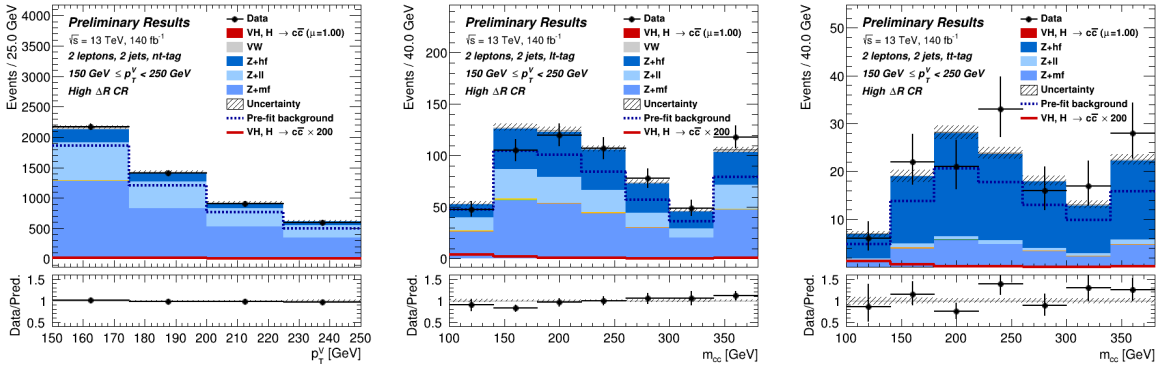



(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

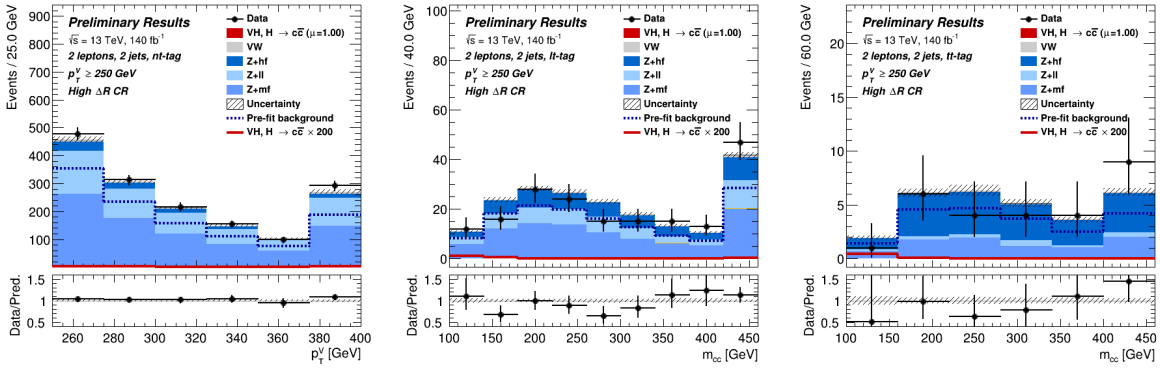
Figure A.24: The 2L signal regions in the 1 $c$-taggedged regions.




(a) $75 < p_T^V < 150$ GeV.

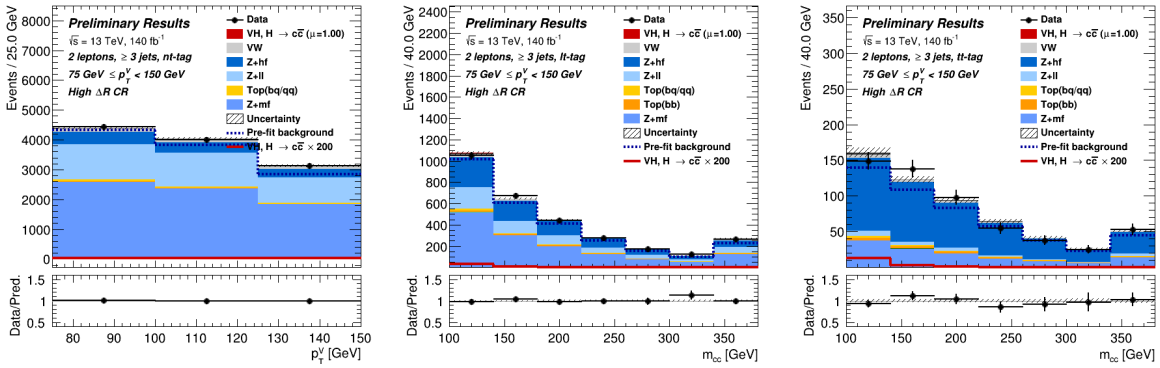



(b) $150 < p_T^V < 250$ GeV.

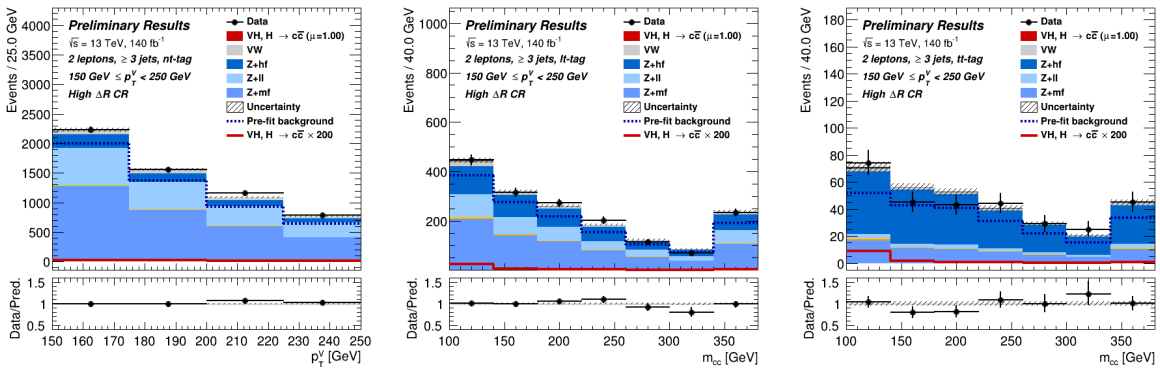



(c) 250 GeV $< p_T^V$.

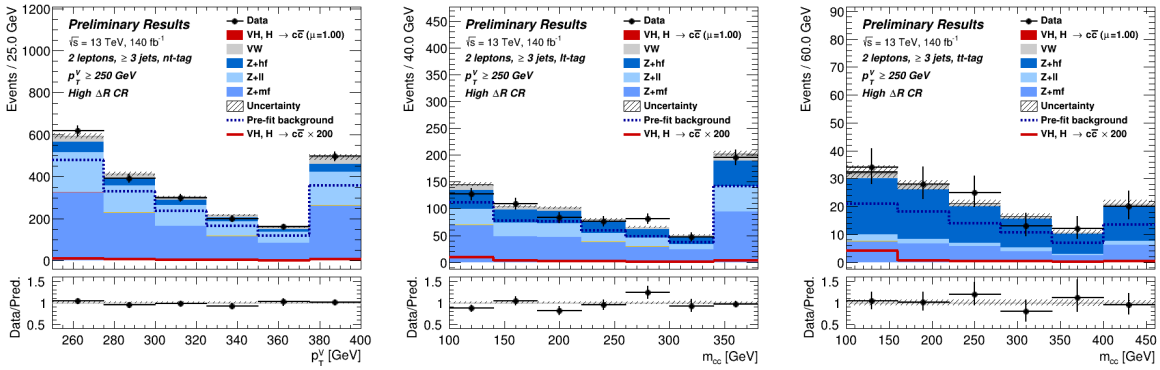
Figure A.25: The 2L High $\Delta R$ CR in the 2-jet $TN$- (left), $LT$- (centre), and $TT$-tagged (right) regions.




(a) $75 < p_T^V < 150$ GeV.

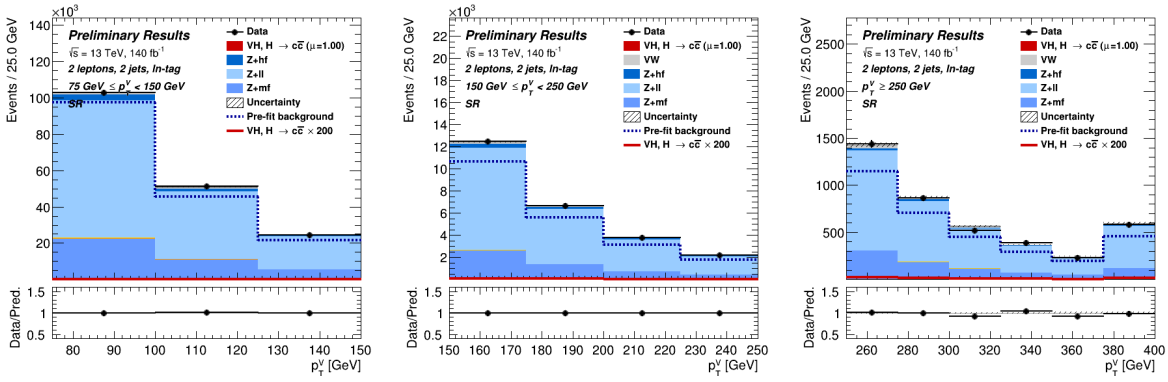



(b) $150 < p_T^V < 250$ GeV.

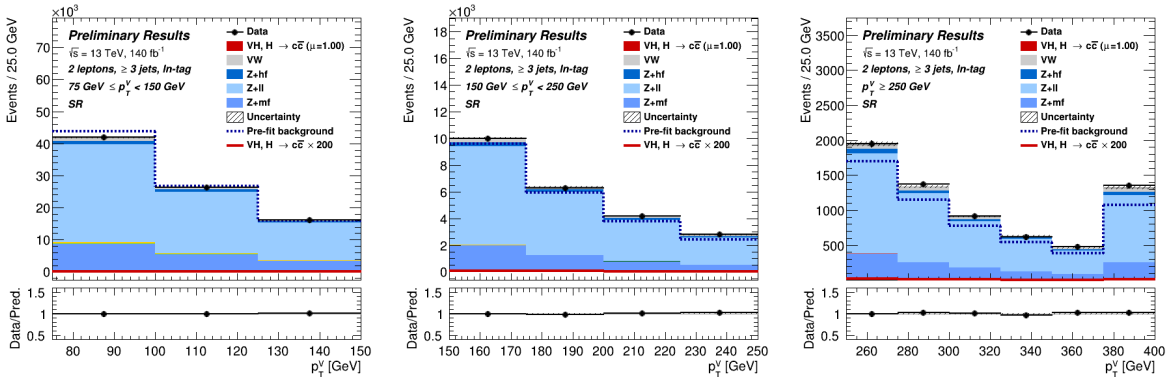



(c) 250 GeV $< p_T^V$.

Figure A.26: The 2L High $\Delta R$ CR in the 3-jet $TN$- (left), $LT$- (centre), and $TT$-tagged (right) regions.

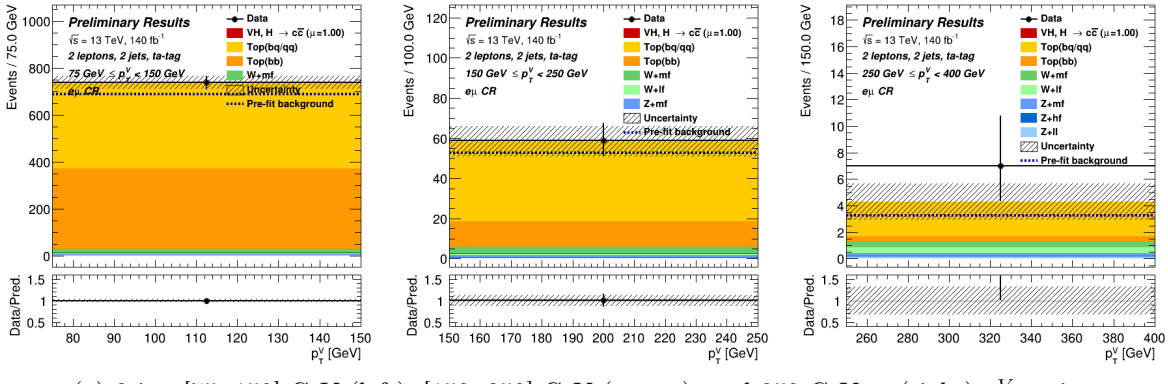



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

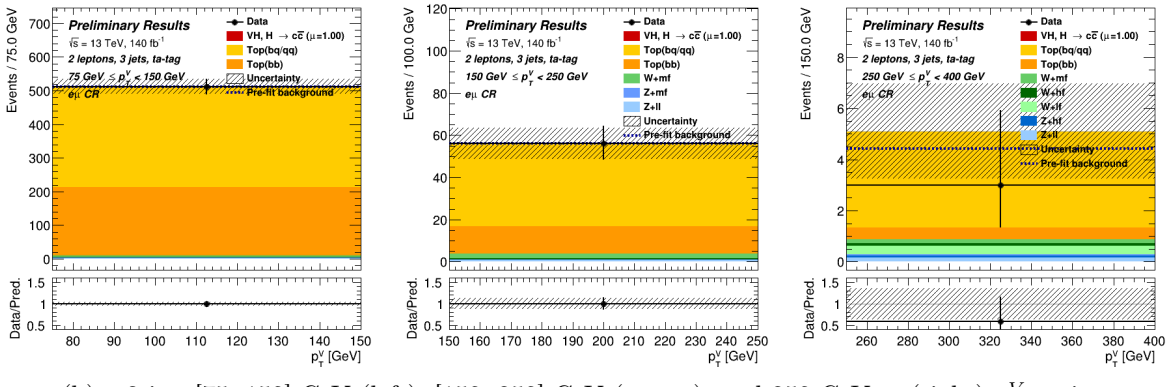



(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

Figure A.27: The 2L $V + l$ CR in the $LN$-tagged regions.

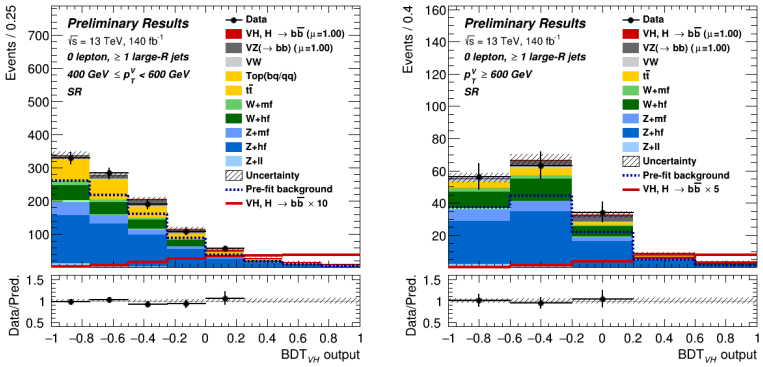



(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

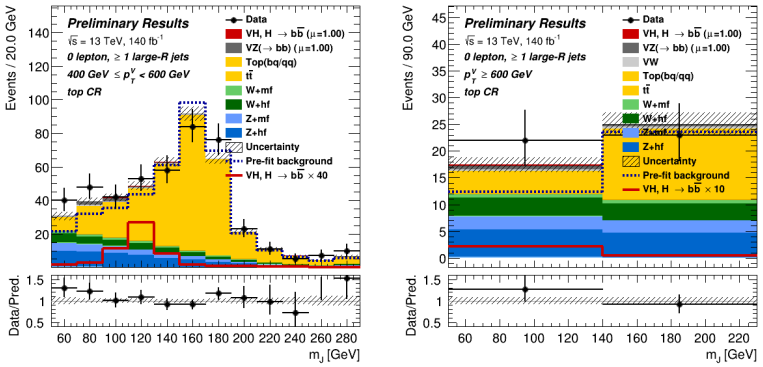



(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) $p_T^V$ regions.

Figure A.28: The 2L Top $e\mu$ CR with ≥1 $T$-tagged regions.

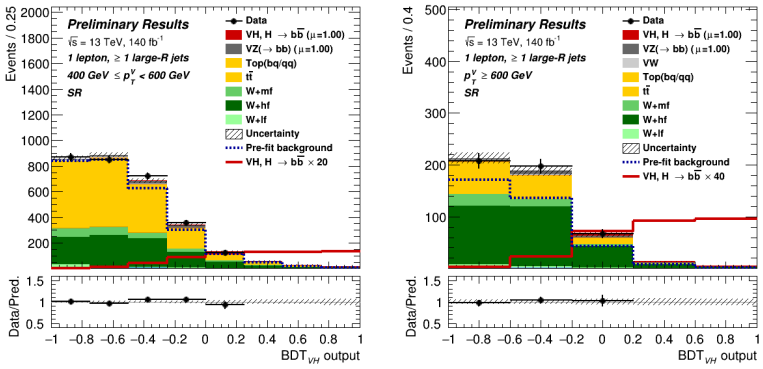



(a) The $p_T^V \in [400, 600]$ GeV (left - combines high- and low-purity combined) and the $p_T^V \geq 600$ GeV (right) signal regions.

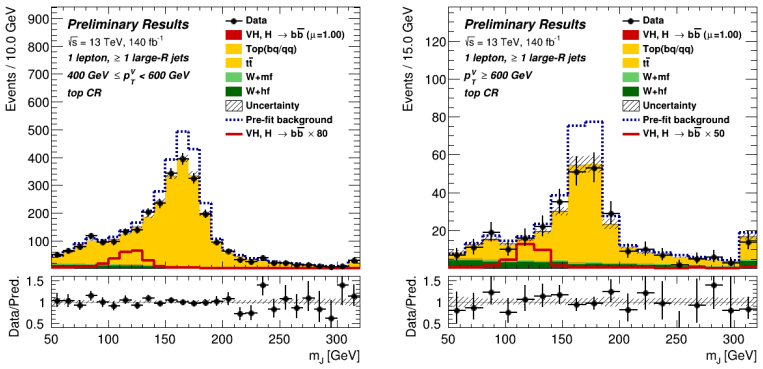



(b) The $p_T^V \in [400, 600]$ GeV (left) and $p_T^V \geq 600$ GeV (right) boosted Top CR.

Figure A.29: The boosted $BB$-tagged 0L regions.




(a) The $p_T^V \in [400, 600]$ GeV (left - combines high- and low-purity combined) and the $p_T^V \geq 600$ GeV (right) signal regions.




(b) The $p_T^V \in [400, 600]$ GeV (left) and $p_T^V \geq 600$ GeV (right) boosted Top CR.

Figure A.30: The boosted $BB$-tagged 0L regions.

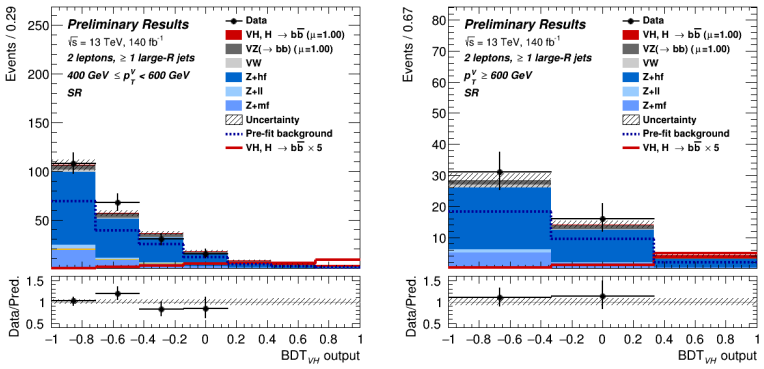



Figure A.31: The boosted $BB$-tagged 2L signal regions, $p_T^V \in [400, 600]$ (left) and $p_T^V \geq 600$ GeV (right).