Search for single production of a Vector-Like partner of the bottom quark in the $bH(b\bar{b})$ final state in $pp$ collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector.

Marco Montella

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of University College London.

Department of Physics and Astronomy
University College London

January 4, 2021
I, Marco Montella, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

All physics results produced within the context of experiments of the size and scope of ATLAS are necessarily borne out of the cooperative efforts of countless individuals overseeing fundamental aspects of the workflow, such as detector maintenance and calibration, data preparation and physics performance optimization. For this reason papers detailing ATLAS results are released to the public under the collective authorship of "The ATLAS Collaboration".

Nevertheless, the actual end-point physical processing of the ATLAS-delivered collisional data is carried out by restricted analysis team working typically under the direction of a senior convener and under the scrutiny of the internal ATLAS peer reviewing chain of command.

The search covered in this thesis, unusually for an ATLAS analysis, was carried out by a team of only two individuals, with the author taking the role of analysis contact, convener and sole analyzer; and with Prof. Konstantinidis serving as paper editor and providing fundamental advice and guidance. Dr. Héctor De La Torre Perez, convener of the ATLAS Heavy Quarks & Top (HQT) subgroup, informally provided an additional measure of invaluable counsel.

As a consequence, all results, tables, inference and figures not presented with an external reference were materially produced by the author, who, together with the aforementioned advisors, personally oversaw and carried out all stages of the physical analysis covered in this work.
Abstract

A search is presented for single-produced vector-like $B$ quark decaying to a Standard Model $b$ quark and a Higgs boson, itself decaying into a $b\bar{b}$ pair.

The $B$ decay system is reconstructed as an association of a large-radius jet and a small-radius jet, respectively taken to reflect the secondary Higgs decay system and the $b$ quark produced in the immediate resonance decay. The search is the first at ATLAS to probe the $bH(b\bar{b})$ final state in Run-2, and aims at extending the region of the $M_B$ phase space the current exclusion limits, set by the pair production combination search at 1220 GeV for a Vector-Like $B$ quark occurring as an isospin singlet.

For this purpose, a newly designed and optimised event selection was implemented for maximal sensitivity in the 1000-2000 GeV invariant mass region, and a particular effort was dedicated in researching and implementing data-driven modelling techniques to provide a reliable estimation of the main, QCD-Driven, background source to the search.

A binned likelihood fit is set up for the statistical interpretation of the results, which are presented in the form of mass- and coupling-dependent 95% $CL_s$ exclusion limits on the production cross section of a VLQ-like signal.
Impact Statement

The Large Hadron Collider was conceived and designed to deliver the unique experimental conditions necessary to complete and refine our understanding of the physical processes predicted by the Standard Model, as well as to uncover evidence of yet undiscovered phenomena beyond our current theoretical picture.

In order to meet such design goals, and in order for LHC-based experiments to provide satisfactory measurements and analysis of the outcome of the delivered collisions, a massive R&D endeavour was launched, spanning multiple decades and hundreds of research groups. As a result, substantial advances have been made in fields such as cryogenics, pixel-based detection and patter recognition techniques, all of which found significant application in recent times outside the academic world.

Analyzing the data produced at the LHC puts the entire research to construction chain to its ultimate academic fruition, bringing about substantial improvements in our knowledge of the panorama of fundamental physics.

To this point, the search covered in this thesis was able, on the grounds of collisional data detected by the ATLAS experiment, to significantly constrain the phenomenological picture for Vector-Like Quarks, a set of exotic particles predicted to occur by a wide range of theoretical extensions to the Standard Model. At the time of the writing, the search results are undergoing the final stages of the internal ATLAS peer reviewing process, in view of an upcoming publication.

The data analysis techniques widely employed in the search and discussed in this thesis, while highly sectorial in their specific definition and application, hinge on
general principles that can be readily extrapolated to any other field of data science within and outside the academic world, from the interpretation of demographic, commercial or ecological trends to the unbiased evaluation of the outcome of clinical trials.
## Contents

1 Introduction 30

2 Theoretical Overview 32

2.1 The Standard Model of Particle Physics 32

2.1.1 Electroweak Unification 34

2.1.2 The Higgs Sector 37

2.1.3 Higgs Coupling to Fermions 40

2.1.4 Open Questions 42

2.1.5 The Higgs Hierarchy Problem 44

2.2 Vector-Like Quarks 46

2.2.1 Model-Independent Phenomenology 48

2.2.2 VLQ production at the LHC 50

2.2.3 Decay Mechanics 53

2.3 VLQ Searches: State of the Art 55

3 The ATLAS Experiment at the LHC 59

3.1 The Large Hadron Collider 59

3.1.1 Luminosity and Run conditions 60

3.2 The ATLAS Detector 63

3.2.1 Inner Detector 65

3.2.2 Pixel Detector 67

3.2.3 SCT 67

3.2.3 TRT 68
## Contents

3.2.4 The ATLAS Calorimetric System .......................... 69  
3.2.5 Muon Spectrometry ........................................ 73  
3.2.6 The ATLAS TDAQ System ................................. 75

### 4 Object Reconstruction at ATLAS  
4.1 Inner Detector Tracks ........................................... 77  
4.2 Hard-Scatter Vertex ............................................. 79  
4.3 Jets ...................................................................... 79  
  4.3.1 Topological Clusters ............................................. 80  
  4.3.2 Particle Flow Algorithm ........................................ 81  
  4.3.3 Jet Reconstruction Procedure ................................. 82  
  4.3.4 Large-R Jets .................................................... 84  
  4.3.5 Track Jets ....................................................... 87  
4.4 Flavour Tagging ..................................................... 88  
4.5 Electrons ............................................................. 91  
4.6 Muons ............................................................... 93  
4.7 Photons ............................................................. 94

### 5 Analysis Overview  
5.1 Search Strategy ................................................... 95  
5.2 Data and Monte Carlo Samples ................................. 99  
  5.2.1 Data samples .................................................. 99  
  5.2.2 Simulated signal samples ................................. 100  
  5.2.3 Coupling Reweighting Mechanism ........................ 104

### 6 Event Reconstruction and Selection  
6.1 Object Selection .................................................. 108  
  6.1.1 Small-R Jets .................................................... 109  
  6.1.2 Large-R Jets .................................................... 110  
  6.1.3 Track Jets ....................................................... 112  
  6.1.4 Flavour Tagging ............................................... 112  
  6.1.5 Leptons and Photons ................................. 114
6.1.6 Trigger Selection ........................................ 115
6.2 High Level Event Reconstruction .......................... 117
  6.2.1 Pre-Selection ........................................ 117
  6.2.2 Higgs Candidate Reconstruction ....................... 120
  6.2.3 Higgs Candidate Categorization ....................... 123
  6.2.4 VLB Candidate Reconstruction ........................ 126
  6.2.5 Event Categorization ................................ 129
6.3 Signal Region Definition .................................. 135
  6.3.1 Event Category Optimization ......................... 135
  6.3.2 Higgs Mass Window Optimization ..................... 139
6.4 Post Reconstruction Kinematic Selection .................. 141
  6.4.1 Optimization Method ................................ 142
  6.4.2 Forward Hadronic Activity ........................... 142
  6.4.3 Higgs Candidate Structure ........................... 143
  6.4.4 s/t-channel kinematics ............................... 146
6.5 Event Selection Summary .................................. 151

7 Background Modelling ....................................... 155
  7.0.1 Data-Driven Modelling Fundamentals .................. 155
7.1 ABCD Regions Definition ................................ 156
  7.1.1 Axes Correlation ................................... 159
7.2 ABCD Modelling Implementation ........................... 165
  7.2.1 Yield Correction .................................... 165
  7.2.2 Shape Correction ................................... 165
  7.2.3 Residual Correlation Correction ...................... 168
7.3 ABCD Closure Plots ....................................... 170
7.4 Model Validation ........................................ 173
  7.4.1 ABCD Modelling Limitations ......................... 173
  7.4.2 Modelling Uncertainties ............................. 174
  7.4.3 Validation Check I: Higgs Mass Sideband ............ 179
  7.4.4 Validation Check II: Low Purity Sample ............. 180
8 Systematic Uncertainties

8.1 Overview

8.2 Experimental Uncertainties on the Signal Simulation

8.2.1 Jet Energy Scale

8.2.2 Jet Energy Resolution

8.2.3 Large-R Jet Mass Uncertainties

8.2.4 Flavour Tagging Uncertainties

8.2.5 Other Uncertainties

8.3 Experimental Uncertainties on the Background Model

8.4 Impact Assessment

9 Statistical Analysis

9.1 Statistical Fit Outline

9.1.1 Significance of an excess

9.1.2 Limit setting setup

9.1.3 Asimov Dataset and Expected Limits

9.2 Interpretation Strategy

9.2.1 Mass-dependent limits on the production cross section

9.2.2 Mass-dependent limits on the coupling

9.2.3 Limits on theoretical model features

10 Results

10.1 Unblinded Signal Region results

10.2 Statistical Fit Results

10.2.1 Background-Only Fit

10.2.2 Unconditional Signal Strength Estimates

10.2.3 Systematics Ranking

10.3 Observed Limits

10.3.1 Exclusion Plots: Cross Section

10.3.2 Exclusion Plots: Phenomenology
<table>
<thead>
<tr>
<th>Contents</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4 Comparison with the state of the art</td>
<td>229</td>
</tr>
<tr>
<td>10.4.1 Single Production Searches</td>
<td>229</td>
</tr>
<tr>
<td>10.4.2 Pair Production Searches</td>
<td>233</td>
</tr>
<tr>
<td>10.5 Summary and discussion</td>
<td>236</td>
</tr>
<tr>
<td>Appendices</td>
<td>238</td>
</tr>
<tr>
<td>A Kernel-based extraction of reweighting functions in the background</td>
<td>238</td>
</tr>
<tr>
<td>estimation</td>
<td></td>
</tr>
<tr>
<td>A.1 Smoothing Technique</td>
<td>238</td>
</tr>
<tr>
<td>A.2 Treatment of Extremes</td>
<td>239</td>
</tr>
<tr>
<td>A.3 Choice of bandwidth parameter</td>
<td>240</td>
</tr>
<tr>
<td>A.4 Uncertainty on reweighting functions</td>
<td>246</td>
</tr>
<tr>
<td>B ABCD Modelling Limitations</td>
<td>247</td>
</tr>
<tr>
<td>B.1 Background homogeneity</td>
<td>247</td>
</tr>
<tr>
<td>B.1.1 Z+b-jets</td>
<td>249</td>
</tr>
<tr>
<td>B.1.2 All Hadronic $t\bar{t}$</td>
<td>250</td>
</tr>
<tr>
<td>B.1.3 Conclusion</td>
<td>254</td>
</tr>
<tr>
<td>B.2 Signal Spillage in Regions B, C, D</td>
<td>255</td>
</tr>
<tr>
<td>B.2.1 Assessment</td>
<td>256</td>
</tr>
<tr>
<td>B.2.2 Correction Scheme</td>
<td>257</td>
</tr>
<tr>
<td>B.2.3 Discussion</td>
<td>258</td>
</tr>
<tr>
<td>Bibliography</td>
<td>259</td>
</tr>
</tbody>
</table>
## List of Figures

2.1 Theoretical branching fractions of a Standard Model Higgs boson for a wide range of masses 2.1a and for a 5 GeV-wide window about the experimentally observed mass. [1] ................................. 43

2.2 First order diagrams representing quantum loop corrections to the Higgs mass arising from the Higgs self-coupling (left) and coupling to gauge bosons (center) and fermions (right). [2] ................................. 44

2.3 Proposed multiplet states for Vector-Like quarks, with the relative electroweak quantum numbers displayed. Standard Model quarks are also included for reference. [3] ................................. 47

2.4 Visual representation of the Bridge Model approach to the interpretation of exotic searches. [4] ................................. 48

2.5 Tree-level Feynman diagrams for single production of a Vector-Like B quark. .................................................. 50

2.6 Leading Order PROTOS cross sections for VLQ single production for several multiplet configurations. The $Q\bar{q}j$ nomenclature refers to the final state products of each process, as visible in Fig. 2.5a-2.5b. Pair production cross section (dashed line) included for reference. [5] ................................. 51

2.7 Theoretical branching ratios of T and B quarks for the main allowed $SU(2)_L$ representation for VLQ. Blue lines show the behaviour of the branching ratios from the low mass (marked by an X) to the high-mass limit (marked by a red dot). The $BR[Q \rightarrow Wq']$ can be inferred as $BR_W = 1 - BR_Z - BR_H$. [5] ................................. 56
List of Figures
2.8

12

Comprehensive exclusion limits drawn in the combination effort of
all ATLAS pair-production searches. For each value of the two independent T and B branching ratio, the combination excludes at
95% CL the occurrence of the relevant Vector-Like quark with mass
represented by the color coding. The overlaid yellow symbols show
popular coupling configurations as outlined in Sec.2.2.3. [6]. The
solid, dotted and dashed white lines follow the paths of constant
excluded resonance mass in the branching ratio phase space. . . . . 57

3.1

Schematic view of the CERN accelerator complex, highlighting the
modular acceleration path from the original sources to the LHC. [7]

3.2

60

Cut-away representation of the ATLAS detector, showing the main
detector sub-systems. [8] . . . . . . . . . . . . . . . . . . . . . . . 64

3.3

Exploded representation of the barrel section of the Inner Detector
and its modules, along with the respective distances from the beam
axis. [9, 10] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

3.4

Schematic representation of a module of the ATLAS LAr calorimeter, highlighting the accordion structure of the sampling unity, also
pictured in the right panel. . . . . . . . . . . . . . . . . . . . . . . 70

3.5

Section view of the ATLAS Muon Spectrometer with the individual
detector technologies labeled. [11]. . . . . . . . . . . . . . . . . . . 74

3.6

Flow chart of the ATLAS TDAQ system during LHC Run 2. [12] . . 76

4.1

Flow chart outlining the Particle Flow iterative reconstruction procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

4.2

Schematic representation of the steps in the "Trimming" procedure
applied to reduce pile-up dependence on large-R jets observables. . 85


4.3 Pictorial representation of the substructure-tagging strategy for a $H \rightarrow b\bar{b}$ initiated large-R jet, here displayed in the unrolled calorimeter plane ($\eta \times \phi$). Red and blue circles represent respectively energy deposits from the actual Higgs decay products and from pile-up interactions. The size of the circle is taken as measure of the cluster energy. In the right-hand side panel, a two-pronged substructure is identified by means of reconstructing and matching Variable-RADIUS Track Jets to the large-R jet.  

4.4 Schematic representation of the feature of the track system originated by a $b$ quark in the final state and reconstructed as a jet. A secondary vertex is produced by virtue of the $b$-hadron decaying into several particles, including a $c$-hadron which in turn decays generating a tertiary vertex. A large impact parameter can also be observed, defined as the closest transverse distance from the primary vertex of the measured tracks as extrapolated to the beam level.  

4.5 Light quark-initiated jet rejection as a function of the $b$-jet efficiency working point for a number of flavour taggers trained on Particle Flow calorimeter jets (top) or Variable-RADIUS track jets (bottom).  

5.1 Feynman diagrams of the main Leading-Order single-production modes for a Vector-Like $B$ quark. The decay chain targeted by the analysis is also displayed.  

5.2 Selected summary plots for the Run 2 data-taking campaigns. Displayed are the integrated luminosity progression (a), the peak instantaneous luminosity delivered in 2018 (b) and the yearly breakdown of the usable fraction of the recorded data (c) as well as the distribution of the average $pp$ interaction per bunch crossing. [13, 14]
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>NLO Cross Section and fractional width $\Gamma/M$ for a Vector-Like B quark resonance with varying pole mass (x axis) and coupling strength with the Standard Model ($\kappa$, on the y axis). 104</td>
</tr>
<tr>
<td>5.4</td>
<td>Particle-level resonance mass for Z and W-Initiated VLB production in the various coupling strength hypotheses. 106</td>
</tr>
<tr>
<td>5.5</td>
<td>VLB Line Shape for resonance of varying pole mass, given a fixed value of the coupling strength $\kappa$. 106</td>
</tr>
<tr>
<td>5.6</td>
<td>Particle-level resonance mass for $B \to bZ$ production in the various coupling strength hypotheses. 107</td>
</tr>
<tr>
<td>5.7</td>
<td>Particle-level Higgs $p_T$ for Z-Initiated signal events. 107</td>
</tr>
<tr>
<td>6.1</td>
<td>Triggering efficiency for various large-R jet triggers with a common High Level Trigger threshold set to $E_T = 460$ GeV. The different curves highlight the different behaviour of the triggering efficiency turn-on observed for three different pile-up mitigation algorithms. The search covered in this thesis makes use Trimmed large-R jets only. 116</td>
</tr>
<tr>
<td>6.2</td>
<td>Individual efficiency of the 6 pre-selection steps as defined and labeled in Table 6.8. The efficiency have been computed for a number of values of the coupling strength $\kappa$. 119</td>
</tr>
<tr>
<td>6.3</td>
<td>Correlation plots displaying, for each event, the DL1 score of the two Higgs-associated track jet in the H2T2B (left) and H2T1B (right) categories. Warmer colors represent higher event occupancy. 124</td>
</tr>
<tr>
<td>6.4</td>
<td>Reconstructed Higgs mass distribution in Z- (top row) and W-initiated samples (bottom row). Candidates are sorted by category. All histograms are normalized to unit area. 126</td>
</tr>
</tbody>
</table>
6.5 Higgs Candidate reconstruction efficiency. In each sub-figure, the red, orange, yellow and green series represent the partial reconstruction efficiencies pertaining respectively to the H2T2B and H2T1B categories, with their combined efficiency displayed as the black series. Note that the VLB samples were generated inclusively in Higgs decays and this influences the proportion of events in the three categories.

6.6 Normalized distributions of selected kinematic variables for simulated signal events belonging to the 1200 GeV, 1600 GeV and 2000 GeV samples. All the events displayed belong to the J1B_H2T2B category and feature at least one reconstructed VLB candidate.

6.7 Signal significance for different Higgs mass Peak Region boundary configurations. The optimal sensitivity point is marked by the superimposed star. Z-initiated (left) and W-initiated (right) samples are treated independently.
6.8 Forward jet multiplicity in signal and background data samples. Background behaviour approximated by the low-signal purity J0B_H2T2B sample. Signal and background are normalized to a common area for easier viewing. All events satisfy the VLB peak condition: $|m - M_B^{pole}| < 150$ GeV, with $M_B^{pole}$ being the nominal VLB sample mass. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

6.9 Increase in signal significance by requiring the event to have a forward jet multiplicity at least equal to the value given by the $x$-coordinate. Multiplicity bins are exclusive, meaning that events falling into a specific bin have exactly the number of forward jets as displayed in the bin label.

6.10 $\log \Delta R^*$ in signal and background data samples. Background behaviour approximated by the low-signal purity J0B_H2T2B sample. Signal and background are normalized to a common value for easier viewing. All events satisfy the VLB peak condition: $|m - M_B^{pole}| < 150$ GeV, with $M_B^{pole}$ being the nominal VLB sample mass. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

6.11 Increase in signal significance by requiring the event to have $\log \Delta R^* > (\log \Delta R^*)_{threshold}$.

6.12 Signal and data distributions of the $|\Delta \eta|$ variable. Arbitrary units. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

6.13 Increase in signal significance by requiring the event to have $|\Delta \eta| < (|\Delta \eta|)_{threshold}$.
6.14 Two possible tree level diagrams regulating the QCD-driven production of $3b$ quarks in the final state. In the left-hand side diagram the momentum exchange occurs via the $s$-channel, whereas the diagram on the right shows a typical $t$-channel process. 148

6.15 Signal and data distributions of the $p_T(HC)/m_B$ variable. Arbitrary units. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal. 149

6.16 Increase in signal significance by requiring the event to have $p_T(HC)/m_B > (p_T(HC)/m_B)_0$. 149

7.1 Schematic representation of the ABCD background estimation method for two variables $X$ and $Y$. 157

7.2 Graphical visualization of the ABCD partition scheme employed in this search. 158

7.3 Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A. 162

7.3 Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A. (Cont.) 163

7.3 Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A. (Cont.) 164
7.4 Comparison of the distributions in the C and D regions for the two variables used to calculate the shape correction factors. The normalization scale factor $k_{FWD}$ is already applied to D events. All events belong to the Higgs mass peak region. The extracted weight function and its uncertainty are shown as a red line and a pink shaded area, respectively, in each of the ratio plots. The plots in the bottom row show the agreement between C and D data after the weight functions are applied to D itself. ....................................................... 169

7.5 Closure test for the per-event shape correction in regions C,D. Event weights are calculated from ratios between templates distribution in C and D, and applied here to the data in D (solid). A comparison is drawn with C events to test closure. All events belong to the Higgs Mass region (SR). ....................................................... 171

7.5 Closure test for the per-event shape correction in regions C,D. Event weights are calculated from ratios between templates distribution in C and D, and applied here to the data in D (solid). A comparison is drawn with C events to test closure. All events belong to the Higgs Mass region (SR). (Cont.) ....................................................... 172

7.6 H2T2B channel. Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband. ....................................................... 181

7.6 H2T2B channel. Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband. (Cont.) ....................................................... 182

7.7 H2T1B channel. Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband. ....................................................... 183
7.7 (Continued from previous page) - H2T1B channel. Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband. . . . . . . . . . . . . . . . . 184

7.8 Distribution of the reconstructed VLB mass in H2T1B data for subsamples 1 to 8. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 186

8.1 Schematic flow diagram outlining the individual steps in the R=0.4 jet energy scale calibration. [15]. . . . . . . . . . . . . . . . . . . . 189

8.2 Summary of the JES uncertainty for R = 0.4 jets (top row) and R = 1.0 jets. The total uncertainty is represented as the blue envelope over a range of both the jet transverse momentum and pseudorapidity. The macro-contributions to the total uncertainty from each class of sources is also displayed. The green envelope in the large-R jet JES uncertainty plot represents the uncertainty level established in the previous set of recommendations, highlighting a sharp improvement in the quality of the measurements. . . . . . . . . . . . . . . . 190

8.3 Impact on the shape of \( m_B \) (reconstructed mass) in a \( M_B = 1.2 \) TeV (resonance mass) (top) and \( M_B = 2.0 \) (bottom), \( \kappa = 0.5 \) simulated signal sample for two of the 30 independent components of the Small-R jet Jet Energy Scale uncertainty. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each \( m_B \) bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework. 191
8.4 Impact on the shape of $m_B$ (reconstructed mass) in a $M_B = 1.2$ TeV (resonance mass) (left) and $M_B = 2.0$ (right), $\kappa = 0.5$ simulated signal sample for two of the 6 independent components of the Small-R jet Jet Energy Resolution. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each $m_B$ bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework.

8.5 Impact on the shape of $m_B$ in a $M_B = 1.2$ TeV (resonance mass) (left) and $M_B = 2.0$ (right), $\kappa = 0.5$ simulated signal of the leading contributor to the JMS uncertainty. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each $m_B$ bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework.

8.6 Monte Carlo driven estimation of the Jet mass Resolution for Higgs (left) or QCD (right) originated $R = 1.0$ jets. The JMR for Higgs jets is provided for jets falling on both sides of the theoretical Higgs peak, while values of the QCD JMR are displayed for three different ranges of $m^{jet}/p_T^{jet}$.

8.7 Impact on the shape of $m_B$ (reconstructed mass) in a $M_B = 1.2$ TeV (resonance mass) (left) and $M_B = 2.0$ (right), $\kappa = 0.5$ simulated signal of uncertainty introduced by the large-R Jet Mass Resolution. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each $m_B$ bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework.
8.8 Impact of the custom systematic uncertainties on the prediction of the Standard Model background contribution in the Signal Region. The shaded area represents the statistical uncertainty on the background model inherited from the original data sample on which the entire yield/shape correcting procedure was applied.

10.1 Distribution of the VLB system invariant mass in the Signal Region for the predicted multi jet contribution (solid black line) and the observed full Run 2 data (dots). The shaded area about the multi jet prediction represents the quadrature sum of all systematic uncertainties affecting the background. The $m_B$ distribution for a 1.3 TeV $\kappa = 0.3$ VLB signal is overlaid to both the main distribution and the ratio plot as a red dashed line.

10.2 Comparison between data and Standard Model prediction for the Signal Region data for a number of key kinematic variables. The comprehensive systematic uncertainty on the background prediction is represented as the shaded band. The expected contribution from a 1.3 TeV, $\kappa = 0.3$ signal is overlaid for comparison.

10.3 Year by year breakdown of the observed data and the predicted Standard Model contribution. The $1.25 \text{ TeV} < m_B < 1.4 \text{ TeV}$ bin is highlighted by the red transparent bar.

10.4 Distribution of $m_B$ for the observed data and the background prediction before (left) and after (right) the fit to the data. The Z-Initiated (leading) and W-Initiated (sub-leading) contributions to a VLB signal with $M_B = 1.3 \text{ TeV}$ and $\kappa = 0.3$ are overlaid for reference.

10.5 Background-only pulls for the nuisance parameters regulating the expected shape and yields of the background contribution.
10.6 Ranking plots for the 10 leading systematics by impact on the signal strength $\mu$ in each resonance mass hypothesis. The blue and teal represent the impact (to be read against the top axis scale) on the value of $\mu$ of respectively a $+1\sigma$ and $-1\sigma$ pull of each nuisance parameter. The nuisance parameter pulls obtained within each conditional fit are represented by the position of the dots, to be evaluated against the bottom axis scale. The $\gamma$ nuisance parameter have a nominal value of 1. ................................. 222

10.7 Exclusion plots on the production cross section times branching ratio for a number of $\kappa$ hypotheses. Expected (dashed line) and observed (solid line) 95% $CL_s$ limits are compared with the theoretical value of $\sigma(pp \rightarrow B \rightarrow bH)$, whose dependence on the resonance mass is conveyed as the red line. ................................. 224

10.8 Expected and observed exclusion limit on the $c_W$ coupling for a VLB singlet. ................................. 226

10.9 95% $CL_s$ exclusion limit on the resonance mass for the generic $c_W,c_Z$ combination. $\tilde{c}_Z = \tilde{c}_H$ is assumed throughout. The beige-coloured areas refer to combinations resulting in a value of $\kappa$ which either lies outside the available range or for which no mass exclusion is. All singlet configurations, which are obtained for $c_Z = \frac{M_Z}{M_W \sqrt{2}} \times c_W$, lie on the overlaid solid black line. ............................. 227

10.10 Expected and observed limit on the VLB mass as a function of the branching ratio into the $tW$ decay channels and the fractional decay width. ................................. 228

10.11 Expected and observed exclusion limits for a Vector-Like B occurring as part of a (B,Y) weak isospin doublet extracted by the ATLAS Run 2 $B \rightarrow bH(\gamma\gamma)$ search. ................................. 231
10.12 Expected and observed 95% CL$_s$ exclusion limit on $\sigma(pp \to B \to bH)$ as a function of the resonance mass and fractional width. The overlaid dashed lines represent contours of constant coupling strength $\kappa$. 233

10.13 Expected and observed 95% CL$_s$ exclusion limit on $\sigma(pp \to B \to bH)$ extrapolated by the CMS $B \to bH(b\bar{b})$ search for two given values of the fractional resonance width [16]. 234

10.14 Mass dependent exclusion limit on the main coupling in the $B$ singlet (left) and $(B,Y)$ doublet scenarios. For a $(B,Y)$ doublet, $c_W = 0$, requiring the limits to be expressed in terms of $c_Z = M_Z/M_W \times \kappa$. In both plots, the exclusion limits on the $B$ resonance mass drawn by the most recent ATLAS and CMS pair-production results are represented by the vertical blue and red lines. 235

A.1 Random Iteration N. 3. Regression of the small-$R$ jet $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth. 243

A.2 Random Iteration N. 3. Regression of the VLB $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth. 244

A.3 Random Iteration N. 128. Regression of the small-$R$ jet $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth. 244
A.4 Random Iteration N. 128. Regression of the VLB $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth. .......................... 245

A.5 Values of the two-fold cross-validation error with respect to the reduced bandwidth $h/b$. Values are averaged over 200 iterations with stochastic splitting of the original data sample in training and testing sets. .................................................. 245

B.1 Value of the bias $\lambda_{t\bar{t}}$ for every bin of the distributions of the two variables involved in the shape correction. The shaded areas represent, for each bin, the systematic uncertainty on the value weight function extracted from the variable under study. The statistical uncertainties in the $t\bar{t}$ sample are not displayed here and are quite large particularly in the high VLB $p_T$ range. .......................... 252

B.2 Monte Carlo prediction for $t\bar{t}$ contribution to the Higgs Peak Region A (dots) and Region B (solid) data. The comparison is provided for $H^2T^2B$ (left) and $H^2T^1B$ Higgs Peak Region Data. $k_{FWD}$, the shape-correcting event weights as calculated on C and D data, and $R_{corr}$ are applied to B-Region events. The error bars and the shaded areas represent the statistical uncertainty on the Monte Carlo simulation. .......................... 253

B.3 Comparison between the Nominal background model derived neglecting the top contribution to the data (black, solid) and handling the subtracting the top before modelling, then re-adding the A-Region top prediction to the result (red dashed line). .......................... 254
List of Tables

5.1 Next-to-Leading Order production cross section for a Z-Initiated standalone Vector-Like B quark given a resonance pole mass ranging from 900 GeV to 2 TeV. The displayed figures have been recomputed from the values generated by MadGraph under the Narrow Width Approximation, and corrected for finite width effects as outlined in the main body. .................................................. 103

5.2 Next-to-Leading Order production cross section for a W-Initiated standalone Vector-Like B quark given a resonance pole mass ranging from 900 GeV to 2 TeV. The displayed figures have been recomputed from the values generated by MadGraph under the Narrow Width Approximation, and corrected for finite width effects as outlined in the main body. .................................................. 103

6.1 Small-$R$ Jet reconstruction criteria. ............................................. 110
6.2 R = 1.0 Jets reconstruction criteria. ............................................. 111
6.3 Summary of selection working points for track jets. ...................... 112
6.4 $b$-tagging selection criteria. .................................................... 113
6.5 Electron and Muon selection criteria. ......................................... 114
6.6 Photon selection criteria. ......................................................... 115
6.7 List of triggers used by the analysis, each associated with the datasets on which they were applied. .................................. 115
6.8 Sequential description of the pre-selection step. A unique identifier is provided for each step for quick reference in the main body and in subsequent tables .................................................. 118
6.9  Step by step efficiency in the Pre-Selection for data and simulated
signal events. The Z-Initiated (\( ZBHb \)) and W-Initiated (\( WBHb \))
production modes are left distinct to better appreciate the differ-
ences. Signal efficiencies refer to datasets generated with \( \kappa = 1.0 \).
6.10  Summary of reconstruction criteria for Higgs Candidates.
6.11  Intermediate Higgs Candidate reconstruction cutflow (top section)
and efficiencies (bottom) for simulated 1200 GeV and 2000 GeV
VLB signal events. Efficiencies are provided for events passing the
Nominal Pre-Selection. Efficiencies pertaining to steps 1:2a and
1:2b are both calculated with respect to the events passing 1:1.
6.12  Naming convention for Higgs Candidates. The 2T part refers to the
number of associated track jets, whereas \( nB \) represents the number
of \( b \)-tags in the candidate.
6.13  Fraction of signal events passing PreSelection (0:6) that fall in the
\( H2T2B \) or \( H2T1B \) categories, or are rejected at the Higgs Candidate
reconstruction stage.
6.14  Summary of reconstruction criteria for VLB Candidates.
6.15  Reconstruction efficiency of the VLB candidate in \( H2T2B \) with re-
spect to three selected reference points in the reconstruction proto-
col, for the Z-initiated signal. For the second and third columns, the
combined efficiency of all previous steps is factored out from the
efficiency figures displayed.
6.16  Naming convention for the event categories. The \( JnB \) part refers to
the number of \( b \)-tags in the small-R jet, whereas the following part
of the name refers to the Higgs candidate reconstruction category.
6.17  Fractional yields for simulated signal events sorted by event cate-
gory. Percentage refer to the number of events with at least one
reconstructed VLB candidate.
6.18 Event yields for 2016 data events with at least one VLB candidate. The Higgs Peak is preliminarily defined as $M_{HC} \in [105, 135]$ GeV, with the off-peak range instead from 135 GeV to 200 GeV.

6.19 Values of $\mu_{QCD}$ for each Higgs Category, evaluated on 2016 data.

6.20 Significance of signal over the 2016 Data in each VLB reconstruction category. Values are relative to the reference signal significance estimated for the Signal Region.

6.21 Significance of signal over the 2016 Data combining each category with the reference $J_{1B_{H2T2B}}$ category. Values are relative to the reference significance.

6.22 Optimal Higgs Peak region boundaries. All figures expressed in units of GeV.

6.23 Optimal cut threshold and signal significance gain for a $\log \Delta R^* > (\log \Delta R^*)_{threshold}$ selection step.

6.24 Comparison of the optimal cut thresholds and signal significance gains for a cut-based selection steps on $|\Delta \eta|$ and $p_T(HC)/M_{VLB}$.

6.25 Event yields for data throughout the reconstruction and event selection stages. The yields marked with a diamond ($\diamond$) are partially blinded and refer to events outside the Higgs Peak Region only.

6.26 Event yields for Z-initiated simulated signal throughout the reconstruction and event selection stages. The yields for the first two steps of the selection represent the number of unweighted events passing each respective steps.

6.27 Event yields for W-initiated simulated signal throughout the reconstruction and event selection stages. The yields for the first two steps of the selection represent the number of unweighted events passing each respective steps.

7.1 Schematic representation of the definition of the four ABCD Regions.
7.2 Event yields in the four ABCD regions in the Validation Sideband for \textbf{H2T2B} data and in the Higgs Peak and Validation Sideband for \textbf{H2T1B} data. The respective values of \( R_{\text{corr}} \) and their uncertainties are also displayed. ............................. 161

7.3 Event yields in the four ABCD regions in the Higgs Peak region \((m_H \in [105, 135] \text{ GeV})\). ................................. 165

7.4 Simulated signal \((M_M = 1.2 \text{ TeV}, \kappa = 1.0)\) and observed data yields in the four ABCD regions built in the Higgs mass sideband range. For comparison, the predicted signal and background yield in the Signal Region are also displayed. ............................. 179

7.5 Summary of the Signal/Background numerical interplay in the main channel and in the average \textit{J1B\_H2T1B} sub-sample. \( S/B \) ratios and signal significance \( S/\sqrt{B} \) are provided for both the theoretical 1200 \text{ GeV} cross section \((\sim 31 \text{ fb})\) and the \( B \rightarrow bH(\gamma\gamma) \) exclusion limit at the same mass, roughly amounting to three times as much at 100 \text{ fb}. ................................. 185

8.1 Bin-wise impact of the statistical uncertainty on ABCD-Region B data propagated onto the background model. ............................. 200

8.2 Relative effect of each group of systematics on the yields of the predicted background and two simulated signals with resonance mass respectively equal to 1.2 and 2.0 \text{ TeV}, \( \kappa = 0.5 \). ............................. 201

10.1 Pre- and post-fit yields for signal and background in the background only hypothesis. The post-fit signal yield is accordingly zero by definition (conditional fit with \( \mu = 0 \)). ............................. 218

10.2 Unconditional best fit estimate for the signal strength \( \mu \) in a number of \( M_B \times \kappa \) configurations. For each configuration, the uncertainty provided on \( \mu \) represents the quadrature sum of the statistical and systematic contributions. ............................. 220
10.3 Summary table displaying the expected and observed mass intervals that are excluded at the 95% $CL_s$ for each value of the signal strength $\kappa$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 223

B.1 Estimated cutflow for the $Z+b$-jets background. . . . . . . . . . . . . 250

B.2 Higgs Peak Region ($M_H \in [105, 135]$GeV) yields for all hadronic $t\bar{t}$ events in the $H2T2B$ category. The displayed uncertainties on the $t\bar{t}$ yields and the $k_{FWD}$ scaling ratios only reflect the statistical uncertainty on the Monte Carlo simulation. . . . . . . . . . . . . . 251

B.3 Higgs Peak Region ($M_H \in [105, 135]$GeV) data yields in the $H2T2B$ category. The Region-A yield (red) represents the final estimate for the number of data events in the Signal Region, including the effects of the shape correction and $R_{corr}$ re-scaling. . . . . . . . . . . . . . 251

B.4 Event yields in the four ABCD regions in the Higgs Peak region ($m_H \in [105, 135]$ GeV). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 256
Chapter 1

Introduction

Ever since the emergence of intellectually modern humans, generations upon genera-
tions of thinkers, philosophers, naturalists and most recently scientists have tire-
lessly struggled to conquer the *golden fleece* of speculative thought: a comprehen-
sive philosophical framework capable of interpreting and predicting every mani-
festation of the natural world, from the frantic workings of the infinitesimal to the
immutable order of the heavens. A theory of Everything.

Although the quest for ultimate understanding has proven and shall likely always
remain elusive, enormous and ever faster progress has been made by broadening
and increasing the depth of our experimental and theoretical reach. To this point,
it is worth reflecting that, in the space of little more than a century, elements have
yielded to atoms, atoms to nucleons and nucleons to subatomic particles as the ulti-
mate known fundamental building units of the physical universe.

Throughout this time, when experimental physics unlocked new and unexpected
realities, theoretical physics would soon follow with an interpretative model, and
whenever new theories were put forward, experiments would soon be carried out to
validate their predictions. In the final decades of the twentieth century, this fruit-
ful interplay of theory and experiment resulted in the formulation of the Standard
Model [17], an overreaching framework describing the genesis and the features of
the 17 fundamental particles and most importantly their interactions through three
of the four known fundamental forces.
More than four decades into its life, the Standard Model has proven a remarkably accurate and durable theory, with the experimental observation of a particle compatible with the Higgs boson [18, 19] at the Large Hadron Collider [20] being often regarded as its crowning achievement. There are, however, both theoretical and experimental clues to the existence of new and undiscovered phenomena beyond the reach of the Standard Model, spurring a worldwide effort from hundreds of research groups to either detect conclusive evidence of new physical processes or produce a credible extension or alternative to the current theoretical standard.

This thesis will provide a detailed account of the search for evidence of production of a Vector-Like partner to the Standard Model bottom quark in proton-proton collisions generated at the Large Hadron Collider and collected by the ATLAS Experiment detector [21].

The first chapter will present an outline of both the Standard Model and the relevant theoretical extension predicting the existence and features of the particle targeted by the search, while the second chapter will provide an overview of the experimental system enabling production and collection of the data by ATLAS. Further chapters will be dedicated entirely to an in-depth presentation of the work carried out for the analysis of the available data, ranging from the outline of the analysis strategy to the ultimate extraction of mass and coupling dependent exclusion limits on the signal.
Chapter 2

Theoretical Overview

2.1 The Standard Model of Particle Physics

The Standard Model (SM) [22] is the most advanced well-tested theoretical framework describing the properties and mutual interactions of all known fundamental particles. It is a quantum field theory (QFT) [23] constructed out of the gauge invariance principle, which requires any physical processes, described by a Lagrangian density function $\mathcal{L}$, to be invariant under a local transformation generated by a unitary group.

In accordance with the quantum field theory structure of the Standard Model, fundamental physical particles are described as excitation states of quantum fields with a well defined and unique set of quantum numbers, such as mass and spin. The value of the latter allows for a most general classification: particles with half-integer spin number are called fermions, while integer spin fields give rise to particles referred to as bosons.

The simplest form of the Lagrangian density function for a spin one-half quantum field $\psi$, known as the Dirac Lagrangian, is:

$$\mathcal{L} = i \bar{\psi} \gamma^\mu \partial_\mu \psi - m \bar{\psi} \psi$$  \hspace{1cm} (2.1)
where $\gamma^\mu$ represents the set of four four-dimensional Dirac matrices, each acting on the spin components of the particle field $\psi$, and $\partial_\mu$ the four-derivative acting on its space-time coordinates.

The Dirac Lagrangian does not meet the fundamental principle of gauge invariance, as seen by applying a simple $U(1)$ local symmetry. The particle field $\psi$ transforms under $U(1)$ as:

$$\psi(x) \rightarrow \psi'(x) = e^{iq\chi(x)} \psi(x) \quad (2.2)$$

with $q$ being a generic constant and $\chi(x)$ an arbitrary local phase. Applying $U(1)_q$ to the Dirac Lagrangian yields:

$$\mathcal{L}' = \mathcal{L} - q\bar{\psi}\gamma^\mu (\partial_\mu \chi) \psi \quad (2.3)$$

where the second term, arising from the gauge transformation of the kinetic term $i\bar{\psi}\gamma^\mu \partial_\mu \psi$, breaks the invariance. To re-instate gauge invariance, the derivative $\partial_\mu$ needs to be replaced with the covariant derivative $D_\mu$, defined as:

$$D_\mu = \partial_\mu + iqA_\mu \quad (2.4)$$

in which $A_\mu$ is a spin-one additional field itself transforming as:

$$A_\mu \rightarrow A'_\mu = A_\mu - \partial_\mu \chi \quad (2.5)$$

The full locally $U(1)$-invariant Lagrangian can therefore be written out as:

$$\mathcal{L} = i\bar{\psi}\gamma^\mu \partial_\mu \psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - m \bar{\psi} \psi - q\bar{\psi}\gamma^\mu \psi A_\mu \quad (2.6)$$

Each term of the gauge-invariant Lagrangian has a precise physical meaning:

- $i\bar{\psi}\gamma^\mu \partial_\mu \psi$ is the free propagation term of the spin one-half field $\psi$;
- $\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$, with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ represents the kinetic term for the spin-one gauge field $A_\mu$;
2.1. The Standard Model of Particle Physics

- \( m\bar{\psi}\psi \) defines the mass of the fermion;

- \( q\bar{\psi}\gamma^\mu \psi A_\mu \) defines the interaction between the fermion and the gauge fields.

To summarize, requiring the free-fermion Lagrangian to be invariant under a local \( U(1) \) symmetry results in the definition of a massless, spin-one gauge field \( A_\mu \) which interacts with the fermion field with a coupling constant \( q \). It is worth remarking that a mass term for the gauge field, such as \( m_A^2 A_\mu A^\mu \), would violate the invariance of the Lagrangian, given the \( U(1) \) transformation paradigm for \( A_\mu \) defined in Eqn. 2.5.

The theoretical setup outlined here, while simple, has nevertheless proven exceptionally effective in describing and predicting the electromagnetic interactions of charged fermions such as massive leptons and quarks, in which the gauge field \( A_\mu \) is interpreted as the electromagnetic potential or a massless mediator boson (photon) according to the preferred quantum mechanical interpretation.

The resulting theory, known as quantum electrodynamics (QED) [24], was the first and arguably most successful building step of the Standard Model, with its predictions being verified to a higher degree of precision than any other theory in physics.

2.1.1 Electroweak Unification

Weak interactions, owing to their unique traits of parity violation [25] and charged-current processes, cannot be framed within a quantum field theory originating from a local \( U(1) \) symmetry, requiring instead a more complex chiral \( SU(2)_L \) gauge symmetry acting on left-handed weak isospin doublets. For leptons:

\[
\varphi_L(x) = \begin{pmatrix} \nu_L(x) \\ \ell(x) \end{pmatrix} \quad \rightarrow \quad \varphi'(x) = e^{iR_W \alpha(x) \cdot \frac{1}{2} \sigma} \varphi_L(x)
\] (2.7)
with $\mathbf{a}(x)$ being an arbitrary local phase vector and $\mathbf{\sigma}$ the three 2-dimensional Pauli matrices representing the generators of the $SU(2)$ group. Quarks transform identically, with two components of the left-handed isospin doublet being a generic up-type quark and an admixture of the three down-type quark fields.

Once again, gauge invariance in the Dirac Lagrangian is ensured by replacing the space-time derivative with a covariant form encompassing the symmetry group generators, acting on a novel spin-one gauge field. As three matrices make up the set of $SU(2)$ generators, three gauge fields arise from the enforcement of gauge invariance:

$$D^\text{weak}_\mu = \partial_\mu - ig \frac{1}{2} \mathbf{\sigma} \cdot \mathbf{W}_\mu$$

(2.8)

The three interaction terms of the $SU(2)_L$ gauge invariant Dirac Lagrangian can be written in a compact form as the dot product of three fermion currents with the gauge fields:

$$j^\mu \cdot \mathbf{W}_\mu \quad \text{with} \quad j^\mu_i = \frac{g}{2} \bar{\psi}_L \gamma^\mu \sigma_i \psi_L$$

(2.9)

The physical interaction terms involved in real-world charged-current weak processes such as the decay of the muon, beta decay or electron-neutrino scattering, are obtained as a linear combination of the first two terms of $j^\mu \cdot \mathbf{W}_\mu$, with the physical $W$ bosons being therefore identified as:

$$W^\pm_\mu = \frac{1}{\sqrt{2}} \left( W^1_\mu \mp W^2_\mu \right)$$

(2.10)

Owing to the diagonal form of the third Pauli matrix, the $j^3$ current does not involve a change in the third component of the weak isospin between the initial and the final fermion, with the relative gauge boson $W^3_\mu$ therefore mediating what is called a weak neutral current interaction.

While the left-handed $SU(2)$ field theory seems to provide a compelling case for the description of weak interactions, strong hints of evidence of its incompleteness also exists. Primarily, while it is tempting to associate the neutral gauge field $W^3_\mu$ to the experimentally detected [26] weak neutral boson $Z$, the latter was observed to
couple with the right handed component of massive fermions\(^1\), while \(W_\mu^3\) does not by definition of the chirality of its underlying gauge symmetry.

Furthermore, the calculated cross section of the \(\ell^+\ell^- \rightarrow W^+W^-\) production, mediated in the t-channel by the exchange of a neutrino and in the s-channel by the exchange of a photon\(^2\) or a Z boson, diverges at high center-of-mass energies unless the couplings between the W, the photon and the physical Z bosons are related to one another.

The experimental evidence for a level of unification between the electromagnetic and the weak sectors of particle physics was interpreted through a theoretical mechanism owing to the work of Glashow, Salam and Wineberg [28–30]. The GSW (after the authors’ initials) field theory defines a composite \(U(1)_Y \otimes SU(2)_L\), with the \(U(1)_Y\) symmetry group replacing the electromagnetic \(U(1)\) symmetry and introducing a new physical quantity, the weak hypercharge \(Y\), and a new gauge field \(B_\mu\). After imposing gauge invariance on the fermion Lagrangian by defining the electroweak covariant derivative:

\[
D^{EW}_\mu = \partial_\mu - i\frac{g'}{2}YB_\mu - i\frac{g}{2}\sigma \cdot W_\mu
\]

the physical neutral boson fields, the photon \(A_\mu\) and the \(Z_\mu\), are defined as linear combination of the \(U(1)_Y \otimes SU(2)_L\) gauge fields responsible for neutral current interactions via a mixing angle \(\theta_W\):

\[
A_\mu = +B_\mu \cos \theta_W + W_\mu^3 \sin \theta_W
\]

\[
Z_\mu = -B_\mu \sin \theta_W + W_\mu^3 \cos \theta_W
\]

with the physical \(Z_\mu\) boson acquiring a coupling to the right-handed fermion component through its non-chiral \(U(1)_Y\) component. Requiring the resulting electroweak Lagrangian to reproduce the results of the individual electromagnetic and weak field theories produces fundamental relations connecting the weak hypercharge to the

\(^1\)This was detected in the context of the observation of an asymmetry in the Z production cross section from left or right-handed polarized electrons at SLAC [27].

\(^2\)to which the charged W bosons couple owing to their carrying electrical charge.
electromagnetic charge$^3$ and third component of the weak isospin:

$$Y = 2\left(Q - I_W^3\right)$$  \hspace{1cm} (2.13)$$

further linking together the electromagnetic, weak and $U(1)_Y$ coupling constants as:

$$e = g_W \sin \theta_W = g' \cos \theta_W$$  \hspace{1cm} (2.14)$$

While the GSW theory introduces an elegant and effective unified framework for electroweak interactions, convincingly passing experimental tests such as the discovery and measurements of the W and Z resonances [31], it failed to provide a theoretically sound mechanism to explain the most striking feature of weak interactions which sets them apart from the electromagnetic and strong forces: the non-zero mass of the weak gauge bosons $W^\pm$ and $Z^0$.

### 2.1.2 The Higgs Sector

The quantum field theory archetype re-defines particles as excitation states of quantum fields $\phi(x)$ about the vacuum expectation value represented by the minimum(a) of a potential as defined in the relevant Lagrangian. For a single free-propagating scalar particle, for instance, where the Lagrangian is:

$$\mathcal{L} = T - V = \frac{1}{2} \left( \partial_\mu \phi \right) \left( \partial^\mu \phi \right) - \frac{1}{2} m^2 \phi^2$$  \hspace{1cm} (2.15)$$

the potential $V = \frac{1}{2} m^2 \phi^2$ has a global minimum for $\phi(x) = 0$, and the interpretation of non-null $\phi(x)$ as a physical particle is straightforward.

Let us consider instead a more articulated potential featuring a quartic term in addition to the quadratic contribution:

$$V(\phi) = \frac{1}{2} \mu^2 \phi^2 + \frac{1}{4} \lambda \phi^4$$  \hspace{1cm} (2.16)$$

$^3$ expressed in units of the electron charge.
It can be seen that the behaviour and critical points of the potential depend on the mutual value and sign of the parameters $\lambda$ and $\mu^2$. Most interestingly, if $\lambda > 0$ and $\mu^2 < 0$, the potential has symmetric global minima for:

$$\phi_{\text{min}} = \pm \nu = \pm \sqrt{-\frac{\mu^2}{\lambda}}$$  \hspace{1cm} (2.17)

It’s worth noticing that under this condition the $\mu^2 \phi^2$ term cannot be interpreted as a mass term, as $\mu^2$ is negative. Having identified particles as perturbations of the field around the vacuum energy level, the Lagrangian can be reworked to be an explicit function of the particle fields by expressing:

$$\phi(x) = \pm \nu + \eta(x)$$  \hspace{1cm} (2.18)

$$\mathcal{L}(\eta) = \frac{1}{2} \partial_{\mu} \eta \partial^{\mu} \eta - \lambda \nu^2 \eta^2 - \lambda \nu \eta^3 - \frac{1}{4} \lambda \eta^4 + \frac{1}{4} \lambda \nu^4$$  \hspace{1cm} (2.19)

The new Lagrangian is exactly the same as that outlined in Eqn. 2.15, and therefore describes an identical set of physical phenomena.

However, expressing the Lagrangian as a function of the particle field $\eta(x)$ around a chosen point of vacuum $\pm \nu$ "hides" away the underlying symmetry of the model, since $\mathcal{L}(\phi) = \mathcal{L}(-\phi)$ holds true while $\mathcal{L}(\eta) = \mathcal{L}(-\eta)$ does not. The mathematical process of collapsing a symmetrical system into an asymmetrical state after choosing among a set of equally valid degenerate vacuum states is referred to as 

**Spontaneous Symmetry Breaking** [32] and plays a role in several familiar phenomena, such as the residual polarization of a magnet below its Curie temperature.

Most remarkably, however, spontaneous symmetry breaking is the gateway to a theoretical mechanism, suggested by Brout, Englert and Higgs [32, 33], justifying the non zero mass of the electroweak gauge bosons within the quantum field theory archetype without waiving the basic principle of gauge invariance.

As the mechanism must be coherently fitted within the $SU(2)_L \times U(1)_Y$ symmetry,
and provide mass to the three weak bosons, the minimal Higgs model requires the introduction of four new degrees of freedom in the form of a weak isospin doublet of complex scalar fields, with hypercharge $Y = 1/2$:

$$
\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i \phi_2 \\ \phi_3 + i \phi_4 \end{pmatrix}
$$

(2.20)

A *mexican hat*-like potential $V(\Phi)$ is defined, with the full free-propagation lagrangian being:

$$
\mathcal{L} = (\partial_\mu \Phi)^\dagger (\partial^\mu \Phi) - \mu^2 \Phi^\dagger \Phi - \lambda (\Phi^\dagger \Phi)^2
$$

(2.21)

In the $\mu^2 < 0$ configuration, the vacuum state is identified by the hyper-circle where the four field components are such that:

$$
\Phi^\dagger \Phi = -\frac{\mu^2}{2\lambda} = \frac{1}{2} \left( \phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2 \right)
$$

(2.22)

The vacuum state can be chosen without loss of generality to be purely along the $\phi_3$ direction, with the field being now expressed through the excitations around it:

$$
\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \eta_1(x) + i \eta_2(x) \\ \eta_3(x) + i \eta_4(x) \end{pmatrix}
$$

(2.23)

This expression can be simplified further by choosing a specific form for the transformations of the gauge fields introduced along with the electroweak covariant derivative. While such a specific choice, known as the *unitary gauge*, effectively removes the $\eta_1$, $\eta_2$, $\eta_4$ fields from the Lagrangian, it does not imply a loss of degrees of freedom, as those originally represented by the gauged away $\eta$ fields are transferred to the longitudinal polarizations of the now massive $W^\pm$ and $Z$ bosons.
Defining the physical Higgs field \( \eta_3(x) = h(x) \), the scalar doublet becomes:

\[
\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}
\]  

(2.24)

Replacing once again the space-time derivative \( \partial_\mu \) with the electroweak covariant derivative and working in the unitary gauge, the Lagrangian now displays both kinetic and mass terms for the Higgs scalar field and the three massive electroweak bosons, with the mass coefficient being:

\[
m_H = \sqrt{-2\lambda v^2}
\]

\[
m_W = \frac{1}{2} g_W v
\]

\[
m_Z = \frac{1}{2} v \sqrt{g_W^2 + g'^2} = \frac{m_W}{\cos \theta}
\]

\[
m_A = 0
\]  

(2.25)

2.1.3 Higgs Coupling to Fermions

In the free propagation Dirac Lagrangian for fermions, masses are defined by quadratic terms of the form:

\[
\mathcal{L} = T - m\bar{\psi}\psi
\]  

(2.26)

While quadratic terms are trivially invariant under local \( U(1) \) transformations, they break the chiral \( SU(2)_L \) symmetry enforced in the electroweak unification mechanism. This can be proved defining projection operators \( P_L \) and \( P_R \) acting on the fermion field such as that

\[
P_L \psi = \psi_L \quad \text{and} \quad P_R \psi = \psi_R, \quad \text{with} \quad P_L + P_R = 1:
\]

\[
\mathcal{L}_{\text{mass}} = -m\bar{\psi}\psi = -m\bar{\psi}\left[ P_L + P_R \right] (\psi_L + \psi_R)
\]

\[
= -m\bar{\psi}\left[ P_L \psi_L + P_R \psi_R \right]
\]

\[
= -m \left[ \bar{\psi}_R \psi_L + \bar{\psi}_L \psi_R \right]
\]  

(2.27)
with the $\bar{\psi}_L \psi_R$ terms breaking the $SU(2)_L$ invariance as left handed fermion field transform as $SU(2)$ doublets, while the right handed component transforms as a singlet. This result implies that, to preserve gauge invariance, the fermion Lagrangian must do without mass terms entirely, in turn requiring an alternative theoretical mechanism at the base of the real-world non-zero masses.

While the Higgs model as originally proposed and outlined above does not immediately provide a justification for the observed fermion masses, those can be derived through a minimal extension relying on an additional direct coupling (known as Yukawa coupling [28]) of the Higgs field with fermions in the form:

$$L_f = -g_f \left( \bar{\psi}_L \Phi \psi_R + h.c. \right)$$

This Lagrangian retains gauge invariance as both left-handed fermions and the scalar $\Phi$ field transform as doublets under $SU(2)_L \times U(1)_Y$. Working in the unitary gauge, and considering for instance a lepton doublet, the Lagrangian becomes:

$$L_\ell = -\frac{g_\ell}{\sqrt{2}} \nu \left( \bar{\ell}_L \ell_R + h.c. \right) - \frac{g_\ell}{\sqrt{2}} h(x) \left( \bar{\ell}_R \ell_L + h.c. \right)$$

Identifying the mass term coefficient as the physical lepton mass as:

$$m_\ell = \frac{g_\ell}{\sqrt{2}} \nu$$

implies that the coupling between fermions and the Higgs field will be proportional to the mass of the fermion itself. For a massive lepton, therefore:

$$L_\ell = -m_\ell \bar{\ell} \ell - \frac{m_\ell}{\nu} h \bar{\ell} \ell$$

This mechanism can be seemlessly applied to generate masses for down-type quarks, with up-type quarks requiring a minimal extension involving an hermitian
conjugate Higgs field $\Phi_c$ where the non-zero vacuum level occurs in the upper component of the $SU(2)_L$ doublet. Both mechanisms result in identical forms for both the mass and Higgs coupling coefficients.

The Higgs mechanism and its minimal extension via the Yukawa coupling imply that a Higgs boson could in principle decay to any pair of charged particles, with the branching fractions determined by the couplings as well as the available phase space.

Figure 2.1 shows theoretical branching fractions for a Standard Model Higgs boson as a function of the Higgs mass, with the right-hand side graph allowing a closer inspection of the mass range where a Higgs-like particle was observed at the LHC [18, 19]. At the observed Higgs boson mass value of 125 GeV, the decay into a pair of bottom quarks, as the heaviest fermion where $2m_f < M_H$, is expected to be the dominant decay mode with a branching fraction estimated to be near 57%. Sub-leading modes include the decay into a pair of off-shell $W^\pm$ or $Z$ bosons, suppressed at matrix element level, and the decay to a pair of photons via a virtual top quark loop. The latter two channels, despite covering less than 3% of the overall Higgs width, result in an exceptionally clean experimental signature in a high-luminosity hadron collider such as the LHC, and were the first channels where a statistically significant Higgs signal was observed.

### 2.1.4 Open Questions

While the observation of a Higgs boson at the LHC was widely publicised as the last missing piece for the completion of a self-consistent theoretical framework for particle physics, several fundamental questions on the nature of the physical world still remain without a clear theoretical answer.

Gravity, for instance, the very force responsible for of the largest structures in the Universe, still refuses to be embedded in a coherent and satisfactory quantum mechanical formalism. Ironically, a large fraction of the inferred gravitational mass of
2.1. The Standard Model of Particle Physics

Figure 2.1: Theoretical branching fractions of a Standard Model Higgs boson for a wide range of masses 2.1a and for a 5 GeV-wide window about the experimentally observed mass [1].

The observable Universe is unaccounted for in searches, with the total light-emitting mass of the Cosmos being found to amount to as little as 15% of the total mass. This disconcerting discrepancy has prompted a wide-ranging theoretical and experimental search for new types matter weakly (if at all) interacting with the Standard Model fields, collectively referred to as "Dark Matter" [34].

Another longstanding issue is the lack of a theoretical mechanism explaining the large matter-antimatter asymmetry in today’s Universe. While electroweak interactions are known to violate the CP symmetry between specific particles such as neutral \( K \) or \( B \) mesons [17,35], the known CP-violating processes are largely insufficient to explain the almost total lack of residual antimatter.

Another class of open questions on the present day understanding of fundamental physics revolves around the concept of naturalness, that is the intuitive expectation that the dimensionless constants regulating the relative strength of the fundamental forces are of the same order of magnitude, and close to unity.

The naturalness principle stands in contrast to the apparent fine tuning of the values such fundamental constants must have to bridge the gap between theoretical predictions and the corresponding experimental evidence. The most visible instance of fine tuning within the Standard Model is known as the Electroweak or Higgs
2.1. The Standard Model of Particle Physics

*Hierarchy Problem* which has been exacerbated by the experimental observation of a Higgs boson with the relatively low mass of 125 GeV.

### 2.1.5 The Higgs Hierarchy Problem

The mass of the Higgs boson arises from a quadratic term in the Standard Model Lagrangian of the form:

$$
L_{H,\text{mass}} = -m_H HH
$$

(2.32)

with $H$ marking the physical Higgs field as defined in Sec. 2.1.2 and $m_H$ the *bare* Higgs mass. The physical and observable value of the Higgs mass is influenced by radiative corrections from fermion and gauge boson-mediated quantum loop diagrams [2], such as those displayed in Fig. 2.2.

![Figure 2.2](image.png)

**Figure 2.2:** First order diagrams representing quantum loop corrections to the Higgs mass arising from the Higgs self-coupling (left) and coupling to gauge bosons (center) and fermions (right). [2]

The contribution to the Higgs mass from such loop diagrams are quadratically divergent and would therefore push the physical Higgs mass to the energy scale where the current description of physics breaks down.

At the present stage of our understanding, the Standard Model has been probed up to the TeV scale without concrete evidence of breakdown. Gravity, conversely, is known to sit outside the Standard Model and its energy scale, the Planck mass $m_P \sim 10^{18}$ GeV, stands as the highest possible energy threshold for the validity of the Standard Model.
To bridge the sixteen orders of magnitude between the expected and observed Higgs mass without assuming an earlier breakdown of the Standard Model, it is necessary to postulate an unlikely reciprocal cancellation of the radiative contributions to the Higgs mass from all the contributing particles, each of which would push the mass towards the Planck scale. This apparent example of fine tuning has been likened to the Higgs having "a snowball’s chance in hell" to have a physical mass as low as the observed value [36]. If, however, new physics exists at a much lower energy than the Planck scale in the form, for instance, of a new generation of particles coupling to the Higgs boson, the quadratic dependence on the energy scale of the radiative Higgs correction can be more easily reworked into matching the observed mass at 125 GeV [2].

In the past decades, a large number of theoretical models for new physics at or near the TeV scale has been put forward with a particular eye to providing a solution to the Higgs hierarchy problem. Among those, supersymmetrical extensions to the Standard Model [37] have spurred a large-scale campaign for experimental observation at hadron collider such as the LHC, so far without success.

Non-supersymmetrical solutions to the hierarchy problem typically revolve around the assumption of the Higgs behaving as a resonance of strongly interacting components around the TeV scale [38], cutting short the radiative contributions and avoiding altogether the issue of fine tuning. Other alternatives involve warped extra dimensions in a theory of partial compositeness of flavour [39] or regard the Higgs as a pseudo-Goldstone boson occurring as a byproduct of a second iteration of global symmetry breaking around the TeV scale [40].

A major point of convergence of these seemingly distant theories is the common prediction of the emergence of new fermion resonances around the TeV scale that, whilst not directly solving the hierarchy problem, are most often required for the consistency of the suggested New Physics mechanisms behind its solution [41–43]. Such particles are generally understood to mix with the Standard Model quarks and are theorized to have non zero coupling between the Standard Model weak sector
and both their left-handed and right-handed components, forgoing the typical $V-A$ couplings of the SM in favour of a purely vectorial interaction, much in the same fashion as the QED interactions of quarks and leptons with the electromagnetic field.

Owing to this major difference with respect to Standard Model quarks, this widely theorized generation of heavy fermion resonances is generally referred to as Vector-Like Quarks.

### 2.2 Vector-Like Quarks

Section 2.1.5 introduced the concept that several theoretically popular models for new physics foresee the emergence of new fermion resonances at or near the TeV scale that mix with the Standard Model quark sector.

Extending the Standard Model beyond its current minimal representation by postulating one or more generations of quark and lepton partners has long been a staple of new physics models [44]. The viability of fourth generation chiral quarks has however been put into serious question [45] as the ATLAS and CMS collaborations failed to detect an enhancement in the $H \rightarrow \gamma\gamma$ decay mode, which is sensitive to the scale of new physics in the form of chiral fermions that obtain mass through a Yukawa coupling with the Higgs boson$^4$. The very occurrence of the Higgs boson at 125 GeV constrains the mass of fourth generation chiral quarks well below the current experimental exclusion limit for such particles [46].

Unlike new chiral quarks, Vector-like quarks are much less constrained by the electroweak and Higgs experimental measurements. This is due to the fact that such particles by definition have both their left and right-handed components transforming in the same way under the Standard Model gauge group, allowing bare mass

---

$^4$The Higgs decay into a pair of photon is mediated by a virtual fermion loop, with relative contributions proportional to the mass of each considered fermion.
2.2. Vector-Like Quarks

terms of the form:
\[ \mathcal{L}_{\text{bare}} = - M Q \bar{Q}_L Q_R + h.c. \] (2.33)
to exist without breaking gauge invariance (see Sec. 2.1.3). As a consequence, Vector-Like quarks do not require a direct Yukawa coupling to the Higgs to acquire mass, and are much less tied to the properties of the Higgs such as the physical mass, as they do not enter the fermion loops causing the Higgs Hierarchy Problem, or its fermion loop-mediated branching ratio into a pair of photons. Nevertheless, most theoretical frameworks for Vector-Like quarks still envision a Yukawa coupling with the Higgs boson as a part of the mixing mechanism with the Standard Model quarks, leading to enhancements in the overall Higgs production via Vector-Like quarks decay [47].

Beyond the common assumption of a vectorial coupling between the new fermion resonances and the SM gauge bosons, the various theoretical models mentioned in Sec. 2.1.5 disagree on the exact number of Vector-Like quarks and their \( SU(2)_L \) representation [48]. Figure 2.3 summarizes the most popular proposal for the weak isospin multiplet state of such quarks [49].

Vector-Like quarks can therefore only occur as weak \textit{up}-type and \textit{down}-type isosinglets, weak isospin doublets and triplets. Given this constraint, at most four new such particles can exist, with electrical charges \( Q = 5/3, 2/3, -1/3 \) and \(-4/3\). The Vector-Like quarks with exotic charges are respectively named \( X \) and \( Y \), while the particles with the same charge as up-type and down-type SM quarks are generally referred to as \textit{top} \((T)\) and \textit{bottom} \((B)\) partners, owing to the common prediction

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{multiplet_states}
\caption{Proposed multiplet states for Vector-Like quarks, with the relative electroweak quantum numbers displayed. Standard Model quarks are also included for reference. [3]}
\end{figure}
that Vector-Like quarks prevalently mix with the third generation of the Standard Model.

2.2.1 Model-Independent Phenomenology

Given the wealth and range of theoretical frameworks predicting the emergence of Vector-Like quarks, converting experimental observations into constraints on the free parameters that regulate each model can be extremely challenging. For this reason, searches for exotic resonances typically adopt an agnostic interpretation of the experimental results based on a parametrization where the mass of the particles and strength of the interactions are left as free parameters of a simplified Lagrangian. The philosophy of this approach to the experimental interpretation of data is known as the "Bridge Method" [4, 50], and is visualized in Figure 2.4.

![Figure 2.4: Visual representation of the Bridge Model approach to the interpretation of exotic searches. [4].](image)

Starting from the central pillar of the bridge, the simplified Lagrangian is expressed as a function of the set of free parameters \( \vec{c} \), on which experimental constraints are later set through a likelihood-based analysis of the data ( \( L(\vec{c}) \) ). Comparing term-wise the simplified Lagrangian with the theory-specific Lagrangians, it is then possible to define an analytic transfer function \( \vec{c}(\vec{p}) \) between the "agnostic" parameters \( \vec{c} \) and the model parameters \( \vec{p} \), on which the experimental constraints are finally transferred.
2.2. Vector-Like Quarks

For Vector-Like quarks, the simplified interaction Lagrangian is given by: [4,51]

\[ L_S = \sum_{Q=X,T,B,Y} \sum_{i=1,2,3} \sum_{\lambda=L,R} c_{i,\lambda}^{QW} \frac{g_W}{\sqrt{2}} [\bar{Q}_\lambda W_\mu \gamma^\mu q_i^\dagger, \lambda] + \]

\[ + \sum_{Q=T,B} \sum_{i=1,2,3} \sum_{\lambda=L,R} c_{i,\lambda}^{QZ} \frac{g_W}{2 \cos \theta_W} [\bar{Q}_\lambda Z_\mu \gamma^\mu q_i^\dagger, \lambda] + \]

\[ + c_{i,\lambda}^{QH} \frac{M_Q}{\sqrt{2}} [\bar{Q}_\lambda H_\mu \gamma^\mu q_i^\dagger, \lambda] + h.c. \]  \hspace{1cm} (2.34)

with the comprehensive coupling constant \( C \) regulating, for the chirality \( \lambda \), the strength of the interaction between the generic Vector-Like quark \( Q \) and the Standard Model quark \( q \) of the \( i \)–th generation, as mediated by the \( W, Z \) or \( H \) bosons. In principle, the values of \( C_L \) and \( C_R \) differ, but discovery searches, such as the one documented in this thesis, are generally not sensitive to the chirality of the targeted resonance. For this reason, the index \( \lambda \) will be omitted from further discussion, re-defining:

\[ c = \sqrt{c_L^2 + c_R^2} \]  \hspace{1cm} (2.35)

The phenomenology of Vector-Like quarks at the LHC can be inferred from the simplified Lagrangian in Eqn. 2.34 through a set of theoretically well motivated assumptions:

- VLQs mix predominantly with 3rd generation quarks [5], therefore: \( c_1^{QV} \sim 0 \), \( c_2^{QV} \sim 0 \). The \( i \) index will therefore be omitted from further discussion;

- The VLQ mass is much greater than the mass of the SM top quark;

- The mass splitting between members of the same multiplet is very small compared to the electroweak scale, forbidding cascade decays such as \( T \rightarrow W^+B \) or \( B \rightarrow W^+Y \). [52];
2.2. Vector-Like Quarks

2.2.2 VLQ production at the LHC

Vector-like quarks are expected to be produced either in quark-antiquark pairs or standalone in association with a Standard Model quark.

Pair production is dominated by QCD processes such as: \( q\bar{q}, \, gg \rightarrow Q\bar{Q} \) which do not involve the mixing vertices with the Standard Model quarks and therefore solely depend on the mass of the produced \( Q \). Consequently, comprehensive pair production measurements are not entirely suitable to set constraints on the phenomenology parameters \( c^{QV} \). Nevertheless, pair production is the dominant production channel at the low end of the \( M_Q \) range accessible at the LHC [5, 53], and has played a major role in discovery searches carried out at the ATLAS and CMS experiments (see Sec. 2.3).

Several processes contribute to the production of a single Vector-Like quark in the final state, with the exchange of a \( t \)-channel vector boson playing a dominant role at tree level [5]. Figure 2.5 shows the tree-level Feynman diagrams regulating the production of a single Vector-Like B quark (VLB) in the final state.

Production via exchange of a Higgs boson in the \( t \)-channel is highly suppressed due to the small Yukawa coupling with the initial state light quarks, and is omitted from consideration.

\[ \text{Figure 2.5: Tree-level Feynman diagrams for single production of a Vector-Like B quark.} \]
2.2. Vector-Like Quarks

Unlike pair-production, the generation of a single VLQ hinges on the interaction terms of the Lagrangian written out in Eqn. 2.34, and is therefore sensitive to the coupling regulating the mixing with the Standard Model, making single-production searches a prime gateway for setting experimental constraints on the simplified parameters $c^{QV}$ and, by extension, the various physics-motivated models for Vector-Like quarks. Furthermore, leading order calculations [5] show that single production cross section, though greatly varying according to the weak isospin representation of choice, overtake pair production around the TeV scale (Fig. 2.6), becoming the altogether best process for investigation for Vector-Like quarks.

![Figure 2.6: Leading Order PROTOS cross sections for VLQ single production for several multiplet configurations. The $Q\bar{q}j$ nomenclature refers to the final state products of each process, as visible in Fig. 2.5a-2.5b. Pair production cross section (dashed line) included for reference. [5]](image)

The dependence of single production cross section on the $c^{QV}$ parameter can be factored out assuming the Narrow Width Approximation (NWA)$^5$ [4, 51]. Therefore, for a Vector-Like $B$ quark produced in association with a Standard Model quark $q$ via the $qVB$ vertex:

$$
\sigma_{NWA}^{Bq}(M_B, c^{B}_{V}) \approx \left( \left[ c^{B}_{V} \right]^2 + k \cdot c^{B}_{V L} c^{B}_{V R} \frac{M_q}{M_B + M_q} \right) \times \sigma_{NWA}^{Bq}(M_B) \tag{2.36}
$$

$^5$The NWA assumes the total decay width $\Gamma$ to be negligible with respect to the resonance mass.
with the second term suppressed by both kinematic $M_q \ll M_B$ and theoretical motivations holding that in each theoretical scenario only one of the Left or Right-handed couplings has non negligible values [5]. Eqn. 2.36 establishes a quadratic dependence of the total production cross section from the phenomenology couplings $c^B_V$, allowing an immediate interpretation of the results in terms of constraints on the $c_V$s.

The search covered in this thesis, however, is not set to be sensitive to the inclusive cross section of a single-produced Vector-Like B quark, targeting instead only the fraction of events pertaining to a specific B decay channel. To provide a meaningful interpretation of the experimental result, it is therefore necessary to gain an understanding of the decay mechanics of Vector-Like quarks and the extent to which they are affected by the critical couplings $c^B_V$. 
2.2. Vector-Like Quarks

2.2.3 Decay Mechanics

Given the heavy kinematic suppression of cascading decays between members of the same isospin multiplet, which share the same bare mass term in the Lagrangian [5], the decay width of Vector-Like quarks is entirely determined by their mixing with the Standard Model, and is therefore heavily affected by the value of $c_{V}^{O}$. The allowed decay modes can therefore be summarized to be:

Vector – Like $X_{5/3}$: $X_{5/3} \rightarrow W^{+}t$

Vector – Like $T$: $T \rightarrow W^{+}b$, $T \rightarrow Zt$, $T \rightarrow Ht$

Vector – Like $B$: $B \rightarrow W^{-}t$, $B \rightarrow Zb$, $B \rightarrow Hb$

Vector – Like $Y_{-4/3}$: $Y_{-4/3} \rightarrow W^{-}b$

Casting aside the $X$ and $Y$ VLQs, which have only one accessible decay mode owing to their exotic charges, the partial widths of the top and bottom partners can be expressed as a function of kinematic quantities and the relevant coupling $c_{V}^{O}$ [51]:

$$\Gamma(Q \rightarrow Wq') = \left[c_{W}^{O}\right]^{2} \times \frac{g_{W}}{128\pi} \cdot \frac{M_{Q}^{3}}{m_{W}^{2}} \times \rho(Wq')$$

$$\Gamma(Q \rightarrow Zq) = \left[c_{Z}^{O}\right]^{2} \times \frac{g_{W}}{128\pi} \cdot \frac{M_{Q}^{3}}{m_{Z}^{2}} \times \rho(Zq)$$

$$\Gamma(Q \rightarrow Hq) = \left[c_{H}^{O}\right]^{2} \times \frac{M_{Q}}{32\pi} \times \rho(Hq)$$

(2.37)

with $\rho(Wq')$, $\rho(Zq)$, $\rho(Hq)$ being kinematic functions of the masses of the particles involved, all evaluating to unity in the limit where $M_{Q} \gg m_{q}$, $m_{V}$.

The coupling constants $c_{V}^{O}$ can be re-written in such a way that the kinematic differences between the partial widths are absorbed into the new definition of "reduced"
coupling $\tilde{c}_V^Q$:

$$
\Gamma(Q \to Wq') = \left[\tilde{c}_W^Q\right]^2 \times \frac{g_W}{128\pi} \frac{M_Q^3}{m_W^3} \times \rho(Wq')
$$

$$
\Gamma(Q \to Zq) = \left[\tilde{c}_Z^Q\right]^2 \times \frac{g_W}{128\pi} \frac{M_Q^3}{m_W^3} \times \rho(Zq)
$$

$$
\Gamma(Q \to Hq) = \left[\tilde{c}_H^Q\right]^2 \times \frac{g_W}{128\pi} \frac{M_Q^3}{m_W^3} \times \rho(Hq)
$$

Under this revised parametrization, the branching ratio for the generic $Q \to Aq'$ decay is:

$$
BR[Q \to Aq] = \frac{\tilde{c}_A^Q \rho(Aq)}{\tilde{c}_W^Q \rho(Wq') + \tilde{c}_Z^Q \rho(Zq) + \tilde{c}_H^Q \rho(Hq)}
$$

Two important conclusions can be inferred from Eqn.2.38 and 2.39:

- As with the total cross section, the relative contribution of the three decay channels is regulated entirely by the reduced couplings $\tilde{c}_W^Q$, canceling effects from possible complex phases in the individual couplings;

- As both the production cross section and the partial decay widths scale with the square of the relevant coupling, given a constant resonance mass $M_Q$ a larger production cross-section will occur together with a larger $\Gamma/M_Q$ ratio, affecting the experimental signature.

In accordance with the "Bridge Method" for the interpretation of exotic searches, the result shown in Eqn.2.39 can be made to describe the individual physics-driven models for Vector-Like quarks by requiring the $\tilde{c}_V^Q$ parameters to have theory-specific values or relations between each other.

In the simplest scenario where highly massive $T$ or $B$ quarks lie in a singlet state, the decay widths are evenly distributed among the four degrees of freedom introduced through the electroweak symmetry breaking, which generates the masses of
the $W^+$, $W^-$, $Z$ and $H$ bosons [5,48,53]. This specific configuration implies:

$$c_Q^Z \simeq c_H^q \simeq \frac{1}{\sqrt{2}} c_W^q$$

T,B singlets \hspace{1cm} (2.40)

with the T, B branching fractions into $Wq'$ approaching 0.5 and $BR[Q \to Zq] \sim BR[Q \to Hq] \sim 0.25$ in the high mass limit.

The phenomenology of Vector-Like quark decays in the doublet and triplet configurations is more complex and relies on model-dependent behaviour of the Yukawa couplings between the chiral components of the Vector-Like quarks and the third generation SM quarks [48]. In all doublet configurations, the VLQs mix only with a single third Standard Model quark. Summarizing the most popular scenarios, in the high-mass limit the branching ratios are:

\[
\begin{align*}
(T, B) \text{ doublet} & \quad \begin{cases} 
BR[T \to Zt] = BR[T \to Ht] = 0.5 \\
BR[B \to Wt] = 1
\end{cases} \\
(B, Y) \text{ doublet} & \quad \begin{cases} 
BR[B \to Zb] = BR[B \to Hb] = 0.5 \\
BR[Y \to W^-b] = 1
\end{cases}
\end{align*}
\]

A more comprehensive overview of the theoretical T and B branching ratio across the several theoretical scenarios is provided in Figure 2.7.

### 2.3 VLQ Searches: State of the Art

Extensive attention has been dedicated by both the ATLAS and CMS collaborations to the search for experimental evidence of Vector-Like quark production in $pp$ collision data generated at the Large Hadron Collider.

The most comprehensive set of experimental constraints on Vector-Like quarks has been recently published by ATLAS in the form of a statistical combination [54] of seven complementary searches targeting a plethora of final states originating
2.3. VLQ Searches: State of the Art

Figure 2.7: Theoretical branching ratios of T and B quarks for the main allowed $SU(2)_L$ representation for VLQ. Blue lines show the behaviour of the branching ratios from the low mass (marked by an X) to the high-mass limit (marked by a red dot). The $BR[Q \rightarrow W q']$ can be inferred as $BR_W = 1 - BR_Z - BR_H$. [5]

from $T\bar{T}$ and $B\bar{B}$ production and decay. Exclusion limits at the 95% Confidence Level [55] were extracted for a range of T and B masses as a function of of the decay branching fractions, therefore indirectly setting mass-dependent constraints on the phenomenology parameters $c^2_{QV}$. Figure 2.8 shows the comprehensive limits on T and B quarks drawn within the pair production combination effort. The combined searches proved especially sensitive to the two decay channels involving a final state top quark, owing to the easily identifiable experimental signature of leptonic top decays. Conversely, the T and B decay modes involving a final state bottom quark offered an intrinsically lower background rejection power, and yielded inferior limits, as evidenced by the $BR[B \rightarrow H b]$ corner.

A large amount of pair-production searches were also run within the CMS collaboration, with exclusion limits on T and B quarks ranging from 900 GeV to 1.3 TeV across the different theoretical benchmarks and targeted decay channels [58–60].

Owing to its comparative larger cross section above the TeV threshold with respect to pair production, single production has been targeted at both ATLAS and CMS

---

6As modelled by PROTOS v2.2 [56] at Leading Order, with the QCD-driven cross-sections then normalized to next-to-next-to leading order with Top++ v2.0. [57].
Figure 2.8: Comprehensive exclusion limits drawn in the combination effort of all ATLAS pair-production searches. For each value of the two independent T and B branching ratio, the combination excludes at 95% CL the occurrence of the relevant Vector-Like quark with mass represented by the color coding. The overlaid yellow symbols show popular coupling configurations as outlined in Sec.2.2.3. [6]. The solid, dotted and dashed white lines follow the paths of constant excluded resonance mass in the branching ratio phase space.

with ever growing interest. Within ATLAS, published searches on early Run 2 (2015+2016) data have explored the following channels:

- $T \rightarrow Zt$ decay channel, excluding $\sigma(p p \rightarrow T) \times BR[T \rightarrow Zt]$ above $\sim 80 - 100 \text{ fb}^{-1}$ across the 1-2 TeV range in the singlet scenario where $C_{T Z} = 0.45$ [61];

- $T/Y \rightarrow W^{\pm} b$ decay channel, setting coupling-dependent limits for a range of T and Y masses between 800 and 1800 GeV in both the T singlet and (B,Y) doublet scenarios [62];

- $T \rightarrow tZ$ and $B \rightarrow bZ$, targeting the opposite sign $ee$ and $\mu\mu$ decays of the Z, with competitive coupling-dependent limit on single T production [63]

- $B \rightarrow Hb$ in the $H \rightarrow \gamma\gamma$ channel, which excluded $\sigma(p p \rightarrow B) \times BR[B \rightarrow Hb]$ above $80 - 100 \text{ pb}$ in the (B,Y) multiplet scenario for B masses between 800 GeV and 1.6 TeV [64].

The search covered in this thesis is set to complement the past and ongoing effort in exploring the complete $C_{QV}$ parameter space by targeting the yet little explored
$B \to Hb$ decay in the final state with the greatest acceptance, $H \to b\bar{b}$, compensating the cleaner experimental signature of the $H \to \gamma\gamma$ mode with a two orders of magnitude increase in potential signal events.

A search on the $B \to bH(b\bar{b})$ final has been carried out within the CMS collaboration on 2015 and 2016 data [16], with mass-dependent exclusion limits ranging from 200 to 80 $fb$ for a singlet B of mass between 1200 and 1800 GeV, and will stand as the closest comparison benchmark for the compatibility and competitiveness of the results of the present search.
Chapter 3

The ATLAS Experiment at the LHC

3.1 The Large Hadron Collider

The Large Hadron Collider [20], or LHC, is a circular collider designed with the purpose of generating proton-proton, proton-ion and ion-ion collisions at an unprecedented center-of-mass energy for human-made particle interactions.

The latest in a lineage of CERN colliders purposed to explore the energy frontier which dates back to the Proton Synchrotron (PS) [65], the LHC is housed in the 27 km long underground tunnel previously built to host the Large Electron-Positron Collider (LEP). Inside the tunnel, an array of superconducting dipole, quadrupole and sextupole magnets respectively operate the tasks of bending the flight path of the particles into the approximately circular shape of the collider and to regulate the properties of the beam to maximise the number of collisions occurring at the beam interaction points.

The LHC is designed to reach the nominal collision energy of 14 TeV, symmetrically split between the two colliding beams. While the LHC posesses standalone means of acceleration in the form of 16 radio-frequency cavities (RF) [66] capable of increasing the particle energy by 485 keV per turn, these are designed to operate on a minimum extant energy of 450 GeV, requiring a series of preliminary acceleration steps before the final injection into the LHC, a task accomplished converting the previous smaller circular colliders as intermediate boosting stages.
3.1. The Large Hadron Collider

Figure 3.1: Schematic view of the CERN accelerator complex, highlighting the modular acceleration path from the original sources to the LHC. [7]

Figure 3.1 offers a schematic view of the complex of accelerators and injection lines involved in the particle boosting process from a rest state all the way up to the LHC.

3.1.1 Luminosity and Run conditions

The LHC is designed to re-create, within a vanishingly small collisional volume, the most extreme energy conditions ever attained in earth-based experiments. This daunting engineering challenge lies at the base of the fulfilment of the stated LHC physical goal to refine the present understanding of the Standard Model processes as well as to unveil a possible emergence of physics yet unexplored at the energy frontier.

Pursuing the two aforementioned goals in a satisfactory manner, however, generates the necessity to also push the event generation rate beyond the accomplishments of all previous hadron colliders, in order to both produce the most precise measure-
ments of the Standard Model to date and the first evidence of an extremely rare exotic process.

The event generation rate in a particle collider is described through a process-independent quantity known as Luminosity, defined as the delivered event rate for a unitary cross section. Given the definition of luminosity, for any time interval $\Delta T$, the number of events for a specific physical process of cross section $\sigma$ is then:

$$N = \sigma \times \int_{\Delta T} \mathcal{L}(t) dt = \sigma \times \mathcal{L}_{\text{INT}}$$  \hspace{1cm} (3.1)

The typical unit measure for the instantaneous luminosity $\mathcal{L}(t)$ is cm$^{-2}$s$^{-1}$, while the total integrated luminosity for a given data taking period is often given in multiples of the inverse femtobarn $fb^{-1}$ to enable a quick estimation of the number of events generated for a specific process whose cross section is expressed in femtobarn $fb$.

The instantaneous luminosity of a particle collider is entirely determined by its construction and working parameters. For a collider, such as the LHC, where interactions occur in the crossing of $n_b$ finite-size bunches of $N$ particles each, rather than through intersecting continuous beams, the maximum luminosity is given by:

$$\mathcal{L}_0 = n_b \frac{N^2 f}{4\pi \sigma_x \sigma_y}$$  \hspace{1cm} (3.2)

where $f$ denotes the bunch revolution frequency and $\sigma_{x,y}$ the bunch projections on the plane transverse to the beam direction [20]. As the LHC does not support bunch refurbishment during collision runs, the instantaneous luminosity $\mathcal{L}(t)$ degrades exponentially over time as the beam population $N$ is depleted, leading to the end of the run once the Luminosity drops below an established threshold.

While necessary for the achievement of the stated LHC physics goals, extremely high luminosity also generates serious experimental challenges in the planning and commissioning of the detector, as well as in the data taking and analysis phases. In 2018, for instance, the final year of the Run 2 data taking campaign, the LHC
3.1. The Large Hadron Collider

delivered a nominal maximum luminosity of \( \mathcal{L}_0 = 2 \cdot 10^{34} \text{ cm}^{-2} \text{s}^{-1} \), resulting in an average\(^1\) of \( \langle \mu \rangle = 40 \) proton-proton interactions per bunch crossing \([13, 67]\) for a total of over 1000 final state particles to be accounted for within the bunch spacing time of 25 ns. The superimposition of detector information generated by the unrelated simultaneous collisions products is referred to as in-time pile-up. The effects of pile-up are countered both at hardware and software level, with high resolution sub-detector components and offline cleaning algorithms mitigating the impact on the physics programme.

Despite the intrinsic issues associated with a high-luminosity hadronic environment, pushing the luminosity frontier forward is the key factor in maximising the reach of the experimental effort at the LHC, and is the focus of the upgrade project for the second phase of the collider life cycle \([68]\), with a projected peak luminosity of \( \mathcal{L}_{HL} = 5 \cdot 10^{34} \text{ cm}^{-2} \text{s}^{-1} \), equivalent to a yearly event volume of 250 fb\(^{-1}\).

Since the beginning of the main LHC research programme in March 2010, there have been two main data taking periods separated by a "Long Shutdown" dedicated to maintenance and upgrade work. The first data taking run (Run 1) spanned from March 2010 to January 2013, for a total of \( 5.46 \text{ fb}^{-1} \) integrated luminosity at a center of mass energy \( \sqrt{s} = 7 \text{ TeV} \) (2010 and 2011) and \( L = 22.8 \text{ fb}^{-1} \) at \( \sqrt{s} = 8 \text{ TeV} \) (2012) \([69]\). Data taking resumed in 2015, with collision energy ramped up to \( \sqrt{s} = 13 \text{ TeV} \) and peak luminosity reaching \( \mathcal{L}_0 = 2 \cdot 10^{34} \text{ cm}^{-2} \text{s}^{-1} \). From May 2015 until the end of the LHC data-taking Run 2 in late 2018, a total of \( 156 \text{ fb}^{-1} \) of collision data was delivered by the LHC, so far the highest event volume ever generated at a hadron collider.

\(^{1}\)the average interactions per bunch crossing is estimated as \( \langle \mu \rangle = \mathcal{L}(t) \times \sigma_{\text{inel}} \times 25 \text{ ns} \), with \( \sigma_{\text{inel}} \) being the inelastic pp cross section, by far the dominant process at the LHC.
3.2 The ATLAS Detector

ATLAS (A Toroidal LHC ApparatuS) [21] is, together with CMS, one of the two general-purpose experiments collecting and analyzing the collisional data generated at the LHC beam interaction points. The ATLAS detector, housed in the underground cavern at the LHC Interaction Point 1, is built in the shape of a 44 metre long barrel, with a diameter of 25 metres.

An in-detector right-handed cartesian coordinate system is generated out of the nominal beam interaction point in order to uniquely and easily identify the position of each detector sub-component. The positive $z$ axis is defined to run along the beam line, as one looks at the LHC from the surface, with the $x$ and $y$ axis respectively pointing towards the LHC ring centre and the surface. The plane generated by the $x$ and $y$ axes is typically referred to as the transverse plane, owing to its orthogonality to the beam direction.

Owing to the cylindrical symmetry of the ATLAS detector, it is often more practical to forego the $x$ and $y$ axes in favour of the azimuthal angle $\phi$, defined right-handedly out of the positive $x$ axis. The polar angle $\theta$ is instead defined as the angular distance from the transverse plane, with $\theta = 0$, $\pi$ respectively identifying directions parallel and anti-parallel to the $z$ axis.

Unlike electron-positron colliders, such as the LEP, the center of mass frame in hadron collider interactions is generally not stationary nor pre-emptively determined, rendering both the $(x, y, z)$ and $(\theta, \phi, z)$ less than optimal to provide a homogeneous description of the events. For this reason, the rapidity is defined as:

\[ y = \frac{1}{2} \ln \left( \frac{E + p_z}{E - p_z} \right) \]  

(3.3)

with $E$ and $p_z$ representing the energy and the momentum projection along the $z$ axis of the particle under consideration. The rapidity has the key advantage over $\theta$ of allowing a Lorentz-invariant event description along the beam axis, as rapidity
differences $\Delta y$ are not affected by boosts along the $z$–axis. In the massless limit where $|\vec{p}| \sim E$, the rapidity reduces to a very convenient analytical function of the polar angle $\theta$ known as pseudorapidity:

$$\eta = -\ln \left[ \tan \left( \frac{\theta}{2} \right) \right]$$  \hspace{1cm} (3.4)

As the large LHC collision energy typically ensures highly boosted final state particles, $\eta$ is nearly universally employed in place of the rapidity to identify the flight direction of a particle relative to the beam axis, retaining Lorentz-invariance at an excellent degree of approximation. Given the boost-invariant $(\eta, \phi)$ picture, the angular distance between any flight directions is given by the immediately Lorentz-invariant $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$. Adding the radius parameter $r$, defined as the distance from the beam axis, every detector component or energy deposit caused by a passing particle can be unambiguously defined through a triplet in the $(\eta, \phi, r)$ space. The ATLAS detector, represented in its entirety in Figure 3.2, is divided in three symmetric regions reflecting the different detecting strategies and requirements for increasing values of $\eta$: the barrel, covering the most central region with small values of $\eta$, the end-cap and finally the *forward* detector region covering the
3.2. The ATLAS Detector

extreme high-\(\eta\) values. All these subsections of the detector share a roughly cylindrical symmetry, with differently purposed instrumentation positioned at different distances from the beam axis. The following sections will provide an overview of the main detector sub-systems, their position, their experimental purpose as well as their detection philosophy and delivered performance.

![Exploded representation of the barrel section of the Inner Detector and its modules, along with the respective distances from the beam axis.](image)

**Figure 3.3:** Exploded representation of the *barrel* section of the Inner Detector and its modules, along with the respective distances from the beam axis. [9, 10]

### 3.2.1 Inner Detector

The Inner Detector (ID) (Fig. 3.3) is the innermost ATLAS sub-system, immediately enveloping the beam pipe. Extending from just 32.25 mm to 1080 mm from the beam axis and covering the region where \(|\eta| < 2.5\), the ID is set to provide the best space point measurement resolution of the entire detector, enabling an accurate and precise reconstruction of the trajectories of charged particles in the final state.

The performance of the ID is of paramount importance to all physics analyses rely-
3.2. The ATLAS Detector

ing on charged particles: an optimal tracking quality stands as a direct requirement of key data quality tasks such as the measurement of the particle transverse momentum ($p_T$), its original position in the ($\eta$, $\phi$) phase space and the identification, via track extrapolation to the beam level, of its original interaction point, providing a mitigation of pile-up effects.

As the closest detector component to the beam pipe, the Inner Detector is required to withstand the highest level of radiation of any other active element in the entire detector [70], which requires the building material to be especially radiation-resistant. On the other hand, the Inner Detector components need to be as "transparent" as possible, in order to minimize the degradation of the tracking measurements brought about by multiple scattering. Similarly, low-mass components reduce the probability of electrons radiating photons by Bremstrahlung and photons converting to an $e^+e^-$ pair already in the Inner Detector, something that would impede calorimetric measurements. Summarising, the Inner Detector was conceived in accordance with the following design goals:

- Maximum granularity compatible with the achievable read-out speed;
- High degree of radiation resistance;
- Low thickness to minimise the energy loss caused by multiple Coulomb scattering [71].

The final design of the Inner Detector [72, 73] consists of three concentric tracking systems, the Pixel Detector (PIXEL), the Semiconductor Tracker (SCT) and Transition Radiation Tracker (TRT), here mentioned by increasing distance from the beam axis. The entire system is immersed in a 2T magnetic field generated by the super-conducting Central Solenoid (CS) [74]. The magnetic field is aligned by design with the $z$-axis and deflects out-going particles along the transverse plane, enabling the measurement of the particle transverse momentum with a combined resolution determined by the following identity:

$$\frac{\sigma_{p_T}}{p_T} = 0.05\% \cdot p_T \oplus 1\%$$  \hspace{1cm} (3.5)
3.2. The ATLAS Detector

The following subsections will provide a brief account of the Inner Detector sub-modules specifics, detection methods and delivered performance.

**Pixel Detector**

The innermost element of the Inner Detector, The PIXEL [75] consists of 1744 sensors arranged in four cylindrical layers in the barrel region and three disks in each detector end-cap. All sensors are identical, housing 47232 pixels, for a grand total of 80 million read-out in the system. The individual pixel size is 50\(\mu m\) on the \(r \times \phi\) plane and 250\(\mu m\) (at the innermost layer) to 400\(\mu m\) (remaining layers) on the \(z\)-direction. A charged particle inside the \(\eta\) acceptance of the ID is expected to produce on average 4 hits in as many layers, providing spatial information with the greatest spatial granularity of the entire ATLAS detector (\(\sigma_{R \times \phi} \times \sigma_z = 10\mu m \times 72\mu m\) for the innermost Pixel layer and 14\(\mu m \times 115\mu m\) for the remaining layers and end-cap wheels).

The innermost of the four Pixel barrel layers, the IBL (Insertable B-Layer) [76], was installed during the first long shutdown between LHC Run 1 and Run 2 between the renovated beam pipe and formerly innermost Pixel layer, with the explicit goal of providing additional tracking robustness compensating for the extant and future irreparable failures and inefficiency of the original layers. The increased hit redundancy, as well as the reduced radial distance between the innermost tracking point and the beam line, proved to be a precious feature in the complex environment determined by the high level of *pile-up* during Run 2.

**SCT**

The Silicon Microstrip Tracker (SCT) extends the tracking coverage of the ATLAS detector in the region immediately outside the PIXEL detector, providing a maximum of four additional particle hits thanks to four cylindrical silicon strip lay-
3.2. The ATLAS Detector

ers in the barrel and 9 disks in each end-cap detector region, totalling 6.3 million strips each 6.5 cm long and 80 $\mu m$ in pitch. Each SCT layer houses two superimposed sub-layers of strips offset by a 40 mrad stereo, determining a grid-like macro-arrangement, with one of the two grid axes aligned with the detector $z$ direction in the barrel and radially in the end-cap disks. The position of a particle producing a hit in the SCT is provided with a nominal resolution of $17 \times 580 \mu m^2$.

3.2.3 TRT

The Transition Radiation Tracker is designed to provide suitable information to discriminate electrons from pions, two particles which in principle leave a very similar experimental signature in small and relatively transparent tracking devices. The TRT is constructed out of approximately 30000 gas-filled Polyimide drift straws arranged in 73 planes in the barrel and 160 wheels in the end-cap regions. Each straw has a diameter of 4mm and contains an axial gold-plated tungsten wire acting as the anode. The straws are filled with an admixture of 70% Xenon, 27% CO$_2$ and 3% Oxygen. Charged particles passing through a tube ionize the gas leaving a trail of electrons and ions collected by the electrodes, producing a measurable electric signal. Owing to the geometry of the straw tubes, the TRT provides little information on the $z$-coordinate of the crossing particles in the, while the position on the $r \times \phi$ plane is given with a resolution of $130 \mu m$. Conversely, no information is provided on the $r$ coordinate in by the end-cap straw tubes assembly. The lower spatial resolution of the TRT with respect to the other Inner Detector components is compensated by the large hit redundancy, with an average of 36 expected hits for an outgoing particle crossing the system.

The space between stab tubes is filled with polymer fibres (barrel) and foils (end-caps) to enhance the production of transition radiation as particles cross the sub-detector [77]. The emitted radiation, consisting of photons in the $5 - 30$ keV range which excite the Xe atoms producing larger electrical signals, scales with the relativistic factor $\gamma = E/m$, providing grounds for electron-pion discrimination.
3.2.4 The ATLAS Calorimetric System

At radii immediately larger than that of the solenoid magnet providing the necessary magnetic field for the track measurement in the Inner Detector lies the ATLAS calorimetric system, the sub-detector purposed with the measurement of the energy of photons, electrons and hadrons.

The founding principle of all calorimetric measurement is to stop the relevant outgoing particles by means of their repeated interactions with a thick layer of matter. Under this picture, the entirety of the kinetic energy of such particles is deposited within the calorimeter, becoming available for measurement. Through a calibration of the detector response, the read-out electrical signal is then converted into a measurement of the original energy of the particle(s) entering the calorimeter.

Among the cardinal design features that the calorimeters of a general-purpose experiment such as ATLAS must abide by is the near-total hermeticity required to produce an accurate measurement of the total energy in each event. This is in turn a key component in the calculation of the missing transverse energy $E_T$, which plays a fundamental role in a large number of measurement and searches for electroweak processes involving neutrinos or exotic long-lived or weakly interacting massive particles beyond the Standard Model.

Due to the intrinsically different nature of the interactions that electrons and photons, as opposed to hadrons, undergo while passing through matter, the ATLAS calorimeter is divided into two separated coaxial sub-systems: the Electromagnetic (EM or LAr) Calorimeter, located just outside the Inner Detector, and the Hadronic Calorimeter, at immediately larger radii.

Both ATLAS calorimeters are conceived as sampling calorimeters, a space and cost-effective design that aims at measuring a known fraction of particle energy, rather than the total. This is achieved alternating layers of active materials, in which the deposited energy produces a read-out signal, to denser, passive materials with the sole purpose of contributing to the full stopping power of the apparatus.
3.2. The ATLAS Detector

Electromagnetic Calorimeter

The Electromagnetic Calorimeter [78] consists of a barrel-shaped component covering the central regions of the detector with $|\eta| < 1.475$ and two End Cap wheels covering the more forward $1.475 < |\eta| < 3.2$ region. The design purpose of the EM Calorimeter is to measure the energy of the incoming electrons, photons and neutral pions, as well as providing an element of discrimination between the experimental signature produced by such particles.

A sampling calorimeter by design, the ATLAS electromagnetic calorimeter employs the inert Argon noble gas in a liquid state as active material and lead as the passive absorber. The alternance between active and passive materials within each calorimeter module is achieved by means of an accordion-shaped lead structure with liquid Argon-filled gaps, as can be appreciated in Figure 3.4.

At the base of the detection principle of the EM calorimeter lay the bremsstrahlung ($e \to e\gamma$) and pair-production ($\gamma \to e^+e^-$) interactions, the principal avenues by which highly energetic electrons and photons respectively lose energy traversing

![Figure 3.4: Schematic representation of a module of the ATLAS LAr calorimeter, highlighting the accordion structure of the sampling unity, also pictured in the right panel.](image)
3.2. The ATLAS Detector

matter. Regardless of whether the original particle was an electron, a primary photon or a photon pair from a $\pi^0$ decay, the two energy-loss processes connect into a runaway chain interaction generating a so-called electromagnetic shower extending radially into the calorimeter until the secondary products are all but stopped.

The radial extension of the shower is measured in units of the radiation length $X_0$, measured in $\text{g} \cdot \text{cm}^{-2}$ and defined as the average distance for unit absorber density in which the energy of a showering electron is reduced by a factor of $e$. The total thickness of the EM calorimeter in the barrel section ranges from 24 to 32 radiation lengths, depending on the $|\eta|$ of the incoming particle, while the End Cap sub-detector is designed with a thickness of 24 to 38 $X_0$.

Each layer of the EM calorimeter is finely segmented in the $\eta \times \phi$ plane in order to provide critical information about the flight path of the incoming particles, as well as to differentiate geometrically close but unrelated energy deposits. The innermost layer of the calorimeter features the finest granularity of the entire system at $\Delta \eta \times \Delta \phi = 0.003125 \times 0.1$ in order to probe the initial phase of the shower for the purpose of particle identification. A thin layer of liquid Argon positioned at radii immediately smaller than the first calorimeter layer is also tasked with an estimate of the energy loss induced by the detector materials encountered prior to the calorimeter. This monitoring device is known as the pre-sampler.

The second layer of the EM calorimeter is designed with transverse granularity $\Delta \eta \times \Delta \phi = 0.025 \times 0.025$, with further layers featuring a progressively coarser granularity.

The overall resolution on energy measurements provided by the LAr calorimeter is given by:

$$\frac{\sigma_E}{E} \simeq \frac{10\%}{\sqrt{E \text{[GeV]}}} \oplus 0.7\%$$  \hspace{1cm} (3.6)$$

with the contribution proportional to the inverse square root of the energy reading highlighting the sampling nature of the calorimeter by which only a stochastic fraction of the total energy is deposited in the active material.
3.2. The ATLAS Detector

Hadronic Calorimeter

Owing to their far larger mass with respect to electrons, hadronic particles, with the notable exception of the neutral pion, do not initiate an electromagnetic shower as they pass through matter.

The principal gateway for the energy loss of particles such as charged pions, $K$ mesons and nucleons is therefore the nuclear interactions with the calorimeter material which give raise to multiple secondary hadrons, which in turn undergo secondary interactions with the detector giving raise to the so-called hadronic shower. The shower extension is expressed in units of the interaction length $\lambda$, the free mean path of an incoming hadron given the nuclear cross-section and the detector material density. The showering process for hadrons starts within the electromagnetic calorimeter and extends all the way through the hadronic calorimeter, requiring optimal inter-calorimetric alignment and combined energy calibration to produce a comprehensive measurement of the incoming particle energy.

Similarly to the electromagnetic counterpart, the hadronic calorimeter is composed of a composite central section in the shape of barrel (TileCal [79], covering $|\eta| < 1.7$) and two wheel-shaped sections (Hadronic End-Cap Calorimeter [80]) covering the immediate forward detector region where $1.5 < |\eta| < 3.2$.

Both hadronic calorimeters are designed as sampling calorimeters, with TileCal alternating tiles of plastic scintillator as active material and steel as passive absorber. The Hadronic End Cap Calorimeter, conversely, uses liquid Argon and copper as active and passive materials respectively.

The full hadronic calorimetric system is designed to provide a relative resolution on the energy of a single, incoming particle as given by:

$$\frac{\sigma_E}{E} = \frac{50\%}{\sqrt{E \ [\text{GeV}]}} \oplus 3\%$$
3.2. The ATLAS Detector

Forward Calorimeter

The ATLAS Forward Calorimeter, known as FCal [81], is designed to complete the hermeticity of the overall calorimetric system, covering the $3.1 < |\eta| < 4.9$ pseudo-rapidity range. By far the closest calorimetry component to the beam line, and consequently receiving the highest radiation dose, the forward calorimeter is designed to meet stringent standards of radiation component heat dissipation and radiation hardiness.

Similarly to the remaining calorimetric systems, FCal is a sampling detector by design, employing stacked copper and tungsten plates as absorber and liquid Argon as active material.

The forward calorimeter system provides a single-particle energy resolution given by:

$$\frac{\sigma_E}{E} = 100\% \sqrt{E \text{ [GeV]}} \oplus 10\%$$

3.2.5 Muon Spectrometry

The ATLAS Muon Spectrometer [11] is tasked with the tracking measurements of muons as they proceed past the electromagnetic and hadronic calorimeters in their outbound flight. This is accomplished by means of a series of diverse tracking devices immersed in a magnetic field located at radii immediately above the outermost calorimeter layer.

The magnetic field employed for the task of muon spectrometry is generated by 8 superconducting toroidal magnets in the central barrel detector region ($|\eta| < 1.0$) and by as many smaller toroids arranged into two end cap wheels for $1.4 < |\eta| < 2.7$. The small rapidity gap between the two magnet systems is nevertheless serviced adequately by the interface of the two generated magnetic fields.

The main tracking operations are carried out by an array of detectors employing
3.2. The ATLAS Detector

Figure 3.5: Section view of the ATLAS Muon Spectrometer with the individual detector technologies labeled. [11].

Diverse technological means and detecting strategy across the pseudorapidity range covered by the system. For $|\eta| < 2.7$, tracking hits are provided by three layers of Monitored Drift Tubes (MDT), each of which consists of two layers of drift tubes, each in turn three tubes thick. Drift tubes are 1 to 6 meters long and 30mm in diameter, and detect the passing of a muon collecting the signal produced through ionization of an Ar/CO$_2$ gas admixture in an axial tungsten-rhenium anodic wire. The MDTs can only provide the $Z$ (longitudinal) and $R$ (radial) coordinates of a hit, thus requiring an additional subset of detectors to measure the azimuthal angle $\phi$ of the hit. In the end-cap region, the MDT chambers are fashioned into triangular shape and arranged into a wheel.

In the end caps ($2.0 < |\eta| < 2.7$), muon spectrometry is carried out by means of four stacked layers of multiwire proportional chambers with segmented cathodes (known as Cathode Strip Chambers, CSC) arranged into two disks to provide a complete set of hit coordinates in the high pseudo-rapidity range. As the particle density is expected to be critically high in the end cap and forward detector regions, the CSCs were designed to meet tighter standards of rate capability and time resolution than those demanded of the MDTs.
Within the $|\eta| < 2.4$ region, tracking information is completed and strengthened by sets of Resistive Plate Chambers (RPC) in the barrel and Thin Gap Chambers (TGC) in the forward regions. The coarse resolution offered by such detectors is compensated by an extremely fast readout response, making them an apt triggering system.

### 3.2.6 The ATLAS TDAQ System

The LHC-delivered bunch spacing of 25ns results in bunches crossing and interacting with a rate of 40 MHz, orders of magnitude above the writing speed of even the fastest data-writing systems available today. Given the practical impossibility of reading out and storing the entirety of the ATLAS detector live output, a dedicated triggering system is required to choose whether to record and save the full event information on the grounds of the presence of key event features signalling the likelihood of the examined event being interesting for physics or performance analysis.

The ATLAS Triggering and Data Acquisition system (TDAQ) [12], represented schematically in Figure 3.6, is a two-level system employing dedicated custom hardware and processing units to operate the first instance of event selection on events observed at ATLAS, resulting in an output event stream clocking at the manageable rate of approximately 1.5 kHz.

The first triggering line of action is the hardware-based Level 1 (L1) system, which combines global coarse-granularity information from the calorimeter and the muon spectrometer to reduce the event rate from the initial 40 MHz to about 100 kHz.

The L1 output is stored by the Read-Out System (ROS) and passed to the software-driven High Level Trigger (HLT) as Regions of Interest (ROIs), each covering the general detector area in which the activity that drove the L1 decision was located. The HLT runs a regional object reconstruction within the provided ROIs, with full access to the nominal detector granularity, on the grounds of which the event
3.2. The ATLAS Detector

Figure 3.6: Flow chart of the ATLAS TDAQ system during LHC Run 2. [12]

may be discarded or selected for global event reconstruction. Finally, globally-reconstructed events that meet the physics criteria defined by any of the HLT trigger streams, are permanently recorded and stored in the CERN data storage facility for further use in offline analyses.
Chapter 4

Object Reconstruction at ATLAS

The low level information recorded by the various components of the ATLAS detector, such as Inner Detector tracking hits and calorimeter energy readings, is not immediately useful to physics analyses aiming to describe the outcome of \( pp \) interactions in terms of final state particles, their decay systems and corresponding kinematic properties. An intermediate reconstruction step is therefore applied centrally in order to combine and condense the detector read-out output into reconstruction-level objects ready for analysis, such as electrons, muons and jets.

The following sections will provide an account of how such objects are defined, reconstructed and calibrated in order to best approximate the physical and kinematic properties of the particle(s) they are taken to represent.

4.1 Inner Detector Tracks

Charged particles generated in the \( pp \) collision at the ATLAS interaction point produce a series of hits in the Inner Detector components as they fly towards the outer regions of the detector. The algorithm to extract the trajectory and momentum of the instigating particle out of such hits takes the name of track reconstruction [82], a pivotal step of the overall ATLAS event reconstruction effort.

The importance of a robust and efficient track reconstruction in ATLAS cannot be
overstated. Inner Detector tracks do not simply provide information on the energy and flight parameters of electrons and muons, but play a major role in key performance tasks such as the rejection of physical objects that do not originate from the hard-scatter vertex and the identification of the flavour content of jets, whose very reconstruction and calibration process is also heavily reliant on tracking information.

The main challenge faced by the tracking reconstruction process is presented by the extremely high average number of final state particles per event, each generating its own set of Inner Detector hits as they travel outwards. This gives rise to the possibility of adverse scenarios such as ambiguous hit to track association and spurious combinatorial track reconstruction, and requires the implementation of a robust track finding and fitting algorithm to produce a reliable representation of the physical particles detected.

The first step of the track reconstruction aims at identifying so called space points, three dimensional representations of the particle position built clustering together adjacent detector cells (either pixels from the Pixel detector or strips from the SCT) with an above threshold energy level. A space point can be representative of the passage of either a single or a collimated group of particles, with the two scenarios yielding slightly different clustering configurations.

Given the full set of reconstructed space points, track seeds are defined as an association of three space point from nearby detector layers. For each of such track seed, a primitive estimation of the impact parameter $d_0$, defined as the minimal transverse distance between the track as extrapolated to the beam level and the beam itself. and the track momentum is computed assuming a perfect helicoidal trajectory (which is uniquely defined for a given set of three points). Track seeds are then grown into full tracks by means of an iterative combinatorial procedure which adds additional space points from the remaining ID layers and re-computes the critical track parameters as well as the chi squared $\chi^2$, which provides a measurement of the track fitting quality.
In the event of tracking ambiguity, by which one or more hits are shared by multiple tracks, shared space points are assigned to the highest ranking track by means of a scoring system based on such elements as the number of hits per track, its chi-squared value and the logarithm of the track momentum.

At the end of the ambiguity solving stage, the lower granularity hits from the outermost Inner Detector component (TRT) are added to the existing tracks through one final track-fitting iteration. The resulting track collection is then saved and stored for further use.

### 4.2 Hard-Scatter Vertex

In the busy environment generated by a high luminosity hadron collider such as the LHC, which delivered in 2018 an average number of 40 \( pp \) interactions per bunch crossing, identification of the hard scatter interaction vertex is paramount. Vertices are reconstructed \[83\] out of the inner detector tracks through an iterative fitting procedure extrapolating the candidate track to the beam level, and are subsequently ranked by the scalar sum of the square transverse momenta of all associated tracks \( \Sigma |p_{T,\text{track}}|^2 \). The highest ranking vertex, provided at least two tracks are associated to it, is then selected as the hard scatter vertex of the event under study.

### 4.3 Jets

By virtue of colour confinement \[84\], an exclusive feature of Quantum Chromodynamics preventing the observation of free coloured particles, final state partons produced as a result of \( pp \) interaction can never be observed as standalone particles by any means of detection. This theoretical constraint translates into a real-world phenomenon known as hadronization, whereby numerous quark-antiquark pairs are generated out of the vacuum between two coloured partons as they become more
and more separated. This ultimately leads to the emergence of several colourless (and therefore observable) bound states, such as mesons and baryons, which, as an ensemble, match the original flight direction and kinematic properties of the initiating parton. Such an assortment of collimated hadrons is collectively referred to as a jet, a nearly ubiquitous class of objects in physics analyses at ATLAS.

A jet of hadrons originating out of quarks or gluons is observed experimentally as a close association of tracks in the Inner Detector and most importantly as a relatively localized energy signature in the electromagnetic and hadronic calorimeters. An outline of the main phases of the ATLAS jet reconstruction scheme will be provided in the following sections.

### 4.3.1 Topological Clusters

The ATLAS calorimeters are segmented by design along the radial axis as well as the $\eta \times \phi$ plane, enabling a continuous description of the lateral and longitudinal evolution of the shower within the detecting system.

In the initial step of the jet reconstruction procedure, three-dimensional topological clusters (also referred to as topoclusters) are built out of calorimeter cells with an energy reading$^1$ greater than four times the expected cell energy noise, defined as the joint contribution of the electronic noise and the expected contribution from pile-up interactions.

Topoclusters are then further grown adding two peripheral layers of cells surrounding the seed, provided they pass less stringent criteria of energy reading significance. An iterative sliding-window algorithm is then applied to seek out local maxima in the topocluster energy profile and consequently split extended clusters into two or more smaller entities.

---

$^1$At this stage, all calorimeter cells are calibrated at the EM scale, meaning that the raw electrical charge readout in each cell of the Hadronic Calorimeter is converted into an energy reading through a constant conversion factor computed through test-beam calibration tests.
4.3.2 Particle Flow Algorithm

During the entire Run 1 and the early Run 2 data analysis campaigns, topological cluster served as the fundamental building blocks for jet reconstruction. More recently, an alternative cluster reconstruction procedure incorporating tracking information to complement the calorimetric readings has been adopted as the new standard input for jet reconstruction algorithms.

The philosophy of the Particle Flow [86] clustering approach follows a "best of both worlds" strategy, combining tracking and calorimeter information into a single composite object ideally more closely reflecting each final state hadron with respect to topological clusters. This combined approach yields a series of experimental advantages over the calorimeter-based clustering:

• Low-energy tracks, which are more heavily deflected by the Inner Detector magnetic field, and may initiate a calorimeter cluster separate from the remaining jet particles, are more immediately accounted for in the overall jet reconstruction;

• The energy of a single particle in the low-boost regime is measured with a far greater accuracy by the Inner Tracker, which exploits the magnetic field-induced track curvature, than by the calorimeter, which at low energy is hampered by the statistical fluctuations on the deposited energy. Conversely, high-energy particles are measured to a greater accuracy by calorimeter, as the stochastic fluctuations become negligible;

• Immediate track-to-calorimeter association allows a prompt rejection of calorimeter deposit not associated with the hard-scatter event.

In order for tracking and calorimetric information to be exploited simultaneously as input to jet reconstruction, special care must be taken to avoid double counting a candidate particle as both a track and a calorimeter deposit. This is accomplished through an iterative procedure defined as follows (Fig. 4.1):
4.3. Jets

Figure 4.1: Flow chart outlining the Particle Flow iterative reconstruction procedure

- Inner Detector tracks are extrapolated to the second layer of the calorimeter and geometrically matched to the corresponding calorimeter topological cluster;

- Given the track momentum and parameters as measured in the Inner Detector, the expected energy to be deposited in the calorimeter is computed. If the actual cluster energy falls below a given fraction of the track-based estimate, neighbouring clusters are iteratively incorporated;

- The expected deposited energy is subtracted from the cluster energy. The cluster remnant is removed if the remaining energy is consistent with 0 given the expected single-particle energy resolution provided by the calorimeter.

Once this combined reconstruction algorithm has been applied to all hard-scatter tracks, the output collection consists of fully-matched track-calorimeter objects, unassociated clusters originating from neutral particles and cluster remnants with significant residual energy, typically arising from a geometric superimposition of a neutral and a charged particle.

4.3.3 Jet Reconstruction Procedure

The analysis-level jets considered in the present analysis are constructed running the sequential anti-$k_T$ recombination algorithm [87] on the output collection of matched tracks and unmatched calorimeters clusters produced as output of the Particle Flow clusterization procedure.

As opposed to fixed-cone reconstruction algorithms which combine all input ob-
jets within a certain radius $R$ into a jet, the underlying philosophy of sequential recombination algorithms is to provide an organic description of the hadronic shower evolution by favouring the association of input elements considered likely to have originated from soft or collinear QCD radiation with high-energy elements nearby.

The anti-$k_T$ algorithm translates this line of reasoning into practice defining an *ad hoc* metric to account for the "distance" between two input objects which takes into account the relative momentum imbalance:

$$d_{ij} = \min \left\{ \frac{1}{k_{T,i}^2}, \frac{1}{k_{T,j}^2} \right\} \times \frac{\Delta R_{i,j}^2}{R^2} \quad \text{input-input distance}$$

and the distance between an input object and the beam:

$$d_i = \frac{1}{k_{T,i}^2} \quad \text{input-beam distance}$$

where $k_{T,i}$ is the transverse momentum of the $i$-th input, $\Delta R_{i,j}$ its geometric separation with the $j$-th input and $R$ a dimensionless radius-like parameter with the default value of 0.4.

At the first anti-$k_T$ iteration, an ordered list of all values of $d_{ij}$ and $d_i$, given the Particle Flow inputs, is compiled. If the smallest distance is between two inputs, those are combined and the resulting *proto-jet* re-added to the input list with its position and momentum re-computed after the merger. If, conversely, the smallest distance of the list is that between an input object and the beam, the cluster is removed from the pool and passed to the output as a standalone jet. The procedure is iterated until all isolated clusters or proto-jets are removed from the input pool.

Employing the inverse squared momentum as a weighting factor in the distance metric ensures that the clusterization between two close, low $k_T$ inputs will be greatly disfavoured over asymmetric energy configuration with greater geometric separation. This is to prevent the jet algorithm from producing soft-induced "spurs" in the jet shape, rather incorporating all the soft radiation within the high-energy cluster from which it likely originated. Clustering algorithms, such as the anti-$k_T$, which
4.3. Jets

prevent the geometric and kinematic properties of output jets from being sensitive to soft or collinear emission are referred to as infrared and collinear safe.

The anti-\(k_T\) algorithm is widely used in ATLAS. A variety of distinct output jet collections, each with a specific purpose and utility in physics analyses, are produced out of different recombination inputs and reconstruction parameter used. The next sections will provide an outline of the distinctive features of jet variants relevant to the analysis covered in this thesis.

4.3.4 Large-R Jets

Several theories for new physics phenomena foresee the emergence of massive unobserved resonances decaying to heavy Standard Model objects such as the top quark and the electroweak and Higgs bosons, which inherit a significant Lorentz boost from the rest mass of the progenitor resonance.

As those intermediate particles in turn decay, the geometric separation of the final state products in the detector frame is inversely proportional to the Lorentz boost of the progenitor. For a boosted Higgs boson decaying to a pair of bottom quarks such as is predicted to occur in the physical processes targeted by this search, the average separation can be shown to be [88]:

\[
\Delta R_{b\bar{b}} \simeq \frac{2M_H}{p_{T,H}}
\]  

which implies that, for progenitor \(p_T\) greater than 600 GeV, a large fraction of \(H \rightarrow b\bar{b}\) events will not be efficiently reconstructed as two separated calorimeter jets, as the \(b\bar{b}\) separation falls below the effective radius parameter \(R = 0.4\) of standard jets.

For the purpose of correctly identifying similar cases of extensive merged calorimeter showers, a collection of large-radius jets is built out of calibrated calorimeter topoclusters [89] setting the effective radius parameter within the anti-\(k_T\) recom-
4.3. Jets

Figure 4.2: Schematic representation of the steps in the "Trimming" procedure applied to reduce pile-up dependence on large-R jets observables.

Combination process to 1.0, as opposed to 0.4 as per the standard ATLAS jet definition. Such jets are ideally set to encompass the entirety of the calorimeter signature of boosted hadronic decays such as the aforementioned $H \to b\bar{b}$, as well as fully hadronic $W$, $Z$ and top quark decays.

Given the stated purpose of large-R jet to produce a comprehensive measurement of the entire decay system of a massive resonance, a major point of focus for the jet calibration procedure is that the jet mass\(^2\) correctly reflects the invariant mass of the decaying resonance. This is accomplished through a sequence of calibration steps [90] exploiting both Monte Carlo truth information and in situ measurement of the $W$ pole in high purity $t\bar{t}$ samples.

Owing to the larger area of $R = 1.0$ jets, both the energy and mass readings are more heavily affected by the contribution from energy clusters originating in pile-up interactions. This dependence is neutralized applying a so-called "grooming" procedure [91] to newly reconstructed large-R jets, a targeted algorithm purposed with the identification and removal of the jet components most likely arising from background pile-up interactions. While several grooming protocols exist, the ATLAS JetEtMiss performance group recommends Trimmed jets for use in standard physics analysis. The Trimming procedure of large-R jets is outlined schematically in Figure 4.2, and can be summarized as follows:

- Ungroomed large-R jets are de-clustered into their respective original cluster

---

\(^2\)Defined as the invariant mass of the four-momentum sum of all involved topoclusters, whose mass is assumed to be zero.
• Each jet is then re-clustered into sub-jets by means of the $k_T$ algorithm\[^3\], with effective radius set to 0.2;

• Sub-jets carrying less than 5% of the total jet momentum are discarded;

• The remaining clusters are then rebuilt into the "trimmed" $R = 1.0$ anti-$k_T$ jet.

While large-R jets are conceived to be wide-reaching reconstruction-level entities capable of accounting for the full hadronic decay of a boosted resonance, nothing prevents narrow, single-parton calorimeter showers from being likewise reconstructed as large-R jets.

In order to identify large-R jets most compatible with originating from boosted body decays, the jet substructure is probed for the presence of a typical two-pronged (H and W/Z decays) or three-pronged (hadronic top) energy signature as opposed to the single-pronged energy profile more commonly associated with QCD-initiated large-R jets.

Among the substructure-sensitive variables for the top or boson tagging of large-R jets, most commonly used are the energy correlation functions $D_2$ and $C_2$ and the $n$-subjettines $\tau$ [93].

The sensitivity of the search covered in this thesis relies heavily on the identification of Higgs-compatible large-R jets as means of rejecting the dominant QCD background, therefore bolstering the visibility of a $B \rightarrow bH(b\bar{b})$ signal. A custom Higgs-tagging procedure, which eschews the aforementioned substructure variables, is defined and implemented as described in Section 6.2.2. Within this custom tagging procedure, a pivotal role is played by a third collection of jets: Track Jets.

\[^3\]The original sequential recombination algorithm, differs from the anti-$k_T$ in such a way that the momentum-weighting in the distance metric is proportional to the minimum squared, rather than inverse squared, cluster momentum [92].
4.3.5 Track Jets

Contrary to the two collections of calorimeter jets discussed thus far, Track Jets are built running the standard anti-\( k_T \) algorithm on the collection of Inner Detector tracks.

The spatial resolution provided by the Inner Detector is exploited in the Track Jet reconstruction by setting a smaller effective radius parameter than the calorimeter granularity can allow for the two standard, calorimeter-based, jet collections mentioned in the earlier section.

The possibility of reconstructing smaller-radius jets makes Track Jets promising tools to probe large-\( R \) jets for compatibility with a boosted decay origin, with separate, high energy Track Jets ideally being reconstructed out of the Inner Detector tracks generated by each decay product initiating the large-\( R \) jet. A schematic representation of this basic substructure-tagging scheme is displayed in Figure 4.3.

On the other hand, Track Jets are inherently less suitable than calorimeter or Particle Flow jets to provide a comprehensive picture of the energy of the initiating partons, as they cannot, by definition, account for the energy fraction carried by neutral hadrons.

Historically, Track Jets used in ATLAS analyses were reconstructed setting the anti-\( k_T \) effective radius to 0.2. More recently, however, an alternative Track Jet definition has been recommended specifically for searches seeking to exploit Track Jets for substructure identification. Such jets, known as Variable-Radius (VR) Track Jets [94], are reconstructed allowing the effective radius to scale with the inverse momentum of the track under examination:

\[
R_{VR} = \min\left[0.4, \max\left[0.02, \frac{30 \text{ GeV}}{\not{p}_T \text{ [GeV]}}\right]\right]
\]  

Introducing a dependence of the radius parameter on the track momentum aims to reflect the greater collimation of tracks originating from a highly boosted jet of
4.4 Flavour Tagging

Identifying the flavour of the parton(s) from which a jet originates is a point of major importance in all analysis involving heavy flavour quarks in the final state.

The conservation of flavour in Strong Interactions implies that the jet of particles arising from the hadronization of a bottom quark shall feature one $b$-hadron. Such hadrons are unstable and decay with mean lifetime $\tau_b \simeq 1.5 \times 10^{-12}$ s, corresponding to decay length $c\tau \simeq 450 \mu m$. In this picture, a boosted $b$-hadron often decays...
while inside the beam pipe\(^4\), resulting in several charged tracks which are then identified as originating from a secondary, displaced vertex rather than from the primary vertex. Other distinctive elements of the experimental signature of final state \(b\)-hadrons are the large impact parameter of the tracks arising from the secondary vertex and the presence of a tertiary vertex generated by the subsequent decay of the \(c\)-hadron generated in the \(b\)-hadron decay. A schematic diagram of the distinctive experimental signatures of a \(b\)-initiated jet is displayed in Figure 4.4.

Each of the aforementioned signatures singling out jets initiated by \(b\)-hadrons is targeted by a specific low-level tagger acting on the measured track parameters:

- the \textbf{ID2P} and \textbf{ID3P} taggers [97] produce likelihood-based estimates of the

\(^4\)Highly energetic \(b\)-hadrons may decay in the innermost ID layers by effect of relativistic length contraction in the hadron’s frame of reference, whereby the decay length of the hadron in the laboratory frame is multiplied by the Lorentz factor \(\gamma\).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.4.png}
\caption{Schematic representation of the feature of the track system originated by a \(b\) quark in the final state and reconstructed as a jet. A secondary vertex is produced produced by virtue of the \(b\)-hadron decaying into several particles, including a \(c\)-hadron which in turn decays generating a tertiary vertex. A large impact parameter can also be observed, defined as the closest transverse distance from the primary vertex of the measured tracks as extrapolated to the beam level.}
\end{figure}
probability of the measured transverse and longitudinal impact parameter being consistent from the decays of $b$-hadrons, $c$-hadrons or light-flavoured particles, comparing the available measurement with simulation-based templates;

- the **SV1** algorithm [98] runs an iterative secondary vertex reconstruction associating a candidate displaced vertex to each compatible pair of tracks, and ranking the resulting output vertex pool by reconstructed $b$ decay length and track-to-vertex fit $\chi^2$;

- The **JETFITTER** algorithm [99] reconstructs the entire decay chain of the $b$-hadron all the way through the $c$-hadron tertiary decay, providing a full picture of the process in terms of $b$ and $c$ decay lengths, track multiplicity and invariant mass of the tracks originating from every displaced vertex found.

The outputs of the low level taggers are subsequently taken as inputs, together with the high-level jet kinematic measurements, to an Artificial Deep Neural Network which produces as output the multi dimensional high-level tagger DL1 [100].

DL1 simultaneously quantifies the compatibility of the jet under exam with originating from a $b$, $c$ or light quark/gluon, with the analysis-level one-dimensional discriminant for the $b$-jet hypothesis defined as:

$$DL1_{\text{score}} = \log \left( \frac{p_b}{f_c \cdot p_c + (1-f_c) \cdot p_{\text{light}}} \right)$$  \hspace{1cm} (4.5)

where $p_b$, $p_c$ and $p_{\text{light}}$ refer to the probabilities of the jet originating from a $b$-quark, a $c$-quark or a light parton, and $f_c$ the charm fraction in the background training sample.

The DL1 neural network can be trained and optimized on any jet collection for which tracking information is available, proving an powerful universal tool for flavour tagging in ATLAS analyses. Figure 4.5 shows the expected rate of rejection of light-quark initiated jets obtained selecting upon the output value of a number of ATLAS-recommended flavour tagging algorithms, as a function of the $b$-tagging
4.5 Electrons

Electrons are reconstructed associating Inner Detector tracks to clusters of cells in the Electromagnetic Calorimeter with an above-threshold energy reading. A likelihood-based identification score [101] is computed for each candidate electron in order to discriminate prompt, legitimate electrons from spurious candidates arising from charged hadrons leaving a compatible experimental signature in the tracker and the EM calorimeter.

Given a candidate electron as described by a large set of discriminating variables $\vec{x}$, a likelihood-based score is built given the expected distribution of such quantities in signal and fake electrons. The variable set $\vec{x}$ incorporates such elements as the track quality, the goodness of track-cluster matching and the longitudinal and transverse profile of the electromagnetic shower. The decision to accept or reject the candidate as a signal electron is taken on the grounds of the composite tagger $d_L$, defined as:

$$d_L = \frac{L_S}{L_S + L_F}$$  \hspace{1cm} (4.6)

where $L_S$ and $L_F$ are the likelihood scores of the candidate respectively in the signal and fake electron hypothesis. Three working points (LOOSELH, MEDIUMLH, TIGHTLH) are provided to implement a selection on the electron identification score, each associated with an increasingly higher purity of the resulting sample in exchange for a diminishing signal efficiency.
Figure 4.5: Light quark-initiated jet rejection as a function of the $b$-jet efficiency working point for a number of flavour taggers trained on Particle Flow calorimeter jets (top) or Variable-Radius track jets (bottom).
4.6 Muons

Muons are reconstructed [102] combining the tracking information gathered in the Inner Detector with Muon candidate tracks reconstructed in the Muon Spectrometer. Occasionally, the reconstruction procedure is supported by information provided by the calorimeter, where muon, while not initiating any meaningful electromagnetic shower, is expected to release a small but detectable energy signature.

The baseline muon reconstruction procedure extrapolates a Muon Spectrometer (MS) track, guaranteed to have been generated by a muon, all the way to the Inner Detector (ID) with the purpose of matching any of the established ID tracks. If a suitable tracking match is found, the combined ID+MS track is re-fitted for a more precise determination of the kinematic properties and vertex association of the muon. Muons where a successful MS-ID extrapolation and matching is performed are referred to as combined muons, and provide optimal identification and momentum resolution.

Owing to the mismatch between the Inner Detector and the Muon Spectrometer, both in terms of pseudorapidity coverage and baseline detector efficiency, it is not guaranteed that a suitable match will be found in the Inner Detector for the extrapolated spectrometer track. Unmatched MS tracks are nevertheless extrapolated to the beam level, the kinematic track properties computed accordingly, and the reconstructed muon passed to the output having met the standalone muon reconstruction criteria.

Finally, muons can also be reconstructed as a result of an Inside-Out (IO) extrapolation of ID tracks to the Muon Spectrometer, where standalone hits fitting the extrapolated tracks are sought and, if present, added to form a muon candidate. Inside-Out muons compensate for the lower hit efficiency of the Muon Spectrometer, whereby a passing muon may not trigger a sufficient number of hits to form a full track to be used for Outside-In extrapolation, but still leave an incomplete
signature to be exploited as complementary reconstruction information.

Similarly to electrons, a number of identification working points are provided for individual analyses to generate a sample of signal muons with the preferred levels of efficiency and purity with respect to fakes.

### 4.7 Photons

Prompt photons are reconstructed out of three-dimensional topological clusters in the electromagnetic calorimeter that are not matched to an Inner Detector track. A cut based identification selection, with the chief purpose of rejecting photon candidates arising from π⁰ decays, is implemented on discriminating shower properties such as the fraction of the shower energy deposited in the Hadronic Calorimeter, the shower width at each layer of the calorimeter and the ratio between the energy deposited in adjacent layers.

Within the context of the photon identification, Tight, Medium and Loose working points, separately optimized for a number |η| and $E_T$ ranges, are provided centrally to the individual analyses.
Chapter 5

Analysis Overview

5.1 Search Strategy

As discussed in Chapter 2, the production and subsequent decay of a single Vector-Like B quark can result in a plethora of final states, depending on both the decay mode of the B and eventual secondary decays of the immediate decay products of the B.

The search covered in this thesis targets the $B \rightarrow bH$ channel in the $H \rightarrow b\bar{b}$ decay mode, which results in 3 $b$ quarks in the final state originating from the Vector-Like B decay cascade.

Two additional spectator final state quarks, not originating from the B decay, are produced in the primary interaction vertex, as can be gathered from Fig. 5.1, which displays the two main single production diagrams for a vector-like B quark. The spectator quarks can be identified as:

- A predominantly light quark recoiling off the Z or W boson involved in the $b\bar{Z}B (t \bar{W}B)$ Feynman vertex;

- A bottom (top) quark produced as the result of the gluon splitting\(^1\) which generates the bottom (top) quark involved in the $b\bar{Z}B (t \bar{W}B)$ Feynman vertex.

---

\(^1\)The initial-state $b$ quark involved in the $b\bar{Z}B$ vertex may also arise from the small but non negligible bottom-flavour PDF of the proton.
5.1. Search Strategy

(a) Z-initiated Production  
(b) W-initiated Production

Figure 5.1: Feynman diagrams of the main Leading-Order single-production modes for a Vector-Like B quark. The decay chain targeted by the analysis is also displayed.

This quark typically does not enter the tracking volume and cannot be readily inspected for heavy-flavour tagging.

The geometry of the standalone generation of a vector-like quark according to the mechanisms depicted in Fig. 5.1 bears a topological resemblance to the production of a Higgs boson or any other massive resonance via Vector-Boson Fusion (VBF), in which two initial-state quarks emit electroweak bosons which then interact in a \( \text{WWH} \) or \( \text{ZZH} \) vertex.

Within the context of the single VLB production, and similarly to VBF processes, the final state recoiling quark emerging from the \( qZq'/qWq' \) emission vertex may be detected in the forward region of the apparatus with a smoothly falling transverse momentum spectrum matching that of the emitted boson.

As the analysis is primarily intended to probe the existence of Vector-Like Quarks compatible with the resonance pole masses around and above the 1 TeV threshold, the particles arising from the initial \( B \rightarrow bH \) decay are expected to carry a significant momentum, which, in the case of the Higgs boson, is transferred onto the products of the \( H \rightarrow b\bar{b} \) decay. The angular separation of the \( b\bar{b} \) pair in the detector \( \eta \times \phi \) plane is inversely proportional to the Higgs system transverse boost, rendering the task of individually reconstructing the two \( b \) quarks as standard R = 0.4 calorimeter jets challenging\(^2\). Consequently, the search seeks to reconstruct the Higgs decay

\(^2\)The angular separation of boosted \( H \rightarrow b\bar{b} \) is approximately given by: \( \Delta R \approx 2m(H)/p_T(H) \), resulting in \( \Delta R(b\bar{b}) \approx 0.5 - 0.25 \) respectively at the lower and higher end of the B investigated mass spectrum.
ensemble through a large-radius ($R = 1.0$) jet, encompassing the calorimeter energy deposits initiated by both $b$ quarks into a single physical object. Conversely, the standalone $b$ quark from the direct $B$ decay is reconstructed as a small-radius ($R = 0.4$) jet.

Identifying the final state reconstructed objects as being initiated by $b$ hadrons is the first and most important handle on rejecting the bulk of recorded data events incompatible with originating from a VLB decay.

Specific care is taken to select as Higgs Candidates only large-$R$ jets displaying a two-pronged energy profile in the $\eta \times \phi$ plane, the typical signature of final state objects initiated by two collimated quarks. This is accomplished exploiting the superior spatial resolution offered by the smaller, Inner Detector-based track jets, which are spatially matched to the selected large-$R$ jet providing essential information on its substructure. Track jets associated to the Higgs candidate are then individually and independently inspected for $b$-tagging, achieving maximal rejection of candidate jets not originated from a pair of bottom quarks.

The sensitivity of the search with respect to the VLB signal hypothesis is further enhanced by applying a cut-based kinematic event selection optimized through dedicated studies on the available simulated benchmark samples of varying mass (see Sec. 6.4). To ensure maximal signal purity, the search is ultimately restricted to events satisfying a Higgs compatibility criterion enforced on the invariant mass of the objects assumed to reconstruct the Higgs candidate decay system.

As expected from a fully hadronic final state, the main source of Standard Model events passing the event selection and serving as background to the search is the QCD-driven production of multiple final state $b$ quarks, with the additional contribution from events where one or more gluon, light or charm quark initiated jets are erroneously tagged as a $b$-initiated jet. As a consequence, this search does not rely on Monte Carlo simulations to establish yield and shape of the expected Standard Model contribution to the reconstructed $m_B$ spectrum, electing instead to implement a custom-made data-driven procedure for the modelling of the background, as de-
tailed in Chapter 7.

Beyond the aforementioned sources of Standard Model background, a 3 $b$-quark final state can also be generated as the result of the allowed $B \to bZ(b\bar{b})$ decay chain, with the $b\bar{b}$ reconstructed invariant mass peaking at lower values given that $M_Z < M_H$. The contribution of $Z$ events in the Higgs mass peak region (as shall be defined in Sec. 6.3) is expected to be small owing to both the mass peak requirement and the smaller $Z \to b\bar{b}$ branching fraction.

To insulate the analysis from any bias arising from a possible $B \to bZ$ signal, data events with $M_{b\bar{b}}$ below the lower end of the Higgs mass peak region have not been examined nor used in any capacity.

Evidence for resonant production of a new particle is sought as a statistically significant local excess over the spectrum of $m_B$, the invariant mass of the objects assumed to encompass the VLB decay system, as predicted for the ensemble of Standard Model processes contributing to the Signal Region data. The statistical interpretation of the search findings is carried out in the form of a binned Maximum-Likelihood fit [55, 104] of the data given the established predictions of the Standard Model contributions, the simulated signal hypothesis and their respective uncertainties.

In the absence of statistically significant evidence of a signal, the search results are presented in the form of exclusion limit at the 95% Confidence Level [104] on the production cross section of a Vector-Like B Quark decaying into the $bH$ channel, as a function of either or both the resonance pole mass and specific combinations of values of the phenomenological parameters determining the coupling strength between the Vector Like B and the Standard Model.
5.2. Data and Monte Carlo Samples

5.2.1 Data samples

The data pool on which the analysis is carried out was recorded by the ATLAS detector during the 2015, 2016, 2017 and 2018 data-taking campaigns, encompassing the entirety of the Run 2 proton-proton collisional programme. Over the four years of data-taking, the Large Hadron Collider delivered a grand total of 156 fb⁻¹ integrated luminosity (Fig. 5.2), out of which 147 fb⁻¹ were successfully recorded.
at ATLAS and $139 \text{ fb}^{-1}$ passed the full Data Quality requirements to be eligible for Physics analysis. A final offline data filtering is carried out at the analysis level rejecting events where certain ATLAS detector sub-systems relevant to the analysis were offline, according to the Data Quality information stored in the Good Run Lists provided centrally [105].

5.2.2 Simulated signal samples

A large majority of ATLAS analyses rely on Monte Carlo simulations to obtain an \textit{a priori} understanding of the behaviour of the physical processes involved and what kind of experimental signatures such processes would leave in the ATLAS detector.

While the analysis covered in this thesis employs a fully data-driven procedure to estimate the contribution of Standard Model phenomena to the data passing the full event selection, Monte Carlo simulations of the targeted signal process are employed within both the sensitivity optimization and the statistical interpretation of the results.

Data samples simulating the production and decay of a Vector-Like B quark with resonance pole mass ranging from 1 TeV to 2 TeV were generated assuming by default the singlet configuration, which implies the following asymptotic branching ratios under the Narrow-Width Approximation:

\[
\text{BR}[B \rightarrow bH] \sim \text{BR}[B \rightarrow bZ] \sim 0.5 \cdot \text{BR}[B \rightarrow tW] \simeq 25\%
\]

All signal samples were generated with MadGraph5_aMC@NLO [106] under the four-flavour scheme, using the NNPDF2.3 Parton Density Function sets [107]. Parton showering was handled by Pythia [108] v8.212 interfaced with MadGraph. The generated final state particles were then run through the \textbf{FULL} simulation of the ATLAS detector layout based on Geant 4 [109] to match the established reconstruction procedure of actual data.
For each of the available pole resonance mass points, ranging from 1.0 TeV to 2.0 TeV with constant 200 GeV interval, separate samples were generated for both the leading production mode, arising from the $bZ \to B$ interaction vertex (from now on referred to as $Z$-Initiated or $ZBHb$ production), as well as the sub-leading contribution from the $tW \to B$ vertex ($W$-Initiated or $WBHb$).

The strength of the coupling between the $B$ vector-like quark and the electroweak and Higgs bosons is modeled through the following parameters [110]:

\[ c_Z = \frac{g_W}{2 \cos \theta_W} \kappa_Z, \quad c_W = \frac{g_W}{\sqrt{2}} \kappa_W, \quad c_H = \frac{M_B}{v} \kappa_H, \]  

(5.1)

with $g_W$ and $\theta_W$ being the weak coupling constant and the Weinberg angle, $v$ the vacuum expectation value and $\kappa_W$, $\kappa_Z$, $\kappa_H$ regulating the coupling strengths of a Vector-Like $B$ quarks to the electroweak and Higgs bosons. The individual kappas are effective coupling constants encompassing the range of possible phenomenological scenarios that arise in the individual theoretical models predicting the emergence of Vector-Like quarks of any kind (eg. occurring as singlet, doublets, etc.).

The $\kappa_V (V = Z, W, H)$ parameters can be further parametrized as scaling with a common, ultimate coupling strength $\kappa$ regulating the generic "strength" of the coupling between the Vector-Like quark under consideration and the Standard Model:

\[ \kappa_V = \kappa \times \xi_{V}^{theo} \]  

(5.2)

where $\xi_{V}^{theo}$ is a dimensionless coefficient regulating the hierarchy of the three $\kappa_V$ in each theoretical scenario.

Under this effective parametrization, the partial decay width of the resonance in any of its three allowed decay modes is proportional to the square of the relevant coupling $\kappa_V$.

The Monte Carlo samples used as benchmark for the present search were generated with a nominal strength $\kappa = 1.0$. In order to facilitate the interpretation of the experimental search results in terms of limits on the VLB coupling array $c_V$,
the sample production included a matrix-element reweighting mechanism allowing, through the application of a set of event weights, to simulate the effect on the kinematic shapes and event yields of different values of the coupling strength, ranging from $\kappa = 0.1$ to $\kappa = 1.6$. The matrix-element reweighting mechanism also allows to map the signal distribution of the considered sample into those relative to a hypothetical signal with a pole mass 100 GeV below the nominal mass, allowing a finer effective spacing of the eventual mass-dependent limits.

All samples are inclusive with respect to the secondary Higgs decay products. The values of the cross sections have been re-computed at Next-To-Leading-Order (NLO) [110, 111] assuming a singlet configuration and enforcing the Narrow Width Approximation (NWA) through an appropriate choice of coupling strength. The Finite-Width cross section for the generated samples can be computed out of the NWA cross sections ($\hat{\sigma}$) exploiting the quadratic scaling of the total cross section with the coupling constants:

$$\sigma_{FW}^{W BHb}(c_W) = \sigma_{NWA}^{W BHb} \times \left[ c_W \right]^2 \times \frac{1}{P_{NWA}}$$

$$\sigma_{FW}^{W BHb}(c_Z) = \sigma_{NWA}^{Z BHb} \times \left[ c_Z \right]^2 \times \frac{1}{P_{NWA}}$$

where $P_{NWA}$ is a correction factor accounting for the imperfect factorization of the coupling dependence of the cross section for large width [51].

For technical reasons pertaining to the coupling reweighting implementation, the generation of the CP states $B$ and $\bar{B}$ in processes involving a neutral boson (Z,H) in both the generation and the decay of the resonance were handled separately and merged at a later stage.

The value of the finite-width production cross section of a single VLB of each considered mass, calculated at Next-to-Leading Order, are summarised in Table 5.1, 5.2 for respectively the $Z BHb$ and $W BHb$ production modes. A graphical rendition of the same result is displayed in Fig. 5.3a-5.3b, while Figure 5.3c provides a vi-
### 5.2. Data and Monte Carlo Samples

<table>
<thead>
<tr>
<th>$\sigma_{FW}(ZBhh) \ [fb]$</th>
<th>$\kappa$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_B = 1000 \text{ GeV}$</td>
<td>$2.39^{+0.07}_{-0.05}$</td>
<td>$21.3^{+0.6}_{-0.4}$</td>
<td>$63.0^{+2.0}_{-1.0}$</td>
<td>$176.0^{+5.0}_{-4.0}$</td>
<td>$282.0^{+8.0}_{-6.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1200 \text{ GeV}$</td>
<td>$1.25^{+0.04}_{-0.03}$</td>
<td>$11.6^{+0.4}_{-0.3}$</td>
<td>$36.6^{+1.1}_{-0.9}$</td>
<td>$108.0^{+3.0}_{-3.0}$</td>
<td>$176.0^{+5.0}_{-5.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1400 \text{ GeV}$</td>
<td>$0.69^{+0.02}_{-0.02}$</td>
<td>$6.8^{+0.2}_{-0.2}$</td>
<td>$23.2^{+0.8}_{-0.7}$</td>
<td>$72.0^{+3.0}_{-2.0}$</td>
<td>$116.0^{+4.0}_{-3.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1600 \text{ GeV}$</td>
<td>$0.39^{+0.01}_{-0.01}$</td>
<td>$4.3^{+0.1}_{-0.1}$</td>
<td>$16.2^{+0.6}_{-0.5}$</td>
<td>$51.0^{+2.0}_{-1.0}$</td>
<td>$80.0^{+3.0}_{-2.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1800 \text{ GeV}$</td>
<td>$0.231^{+0.007}_{-0.007}$</td>
<td>$2.92^{+0.1}_{-0.09}$</td>
<td>$12.2^{+0.4}_{-0.4}$</td>
<td>$36.0^{+1.0}_{-1.0}$</td>
<td>$56.0^{+2.0}_{-2.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 2000 \text{ GeV}$</td>
<td>$0.142^{+0.006}_{-0.006}$</td>
<td>$2.18^{+0.09}_{-0.09}$</td>
<td>$9.8^{+0.4}_{-0.4}$</td>
<td>$28.0^{+1.0}_{-1.0}$</td>
<td>$43.0^{+2.0}_{-2.0}$</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.1:** Next-to-Leading Order production cross section for a $Z$-Initiated standalone Vector-Like B quark given a resonance pole mass ranging from 900 GeV to 2 TeV. The displayed figures have been recomputed from the values generated by MadGraph under the Narrow Width Approximation, and corrected for finite width effects as outlined in the main body.

<table>
<thead>
<tr>
<th>$\sigma_{FW}(WBhh) \ [fb]$</th>
<th>$\kappa$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_B = 1000 \text{ GeV}$</td>
<td>$0.69^{+0.05}_{-0.06}$</td>
<td>$5.8^{+0.4}_{-0.5}$</td>
<td>$18.0^{+1.0}_{-2.0}$</td>
<td>$46.0^{+3.0}_{-4.0}$</td>
<td>$72.0^{+5.0}_{-6.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1200 \text{ GeV}$</td>
<td>$0.38^{+0.04}_{-0.04}$</td>
<td>$3.3^{+0.3}_{-0.3}$</td>
<td>$10.2^{+1.0}_{-1.0}$</td>
<td>$28.0^{+3.0}_{-3.0}$</td>
<td>$44.0^{+4.0}_{-4.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1400 \text{ GeV}$</td>
<td>$0.21^{+0.02}_{-0.02}$</td>
<td>$2.0^{+0.2}_{-0.2}$</td>
<td>$6.4^{+0.7}_{-0.7}$</td>
<td>$18.0^{+2.0}_{-2.0}$</td>
<td>$29.0^{+3.0}_{-3.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1600 \text{ GeV}$</td>
<td>$0.12^{+0.01}_{-0.01}$</td>
<td>$1.3^{+0.1}_{-0.1}$</td>
<td>$4.3^{+0.5}_{-0.5}$</td>
<td>$12.0^{+1.0}_{-1.0}$</td>
<td>$20.0^{+2.0}_{-2.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 1800 \text{ GeV}$</td>
<td>$0.073^{+0.009}_{-0.009}$</td>
<td>$0.8^{+0.1}_{-0.1}$</td>
<td>$3.1^{+0.4}_{-0.4}$</td>
<td>$9.0^{+1.0}_{-1.0}$</td>
<td>$14.0^{+2.0}_{-2.0}$</td>
<td></td>
</tr>
<tr>
<td>$M_B = 2000 \text{ GeV}$</td>
<td>$0.045^{+0.006}_{-0.006}$</td>
<td>$0.6^{+0.08}_{-0.07}$</td>
<td>$2.4^{+0.3}_{-0.3}$</td>
<td>$6.7^{+0.8}_{-0.8}$</td>
<td>$10.0^{+1.0}_{-1.0}$</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.2:** Next-to-Leading Order production cross section for a $W$-Initiated standalone Vector-Like B quark given a resonance pole mass ranging from 900 GeV to 2 TeV. The displayed figures have been recomputed from the values generated by MadGraph under the Narrow Width Approximation, and corrected for finite width effects as outlined in the main body.

Visualization of how the width $\Gamma/M$ is affected by the choice of coupling strength $\kappa$. 
5.2.3 Coupling Reweighting Mechanism

Prior to the large-scale generation of the Vector-Like B signal samples, a small-scale private production was undertaken to examine the macroscopic features of the signals as well as to validate the coupling and mass reweighting mechanism. A total of 40000 events for a select number of resonance mass points were generated (20000 for Z-Initiated and W-Initiated production each) with the nominal coupling strength $\kappa = 1.0$. 

Figure 5.3: NLO Cross Section and fractional width $\Gamma/M$ for a Vector-Like B quark resonance with varying pole mass (x axis) and coupling strength with the Standard Model ($\kappa$, on the y axis).
Figure 5.4 shows the coupling-reweighted resonance line shape at particle level for simulated samples with pole mass set to 1.2 TeV (top row) and 2.0 TeV (bottom row) for several values of $\kappa$ in the available range.

It is immediately evident how larger values of $\kappa$, corresponding to a larger relative width $\Gamma_B/M_B$, result in a very sizeable enhancement of the production in the region where the actual particle level resonance mass $M_B$ is lower than the pole mass $M_B^{\text{pole}}$, to the point of losing completely the peaking resonance shape for the highest values of $\kappa$. The low-mass enhancement effect is observed to be particularly severe in the high end the resonance mass spectrum, as made evident in Fig. 5.5, which displays the VLB line shapes for various pole mass hypotheses, given a fixed value of the coupling $\kappa$.

The low-mass production enhancement is an exclusive finite-width feature of the $pp \to B \to bH$ production/decay chain (hence the observed $\kappa$ dependence), and is understood to originate from the interplay between the fermionic (B) and the scalar (H) propagators in the matrix element. $pp \to B \to tW$ and $pp \to B \to bZ$ do not show such a feature and are therefore observed to display the more classic peaking shape (with a small drift of the resonance peak towards larger mass values for large-width samples) throughout the available range of $\kappa$ values, with a lower total cross section (Fig. 5.6).

The impact on the analysis of the Higgs-specific production enhancement on the low $M_B/M_B^{\text{pole}}$ region is two-fold. At moderate to large values of $\kappa$, the VLB production takes on a fully non-resonant shape in the natural observable provided by the VLB decay invariant mass, leaving little kinematic difference between the various pole mass hypotheses and therefore turning the search into a counting experiment.

On the other hand, most of the events belonging to the low $M_B/M_B^{\text{pole}}$ tail owed to the production enhancement fall outside the acceptance of the analysis, which is engineered to be sensitive to the region of the phase space where the immediate B decay products have transverse momentum around or above 500 GeV. Figure 5.7,
5.2. Data and Monte Carlo Samples

Figure 5.4: Particle-level resonance mass for Z and W-Initiated VLB production in the various coupling strength hypotheses.

Figure 5.5: VLB Line Shape for resonance of varying pole mass, given a fixed value of the coupling strength $\kappa$. 

(a) ZBHb, 1200 GeV, Linear Scale

(b) WBHb, 1200 GeV, Linear Scale

(c) ZBHb, 2000 GeV, Linear Scale

(d) WBHb, 2000 GeV, Linear Scale
5.2. Data and Monte Carlo Samples

![Graphs](image)

(a) ZBZb, 1200 GeV, Linear Scale  
(b) ZBZb, 1200 GeV, Log Scale

**Figure 5.6:** Particle-level resonance mass for $B \to bZ$ production in the various coupling strength hypotheses.

![Graphs](image)

(a) ZBHb, 1200 GeV, Higgs $p_T$  
(b) ZBHb, 2000 GeV, Higgs $p_T$

**Figure 5.7:** Particle-level Higgs $p_T$ for Z-Initiated signal events.

displaying the $p_T$ of the Higgs for various values of $\kappa$, shows how most of the signal events will be lost to the aforementioned condition, enforced both online and offline by the choice of trigger (see Sec. 6.1.6) and by a matching $p_T$ requirement on the large-R jets in the event (Sec. 6.2.1).
Chapter 6

Event Reconstruction and Selection

Chapter 4 provided an extensive account of how the plethora of experimental signatures recorded by the ATLAS detector is interpreted and processed into physical objects reflecting the nature and properties of the final state particle originating them.

The present chapter will first detail the analysis-specific identification and selection criteria of the aforementioned objects (Sec. 6.1) and subsequently outline (Sec. 6.2) how those are progressively assembled into custom-defined composite objects taken to reflect the particles involved in a hypothetical signal event, such as the Higgs boson produced in the Vector-Like Quark decay or the vector-like resonance itself.

Finally, the last section (Sec. 6.4) of this chapter will focus on the discussion of the expected properties of the high-level objects in signal and background events, and how those can be exploited to define a series of event selection criteria that maximise the sensitivity of the search.

6.1 Object Selection

The selection of physical objects is performed at the user code level on centrally produced datasets for both simulated signal events and actual data. The parameters, tools and working points used in the object definition have been chosen in
accordance with the latest recommendation by the respective ATLAS performance groups, as detailed below.

### 6.1.1 Small-R Jets

As the present search targets a fully hadronic final state, a specific focus has been dedicated to the definition and the selection criteria for a specific jet to be included in the selected events. Throughout the search, three jet collections were used, each with a distinctive set of inputs, reconstructed parameters and eventual purpose in the analysis.

Small-\(R\) jets are the physical objects most commonly used within ATLAS to account for hadronic activity in the calorimetric region of the detector.

Several hardware issues, such as coherent noise or noise spikes in the end-cap region, as well as non-collision or cosmic sources of background, can give rise to spurious energy clusters which are then reconstructed into fictitious jets commonly referred to as "bad jets" [112, 113]. Event cleaning requirements are enforced according to the ATLAS recommendation [114] in order to single out such jets. A veto is applied on events featuring at least one jet failing the quality criteria.

Jets passing the cleaning stage are tested for compatibility with originating from the primary vertex via a fixed-value cut on the Jet-Vertex Tagger (JVT) [115] multivariate algorithm output, a procedure designed to reduce the impact of in-time pile up on the analysis. Jets falling outside the tracking acceptance (with values of pseudorapidity \(|\eta| > 2.4\)) undergo a tighter selection via the specifically designed forward JVT (fJVT) [116] along with stricter momentum requirement \(p_T > 40\) GeV compared to the \(p_T > 25\) GeV threshold for central jets. The forward Jet-Vertex-Tagger fJVT evaluates the compatibility of the forward jet under examination with originating from each of the reconstructed pile-up vertices. This is done by projecting the jet transverse four-momentum \(p_T\) along the direction of the missing transverse momentum computed summing all central tracks and jets associated to a specific
6.1. Object Selection

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>anti-$k_T$</td>
</tr>
<tr>
<td>$R$-parameter</td>
<td>0.4</td>
</tr>
<tr>
<td>Input constituent</td>
<td>Particle Flow objects</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable</td>
</tr>
<tr>
<td>( p_T )</td>
</tr>
<tr>
<td>( p_T ) (forward jets)</td>
</tr>
<tr>
<td>(</td>
</tr>
<tr>
<td>(</td>
</tr>
</tbody>
</table>
| JVT               | \( \text{JVT > 0.2 in } |\eta|<2.4 , \ p_T \in [20, 60] \text{ GeV} \)
|                    | \( \text{JVT > 0.11 in } |\eta| \in [2.4, 2.5] \) |
|                    | \( \text{fJVT < 0.4 in } |\eta| \in [2.4, 4.5], \ p_T \in [20, 120] \text{ GeV} \) |

Table 6.1: Small-$R$ Jet reconstruction criteria.

pile-up vertex. If the forward jet does indeed belong to any of the pile-up vertices, the relevant projection will take values close to unity, highlighting how the examined forward jet would restore momentum balance on the transverse plane of all objects originating from the pile-up vertex.

A comprehensive summary of the algorithm, parameters and working points used in the small-$R$ jet reconstruction, cleaning and selection is presented in Table 6.1.

6.1.2 Large-$R$ Jets

Large-$R$ Jets [117] are wider-reaching physical objects conceived to encompass diffuse, rather than localized as with small-$R$ jets, hadronic activity. They are especially favoured to seamlessly reconstruct the kinematic properties of a hadronically decaying object in moderate to highly boosted configurations.

Large-$R$ jets are reconstructed by applying the anti-$k_T$ algorithm to calorimeter
6.1. Object Selection

### Feature | Criterion
--- | ---
Algorithm | AntiKt
R-parameter | 1.0
Input constituent | LCTopo calorimeter clusters
Grooming | Trimming, $f = 5\%$, $R_{sub} = 0.2$

| Observable | Requirement |
--- | --- |
$p_T$ | $> 480$ GeV
$|\eta|$ | $< 2.0$
$|\eta|$ | $< 2.0$
Mass | $> 50$ GeV

**Table 6.2:** $R = 1.0$ Jets reconstruction criteria.

clusters and setting radius parameter to $R = 1.0$. These are referred to as AntiKt10LCTopo jets.

Due to the larger catchment area [118] of large-$R$ jets, energy clusters not associated with the hard scatter event fall in greater number into the reach of the clustering algorithm. This causes the key kinematic features of the jets, such as its mass and transverse momentum, to be strongly affected by pile-up interactions.

Out of the several algorithms that were developed to mitigate the effects of the non-hard-scatter energy deposits, collectively known under the name of *grooming* [91], the ATLAS collaboration recommends for analysis to use large-$R$ jets that have undergone *trimming* [119], with reclustering radius set to $R_{sub} = 0.2$ and trimming fraction $f_{sub} = 5\%$.

As summarised in Table 6.2, requirements for large-$R$ jets to have large enough transverse momentum and to fall within the central region of the detector are enforced in compliance with the current recommendations [120] of the ATLAS JetET-miss Group.
6.1.3 Track Jets

This search exploits the superior spatial resolution of track jets to identify features in the substructure of the existing large-R jets, greatly improving the rejection of any such candidate jets incompatible with originating from the two-body decay of a boosted Higgs boson into a bottom quark pair. To achieve optimal Higgs identification and flavour tagging efficiencies, Variable Radius track jets as defined in Sec. 4.3.5 are used throughout the analysis.

The identification of local maxima in the internal energy profile of the large-R jet, suggesting compatibility with the $H \rightarrow b\bar{b}$ hypothesis, is achieved matching the $R = 1.0$ jet under exam with a number of track jets. Following the ATLAS JetEtMiss group recommendation, the matching between is performed through an area-based procedure known as *ghost association* [93, 121, 122]. Table 6.3 summarises the selection criteria for track jets employed in the analysis.

<table>
<thead>
<tr>
<th>Selection</th>
<th>Working Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>anti-$k_T$</td>
</tr>
<tr>
<td>Radius</td>
<td>$\max\left[0.02, \min\left(0.4, \frac{30}{\text{track } p_T \text{[GeV]}}\right)\right]$</td>
</tr>
<tr>
<td>Input Collection</td>
<td>Inner Detector Tracks</td>
</tr>
<tr>
<td>Transverse Momentum</td>
<td>$p_T &gt; 10 \text{ GeV}$</td>
</tr>
<tr>
<td>Large-R Jet association</td>
<td>Ghost-Association</td>
</tr>
</tbody>
</table>

*Table 6.3: Summary of selection working points for track jets.*

6.1.4 Flavour Tagging

As this search targets a decay channel eventually resulting in 3 bottom quarks in the final state, implementing a procedure to reliably identify the physical objects initiated by such quarks, as opposed to other partons, results in a major boost in the signal-to-background ratio and consequently in the overall reach of the search.
As summarised in Table 6.4, this analysis estimates the $b$-quark content in selected large-R jets performing $b$-tagging on associated track jets through a fixed-level cut on the value of the DL1 tagger as specifically trained on the variable-radius track jet collection. The chosen working point on the DL1 selection ensures a theoretical 70% tagging probability on genuine $b$-initiated track jets in a high-purity $t\bar{t}$ simulated sample, and a rejection factor of 9 and 356 respectively on charm and light quark initiated jets.

The $b$-tagging of $R = 0.4$ jets is instead performed on the new generation tagger DL1r [123], an alternative tagger re-trained through a recurring neural network and providing the state of the art in light and c-jet rejection, with rejection factors respectively of the order of 800 and 10.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable-R TrackJets</strong></td>
<td></td>
</tr>
<tr>
<td>Jet collection</td>
<td>Variable-R Track Jets</td>
</tr>
<tr>
<td>Jet selection</td>
<td>$p_T &gt; 10$ GeV</td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>Algorithm</td>
<td>DL1</td>
</tr>
<tr>
<td>Operating point</td>
<td>Fixed-Value</td>
</tr>
<tr>
<td></td>
<td>Eff = 70%</td>
</tr>
<tr>
<td><strong>R=0.4 Jets</strong></td>
<td></td>
</tr>
<tr>
<td>Jet collection</td>
<td>$R = 0.4$ Particle Flow Jets</td>
</tr>
<tr>
<td>Jet selection</td>
<td>$p_T &gt; 25$ GeV</td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>Algorithm</td>
<td>DL1r</td>
</tr>
<tr>
<td>Operating point</td>
<td>Fixed-Value</td>
</tr>
<tr>
<td></td>
<td>Eff = 70%</td>
</tr>
</tbody>
</table>

*Table 6.4: b-tagging selection criteria.*
6.1.5 Leptons and Photons

As the final state targeted by the search features no prompt leptons, events are vetoed if they are found to contain well reconstructed electrons or muons at the pre-selection stage. Identification criteria [124,125] are enforced on candidate veto leptons according to the latest recommendations by the ATLAS EGamma and Muon Combined Performance group [126], which are implemented in the analysis code as displayed in Fig. 6.5.

No photons originating from the primary vertex are expected to be present in the final state. In order to preserve orthogonality with ongoing and future searches targeting the $H \rightarrow \gamma\gamma$ decay channel, a vetoing selection step is implemented on $\gamma\gamma$ pairs with reconstructed invariant mass compatible with the Higgs boson mass. The reconstruction criteria for a veto photon are in accordance with the custom criteria defined by the ongoing ATLAS $B \rightarrow bH(\gamma\gamma)$ search, as detailed in Table 6.6.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Electrons</strong></td>
<td></td>
</tr>
<tr>
<td>Pseudorapidity range</td>
<td>$</td>
</tr>
<tr>
<td>Transverse momentum</td>
<td>$&gt; 20$ GeV</td>
</tr>
<tr>
<td>Track to vertex association</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>Identification WP</td>
<td>LOOSE/LH</td>
</tr>
<tr>
<td><strong>Muons</strong></td>
<td></td>
</tr>
<tr>
<td>Pseudorapidity range</td>
<td>$</td>
</tr>
<tr>
<td>Transverse momentum</td>
<td>$&gt; 20$ GeV</td>
</tr>
<tr>
<td>Track to vertex association</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>Identification WP</td>
<td>MEDIUM</td>
</tr>
</tbody>
</table>

*Table 6.5: Electron and Muon selection criteria.*
### 6.1. Object Selection

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Photons</strong></td>
<td></td>
</tr>
<tr>
<td>Pseudorapidity range</td>
<td>$</td>
</tr>
<tr>
<td>Transverse energy</td>
<td>$E_T &gt; 25$ GeV</td>
</tr>
<tr>
<td>Identification Quality</td>
<td>Tight</td>
</tr>
<tr>
<td>$H \rightarrow \gamma \gamma$ veto</td>
<td>$m_{\gamma \gamma} \in [105, 160]$ GeV</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.6:** Photon selection criteria.

### 6.1.6 Trigger Selection

A trigger selection requirement was applied on both data and simulated signal events, corresponding to the lowest unprescaled large-$R$ jet trigger recommended for each year of data taking, as summarized in Table 6.7. All the three trigger streams used record events with an $R = 1.0$ jet reconstructed at the High Level Trigger stage, with a transverse momentum threshold ranging from 420 GeV (Early Run 2) to 460 GeV (2017 and 2018 data taking campaigns).

As displayed in Figure 6.1, the triggering efficiency for large-R jet as a function of their offline reconstructed $p_T$ reaches approximately 100% at 480 GeV. In order to avoid selecting jets that fall into the turn-on region of the trigger efficiency curve, which in principle may be incorrectly modelled in Monte Carlo simulations, large-R jets are required to have $p_T > 480$ GeV to be eligible for further use.

<table>
<thead>
<tr>
<th>Year</th>
<th>Trigger</th>
<th>samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015+2016</td>
<td>HLT_j420_a10r_L1J100</td>
<td>Data, MC16a Signal</td>
</tr>
<tr>
<td>2017</td>
<td>HLT_j460_a10t_lcw_jes_L1J100</td>
<td>Data, MC16d Signal</td>
</tr>
<tr>
<td>2018</td>
<td>HLT_j460_a10t_lcw_jes_L1J100</td>
<td>Data, MC16e Signal</td>
</tr>
</tbody>
</table>

**Table 6.7:** List of triggers used by the analysis, each associated with the datasets on which they were applied.
Figure 6.1: Triggering efficiency for various large-R jet triggers with a common High Level Trigger threshold set to $E_T = 460$ GeV. The different curves highlight the different behaviour of the triggering efficiency turn-on observed for three different pile-up mitigation algorithms. The search covered in this thesis makes use Trimmed large-R jets only.
6.2 High Level Event Reconstruction

6.2.1 Pre-Selection

A series of selection steps is implemented at the earliest stage of the workflow to define a sub-section of the phase space loosely compatible with the expected features of the signals, namely the presence of a high-$p_T$ large-$R$ jet opposite to a similarly boosted small-$R$ jet.

Applying a pre-selection at the earliest stage of the analysis sharply reduces the number of events that undergo the more computationally costly stages of the event reconstruction framework, allowing to run the reconstruction, selection and background modelling stages locally for a more agile and flexible analysis workflow.

The set of pre-selection kinematic and geometric requirements is outlined in Table 6.8, with Table 6.9 instead providing an approximate step-wise efficiency for data and two simulated signal mass points with the nominal coupling strength value $\kappa = 1.0$.

Requiring large-$R$ jets to satisfy the trigger plateau $p_T$ threshold, which is achieved accepting only $R = 1.0$ jets with $p_T > 480$ GeV, is at the same time the greatest source of signal loss and background rejection in the pre-selection.

The low signal efficiency of the large-$R$ jet $p_T$ cut is understood to originate from the cross section enhancement mechanism outlined in Sec. 5.2.3 and visualized in Fig. 5.4. This mechanism results in the bulk of high-$M_B^{pole}$, high-$\kappa$ resonance events being produced off shell, far below the "nominal" resonance mass $M_B^{pole}$, and consequently generate decay products that fall outside the $p_T$ acceptance.

This interpretation is substantiated by the information displayed in Figure 6.2, which summarises the signal efficiency of every pre-selection step for a number of values of $\kappa$ under various VLB mass hypotheses. The efficiency of the require-
6.2. High Level Event Reconstruction

<table>
<thead>
<tr>
<th>ID</th>
<th>Selection Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:1</td>
<td>$\geq 1$ R=1.0 Jet, $p_T &gt; 480$ GeV</td>
</tr>
<tr>
<td>0:2</td>
<td>0 Leptons</td>
</tr>
<tr>
<td>0:3</td>
<td>$0 H \rightarrow \gamma\gamma$ compatible pairs</td>
</tr>
<tr>
<td>0:4</td>
<td>$\geq 2$ ass. Track Jets to large-R jet, 1 $b$-tagged</td>
</tr>
<tr>
<td>0:5</td>
<td>small-R jet with $p_T &gt; 300$ GeV</td>
</tr>
<tr>
<td>0:6</td>
<td>$\Delta R(\text{small} - R, \text{large} - R) &gt; 2.0$</td>
</tr>
</tbody>
</table>

**Table 6.8:** Sequential description of the pre-selection step. A unique identifier is provided for each step for quick reference in the main body and in subsequent tables.

<table>
<thead>
<tr>
<th>ID</th>
<th>Data</th>
<th>Sig. 1200 GeV</th>
<th>Sig. 2000 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ZBHb</td>
<td>WBHb</td>
</tr>
<tr>
<td>0:1</td>
<td>4.1%</td>
<td>21.2%</td>
<td>37.3%</td>
</tr>
<tr>
<td>0:2</td>
<td>96.2%</td>
<td>95.2%</td>
<td>81.7%</td>
</tr>
<tr>
<td>0:3</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>0:4</td>
<td>9.1%</td>
<td>56.9%</td>
<td>62.0%</td>
</tr>
<tr>
<td>0:5</td>
<td>96.5%</td>
<td>95.9%</td>
<td>94.7%</td>
</tr>
<tr>
<td>0:6</td>
<td>91.5%</td>
<td>85.5%</td>
<td>82.2%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>0.31%</td>
<td>9.4%</td>
<td>14.7%</td>
</tr>
</tbody>
</table>

**Table 6.9:** Step by step efficiency in the Pre-Selection for data and simulated signal events. The Z-Initiated (ZBHb) and W-Initiated (WBHb) production modes are left distinct to better appreciate the differences. Signal efficiencies refer to datasets generated with $\kappa = 1.0$. 


6.2. High Level Event Reconstruction

Figure 6.2: Individual efficiency of the 6 pre-selection steps as defined and labeled in Table 6.8. The efficiency have been computed for a number of values of the coupling strength $\kappa$.

The difference in signal efficiencies for pre-selection steps 0:1, 0:2, 0:4 (and to a smaller degree 0:6) between Z-Initiated and W-Initiated mode are all understood to originate from the presence of a final state top quark, an exclusive feature of W-Initiated events (see Fig. 2.5b). The high mass of the final state top quark from gluon splitting (as opposed to the $b$ quark similarly generated in ZBhb events)
causes the rest of the final state system to carry, on average, a larger transverse momentum to ensure four-momentum balance on the transverse plane. This is reflected on the momentum distributions of the VLB decay products, which therefore satisfy the plateau $p_T$ requirement more frequently.

Conversely, the presence of a final state top quark in W-Initiated events negatively affects the signal efficiency of W-Initiated events in the presence of a lepton veto as enforced in step 0:2, as a fraction close to 20% of $t \rightarrow bW$ decays will result in either a high energy lepton or muon in the final state.

### 6.2.2 Higgs Candidate Reconstruction

The first high-level object to be identified en route to the full VLB candidate reconstruction is the Higgs boson generated in the immediate VLB decay, itself in turn decaying into a $b\bar{b}$ quark pair.

This search defines as Higgs Candidate any large-$R$ jet in the event that passes the basic selection criteria outlined in Sec. 6.1.2 and with at least two selected track jets with $p_T > 50$ GeV associated to it, as described in Sec. 6.1.3. The requirement on the associated track jets is intended to provide information about the compatibility of the Higgs Candidate substructure with the two-body decay hypothesis.

Only large-$R$ jets with transverse momentum above 480 GeV are eligible for the Higgs Reconstruction phase, so that any potential Higgs Candidate lies above the threshold for the 100% trigger efficiency plateau in signal and data events.

The $b$-tagging content of the Higgs Candidate is evaluated on the grounds of the number of $b$-tagged track jets associated to the large-$R$ jet under study. Track Jets associated to a large-$R$ jet are ranked by decreasing values of transverse momentum, with only the two leading ones being subsequently inspected for $b$-tagging.

Following the FlavourTagging group [127] recommendations, large-$R$ jets are rejected if the $\Delta R$ between any of its associated track jets and any other track jet in
the event is less than the smaller of the two track jet radii:

\[
\Delta R_{ij} < \min \left[ R_i, R_j \right] \tag{6.1}
\]

with the \( i \) index running over all track jets inspected for \( b \)-tagging (and therefore solely any of the two track jets associated to the large-\( R \) jet in exam) and the \( j \) index running over the wider collection of all track jets in the event with \( p_T > 10 \text{ GeV} \).

This procedure (henceforth referred to as "Collinear Veto") removes from the analysis the cases of collinear track jet reconstruction that the Variable Radius jet clustering algorithm can lead to, a scenario that can produce ambiguous and questionable track-to-track jet associations within the \( b \)-tagging procedure.

<table>
<thead>
<tr>
<th>Selection</th>
<th>Higgs Candidate</th>
<th>Working Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object ( p_T )</td>
<td>( R = 1.0 ) Jets</td>
<td>( p_T &gt; 480 \text{ GeV} )</td>
</tr>
<tr>
<td>Associated track jets</td>
<td>( p_T \geq 2 )</td>
<td></td>
</tr>
<tr>
<td>Track jets ( p_T )</td>
<td>( p_T \geq 50 \text{ GeV} )</td>
<td></td>
</tr>
<tr>
<td>Collinear Veto</td>
<td>( \Delta R(t\text{Jet}_i,t\text{Jet}<em>j) \leq \min \left[ R</em>{t\text{Jet}<em>i}, R</em>{t\text{Jet}_j} \right] )</td>
<td>( p_T \geq 50 \text{ GeV} )</td>
</tr>
<tr>
<td>Inspected Track Jets (i-range)</td>
<td>( p_T \geq 50 \text{ GeV} )</td>
<td></td>
</tr>
<tr>
<td>Triggering Track Jets (j-range)</td>
<td>( p_T \geq 10 \text{ GeV}, n\text{Tracks} &gt; 1 )</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.10:** Summary of reconstruction criteria for Higgs Candidates.

Table 6.11 provides a comprehensive summary of the Higgs Reconstruction procedure, displaying both the number of events passing each sub-step in two simulated signal samples as well as the corresponding step efficiency. The two signal production modes are once again left separate.
### 6.2. High Level Event Reconstruction

<table>
<thead>
<tr>
<th>ID</th>
<th>Reco. Step</th>
<th>VLB, 1200 GeV</th>
<th>VLB, 2000 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WBHb</td>
<td>ZBHb</td>
</tr>
<tr>
<td></td>
<td>Higgs Reco. Start</td>
<td>865.8</td>
<td>4476.6</td>
</tr>
<tr>
<td>1:0</td>
<td>&gt;0 1.0 Jet, ( p_T &gt; 480 ) GeV</td>
<td>738.2</td>
<td>3429.2</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \geq 2 ) TJ with ( p_T &gt; 50 ) GeV</td>
<td>699.5</td>
<td>3212.4</td>
</tr>
<tr>
<td>1:1</td>
<td>pass Collinear Veto</td>
<td>628.3</td>
<td>2884.9</td>
</tr>
<tr>
<td>1:2a</td>
<td>( \geq 1 ) ( b )-tagged TJ</td>
<td>202.2</td>
<td>1199.9</td>
</tr>
<tr>
<td>1:2b</td>
<td>( \geq 2 ) ( b )-tagged TJ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Efficiencies

<table>
<thead>
<tr>
<th>ID</th>
<th>Reco. Step</th>
<th>VLB, 1200 GeV</th>
<th>VLB, 2000 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:0</td>
<td>&gt;0 1.0 Jet, ( p_T &gt; 480 ) GeV</td>
<td>85.5%</td>
<td>76.6%</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \geq 2 ) TJ with ( p_T &gt; 50 ) GeV</td>
<td>94.8%</td>
<td>93.7%</td>
</tr>
<tr>
<td>1:1</td>
<td>pass Collinear Veto</td>
<td>89.8%</td>
<td>89.8%</td>
</tr>
<tr>
<td>1:2a</td>
<td>( \geq 1 ) ( b )-tagged TJ</td>
<td>28.9%</td>
<td>37.3%</td>
</tr>
</tbody>
</table>

**Table 6.11:** Intermediate Higgs Candidate reconstruction cutflow (top section) and efficiencies (bottom) for simulated 1200 GeV and 2000 GeV VLB signal events. Efficiencies are provided for events passing the Nominal Pre-Selection. Efficiencies pertaining to steps 1:2a and 1:2b are both calculated with respect to the events passing 1:1.
6.2.3 Higgs Candidate Categorization

The Higgs reconstruction procedure outlined above allows for more than a single candidate to be identified within an event. The output candidates are therefore sorted into several categories depending on the number of associated track jets and the overall $b$-tagging content. The naming convention for the Higgs categorization is presented in Table 6.12.

As only one candidate per event is eventually used in the VLB candidate reconstruction, the categories are ranked by the $b$-tagging content, with the defined priority order being: $2T2B > 2T1B > \text{REJECT}$. The Higgs Candidate belonging to the highest-ranking category is then selected for further use, while candidates with fewer $b$-tags, if present, are ignored. If more than one Higgs candidate is found in the highest-ranked non-empty category, the candidate with the higher mass is selected. This simple ordering criterion is imposed to solve the limited number of ambiguous cases without sculpting the background into possibly producing spurious peaks, which might have arisen had the search elected to choose the Higgs candidate with invariant mass closest to the actual Higgs boson mass as a tie-breaking criterion.

Table 6.13 summarizes, for simulated samples of each signal mass point with one eligible Higgs Candidate passing selection step 0:6, the fraction of events where the highest ranking Higgs Candidate belongs to the category considered. Out of the two relevant Higgs categories, $2T2B$ and $2T1B$, the latter category displays a larger reconstruction "efficiency", owing to both the mis-tagging probability and the fraction of mis-reconstructed Higgs decay systems (contributing to the low-mass tail visible

<table>
<thead>
<tr>
<th>2 Track Jet</th>
<th>1 $b$-tag</th>
<th>2 $b$-tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>2T1B</td>
<td>2T2B</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.12:** Naming convention for Higgs Candidates. The 2T part refers to the number of associated track jets, whereas nB represents the number of $b$-tags in the candidate.
in Fig. 6.4 for the H2T1B category). However, owing to the looser $b$-tagging multiplicity requirement, the 2T1B category is also expected to provide a much lower rejection of background events, leading to an overall lower expected sample purity and sensitivity.

Figure 6.3 provides a visualization of the interplay of the $b$-tagging score DL1 in reconstructed Higgs candidates belonging to the H2T2B and H2T1B categories for a 1200 GeV signal sample. In the H2T2B category, which requires a doubly $b$-tagged Higgs Candidate, values of DL1 on both axes range from 1.81, the 70% Efficiency Working Point on VR-TrackJets, upwards. Conversely, on the H2T1B category the ranges of the two axes extend on the opposite sides of the threshold 1.81. The composite origin of the signal events falling into the H2T1B category is made clear by the multiple occupancy maxima along the non-$b$-tagged jet DL1 axis: the compact cluster near DL1 = 1.81 houses legitimate $b$-originated track jets belonging to the 30% tagging inefficiency of the selected working point, while the looser and wider maximum at negative DL1 values likely arises from mis-reconstructed $b$-originated track jets or altogether spurious track jets associated to the Higgs Candidate.

Figures 6.4a-6.4b show the reconstructed Higgs mass distributions for candidates belonging to each category in the 1200 GeV and 2000 GeV VLB simulated sam-
6.2. High Level Event Reconstruction

Table 6.13: Fraction of signal events passing PreSelection (0:6) that fall in the $H_{2T2B}$ or $H_{2T1B}$ categories, or are rejected at the Higgs Candidate reconstruction stage.

<table>
<thead>
<tr>
<th>VLB, $M$ (GeV)</th>
<th>$2T2B$</th>
<th>$2T1B$</th>
<th>REJECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>24.5%</td>
<td>36%</td>
<td>39%</td>
</tr>
<tr>
<td>1200</td>
<td>26.5%</td>
<td>36.5%</td>
<td>36%</td>
</tr>
<tr>
<td>1400</td>
<td>27.5%</td>
<td>37%</td>
<td>34%</td>
</tr>
<tr>
<td>1600</td>
<td>28%</td>
<td>39%</td>
<td>34%</td>
</tr>
<tr>
<td>1800</td>
<td>27%</td>
<td>40%</td>
<td>33%</td>
</tr>
<tr>
<td>2000</td>
<td>27.5%</td>
<td>39%</td>
<td>33.5%</td>
</tr>
</tbody>
</table>

Table 6.13: Fraction of signal events passing PreSelection (0:6) that fall in the $H_{2T2B}$ or $H_{2T1B}$ categories, or are rejected at the Higgs Candidate reconstruction stage.

In either case, the channel with the highest $b$-tagging content, $2B2T$, is found to yield the best reconstructed Higgs mass line shape. The peaking structure visible in $W$-initiated $H_{2T1B}$ events around 170 GeV arises from an erroneous reconstruction of the fully hadronic decay system of the final state top quark produced in association with the VLB as the eventually selected Higgs Candidate (Sec. 5.1). The effect is much less pronounced in $H_{2T2B}$ events, whose double $b$-tagging requirement provides a better rejection of spurious top-induced Higgs Candidates, and is completely absent in the dominant, $Z$-initiated production, where a final state bottom quark is produced in association to the VLB instead of a top.

The Higgs reconstruction efficiency with respect to selection step 0:6 in Table 6.8 is shown in Fig. 6.5 (left), as well as the breakdown in various Higgs categories. There is little difference between the $Z$- and $W$-initiated samples. The plots on the right hand side of Fig. 6.5 provide an insight to the overall Higgs Candidate reconstruction efficiency on the full MC signal samples, which are inclusive in terms of the Higgs decay modes. The turn-on behaviour at low VLB masses, due to the kinematic selection criteria, is evident. The somewhat higher comprehensive efficiency observed in $W$-Initiated events with respect to $Z$-Initiated events can be traced back to the higher acceptance of the kinematic requirement on the large-R jet transverse momentum, enforced in the pre-selection (see step 0:1 in Table 6.9).
6.2. High Level Event Reconstruction

Figure 6.4: Reconstructed Higgs mass distribution in \(Z\)- (top row) and \(W\)-initiated samples (bottom row). Candidates are sorted by category. All histograms are normalized to unit area.

6.2.4 VLB Candidate Reconstruction

A VLB candidate is defined as the association of the selected Higgs Candidate and a small-\(R\) jet assumed to originate from the \(b\) quark produced in the immediate VLB decay. A minimum transverse momentum of 480 GeV is required of the small-\(R\) jet to be eligible for the VLB reconstruction, as well as a large geometric separation from the Higgs Candidate, \(\Delta R(j, HC) > 2.0\).

The \(\Delta R\) requirement renders totally unnecessary any overlap removal between the large-\(R\) and small-\(R\) jets reaching this stage of the selection. The fact that overlap
6.2. High Level Event Reconstruction

(a) Efficiency with respect to the previous selection step, Z-initiated production.  

(b) Efficiency with respect to the full signal sample, Z-initiated production.  

(c) Efficiency with respect to the previous selection step, W-initiated production.  

(d) Efficiency with respect to the full signal sample, W-initiated production.

Figure 6.5: Higgs Candidate reconstruction efficiency. In each sub-figure, the red, orange, yellow and green series represent the partial reconstruction efficiencies pertaining respectively to the $H_{2T2B}$ and $H_{2T1B}$ categories, with their combined efficiency displayed as the black series. Note that the VLB samples were generated inclusively in Higgs decays and this influences the proportion of events in the three categories.
removal between large-$R$ and small-$R$ jets is unnecessary and has no effect on the analysis is also evident in fig. 6.4, which shows that in events with a Higgs candidate, the Higgs candidate mass distribution peaks around the nominal Higgs mass, implying that the correct jet is selected as Higgs Candidate.

The small-$R$ jet momentum requirement is set to match the 480 GeV eligibility threshold of the Higgs reconstruction procedure, as the two objects are taken to reconstruct the two-body decay system of an extremely massive resonance. For a 1200 GeV $Z(W)$-initiated signal sample, the efficiency of this selection requirement in events with a Higgs Candidate already identified is approximately 80(77)%, increasing to 86(84)% for a 1400 GeV sample as the momentum distribution of the VLB decay products shifts towards higher values owing to the larger resonance mass$^1$.

Once again a tie-breaking criterion is necessary to handle the small number of events where more than one eligible VLB candidate exists: as the VLB is assumed to be a highly massive resonance generally produced in a low-boost environment on the plane transverse to the beam axis, the candidate with the lowest ratio $p_T(B)/m_B$ is selected. The selection criteria for the final VLB selection are summarised in Table 6.14.

<table>
<thead>
<tr>
<th>VLB Candidate</th>
<th>Selection</th>
<th>Working Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
<td>Higgs Candidate + R=0.4 Jet</td>
<td></td>
</tr>
<tr>
<td>Small-$R$ Jet $p_T$</td>
<td>$p_T &gt; 480$ GeV</td>
<td></td>
</tr>
<tr>
<td>Higgs Candidate $p_T$</td>
<td>$p_T &gt; 480$ GeV</td>
<td></td>
</tr>
<tr>
<td>Geometry</td>
<td>$\Delta R(j, HC) \geq 2$</td>
<td></td>
</tr>
<tr>
<td>Tie-breaker</td>
<td>$\min \frac{p_T(VLB)}{M(VLB)}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.14: Summary of reconstruction criteria for VLB Candidates.

Table 6.15 provides, for a select number of signal samples in either production modes, a numerical understanding of the comprehensive reconstruction efficiency

$^1$For more details refer to the event selection summary tables 6.26-6.27
given a number of reference points along the selection procedure. For a 1200 GeV VLB quark in the ZBHb production mode with \( \kappa = 1.0 \), for instance, a successful VLB candidate with a H2T2B Higgs candidate is reconstructed only in the 2.8\% of the entire dataset, with limiting factors including the \( H(b\bar{b}) \) branching ratio, the double \( b \)-tagging efficiency in the Higgs Candidate and most importantly the low acceptance of the selection for events falling into the conspicuous low-mass tail described and discussed in Sec. 5.2.3.

Considering instead the sub-set of pre-selected events (see Tab. 6.8) as the reference point, a Z-Produced 1200 GeV VLB candidate with a maximally tagged Higgs is found in 23.1\% of all cases. Finally, out of the events belonging to the same sample where a H2T2B Higgs candidate is already found, a full VLB candidate is reconstructed with efficiency above 85\%. The reconstruction efficiency, whatever the reference points, increases in the higher mass samples reflecting the relevance of the kinematic selection on the large-\( R \) and small-\( R \) jets.

### 6.2.5 Event Categorization

Events with at least one VLB candidate are sorted into categories of decreasing signal purity on the grounds of the number of \( b \)-tags and their distribution across the physical objects making up the VLB candidate. The naming convention for the resulting categories is exemplified in Table 6.16.

Table 6.17 shows how the events with a successfully reconstructed VLB candidate are split between the aforementioned categories.

In order to evaluate the shifting behaviour of the signal across the available VLB mass spectrum, the shapes of a selection of kinematic variables are examined for three benchmark mass points, as displayed in Figure 6.6. All shapes displayed here pertain to the coupling strength \( \kappa = 0.5 \).

The events featured in Fig. 6.6 are required to pass pre-selection, object selection and have at least one reconstructed VLB candidate as defined in Sec. 6.2.4.
6.2. High Level Event Reconstruction

<table>
<thead>
<tr>
<th>Mass Point</th>
<th>Full Sample</th>
<th>Pre-Selection</th>
<th>Higgs Reco.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200 GeV</td>
<td>2.2%</td>
<td>23.1%</td>
<td>86.3%</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>2.4%</td>
<td>24.9%</td>
<td>89.8%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>2.1%</td>
<td>24.7%</td>
<td>91.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mass Point</th>
<th>Full Sample</th>
<th>Pre-Selection</th>
<th>Higgs Reco.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200 GeV</td>
<td>2.8%</td>
<td>19.1%</td>
<td>82.0%</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>3.2%</td>
<td>21.3%</td>
<td>87.3%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>2.9%</td>
<td>21.6%</td>
<td>87.8%</td>
</tr>
</tbody>
</table>

Table 6.15: Reconstruction efficiency of the VLB candidate in \textbf{H2T2B} with respect to three selected reference points in the reconstruction protocol, for the \(Z\)-initiated signal. For the second and third columns, the combined efficiency of all previous steps is factored out from the efficiency figures displayed.

<table>
<thead>
<tr>
<th>HC Category</th>
<th>Small-(R) Jet</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{J0B_H2T2B}</td>
<td>\textbf{J1B_H2T2B}</td>
</tr>
<tr>
<td>\textbf{J0B_H2T1B}</td>
<td>\textbf{J1B_H2T1B}</td>
</tr>
</tbody>
</table>

Table 6.16: Naming convention for the event categories. The \textit{JnB} part refers to the number of \(b\)-tags in the small-\(R\) jet, whereas the following part of the name refers to the Higgs candidate reconstruction category.

For the sake of a clearer presentation, only events belonging to the event category \textbf{J2B\_H2T2B} are displayed.

The existence of two production modes for the signal ultimately results in a series of differences in the kinematic features of the event, as can be appreciated comparing the plots on the left (\(Z\)-initiated) and right-hand (\(W\)-initiated) side. This is particularly evident in variables such as the transverse momentum of the VLB candidate.
6.2. High Level Event Reconstruction

<table>
<thead>
<tr>
<th>Z-Initiated, $\kappa = 1.0$</th>
<th>(J_{1B_H2T2B})</th>
<th>(J_{0B_H2T2B})</th>
<th>(J_{1B_H2T1B})</th>
<th>(J_{0B_H2T1B})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200 GeV</td>
<td>23.1%</td>
<td>20.1%</td>
<td>27.1%</td>
<td>29.6%</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>20.6%</td>
<td>21.6%</td>
<td>25.5%</td>
<td>32.3%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>19.5%</td>
<td>21.7%</td>
<td>24.9%</td>
<td>33.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W-Initiated, $\kappa = 1.0$</th>
<th>(J_{1B_H2T2B})</th>
<th>(J_{0B_H2T2B})</th>
<th>(J_{1B_H2T1B})</th>
<th>(J_{0B_H2T1B})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200 GeV</td>
<td>16.7%</td>
<td>16.8%</td>
<td>28.9%</td>
<td>37.6%</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>16.0%</td>
<td>18.8%</td>
<td>26.8%</td>
<td>38.3%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>15.8%</td>
<td>18.9%</td>
<td>25.5%</td>
<td>39.3%</td>
</tr>
</tbody>
</table>

Table 6.17: Fractional yields for simulated signal events sorted by event category. Percentage refer to the number of events with at least one reconstructed VLB candidate.

and the $\Delta R$ between the Higgs candidate and the small-$R$ jet, which both display a narrower shape in the Z-initiated production.
Figure 6.6: Normalized distributions of selected kinematic variables for simulated signal events belonging to the 1200 GeV, 1600 GeV and 2000 GeV samples. All the events displayed belong to the $J1B_H2T2B$ category and feature at least one reconstructed VLB candidate.
Figure 6.6: Normalized distributions of selected kinematic variables for simulated signal events belonging to the 1200 GeV, 1600 GeV and 2000 GeV samples. All the events displayed belong to the J1B_H2T2B category and feature at least one reconstructed VLB candidate.
Figure 6.6: Normalized distributions of selected kinematic variables for simulated signal events belonging to the 1200 GeV, 1600 GeV and 2000 GeV samples. All the events displayed belong to the $J1B_{-H2T2B}$ category and feature at least one reconstructed VLB candidate.
6.3 Signal Region Definition

The reconstruction protocol leading to the per-event identification of a VLB candidate, as described in the previous sections, operates a relatively sharp selection of events on the grounds of their compatibility with the predicted features of the $B \to bH(b\bar{b})$ final state, with events not compatible with the presence of a VLB candidate being rejected.

Given that the sensitivity of a search is primarily driven by the possibility of obtaining the maximum rejection of background events while keeping an acceptable selection efficiency on a possible signal, it is possible to further act on the properties of the VLB candidate to define a volume where the sensitivity is maximal, typically referred to as Signal Region.

The criteria leading up to the definition of a Signal Region for the present search are conceived with the purpose of reflecting the two capital features of the VLB decay cascade: the presence of three bottom quarks in the final state, and the fact that two of them originate from a Higgs boson decay. This section will describe how an optimal sensitivity region was defined out of the features of the reconstructed event that mirror the aforementioned final state properties: the number and configuration of $b$-tags in the VLB candidate and the invariant mass of the Higgs Candidate.

6.3.1 Event Category Optimization

The naming system introduced in Sec. 6.2.5 provides a natural gateway into investigating the inherent sensitivity of the various event categories differing by their $b$-tagging multiplicity. Ultimately, the search will be restricted to run on events belonging to the most sensitive category only, while the remaining events will be either discarded or employed as an orthogonal data sample for validation purposes.
To evaluate the sensitivity of each category, it is first necessary to estimate the relative number of expected signal and background events. Given that no Monte Carlo simulation of background processes is used in the analysis, it is necessary to rely on the data itself to estimate the relative background contributions. While data-driven estimations eschew all modelling issues originating from incorrect or incomplete description of the process or the detector, they require additional level of sophistication to avoid the introduction of a bias in the background prediction should a signal actually occur.

For this reason, a preliminary blinding\(^2\) is imposed on data fitting a tentative \textit{a priori} definition of signal region, corresponding to the most restricting \(b\)-tagging category \textbf{J1B\textunderscore H2T2B} and a Higgs mass window cut ranging from 105 GeV and 135 GeV. The aforementioned preliminary criteria function simply as an \textit{ansatz} to kickstart the optimization procedure, and are meant to be ultimately amended according to the eventual findings.

As further insurance against the possible signal-induced bias, only a fraction of the total event pool, corresponding to data collected in 2016 only, is used within this study.

Table 6.18 displays the per category data yields for the full 2016 dataset, with the yield for the \textbf{J1B\textunderscore H2T2B} category in the Higgs peak region blinded as previously stated. The ratio between the event yields of two categories sharing the same Higgs type and only differing in the tagging status of the small-\(R\) jet (e.g. \textbf{J1B\textunderscore H2T1B})

<table>
<thead>
<tr>
<th></th>
<th>J1B\textunderscore H2T2B</th>
<th>J0B\textunderscore H2T2B</th>
<th>J1B\textunderscore H2T1B</th>
<th>J0B\textunderscore H2T1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higgs Peak</td>
<td>(\lesssim)</td>
<td>14428</td>
<td>5438</td>
<td>133529</td>
</tr>
<tr>
<td>Off-Peak</td>
<td>2905</td>
<td>78195</td>
<td>31599</td>
<td>736120</td>
</tr>
</tbody>
</table>

\textbf{Table 6.18:} Event yields for 2016 data events with at least one VLB candidate. The Higgs Peak is preliminarily defined as \(M_{HC} \in [105, 135]\) GeV, with the off-peak ranging instead from 135 GeV to 200 GeV.

\(^2\)A set of events is said to have been \textit{blinded} by the search if temporarily removed from any possibility of user inspection. This operation is most commonly performed on signal-enriched datasets in the early stages of an analysis, in order to prevent any possible bias in the structuring of the search generated by the foreknowledge of the behaviour of the actual signal-sensitive data.
6.3. Signal Region Definition

71B H2T1B), referred to as $\mu_{QCD}$, can be used to estimate the predicted data yield in the blinded area, and is computed in both the Higgs mass window (for the H2T1B category only) and outside (for both H2T1B and H2T2B), as displayed in Table 6.19. Observing that the values for $\mu$ in each Higgs category differ by less than 10\% when evaluated in either the Higgs peak or off-peak regions, $\mu_{H2T2B}^{\text{off}}$ can be used to approximate the data yield in the blinded region as follows:

$$N_{J1B,\text{H2T2B}}^{\text{peak}} \simeq N_{J1B,\text{H2T2B}}^{\text{peak}} = \mu_{H2T2B}^{\text{off}} \times N_{J0,\text{H2T2B}}^{\text{peak}} = \frac{N_{J1,\text{H2T2B}}^{\text{off}}}{N_{J0,\text{H2T2B}}^{\text{off}}} \times N_{J0,\text{H2T2B}}^{\text{peak}} \quad (6.2)$$

assuming that: $\mu_{H2T2B}^{\text{peak}} \simeq \mu_{H2T2B}^{\text{off}} \quad (6.3)$

At this point, the relative sensitivity of an event category with respect to the most restrictive category (J1B_H2T2B) is evaluated via the $z = \frac{s}{\sqrt{b}}$ significance estimator, with $s$ and $b$ being the signal and background yields for the category under scrutiny. Therefore, using J1B_H2T2B as reference:

$$\frac{z}{z_0} = \frac{s}{s_0} \times \sqrt{\frac{b_0}{b}} = \frac{s_{\text{cat}}}{s_0} \times \sqrt{\frac{N_{J1B,\text{H2T2B}}^{\text{peak}}}{\tilde{N}_{\text{cat}}^{\text{peak}}}} \quad (6.4)$$

It is worth noticing that in the $z/z_0$ ratio, contrary to the absolute significance $z$, the luminosity factor cancels out, allowing the estimate to stand as the overall data volume increases by inserting the 2017 and 2018 datasets into the analysis.

The per-category signal yield ratios can be computed from the fractional yields shown in Table 6.17. Table 6.20 displays the values of $z/z_0$ evaluated for two reference signal mass points for all event categories.

<table>
<thead>
<tr>
<th></th>
<th>H2T2B</th>
<th>H2T1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{\text{peak}}$</td>
<td>\hspace{1cm}</td>
<td>\hspace{1cm}</td>
</tr>
<tr>
<td>$\mu_{\text{off}}$</td>
<td>0.037</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Table 6.19: Values of $\mu_{QCD}$ for each Higgs Category, evaluated on 2016 data.
None of the remaining event categories are found to yield an expected sensitivity comparable to that of $J1B_H2T2B$. Table 6.21 shows an estimate of the improvement in signal significance achievable by combining each event category individually with $J1B_H2T2B$ at the statistical fit stage. The estimator used assumes a complete absence of correlation between the individual channels under consideration, and is defined as:

$$
\frac{z_+}{z_0} = \frac{1}{z_0} \sqrt{\frac{z_0^2}{z_0^2 + z_i^2}} = \sqrt{1 + \left(\frac{z_i}{z_0}\right)^2}
$$

(6.5)

The values of $z_+/z_0$ show that even combining the second most sensitive category, $J1B_H2T1B$, with $J1B_H2T2B$ events would yield at most a 5% improvement in the sensitivity.

As no further significant gain in sensitivity would be achieved with further combination of lower purity categories, only $H2T2B$ and $H2T1B$ events will be taken into account in the remainder of the search, albeit with different roles. The $J1B_H2T2B$, with by far the best estimated sensitivity, will serve as the basis of the Signal Region definition, while the low sensitivity, but kinematically similar $H2T1B$ will serve as a testing and validation sample for the background modelling and the statistical setup (see Sec 7.4.4).
6.3. Signal Region Definition

### Table 6.21: Significance of signal over the 2016 Data combining each category with the reference J1B_H2T2B category. Values are relative to the reference significance.

<table>
<thead>
<tr>
<th></th>
<th>J0B_H2T2B</th>
<th>J1B_H2T1B</th>
<th>J0B_H2T1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_+/z_0$ (1200 GeV)</td>
<td>1.01</td>
<td>1.04</td>
<td>1.001</td>
</tr>
<tr>
<td>$z_+/z_0$ (2000 GeV)</td>
<td>1.02</td>
<td>1.04</td>
<td>1.002</td>
</tr>
</tbody>
</table>

6.3.2 Higgs Mass Window Optimization

The optimal boundaries of the Higgs mass Peak region are identified by evaluating the signal significance $z = s/\sqrt{b}$ over a lattice of possible choices of region limits $M_{LOW}$ and $M_{HIGH}$, where $s$ and $b$ represent the number of signal and data events falling in the $M_{HC} \in [M_{LOW}, M_{HIGH}]$ interval. Only events with a successfully reconstructed VLB candidate are considered for the present study, belonging to the signal-rich J1B_H2T2B category for simulated signal events and to the signal-depleted but kinematically close J0B_H2T2B category for the data-driven approximation of the background.

As the purpose of this study is to identify the optimal values of $M_{LOW}$ and $M_{HIGH}$, rather than providing absolute and realistic values of the signal significance, the calculated value of $z$ for each combination is displayed in rational form relative to the value of $z$ in the highest sensitivity configuration ($z_{max}$).

Figure 6.7 shows the behaviour of $z/z_{max}$ for all the considered boundary combinations. W-initiated and Z-initiated production are considered separately. A summary of the optimal Peak region limits for each production mode and mass point is provided in Table 6.22.

Since the significance heat maps computed for each considered mass sample show sufficiently broad maxima, the choice of a fixed range for Higgs mass peak region does not compromise the sensitivity of the search in any region of the resonance mass phase space. Ultimately, the original definition for a Higgs Peak region was
6.3. Signal Region Definition

Figure 6.7: Signal significance for different Higgs mass Peak Region boundary configurations. The optimal sensitivity point is marked by the superimposed star. Z-initiated (left) and W-initiated (right) samples are treated independently.

\[ M_{HC} \in [105, 135] \text{ GeV} \]  \hspace{1cm} (6.6)
6.4 Post Reconstruction Kinematic Selection

<table>
<thead>
<tr>
<th></th>
<th>Z-initiated</th>
<th></th>
<th>W-initiated</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>LOW</td>
<td>HIGH</td>
<td>LOW</td>
<td>HIGH</td>
</tr>
<tr>
<td>M = 1000 GeV</td>
<td>111</td>
<td>137</td>
<td>109</td>
<td>139</td>
</tr>
<tr>
<td>M = 1200 GeV</td>
<td>107</td>
<td>137</td>
<td>109</td>
<td>137</td>
</tr>
<tr>
<td>M = 1400 GeV</td>
<td>107</td>
<td>135</td>
<td>109</td>
<td>139</td>
</tr>
<tr>
<td>M = 1600 GeV</td>
<td>109</td>
<td>133</td>
<td>113</td>
<td>135</td>
</tr>
<tr>
<td>M = 1800 GeV</td>
<td>109</td>
<td>141</td>
<td>113</td>
<td>133</td>
</tr>
<tr>
<td>M = 2000 GeV</td>
<td>113</td>
<td>135</td>
<td>105</td>
<td>135</td>
</tr>
</tbody>
</table>

Table 6.22: Optimal Higgs Peak region boundaries. All figures expressed in units of GeV.

6.4 Post Reconstruction Kinematic Selection

On top of the definition of a Signal Region, the sensitivity of the search is further enhanced through the application of a number of selection cuts on those event features showing the greatest discrepancy between the signal, as predicted by the available Monte Carlo simulated samples, and the background.

For the purpose of the event selection optimization, the behaviour of the blinded J1B_H2T2B data in the Higgs peak region is once again assumed to be reasonably compatible with that of the J0B_H2T2B sample in the same mass region. Under this assumption, the features of the background can be studied in the signal poor category as a proxy, providing the necessary, if approximate, information for the optimization of the various selection steps. Pre-background modelling comparison studies between the J1B_H2T2B and J0B_H2T2B samples outside the Higgs peak region show this assumption to be reasonable.

3For further reference, compare left and right-hand side histograms in Figure 7.3.
6.4.1 Optimization Method

For every available resonance mass hypothesis, the distribution of the physical variables under study are examined in the simulated signal and approximated background samples in Signal Region events belonging to a symmetric 150 GeV-wide window centered on the pole resonance mass considered. This last requirement is enforced to highlight any mass-dependent behaviour in data that may otherwise be rendered invisible by the statistical prominence of low-mass events expected from the non-resonant processes dominating the background.

For each variable, several possible cut thresholds are tested, evaluating for each the ratio of the signal significance $z$ before and after the cut:

$$\frac{z}{z_0} = \frac{s}{s_0} \times \sqrt{\frac{b_0}{b}}$$  \hspace{1cm} (6.7)

6.4.2 Forward Hadronic Activity

As introduced in Sec. 5.1 and displayed in Fig. 5.1, two more quarks beyond those collected in the VLB candidate are produced in the final state of the hard-scatter event, one recoiling off the $Z$ (or $W$) boson involved in the $B$ production, the other arising from the gluon splitting responsible for the production of the Standard Model $b$ quark (or $t$ quark) entering the $bZB$ (or $tWB$) production vertex.

Owing to the VBF-like process configuration, these quarks are expected to be generally produced in the forward region of the detector, where they can be reconstructed as jets according to the specific criteria outlined in Sec. 6.1.1.

An excess of relatively hard ($p_T > 40$ GeV) jets in the $|\eta| \in [2.5, 4.5]$ area over the average in data should therefore be visible for signal events. Fig. 6.8 confirm the expectations, displaying the forward jet multiplicity for the 1200 GeV (Fig. 6.8a) and 2000 GeV (Fig. 6.8b) samples. All events shown satisfy the VLB peak crite-
Figure 6.8: Forward jet multiplicity in signal and background data samples. Background behaviour approximated by the low-signal purity J0B_H2T2B sample. Signal and background are normalized to a common area for easier viewing. All events satisfy the VLB peak condition: $|m - M_{B}^{\text{pole}}| < 150 \text{ GeV}$, with $M_{B}^{\text{pole}}$ being the nominal VLB sample mass. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

Another possible point of discrimination between signal and background is the internal structure of the large-R jet selected as the Higgs Candidate, which originates from a legitimate Higgs boson in the signal hypothesis and from QCD activity, likely $g \rightarrow b\bar{b}$ splittings, in the dominant Standard Model background source. A level of discrepancy is therefore to be expected between the two event types even among large-$R$ jets with two associated $b$-tagged track jets.
Figure 6.9: Increase in signal significance by requiring the event to have a forward jet multiplicity at least equal to the value given by the $x$-coordinate. Multiplicity bins are exclusive, meaning that events falling into a specific bin have exactly the number of forward jets as displayed in the bin label.

Figure 6.10 shows the distribution of the structure variable $\log \Delta R^*$, defined as:

$$
\log \Delta R^* = \log \left[ \frac{\Delta R(H_{j_0}, H_{j_1})}{\min [R_{H_{j_0}}, R_{H_{j_1}}]} \right]
$$

(6.8)

where $H_{j_0}$ and $H_{j_1}$ are the two track jets associated to the Higgs Candidate and $R_{H_{j_0}}$, $R_{H_{j_1}}$ their $p_T$-dependent radii (see Sec. 6.1.3). Using $\log \Delta R^*$ over the simple $\Delta R(H_{j_0}, H_{j_1})$ allows for a more comprehensive picture of the energy distribution within the Higgs Candidate, as the radius parameters of the track jets incorporate information about the momenta of their constituents.

As the distributions in Fig. 6.10 show a considerably different behaviour in signal and background, a tentative selection step was established and examined. Figure 6.11 shows, for each presumptive cut threshold represented by the value of the $x$-coordinate, the relative variation in signal significance obtained by requiring events to have a greater value of $\log \Delta R^*$ than the relevant threshold.

Table 6.23 shows the cut values yielding the highest significance enhancement for each available signal sample as well as an approximate measure of the resulting rel-
6.4. Post Reconstruction Kinematic Selection

Figure 6.10: \( \log \Delta R^* \) in signal and background data samples. Background behaviour approximated by the low-signal purity J0B_H2T2B sample. Signal and background are normalized to a common value for easier viewing. All events satisfy the VLB peak condition: \( |m - M_B^{\text{pole}}| < 150 \text{ GeV} \), with \( M_B^{\text{pole}} \) being the nominal VLB sample mass. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

Implementing a selection on the value of \( \log \Delta R^* \) therefore results in a consistent, if

<table>
<thead>
<tr>
<th>Mass Point</th>
<th>Optimal Cut Threshold</th>
<th>((z - z_0)/z_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 GeV</td>
<td>0.7</td>
<td>7.5%</td>
</tr>
<tr>
<td>1200 GeV</td>
<td>0.6</td>
<td>7.5%</td>
</tr>
<tr>
<td>1400 GeV</td>
<td>0.7</td>
<td>7%</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>0.5</td>
<td>4%</td>
</tr>
<tr>
<td>1800 GeV</td>
<td>0.4</td>
<td>1%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>0.6</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 6.23: Optimal cut threshold and signal significance gain for a \( \log \Delta R^* > (\log \Delta R^*)_{\text{threshold}} \) selection step.

limited, increase in the overall significance of a VLB-like signal over the data. Although the optimal cut threshold is shown to drift towards lower values in the higher end of the mass spectrum, the selection working point is fixed to 0.67. This choice
privileges the lower end of the spectrum, where the significance gain is maximal, and does not compromise the effectiveness of the selection at higher masses, where the significance gain is not as sensitive to small variations in the working point (see Fig. 6.11b).

6.4.4 \( s/t\)-channel kinematics

Another variable found to be sensitive to the signal hypothesis is the Lorentz-invariant pseudorapidity gap between the Higgs candidate and the small-\(R\) jet, the two objects taken to reconstruct the primary VLB decay products:

\[ |\Delta \eta| = |\eta_{HC} - \eta_{jet}| \]

Figure 6.12 shows that while signal events consistently favour small values for \(\Delta \eta\), the behaviour in data is highly sensitive to the region of the reconstructed \(m_B\) considered, increasingly favouring larger values at the higher end of the range studied.

Figure 6.13 shows the significance variation for a tentative cut on \(\Delta \eta\) confirming its effectiveness in the high-mass regime.

It is worth mentioning that this seemingly peculiar behaviour in the data is under-
stood to originate from the shifting contributions of the two competing diagrams for the QCD-driven production of multiple jets in the final state, as displayed in Figure 6.14. In the regime where the invariant mass of the $3b$ quarks (the proxy variable hypothetical VLB resonance mass) is of the order of 1 TeV, neither diagram dominates at the center-of-mass energy of the LHC. At the highest end of the investigated resonance mass range, on the other hand, the $s$-channel cross section, which scales with the inverse square of the invariant mass of the final state, be-
6.4. Post Reconstruction Kinematic Selection

Figure 6.14: Two possible tree level diagrams regulating the QCD-driven production of 3 $b$ quarks in the final state. In the left-hand side diagram the momentum exchange occurs via the $s$-channel, whereas the diagram on the right shows a typical $t$-channel process.

comes strongly sub-leading, as the $t$-channel propagator scales with the momentum difference between the initial and final state particles:

$$\sigma_S(3b) \sim \frac{1}{m_{3b}^2} \ll \sigma_T(3b) \sim \frac{1}{(p_{\text{out}} - p_{\text{in}})^2} \quad \text{at } m_{3b} \sim 2 \text{ TeV} \quad (6.9)$$

From the qualitative scaling relations introduced in Eqn. 6.9 it can be gathered that $t$-channel production favours a configuration where the two final state systems (the standalone $b$ quark and the $b\bar{b}$ pair) are preferentially produced with medium-to-large pseudorapidity $|\eta|$, a regime where the critical $(p_{\text{out}} - p_{\text{in}})^2$ quantity is relatively small. This in turn results in a larger $|\Delta \eta|$ in high-invariant mass background events than in the signal, as the latter is always produced through a $s$-channel process (see Fig.5.1).

Taking into consideration a limited range of values of $m_B$, it can therefore be assumed that in data events, owing to the larger average values of $\eta_{HC}$ and $\eta_{\text{jet}}$, the Higgs candidate and the small-$R$ jet will have a lower $p_T$ than in signal events in the corresponding $m_B$ mass range.

Figure 6.15 show distributions of the $p_T(\text{HC})/m_B$ ratio in signal and background (where $p_T(\text{HC})$ is the Higgs candidate transverse momentum), while fig. 6.16 displays the significance enhancement achieved by selecting events with values of the ratio greater than the reference value. As $|\Delta \eta|$ and $p_T(\text{HC})/m_B$ are strongly cor-
Figure 6.15: Signal and data distributions of the $p_T(HC)/m_B$ variable. Arbitrary units. The shaded error bars on each histogram refer to the statistical uncertainty only for data and the Monte Carlo statistical uncertainty for the signal.

Figure 6.16: Increase in signal significance by requiring the event to have $p_T(HC)/m_B > (p_T(HC)/m_B)_0$

related, it is reasonable to include only one in the event selection, choosing the one yielding the greater signal prominence. Table 6.24 allows a comparison of the effect of selecting on either variable for each simulated samples, along with the optimal working points. The tentative cut on $p_T(HC)/m_B$ is observed to produce a greater expected effect on $z/z_0$ with a more stable working point, and is therefore favoured over the pseudorapidity gap $|\Delta \eta|$. The cut threshold is set at $p_T(HC)/m_B = 0.4$. 
### 6.4. Post Reconstruction Kinematic Selection

#### Table 6.24: Comparison of the optimal cut thresholds and signal significance gains for a cut-based selection steps on $|\Delta \eta|$ and $p_T(HC)/M_{VLB}$.

<table>
<thead>
<tr>
<th>Mass Point</th>
<th>Optimal W.P.</th>
<th>$(z - z_0)/z_0$</th>
<th>Optimal W.P.</th>
<th>$(z - z_0)/z_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 GeV</td>
<td>1.33</td>
<td>0 %</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>1200 GeV</td>
<td>1.17</td>
<td>2 %</td>
<td>0.43</td>
<td>2 %</td>
</tr>
<tr>
<td>1400 GeV</td>
<td>1.17</td>
<td>3 %</td>
<td>0.43</td>
<td>5 %</td>
</tr>
<tr>
<td>1600 GeV</td>
<td>1.17</td>
<td>10 %</td>
<td>0.40</td>
<td>10 %</td>
</tr>
<tr>
<td>1800 GeV</td>
<td>1.50</td>
<td>12%</td>
<td>0.40</td>
<td>18%</td>
</tr>
<tr>
<td>2000 GeV</td>
<td>1.83</td>
<td>15 %</td>
<td>0.37</td>
<td>30%</td>
</tr>
</tbody>
</table>
Table 6.25, 6.26 and 6.27 offer a comprehensive picture of the data and simulated signal yield throughout the various reconstruction and selection steps. Fully selected Higgs Peak region data belonging to the $H2T2B$ category (the definition of Signal Region) is left blinded. Given the event categorization (Sec. 6.2.5) determined on the grounds of the $b$-tag multiplicity within the Higgs Candidate, yields are provided separately for the $H2T2B$ and $H2T1B$ categories.
### Table 6.25: Event yields for data throughout the reconstruction and event selection stages. The yields marked with a diamond (●) are partially blinded and refer to events outside the Higgs Peak Region only.
### Table 6.26: Event yields for Z-initiated simulated signal throughout the reconstruction and event selection stages. The yields for the first two steps of the selection represent the number of unweighted events passing each respective steps.

<table>
<thead>
<tr>
<th>SIGNAL, Z-initiated, $\kappa = 1.0$</th>
<th>1000 GeV</th>
<th>1200 GeV</th>
<th>1400 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Events in sample</td>
<td>199000</td>
<td>200000</td>
<td>200000</td>
</tr>
<tr>
<td>1 $R=1.0$ $p_T &gt; 480$ GeV</td>
<td>31513</td>
<td>42586</td>
<td>42755</td>
</tr>
<tr>
<td>0 Leptons</td>
<td>29984</td>
<td>40544</td>
<td>40671</td>
</tr>
<tr>
<td>0 $\gamma\gamma$ pairs</td>
<td>29984</td>
<td>40544</td>
<td>40671</td>
</tr>
<tr>
<td>$\geq 2$ TrackJets in $R=1.0$ Jet, $\geq 1$ b-tagged</td>
<td>15698</td>
<td>23103</td>
<td>23934</td>
</tr>
<tr>
<td>$\geq 0$ $R=0.4$ jet, $p_T &gt; 300$ GeV</td>
<td>14779</td>
<td>22151</td>
<td>23224</td>
</tr>
<tr>
<td>$\Delta R(0.4$ Jet, $1.0$ Jet) $&gt; 2$</td>
<td>11756</td>
<td>18954</td>
<td>20437</td>
</tr>
<tr>
<td>Weighted</td>
<td>4401.73</td>
<td>4476.62</td>
<td>3250.57</td>
</tr>
<tr>
<td>$\geq 0$ Higgs Candidates</td>
<td>2659.47</td>
<td>2833.520</td>
<td>2125.730</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>H2T2B</th>
<th>H2T1B</th>
<th>H2T2B</th>
<th>H2T1B</th>
<th>H2T2B</th>
<th>H2T1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt; 0$ H.C. In Category</td>
<td>1073.38</td>
<td>1586.09</td>
<td>1199.87</td>
<td>1633.64</td>
<td>893.19</td>
<td>1232.54</td>
</tr>
<tr>
<td>$&gt; 1$ H.C. In Category</td>
<td>16.98</td>
<td>125.04</td>
<td>15.29</td>
<td>144.92</td>
<td>14.68</td>
<td>126.67</td>
</tr>
<tr>
<td>$&gt; 0$ VLB Candidate</td>
<td>856.36</td>
<td>1185.26</td>
<td>1036.44</td>
<td>1360.44</td>
<td>790.58</td>
<td>1069.19</td>
</tr>
<tr>
<td>J1B</td>
<td>468.72</td>
<td>584.88</td>
<td>552.55</td>
<td>650.04</td>
<td>416.67</td>
<td>481.73</td>
</tr>
<tr>
<td>$M_{H.C.} \in [105, 200]$ GeV</td>
<td>295.15</td>
<td>300.21</td>
<td>364.28</td>
<td>381.73</td>
<td>272.60</td>
<td>324.39</td>
</tr>
<tr>
<td>$\log \Delta R^\star &gt; 0.67$</td>
<td>264.95</td>
<td>229.54</td>
<td>316.91</td>
<td>288.03</td>
<td>236.35</td>
<td>246.92</td>
</tr>
<tr>
<td>$p_G(H.C.)/m_B &gt; 0.4$</td>
<td>245.98</td>
<td>212.55</td>
<td>287.70</td>
<td>259.52</td>
<td>207.35</td>
<td>210.97</td>
</tr>
<tr>
<td>Triggered</td>
<td>245.98</td>
<td>212.55</td>
<td>287.70</td>
<td>259.52</td>
<td>207.35</td>
<td>210.97</td>
</tr>
<tr>
<td>$&gt; 0$ Forward Jets</td>
<td>177.82</td>
<td>157.79</td>
<td>214.23</td>
<td>156.23</td>
<td>146.39</td>
<td>123.57</td>
</tr>
<tr>
<td>Higgs Peak, $M_{HC} \in [105, 135]$ GeV</td>
<td>145.17</td>
<td>115.93</td>
<td>180.32</td>
<td>108.67</td>
<td>123.93</td>
<td>92.12</td>
</tr>
</tbody>
</table>
### Table 6.27: Event yields for W-initiated simulated signal throughout the reconstruction and event selection stages. The yields for the first two steps of the selection represent the number of unweighted events passing each respective steps.

<table>
<thead>
<tr>
<th>SIGNAL, W-initiated, ( \kappa = 1.0 )</th>
<th>1000 GeV</th>
<th>1200 GeV</th>
<th>1400 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Events in sample</td>
<td>200000</td>
<td>200000</td>
<td>200000</td>
</tr>
<tr>
<td>1 R=1.0 ( p_T &gt; 480 ) GeV</td>
<td>64554</td>
<td>74575</td>
<td>74081</td>
</tr>
<tr>
<td>0 Leptons</td>
<td>52804</td>
<td>60943</td>
<td>60492</td>
</tr>
<tr>
<td>0 ( \gamma \gamma ) pairs</td>
<td>52804</td>
<td>60943</td>
<td>60492</td>
</tr>
<tr>
<td>( \geq 2 ) TrackJets in R=1.0 Jet, ( \geq 1 ) b-tagged</td>
<td>31335</td>
<td>37815</td>
<td>38493</td>
</tr>
<tr>
<td>( \geq 0 ) R=0.4 jet, ( p_T &gt; 300 ) GeV</td>
<td>29151</td>
<td>35826</td>
<td>36721</td>
</tr>
<tr>
<td>( \Delta R(0.4 \text{ Jet}, 1.0 \text{ Jet}) &gt; 2 )</td>
<td>22100</td>
<td>29470</td>
<td>30983</td>
</tr>
<tr>
<td>Weighted</td>
<td>1040.50</td>
<td>865.77</td>
<td>605.35</td>
</tr>
<tr>
<td>( \geq 0 ) Higgs Candidates</td>
<td>712.82</td>
<td>602.380</td>
<td>422.240</td>
</tr>
<tr>
<td>( H^2 T^2 B )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H^2 T^1 B )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%2T2B</td>
<td>228.78</td>
<td>484.04</td>
<td>202.20</td>
</tr>
<tr>
<td>%2T1B</td>
<td>5.39</td>
<td>69.02</td>
<td>4.98</td>
</tr>
<tr>
<td>%0 VLB Candidate</td>
<td>168.14</td>
<td>372.57</td>
<td>166.42</td>
</tr>
<tr>
<td>%J1B</td>
<td>81.52</td>
<td>162.11</td>
<td>82.65</td>
</tr>
<tr>
<td>%H.C. ( M_{H.C.} \in [105, 200] ) GeV</td>
<td>52.67</td>
<td>110.65</td>
<td>53.90</td>
</tr>
<tr>
<td>%log( \Delta R^* &gt; 0.67 )</td>
<td>46.08</td>
<td>88.14</td>
<td>47.20</td>
</tr>
<tr>
<td>%( p_G(H.C.)/m_B &gt; 0.4 )</td>
<td>40.74</td>
<td>74.01</td>
<td>41.15</td>
</tr>
<tr>
<td>Triggered</td>
<td>40.74</td>
<td>74.01</td>
<td>41.15</td>
</tr>
<tr>
<td>%&gt; 0 Forward Jets</td>
<td>25.59</td>
<td>35.56</td>
<td>25.90</td>
</tr>
<tr>
<td>%Higgs Peak, ( M_{HC} \in [105, 135] ) GeV</td>
<td>18.31</td>
<td>14.28</td>
<td>18.18</td>
</tr>
</tbody>
</table>
Chapter 7

Background Modelling

Establishing an algorithmic procedure (discussed in chapter 6) capable of singling out data events sharing a specific set of attributes with the predicted experimental signature of a signal is a necessary but by no means sufficient step along the way to produce a statistically robust evaluation of the signal hypothesis given the available set of data.

For any such statistical inference to be meaningful, the procedure of selecting data events as belonging to a Signal Region must be complemented by an estimate of the number of event generated by established sources of Standard Model background that fit those same criteria. Only then the analyzer, by comparing the Signal Region data with the relative background prediction, can make any statistically substantiated statement on the signal hypothesis.

This chapter details the definition, implementation and validation of a data-driven background modelling procedure based on the common ABCD scaling principle for two uncorrelated variables.

7.0.1 Data-Driven Modelling Fundamentals

In the specific context of the present search, the event topology criteria enforced in both the pre-selection and the high-level reconstruction of a VLB candidate object (see Sec. 6.2.1, 6.2), which require any selected event to feature two or more highly
energetic hadronic objects and no leptons in the final state, effectively restrict the sources of background relevant to this search to mostly the QCD-driven production of multiple jets in the final state.

As the current theoretical understanding of such processes is not sufficient to generate reliable Monte Carlo simulations, the search described in this thesis established an entirely data-driven procedure to predict the expected shape and yield of the background in the blinded Signal Region.

The founding principle of all data-driven techniques resides, as the name itself suggests, in producing an estimate for the behaviour of the sources behind the generation of data events in a specific target sub-region of the phase space (such as a blinded Signal Region) relying on the information provided by one or many orthogonal sub-sets of the data itself. For the purpose of obtaining a meaningful and unbiased prediction of the features of the target sample, the data sub-sample(s) employed as the support for the modelling procedure should be as close as possible to the target sample in terms of event kinematics and topology, while at the same time minimizing the level of contamination from a possible signal occurring chiefly in the target sample. These two cardinal conditions are typically antagonistic, and identifying an optimal trade-off is often necessary.

### 7.1 ABCD Regions Definition

Among the existing data-driven estimation techniques, the ABCD method holds a popular position due to its straightforward concept and implementation. The core principle of the ABCD method is that, given any four-way partition of the phase space defined by the values of two uncorrelated variables, the number of events in any of the four regions can be estimated from the yields in the other three regions. In its most basic setup, exemplified by Fig.7.1, the number of events in the assumed
target region A can be estimated as:

\[ N_A = N_B \times \frac{N_C}{N_D} \]  

(7.1)

The ABCD estimation can either be performed globally, therefore providing solely an estimation for the event yield in the target region, or alternatively be applied to every bin of a histogram template of a certain variable of interest, in which case Eqn. 7.1 is more correctly expressed as:

\[ n_i^A = n_i^B \times \frac{n_i^C}{n_i^D} \]  

(7.2)

with the \( i \)-subscript representing the \( i \)-th bin of the considered distribution. In either case, the lack of any sizable correlation between the two variables generating the ABCD partition is essential to ensure that the background estimate is free from any bias resulting from \( \frac{n_i^C}{n_i^D} \neq \frac{n_i^A}{n_i^B} \).

As detailed in Sec. 6.3, the search is carried out in a sub-sample of the dataset defined, on top of the kinematic selections steps defined in Sec. 6.4, by a Higgs mass window \( (M_H \in [105, 135] \text{ GeV}) \) and an event category requirement \((J1B\_H2T2B)\). Within the context of the ABCD-estimation framework, this blinded region of ultimate sensitivity shall be henceforth referred to as ABCD-Region A.

Having identified the ABCD-Region A with the blinded Signal Region, the three
7.1. ABCD Regions Definition

Table 7.1: Schematic representation of the definition of the four ABCD Regions.

<table>
<thead>
<tr>
<th>Regions</th>
<th>Selection</th>
<th>b-tag, R=0.4 jet</th>
<th>≥ 1 fwd jets</th>
</tr>
</thead>
<tbody>
<tr>
<td>A [Blinded]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>D</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 7.2: Graphical visualization of the ABCD partition scheme employed in this search.

remaining ABCD regions are defined loosening or inverting the previously established requirements on two event variables, which then take the role of the two axes of the ABCD plane.

Out of the implemented cuts, the most likely to be acting on uncorrelated features of the event are those requiring the small-R jet to be b-tagged and the presence of at least one jet in the forward region. The criteria defining the four ABCD regions are summarised in Table 7.1 and visualized in Figure 7.2. The ABCD partition scheme outlined above was purposely conceived so that all four regions still satisfy the Higgs mass window requirement and belong to the H2T2B category, with the intention of preserving a suitably high level of compatibility between the target (A) and the support (BCD) samples in both kinematics and background processes.

This also offers the precious opportunity to instate two additional and completely orthogonal sets of ABCD regions following the same internal logic as the main set,
but only accepting events that fail one of the following Signal Region requirements:

- Regions $A_V, B_V, C_V, D_V$ including events with $M_{HC} \in [135, 200]$ GeV, a signal-depleted mass sideband;

- Regions $A_{1B}, B_{1B}, C_{1B}, D_{1B}$ including only events belonging to the **h2t1b** category, where only one of the two track jets associated to the Higgs candidate is $b$-tagged.

It follows from the definitions above that none of the eight aforementioned regions is blinded, rendering the two novel ABCD sets optimal grounds for testing and validating the modelling procedure before the eventual unblinding of the Signal Region.

### 7.1.1 Axes Correlation

The lack of correlation, and therefore the suitability of the two event features (tagging status of the $R = 0.4$ jet and presence of forward jets) elected to serve as the axes of the ABCD plane is investigated starting from the founding principle stating that:

$$\frac{N_A}{N_B} = \frac{N_C}{N_D} \tag{7.3}$$

if the axes are uncorrelated. If such condition is met, equation 7.3 should hold both globally and differentially with respect to any other kinematic variable describing the event, which can be investigated by comparing kinematic distributions across the four ABCD regions in the Validation Sideband $M_{HC} \in [135, 200]$ GeV.

It is worth remarking that observing shape discrepancies between the C,D and A,B region pairs does not imply the presence of any sizable correlation between the plane axes, as long as such discrepancies are compatible across pairings. This can be ascertained by inspecting, for each variable under scrutiny, the normalized C/D and A/B template ratios, and ensuring that whatever trend or lack thereof is consistently observed in both ratios.
Figure 7.3 shows the C↔D and A↔B comparisons plot for the main physical variables describing the event. Upon examining the relative ratios, the following conclusions can be drawn:

- The lack of forward jets biases the spectra of the VLB $p_T$, small-$R$ jet $p_T$ and the $\Delta R$ between the small-$R$ jet and the leading Higgs Candidate track jet. This is understood to result from the fact that data events in the 0 forward jets sample have, on average, one fewer jet to recoil against;

- the shape ratios of variables which are theoretically unaffected by the VLB system boost (such Higgs Candidate mass, VLB mass) are compatible with unity within the statistical uncertainty of either sample;

- The ratio plots in the biased variables show deviations that are consistent between $J0B$ (left-hand side plots) and $J1B$ events (right-hand side).

The latter observation implies that, while kinematic differences exist between the 0 and $\geq 1$ forward jet samples, these are not affected by the $b$-tagging status of the small-$R$ VLB jet, as would be expected from two uncorrelated event features. While the previous normalized shape ratio comparison ruled out sizable local effects, a global correlation between the two ABCD axes can still exist, potentially leading to an overestimation or underestimation of the total A-Region background yield. The intuitive estimator to quantify this effect is the correlation index $R_{corr}$, defined as:

$$R_{corr} = \frac{N_A \cdot N_D}{N_C \cdot N_B}$$

(7.4)

If no correlation exists, $R_{corr}$ is statistically compatible with unity by virtue of the ABCD principle. The physical significance of $R_{corr}$ is made evident interpreting it as a ratio between the events in the target region A and the corresponding ABCD-driven prediction:

$$R_{corr} = \frac{N_A \cdot N_D}{N_B \cdot N_C} = \frac{N_A}{N_B \cdot \frac{N_C}{N_D}} \equiv \frac{A}{\tilde{A}}$$

(7.5)
from which it follows that the A-Region yield is underestimated when $R_{\text{corr}} > 1$ and conversely overestimated if $R_{\text{corr}} < 1$.

The value of $R_{\text{corr}}$ is examined in the Validation Sideband for H2T2B data and in both the Higgs peak and its sideband for the low-purity H2T1B dataset. In each case, the statistical uncertainty on $R_{\text{corr}}$ is estimated propagating simple Poisson uncertainties on the yields involved. Results are shown in Table 7.2. A statistically significant deviation from unity of about 10% in magnitude is consistently observed in all three data samples probed, suggesting existence of a possible, minor, correlation between the sub-samples making up the ABCD plane. This is considered acceptable for real-world applications of the ABCD estimation method. A yield correction, compensating for the expected underestimation of the expected background, is contextually applied as detailed in Sec. 7.2.3.

<table>
<thead>
<tr>
<th>H2T2B, VS</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ fwd jets</td>
<td>9232</td>
<td>483</td>
</tr>
<tr>
<td>$=0$ fwd jets</td>
<td>41556</td>
<td>1893</td>
</tr>
<tr>
<td>$R_{\text{corr}}$</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>$\Delta R_{\text{corr}}$</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H2T1B, Peak</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ fwd jets</td>
<td>45845</td>
<td>2315</td>
</tr>
<tr>
<td>$=0$ fwd jets</td>
<td>223488</td>
<td>10215</td>
</tr>
<tr>
<td>$R_{\text{corr}}$</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>$\Delta R_{\text{corr}}$</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H2T1B, VS</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ fwd jets</td>
<td>88901</td>
<td>5342</td>
</tr>
<tr>
<td>$=0$ fwd jets</td>
<td>420276</td>
<td>22849</td>
</tr>
<tr>
<td>$R_{\text{corr}}$</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td>$\Delta R_{\text{corr}}$</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.2: Event yields in the four ABCD regions in the Validation Sideband for H2T2B data and in the Higgs Peak and Validation Sideband for H2T1B data. The respective values of $R_{\text{corr}}$ and their uncertainties are also displayed.
Figure 7.3: Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A.
Figure 7.3: Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A. (Cont.)
Figure 7.3: Validation Sideband comparison between data in region A (right, dots) vs B (right, solid) and C (left, dots) vs D (left, solid). The data samples with 0 forward jets (regions B and D) are scaled to the yield respectively of regions C and A. (Cont.)
7.2 ABCD Modelling Implementation

7.2.1 Yield Correction

The number of background events predicted to occur in the blinded Signal Region (corresponding to ABCD-Region A) is estimated out of the B-region yield applying a corrective global scale factor according to basic ABCD principle.

The global scale-factor, henceforth referred to as $k_{FWD}$, is defined as:

$$k_{FWD} = \frac{N_C}{N_D}$$  \hspace{1cm} (7.6)

and the A-Region yield is accordingly obtained as:

$$\tilde{N}_A = k_{FWD} \times N_B$$

The yields of the three unblinded ABCD regions in the Higgs mass peak, as well as the observed value of $k_{FWD}$, are displayed in Table 7.3.

<table>
<thead>
<tr>
<th>H2T2B SR</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ fwd jets</td>
<td>5315</td>
<td>BLIND</td>
</tr>
<tr>
<td>$=0$ fwd jets</td>
<td>22831</td>
<td>972</td>
</tr>
</tbody>
</table>

$k_{FWD} = 0.221 \pm 0.003$

Table 7.3: Event yields in the four ABCD regions in the Higgs Peak region ($m_H \in [105, 135]$ GeV).

7.2.2 Shape Correction

Figure 7.3 shows a comparison of distributions of some of the kinematic features of the event as they occur in data events classified as belonging to regions C vs. D (left)
and A vs. B (right). It can be seen that the distributions of variables like the mass of the Higgs candidate and the $\Delta R$ between the Higgs-matched track jets, which are not affected by the presence of or lack of forward jets, do not display any meaningful discrepancies across the $A \leftrightarrow B$ and $C \leftrightarrow D$ transition. Conversely, the forward jet multiplicity is shown to produce macroscopic differences in the behaviour of other variables, such as the VLB $p_T$, which are intuitively sensitive to the presence of additional, high-$p_T$ objects in the final state owing to the preservation of the balance in the total momentum on the transverse plane before and after the $pp$ interaction.

Even in those cases, however, the ratio plots in fig. 7.3 show that the differential bias on the relevant distributions is the same in both the $C \leftrightarrow D$ transition (left-hand side plots) and in $A \leftrightarrow B$ (right-hand side).

This observation, which validates the choice of the variables defining the ABCD axes, also inspired a refinement of the background modelling procedure: the simple ABCD estimation outlined in the previous sections does not operate any targeted correction to the shapes of ABCD-Region B, which are then simply transferred onto the final model for region A data as the global scale factor $k_{FWD}$ is applied. In order to produce a fully consistent model of the A-Region data, a shape correction aimed at nullifying any $A \rightarrow B$ shape discrepancies is applied in the form of event-by-event weights on the grounds of the observed $C \rightarrow D$ discrepancies, which we now know to be consistent with the former (again see fig. 7.3).

The form of the shape-correcting factors can be derived rewriting the traditional ABCD estimation formula (Eqn. 7.1) so that the probability density function of data in the four ABCD regions are explicitly visibile:

$$N_A = N_B \cdot \frac{N_C}{N_D} \quad \rightarrow \quad \int f_A(\vec{x}) \, d\vec{x} = \int \left[ f_B(\vec{b}) \cdot \frac{f_C(\vec{x})}{f_D(\vec{x})} \right] \, d\vec{x} \quad (7.7)$$

Where $f_i$ represents the normalized probability density function for in the generic region $i$, dependent on the physical variables $\vec{x}$ describing the event.

Eqn. 7.7 can be once again re-written locally, ideally applying the ABCD estimation
to an $n$-dimensional, infinitesimal bin about a specific point $\vec{x}_0$:

$$f_A(\vec{x}_0) \, d\vec{x}_0 = \left[ f_B(\vec{x}_0) \, d\vec{x}_0 \right] \cdot \frac{f_C(\vec{x}_0)}{f_D(\vec{x}_0)} \equiv \left[ f_B(\vec{x}_0) \, d\vec{x}_0 \right] \cdot W(\vec{x}_0) \quad (7.8)$$

In this equation, $f_B(\vec{x}_0) \, d\vec{x}_0$ can be recognized as the expected number of events of region B falling within the examined bin $d\vec{x}_0$. Out of this, the estimate for the number of A-region events falling in the same kinematic bin $d\vec{x}_0$ can simply be obtained multiplying by the ratio the probability density functions (henceforth pdf’s) of data in region C and D, calculated in $\vec{x}_0$.

This line of reasoning can be safely extrapolated to the limit segmentation scenario where $d\vec{x}_0$ in region B contains either 0 or 1 event. In this configuration, the estimate for the data shape in A is built assigning to every event in region B a kinematic weight equal to the value of $W(\vec{x})$ for the value of $\vec{x}$ defining the event under study.

Ideally, the ratio $W(\vec{x})$ should be evaluated experimentally from the ratio of data distributions over a grid of $n$-dimensional bins in order to properly account for cross-variable correlations. This, however, cannot be practically implemented due to the so-called curse of dimensionality, which would quickly render a multi-dimensional characterization of $W(\vec{x})$ dominated by statistical fluctuations and therefore of little use. Nevertheless, the comparison plots in Fig. 7.3 show that the ratio histograms for most variables are compatible with 1, with a few notable and understood exceptions. This allows to approximate $W(\vec{x})$ through a product of the one dimensional ratio between the marginal pdf’s relative to each of the variables showing statistically significant discrepancies:

$$W(\vec{x}) \simeq \prod_k w^j(x_j) \quad (7.9)$$

with $j$ running on the aforementioned variables. In practice, the $w_j$ are estimated through the ratio of the $k_{FWD}$-corrected histograms for the variable $i$ in the ABCD region C and D.

In order to reduce the impact on the reweighting functions of the statistical fluc-
tuations on the bin-wise data content of regions C and D, $W(\bar{x})$ is extrapolated by fitting the relevant ratio histograms. As no \textit{a priori} information exists on the functional form of $W(\bar{x})$, a non-parametric \textit{Kernel Regression} technique, based on the Nadaraya-Watson [128] estimator, is used.

Kernel-based regressions define $W(\bar{x})$ in every point of the $x$-space as a weighted average of the experimental points $(\bar{x}_i, \bar{y}_i)$, with the summation coefficients that regulate the relative contribution of each $\bar{y}_i$ to the average depending solely on the distance $|\bar{x} - \bar{x}_i|$ between the sampling point and each experimental point. A more comprehensive description of the principle and the operative steps taken to optimally implement the Kernel-based extraction of the reweighting function is provided in Appendix A.

Given that sizable $C \leftrightarrow D$ discrepancies are only observed in the spectra of the VLB system transverse momentum and in the $p_T$ of the $R = 0.4$ jet within it, these will be the only variables on which reweighting functions $w_i(x_i)$ are extracted.

Figure 7.4a-7.4b displays the normalized shapes of respectively the VLB system and the small-R jet $p_T$ in either region C (dots) or D (solid yellow). The reweighting function associated to each of these variables is overlaid to the relevant ratio plot. Figure 7.4c-7.4d shows the agreement between C and D data once the weight functions have been applied on the D event sample itself, highlighting a good closure of the reweighting method.

### 7.2.3 Residual Correlation Correction

The yield summary in the Validation Sideband displayed in Table 7.2 for both \textbf{H2T2B} and \textbf{H2T1B}, as well as in the signal region for \textbf{H2T1B} in Table 7.3, raises the issue of $R_{corr}$ not being compatible with 1 within one sigma as defined through statistical uncertainty.

The availability of a Validation Sideband completely orthogonal to the Signal Region suggests a simple correction scheme of the estimate $\hat{N}^{SR}_A$ by re-scaling $t^{SR}_{FWD}$...
Figure 7.4: Comparison of the distributions in the C and D regions for the two variables used to calculate the shape correction factors. The normalization scale factor $k_{FWD}$ is already applied to D events. All events belong to the Higgs mass peak region. The extracted weight function and its uncertainty are shown as a red line and a pink shaded area, respectively, in each of the ratio plots. The plots in the bottom row show the agreement between C and D data after the weight functions are applied to D itself.
by the residual correlation factor $R_{corr}$:

$$\tilde{N}_A = k'_{FWD} \times N_B = k_{FWD} \times R_{corr} \times N_B \quad \text{where} \quad R_{corr} = \frac{N_{AV} \cdot N_{DV}}{N_{BV} \cdot N_{CV}} \quad (7.10)$$

Tables 7.2 also show that the values of $R_{corr}$ computed in either the validation side-band for the $H2T2B$ category or in both the Higgs Peak and the sideband for $H2T1B$ data region are statistically compatible, given the statistical uncertainty on $R_{corr}$ in the three cases:

$$\Delta R_{corr} = R_{corr} \times \sqrt{\frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_C} + \frac{1}{N_D}} \quad (7.11)$$

This lends credibility to the application of the value of $R_{corr}$ from the $H2T2B$ side-band as a yield correction of the main background estimation in the Higgs mass peak region.

The established uncertainty on the correction factor, $\Delta R_{corr}$, is then set to serve as a normalization uncertainty on the background yield prediction in the Signal Region (see Sec. 7.4.2).

### 7.3 ABCD Closure Plots

Figure 7.5 shows the shape closure in a number of key variables between the data samples in ABCD region C (dots) and D (solid histogram) after both the normalization factor $k_{FWD}$ and the per-event kinematic weights, extracted as described above, have been applied to the events in D. In all cases, a good level of agreement is observed.
Figure 7.5: Closure test for the per-event shape correction in regions C,D. Event weights are calculated from ratios between templates distribution in C and D, and applied here to the data in D (solid). A comparison is drawn with C events to test closure. All events belong to the Higgs Mass region (SR).
Figure 7.5: Closure test for the per-event shape correction in regions C,D. Event weights are calculated from ratios between templates distribution in C and D, and applied here to the data in D (solid). A comparison is drawn with C events to test closure. All events belong to the Higgs Mass region (SR). (Cont.)
7.4 Model Validation

Before the blinding on Signal Region data can be lifted and the unveiled data compared with the background prediction, it is necessary to ensure, with an acceptable degree of confidence, that the established background modelling procedure is able to produce an accurate and unbiased prediction.

The first step en route to validating the model is to lay out and account for the possible sources of uncertainty on background prediction originating from the various steps of the modelling procedure itself.

Once a rule is set in place to identify a range of uncertainty for any data prediction, it is possible to put the procedure itself to test by running mock-up estimations for alternative regions of the phase space where the data is already unblinded, such as ABCD-Region A in both the Higgs mass sideband $M_{HC} \in [135, 200] \text{ GeV}$ and the Higgs peak for the low purity $\text{H2T1B}$ dataset. The quality of such predictions is then assessed evaluating the agreement with the actual data, given the established uncertainties on the mock-up models and the statistical uncertainty on the data.

7.4.1 ABCD Modelling Limitations

An ABCD-driven background prediction can in principle suffer from biases introduced by a number of known pitfalls affecting simple data-driven techniques, such as a composite origin of the background or the leakage of the hypothetical signal outside the targeted, blinded area.

The presence non-negligible secondary sources of background can potentially introduce a bias in an ABCD-driven modelling procedure, provided that the proportions and/or kinematics of events from this source in the ABCD categories are different to those from the main source of background.

While QCD multi jet events are expected to be by far the dominant source of back-
ground in this analysis, the potential contributions from two other processes, $t\bar{t}$ decaying hadronically and $Z(bb) + b$-jets, were also investigated. This investigation highlighted how the $Z(bb) + b$-jets contribution is at at the per-cent level, so it can safely be ignored. Whilst the $t\bar{t}$ contribution was estimated via Monte Carlo simulations to be $\sim 7\%$, it was found that any potential bias from this contribution is much smaller than the background modelling systematic uncertainties, further validating the established modelling procedure with respect to the background homogeneity hypothesis.

On the other hand, the leakage of a fraction of the signal into ABCD-Region B was estimated to ultimately yield a 13-15% unavoidable loss in the observed significance of a possible signal, with no a priori correction scheme possible.

A more comprehensive treatment of the two modelling issues outlined above can be found in Appendix B.

### 7.4.2 Modelling Uncertainties

The main sources of uncertainty on the behaviour of the model arise from the reweighting procedure employed in the calculation of the per-event shape correction factor, from residual correlations between the two axes of the ABCD plane and from a possible dependence of the reweighting functions on the Higgs candidate mass. A final source of uncertainty is caused by the statistical fluctuations on the ABCD-Region B dataset as propagated through the entire modelling procedure onto the final estimate for the background. These sources of uncertainty will be discussed individually in the following subsections.

**Uncertainty on $R_{corr}$ correction**

The necessity for a correction of the residual correlation in the ABCD yield estimation, as well as its implementation, are described in Section 7.2.3. A normalization
uncertainty is generated by substituting $R_{\text{corr}}$ with $R_{\text{corr}} \pm \Delta R_{\text{corr}}$ at the modelling stage. Table 7.2 summarises yields, $R_{\text{corr}}$ and $\Delta R_{\text{corr}}$ in the Validation Sideband for both the H2T2B and H2T1B channels. From the information displayed, it can be gathered that the impact of the uncertainty on the $R_{\text{corr}}$ correction is 5.2% for the H2T2B channel.

**Uncertainty on Shape Correction**

The non-parametric regression method employed to extract the continuous weight function (as described in detail within Appendix A) comes with a natural $1\sigma$ band that can be used as a handle to establish a systematic uncertainty on shape correction. The main factors impacting the variance of the extracted weight functions are the statistical uncertainty of the samples involved in the source ratios and the choice of the critical bandwidth parameter regulating the regression of the ratio.

Given this premise, an uncertainty representing the 68% confidence level bands for the ratio estimation is associated to each weight function, allowing to produce both a nominal and a "variated" value for the kinematic weight to be assigned to each event. By construction, the uncertainty on the shape correction is symmetric about the central value.

Given that the full shape-correcting weight is defined as the product of the individual weights pertaining to each variable to which reweighting is applied (see Sec. 7.2.2), as many variated templates can be produced as variables are used in the procedure:

$$ W_j^\pm = \frac{w_j(x_j) \pm \Delta w_j(x)}{w_j(x_j)} \cdot \prod_i w_i(x_i) \quad (7.12) $$

where $i$ runs on all variables reweighted upon, $j$ represents the variable for which the specific systematic variation is calculated and $\Delta w_j(x_j)$ the uncertainty on its individual-weight as extrapolated from Eqn. A.2.
Uncertainty on modelling stability

A source of uncertainty for the model can be traced back to the fact that, in principle, the per-event weights and more importantly the global scale factor $k_{FWD}$ could have a weak dependence on the Higgs Candidate mass, the variable defining the boundaries of the Signal Region. As $k_{FWD}$ is calculated globally, and the combined event weight is by construction not a function of $M_{HC}$, it is conceivable that a small bias could be introduced by integrating over the mass dependency in the Signal Region. An estimate of such a possible bias is provided by creating alternative templates for the final observable by means of alternative sets of weights and scale factor extrapolated from ABCD-region C and D data in the validation sideband:

$$k_{FWD}^{\text{var}} = \frac{N_{CV}}{N_{DV}}$$  \hspace{1cm} (7.13)

The variated template is taken as a one-sided systematic uncertainty, which is then symmetrized and modelled in the statistical analysis through a single nuisance parameter (Sec. 9.1).

Propagated Statistical Uncertainty from ABCD-Region B data

The final source of uncertainty on the background model is estimated propagating the statistical fluctuations on the ABCD-Region B dataset, out of which the prediction is ultimately constructed, through the full ABCD/shape correction procedure.

Let $A_j$ be the number of predicted background events in the $j$-th bin of the $m_B$ template, which has been computed out of the number of events in the same bin for the ABCD-Region B dataset ($B_j$) via the application of the set of event weights $W$:

$$A_j = \sum_{k=0}^{B_j} w_k \quad w_k \in W$$  \hspace{1cm} (7.14)
Working under the assumption that the generic event weight \( w_k \) (with the \( k \)-index running on the unweighted bin events) can only take a discrete and finite range of values \( w_\alpha \) (with the \( \alpha \)-index running on the possible weight values), the equation above can be rewritten as:

\[
A_j = \sum_{\alpha} b_j^\alpha \cdot w_\alpha
\]  

(7.15)

where \( b_j^\alpha \) is the number of unweighted region B events of the \( j \)-th bin that are multiplied by the weight value \( w_\alpha \). As the \( w_\alpha \) are in this picture merely multiplicative factors to the Poisson-distributed \( b_j^\alpha \), the fluctuation on \( A_j \) can be written as:

\[
\delta A_j = \sum_{\alpha} \delta b_j^\alpha \cdot w_\alpha
\]  

(7.16)

And the variance on estimated background content \( A_j \) can then be obtained autocorrelating the fluctuations:

\[
\text{var}[A_j] = \sum_{\alpha} \sum_{\beta} \text{cov}[\delta b_j^\alpha, \delta b_j^\beta] w_\alpha w_\beta
\]  

(7.17)

In the reasonable assumption that the number of events \( b_j^\alpha \) that, within the bin \( j \), are multiplied by the specific weight value \( w_\alpha \) is Poisson-distributed and uncorrelated from the number of events \( b_j^\beta \) in the same bin that take the different weight value \( w_\beta \), the covariance matrix is diagonal, with the non zero element following the definition of variance for a Poisson variate:

\[
\text{cov}[\delta b_j^\alpha, \delta b_j^\beta] = \delta^{\alpha, \beta} \sqrt{b_j^\alpha} \sqrt{b_j^\beta}
\]  

(7.18)

and the standard deviation on \( A_j \) can then be rewritten into the well known result:

\[
\text{var}[A_j] = \sum_{\alpha} \sum_{\beta} \delta^{\alpha, \beta} \sqrt{b_j^\alpha} \sqrt{b_j^\beta} \cdot w_\alpha w_\beta = \sum_{\alpha} b_j^\alpha w_\alpha^2 = \sum_{k=0}^{B_j} w_k^2
\]  

(7.19)
whereby the variance on the content of the generic bin of a weighted histogram is given by the sum of the squared weights acting on the unweighted bin events.

Given the derivation above, each bin of the histogram of the final observable $m_B$ is assigned an uncertainty that is assumed to be uncorrelated with the corresponding propagated statistical uncertainty on the remaining bins as well as the other established modelling uncertainties introduced in the previous sections.
7.4.3 Validation Check I: Higgs Mass Sideband

The availability of a low-purity high mass sideband is a key asset in the background validation strategy, as the full modelling procedure can be run seamlessly on the alternative ABCD region set $A_V$, $B_V$, $C_V$ and $D_V$, defined through same criteria as the main ABCD plane, but in the low sensitivity Higgs candidate mass range $M_{HC} \in [135, 200] \text{ GeV}$.

Table 7.4 displays the data and signal ($M_B = 1.2 \text{ TeV}$, $\kappa = 0.5$) yield in the validation sideband, along with the corresponding values of $S/B$. For better reference, the predicted background and signal yield in the blinded Signal Region is also displayed, highlighting the suitability of the established sideband to be employed as a background-enriched validation region.

Figure 7.6 (H2T2B channel) and 7.7 (H2T1B channel) show a comparison between low-purity data in the Higgs mass sideband ABCD-region A and the corresponding ABCD-driven prediction for a large number of kinematic variables. The alternative set of event weights calculated out of C/D ratios in the Validation Sideband is used at a later stage to define a systematic uncertainty on the background model (see Sec. 7.4.2).

<table>
<thead>
<tr>
<th>Region</th>
<th>Sample</th>
<th>$S$</th>
<th>Data</th>
<th>$S/B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_V$</td>
<td>16.3</td>
<td>483</td>
<td>3.4%</td>
<td></td>
</tr>
<tr>
<td>$B_V$</td>
<td>14.2</td>
<td>1893</td>
<td>0.75%</td>
<td></td>
</tr>
<tr>
<td>$C_V$</td>
<td>15.3</td>
<td>9232</td>
<td>0.16%</td>
<td></td>
</tr>
<tr>
<td>$D_V$</td>
<td>8.9</td>
<td>41556</td>
<td>&lt; 0.1%</td>
<td></td>
</tr>
<tr>
<td>A (Blind)</td>
<td>80.0</td>
<td>256 (est.)</td>
<td>31%</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.4: Simulated signal ($M_M = 1.2 \text{ TeV}$, $\kappa = 1.0$) and observed data yields in the four ABCD regions built in the Higgs mass sideband range. For comparison, the predicted signal and background yield in the Signal Region are also displayed.
All Validation Sideband plots display shaded error bands corresponding to the total error on the model, comprehensive of the systematic derived as explained in Sec. 7.4.2, as well as the native statistical uncertainty on the support dataset of the model (ABCD-region B in either the Validation Sideband or the Signal Region). All contributions are summed in quadrature.

The prediction for the sideband data displayed in the plots was not up-scaled to compensate for the residual correlation observed in Sec. 7.2.3, as the correction is itself derived in the very same mass range currently under scrutiny. Consequently, as expected, the data/prediction comparison highlights a 5 – 10% underestimation of the A-Region yield. No significant shape disagreement between the A-Region sideband data and its prediction is observed.

### 7.4.4 Validation Check II: Low Purity Sample

According to the significance estimates discussed in Sec. 6.3.1, the H2T1B channel has significantly lower expected sensitivity than the main H2T2B channel, owing to the looser requirement on the Higgs candidate tagging status and the consequently poorer background rejection.

The availability of a low-purity channel with comparatively large data statistics gives the opportunity to run an ultimate test of the soundness and robustness of the background modelling procedure as a preliminary step to the eventual unblinding of the sensitive H2T2B channel. The validation procedure is defined as follows:

- Each signal and data event belonging to the H2T1B channel is randomly sorted into sub-groups roughly matching the expected H2T2B data yield in the Signal Region;

- The full background modelling procedure, including the definition of the relevant modelling systematics, is run independently on each H2T1B sub-sample.

- The independent predictions are compared with the actual A-Region data in
7.4. Model Validation

![Graphs](image)

(a) Higgs Candidate $p_T$.
(b) Small-$R$ Jet $p_T$.
(c) Leading Higgs track jet $p_T$.
(d) Sub-Leading Higgs track jet $p_T$.
(e) VLB $p_T$.
(f) VLB Mass

**Figure 7.6: H2T2B channel.** Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband.
7.4. Model Validation

The $H_2T_2B$ sub-sample and the agreement evaluated within the established uncertainty on the predictions.

As the statistic of the full undivided $H_2T_1B$ data pool is roughly eight times larger than that of the $H_2T_2B$ channel, eight sub-samples were created.

Given the full selection results displayed in Table 6.26 and 6.27, the expected average Signal over Background ratio (and signal significance $S/\sqrt{B}$) in the $J_{1B_H2T1B}$ sub-sample can be easily calculated and compared against the values expected in the high sensitivity sample $J_{1B_H2T2B}$. All signal figures here pertain to a signal with 1200 GeV resonance mass.


7.4. Model Validation

Figure 7.7: H2T1B channel. Comparison between the predicted background and the actual data in the A-region of the Validation Sideband. Event weights and scale factor were calculated in the Validation Sideband.
As a further item of insurance against a possible signal sensitivity in each of the eight 1B sub-samples, the average $S/B$ and $S/\sqrt{B}$ are also calculated assuming a larger signal cross section than predicted by the benchmark theoretical model, the choice falling on the lowest cross-section times branching ratio excluded by the $B \rightarrow bH(\gamma\gamma)$ ATLAS analysis [64], at roughly 100 fb. The findings are detailed in Table 7.5.

Figure 7.8 shows, for the final observable $m_B$, the agreement between each H2T1B data sub-sample and the established background model in the Higgs peak region $M_H \in [105, 135]$ GeV. The shaded error bars refer to the systematic uncertainties on
7.4. Model Validation

<table>
<thead>
<tr>
<th>Sample</th>
<th>S</th>
<th>B</th>
<th>$S/B$</th>
<th>$S/\sqrt{B}$</th>
<th>$S(100\text{fb})/B$</th>
<th>$S(100\text{fb})/\sqrt{B}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{1B_H2T2B}$</td>
<td>80.0</td>
<td>256.8</td>
<td>18%</td>
<td>5</td>
<td>59%</td>
<td>9.5</td>
</tr>
<tr>
<td>Av. $1B$ sub.</td>
<td>6.3</td>
<td>289.3</td>
<td>2.2%</td>
<td>0.4</td>
<td>4%</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 7.5: Summary of the Signal/Background numerical interplay in the main channel and in the average $J_{1B\_H2T1B}$ sub-sample. $S/B$ ratios and signal significance $S/\sqrt{B}$ are provided for both the theoretical 1200 GeV cross section ($\sim 31$ fb) and the $B \to bH(\gamma\gamma)$ exclusion limit at the same mass, roughly amounting to three times as much at 100 fb.

the model generated independently for each sub-sample. No discrepancy between data and model, either statistically significant or consistently observed through the eight data samples, is observed.

7.4.5 Validation Outcome

Neither of the two validation protocols described in the previous sections have brought to light evidence of any statistically significant mismodelling by the established modelling procedure. The prediction for the Standard Model background in the Signal Region is therefore judged to be sufficiently reliable to be compared to the high-sensitivity data.
Figure 7.8: Distribution of the reconstructed VLB mass in H2T1B data for subsamples 1 to 8.
Chapter 8

Systematic Uncertainties

8.1 Overview

Predicting the expected kinematic shapes and event yields for signal and background in the Signal Region is a necessary but by no means sufficient step to produce a statistically robust understanding of the observed data with respect to the selected benchmark for New Physics phenomena.

Indeed, the strength of any observation, or the credibility of any measurement, is intimately connected to the degree of accuracy to which the behavior of the background sources and the properties of the signal are understood.

Typical factors limiting the accuracy of the relevant predictions include incomplete or imperfect knowledge of the response of the detector, uncertainties on the theoretical and experimental parameters employed in the Monte Carlo generation, differences in the performance of tagging algorithms between simulations and data and lastly any uncertainty introduced by custom analysis operations such as the estimation procedure described in Chapter 7.

Within the context of an ATLAS analysis, technical systematic uncertainties relating to the modelling of the simulated physical processes or the response of the detector are typically provided centrally in the form of "up" and "down" variations of the physical quantities affected by each source. Conversely, systematic uncertainties
introduced by custom manipulation of the data (or Monte Carlo if used) en route to the estimation of the background are to be defined and computed entirely within the individual analysis workflow.

The purpose of this chapter is to provide an outline of the various sources of systematic uncertainties playing a role in the present search, assess their impact and ultimately identify the chief contributors to the total uncertainty.

8.2 Experimental Uncertainties on the Signal Simulation

A major component of the total uncertainty on the expected properties of the signal arises from our incomplete understanding of experimental items such as the precise response of the detector with respect to the hadronic energy deposited in the calorimeters, the inherent resolution on all momentum measurements and the total integrated luminosity of the data pool to which the prediction should be normalized.

8.2.1 Jet Energy Scale

The raw energy readings of the ATLAS calorimeters undergo a sophisticated sequence of calibrating operations [15, 90], outlined in Fig. 8.1, before a final determination on the energy of each jet is provided to the analyzer. The energy of each reconstructed jet, be it physical or simulated in Monte Carlo, is first corrected to curtail its dependence on the average number of $pp$ interactions per bunch crossing, then scaled in such a way that the jet 4-momentum matches that of the final-state "stable" particles generated in the corresponding Monte Carlo simulation. This latter stage accounts for degrading factors such as reconstruction biases in different $\eta$ regions or caused by the transition between the different technologies across the calorimetric system.
Figure 8.1: Schematic flow diagram outlining the individual steps in the $R=0.4$ jet energy scale calibration. [15].

The response of the calorimeter is studied in Monte Carlo inclusive jet simulated events [129], evaluating the ratio between truth-level and reconstruction-level jet momentum in bins of the jet $p_T$ and pseudo-rapidity. The detector response to heavy flavour-initiated jets, which is in principle affected by the possibility of a semi-leptonic decay, is studied and parametrized separately and individually for each flavour by requiring the presence of either a $b$ or $c$ generator-level hadron within the jet effective radius. The end-point energy calibration of Monte Carlo jets is then applied at the analysis level by means of a centrally provided calibration tool on the grounds of the available Monte Carlo truth information.

At the end of the jet energy calibration procedure outlined above, outstanding differences in the data / Monte Carlo jet energy scales may still exist. To ensure a full compatibility of the energy scales, the jet energy is corrected *in data events only* according to a series of *in situ* measurements transverse of the momentum balance in specific event configurations such as $Z$+jets and $\gamma$+jets.

The energy scale calibration sequence for large-$R$ jets follows a largely similar scheme, with the critical addition of the *grooming* procedure (see Sec. 4.3.4) applied to reduce the contribution of soft radiation from the underlying event$^1$ or pile-up interactions.

$^1$ *Underlying event* refers to the additional physical processes, typically soft inelastic parton scattering, occurring in the same proton-proton collision that gives rise to the hard-scatter event.
Experimental Uncertainties on the Signal Simulation

8.2. Experimental Uncertainties on the Signal Simulation

Figure 8.2: Summary of the JES uncertainty for R = 0.4 jets (top row) and R = 1.0 jets. The total uncertainty is represented as the blue envelope over a range of both the jet transverse momentum and pseudorapidity. The macro-contributions to the total uncertainty from each class of sources is also displayed. The green envelope in the large-R jet JES uncertainty plot represents the uncertainty level established in the previous set of recommendations, highlighting a sharp improvement in the quality of the measurements.

Each module in the calibration sequence comes with an associated uncertainty originating from factors such as residual pile-up contribution, mis-modelling in the Monte Carlo samples employed, uncertainties affecting the in situ measurements and non closure in the correction itself. Ultimately, the full range of Jet Energy Scale (JES) uncertainties comprises 125 individually recognized sources, each of which can be treated as an individual uncertainty on the energy of each jet in the event.

The comprehensive JES uncertainty is typically parametrized in terms of the jet $p_T$ and pseudo-rapidity, as displayed in Figure 8.2 for R = 0.4 jets.

As high-$p_T$ jets play a major role in the reconstructed final state, with their energy correlating heavily with the final observable $m_B$, the Jet Energy Scale uncertainty is
expected to play a significant role in the overall accuracy to which the expected signal contribution (the only instance of Monte Carlo usage) is known. The small-R jet uncertainty is parametrized through 30 ideally uncorrelated components according to the existing ATLAS recommendation, with as many sources also identified for large-R jets. The impact of two of the principal components on the shape of the $m_B$ observable in either jet collection is displayed in Figure 8.3 for a simulated signal with pole resonance mass of 1.2 TeV and coupling strength $\kappa = 0.5$. It can be gathered how none of the displayed sub-sources of JES uncertainty has comparable or greater impact on the signal prediction than the statistical uncertainty on the Monte Carlo itself.
8.2.2 Jet Energy Resolution

The study of the energy signature of a stream of hadrons entering the ATLAS calorimeters is a real-world measurement, with an associated mean response - the Jet Energy Scale - and a corresponding standard deviation, identified as the resolution on the jet energy provided by the measuring apparatus.

The Jet Energy Resolution (JER) is typically parametrized as originating from three independent contributions, so that given a jet of generic $p_T$, the energy resolution $\sigma(p_T)$ is given by [15]:

$$\frac{\sigma(p_T)}{p_T} = \frac{N}{p_T} + \frac{S}{\sqrt{p_T}} + C$$  \hspace{1cm} (8.1)

The first term in Eqn. 8.1 identifies the contribution from noise in the detector, which does not depend on the $p_T$ of the examined jet, and whose relative contribution $\sigma(p_T)_{\text{noise}}/p_T$ therefore scales with $1/p_T$. The second term is stochastic in nature, which is reflected in the fact that, being the determination of the energy ultimately dependent on the Poisson-distributed number of calorimeter cells with an energy deposit, the contribution $\sigma_{\text{stoch.}}$ is proportional to the square root of the total momentum. The third, constant term in Eqn. 8.1 represents the contribution from technology-driven response fluctuations which are generally understood to be proportional to the jet momentum itself.

The Jet Energy Resolution is measured both in Monte Carlo events and in situ chiefly evaluating the width of the momentum asymmetry distribution in di-jet events for a number of bins in the $(p_T \times \eta)$ phase space.

As cuts on the small-R and large-R jet momentum are established as part of the analysis workflow, the Jet Energy Resolution must be accounted for as a factor generating an uncertainty on the number of predicted signal events that satisfy those requirements. The eventual impact on the search is therefore estimated by smearing the jet energy by the corresponding value of the JER, and propagating the resulting
sample through the full event selection, ultimately obtaining a lower bound to the number of simulated events entering the Signal Region. The various components determining the Jet Energy Resolution are parametrised as 6 independent smearings, in according to the ATLAS recommendation.

Figure 8.4 shows the impact on the signal prediction of one of the JER components in two signal samples with resonance mass respectively 1.2 TeV and 2.0 TeV.

![Figure 8.4: Impact on the shape of $m_B$ (reconstructed mass) in a $M_B = 1.2$ TeV (resonance mass) (left) and $M_B = 2.0$ (right), $\kappa = 0.5$ simulated signal sample for two of the 6 independent components of the Small-R jet Jet Energy Resolution. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each $m_B$ bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework](image)

### 8.2.3 Large-R Jet Mass Uncertainties

As the large-R jet invariant mass features predominantly in the present analysis, both directly among the Signal Region definition criteria and indirectly as part of the construction of the observable $m_B$, the impact on the analysis of systematic uncertainties on the large-R jet mass scale (JMS) and resolution (JMR) has to be evaluated.

Similarly to how the energy of a jet needs to be corrected to reflect that of the instigating particle(s), the large-R Jet Mass Scale is determined through both tar-
8.2. Experimental Uncertainties on the Signal Simulation

Figure 8.5: Impact on the shape of $m_B$ in a $M_B = 1.2$ TeV (resonance mass) (left) and $M_B = 2.0$ (right), $\kappa = 0.5$ simulated signal of the leading contributor to the JMS uncertainty. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each $m_B$ bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework.

generated MC-based correction and in situ, with calibrating methods featuring track to calorimeter mass ratios in data and simulation and mass peak fitting in a high purity sample of simulated and physical $t\bar{t}$ events [90].

Similarly to the JES uncertainty, the Jet Mass Scale uncertainty is provided in the form of 5 independent contributions, three of which account for the tracking uncertainty, and one each for the statistical and modelling uncertainties of the Monte Carlo samples involved in the mass calibration.

Figure 8.5 shows the impact of the leading contribution for the total JMS uncertainty in a 1.2 TeV and a 2.0 TeV signal at the end of the full selection process. In both cases the principal JMS component stems from the modelling uncertainties on the calibration Monte Carlo samples. The large-R Jet Mass Resolution is derived as a byproduct of the forward folding method [130] for the jet mass calibration, and is intimately related to the measured width of the top and $W$ boson mass line shapes in $t\bar{t}$ events where the decay products of either object are collected into a single $R = 1.0$ jet.

The mass resolution of a large-R jet depends on a number of factors including the jet...
Experimental Uncertainties on the Signal Simulation

Figure 8.6: Monte Carlo driven estimation of the Jet mass Resolution for Higgs (left) or QCD (right) originated $R = 1.0$ jets. The JMR for Higgs jets is provided for jets falling on both sides of the theoretical Higgs peak, while values of the QCD JMR are displayed for three different ranges of $m_{\text{jet}}/p_T$, mass range and the physical origin of the jet itself. For the purpose of reaching a clearer understanding of the JMR in different mass ranges and physics scenarios, including those involving boosted Higgs decays for which in situ measurements have not yet been established, the JMR is further estimated evaluating the variance of the reconstructed-to-truth jet mass in Monte Carlo simulation of the relevant processes [131]. Figure 8.6 shows the large-R JMR for a wide range of jet $p_T$ and mass for both $H \rightarrow b\bar{b}$ and QCD events.

The uncertainty on the the shape and yield of the observable $m_B$ in the signal introduced as a result of the JMR is displayed in Fig. 8.7. From an inspection of the ratio plot, which shows a 10% (1200 GeV) to 17% (2000 GeV) impact on the signal prediction, it can be gathered that the JMR uncertainty is the main contribution to the total uncertainty on the signal, and is therefore expected to be the primary MC-driven degrading factor in the sensitivity of the search.
8.2. Experimental Uncertainties on the Signal Simulation

Figure 8.7: Impact on the shape of \( m_B \) (reconstructed mass) in a \( M_B = 1.2 \) TeV (resonance mass) (left) and \( M_B = 2.0 \) (right), \( \kappa = 0.5 \) simulated signal of uncertainty introduced by the large-R Jet Mass Resolution. The shaded area represents the statistical uncertainty on the Monte Carlo, the dashed points in the ratio plots the original up and down variations for each \( m_B \) bin and in solid red and blue the up and down variations after the in-built smoothing applied by the statistical analysis framework.

8.2.4 Flavour Tagging Uncertainties

As detailed in Chapter 4, the difference in the efficiency of flavour tagging algorithms in data and Monte Carlo simulation is accounted for and corrected via the application of dedicated scale factors specific to the jet collection and algorithm in use, and dependent on the position of the inspected object in the \( (p_T \times |\eta|) \) plane.

A large number of theoretical and experimental factors introduce an uncertainty on the reciprocal data vs Monte Carlo efficiency, which is propagated through physical analyses as uncertainties on the flavour tagging efficiency scale factors. Among those, relevant contributors are the statistical uncertainty on the Monte Carlo simulations used in the scale factor computation, the modelling of flavour composition, PDF and showering in simulated \( t\bar{t} \) events and the propagated effect of the uncertainty on the involved jet energy scale [132].

The overall impact of the systematic uncertainties originating from the \( b \)-tagging of the jets involved in the analysis is of the order of 3%, observed consistently throughout both the resonance mass and coupling strength range under study. As
such, the contribution of flavour tagging uncertainties is expected to play a minor role in defining the overall sensitivity of the search.

**8.2.5 Other Uncertainties**

Beyond the sources enumerated and discussed in the previous sections, the description of the predicted signal properties is affected, albeit with highly sub-leading contributions, by factors such as the uncertainty on the jet-to-vertex association tagger for central (JVT) and forward jets (fJVT) and the determination of the total Run 2 integrated luminosity to which all Monte Carlo prediction are normalized.

The uncertainty in the combined 2015-2018 integrated luminosity is 1.7% [133], amounting to approximately $2.4 \text{ fb}^{-1}$ out of a total of $139 \text{ fb}^{-1}$, a value obtained using the LUCID-2 detector [134].

Conversely, the impact of the jet-vertex-tagger on the signal yield is well below the 0.1% threshold, and such uncertainties therefore play no role whatsoever in the analysis.

**8.3 Experimental Uncertainties on the Background Model**

As extensively discussed in the previous chapters, no Monte Carlo simulation was employed en route to the construction of a prediction of the background contribution in the signal region. As a consequence, the traditional sources of uncertainties associated to simulations and briefly outlined above in the context of the signal prediction uncertainty are altogether absent.

A comprehensive description of the uncertainties of the background model introduced by the custom data-driven procedure established in the analysis was provided
8.3. Experimental Uncertainties on the Background Model

in Sec. 7.4.2. This section will only account for the relative contribution of such sources on the predicted number of background events and the expected shape of the observable $m_B$, as represented in Figure 8.8.

Examining first the two sources of uncertainties introduced by the shape reweighting, the one related to the VLB $p_T$-driven correction (Fig. 8.8a) displays a relatively flat profile ranging from 3.5% at in the low-$m_B$ bins to about 4.5% at $m_B \sim 2.2$ TeV, with an overall impact on the expected background yield of $\pm 3.6\%$.

The reweighting on the small-R jet $p_T$ (Fig. 8.8b), conversely, introduces a systematic uncertainty with a more marked dependence of $m_B$, a feature that can be traced back to the far greater correlation between small-R $p_T$ and $m_B$ than it is expected to exist between the VLB $p_T$ and $m_B$. The global yield uncertainty introduced by

![Figure 8.8](image)

**Figure 8.8:** Impact of the custom systematic uncertainties on the prediction of the Standard Model background contribution in the Signal Region. The shaded area represents the statistical uncertainty on the background model inherited from the original data sample on which the entire yield/shape correcting procedure was applied.
this source is computed to be ±6.3%, with the local impact ranging from around 5% in the 950 GeV < \( m_B \) < 1050 GeV bin to 12% in the last bin of the invariant mass range.

The uncertainty introduced to account for the stability of the ABCD and shape correction procedure within the Higgs mass range covered in the signal region (Fig. 8.8c) is observed to be negligible below \( m_B \sim 1.4 \) TeV, and a sub-leading contribution in the high-mass end, with impact reaching approximately 4%.

The fourth source of uncertainty on the background model is introduced to account for the statistical uncertainty on the the computation of \( R_{\text{corr}} \), the factor correcting for the underestimation consistently observed in the validation regions and arising from the residual correlation between the ABCD axes. The definition of \( R_{\text{corr}} \) and its statistical uncertainty \( \Delta R_{\text{corr}} \) can be found in Sec. 7.4.2. It is by definition an uncertainty on the total predicted yield, with no dependence on the reconstructed VLB mass, with a relative impact amounting to ±5.1%.

The final source of uncertainty on the background model is given by the statistical uncertainty on the ABCD-Region B dataset propagated through the yield and shape reweighting applied within the background modelling procedure. As shown in Sec. 7.4.2, uncorrelated uncertainties are assigned to each \( m_B \) bin \( i \), given by the sum of the weights squared:

\[
\sigma(n_i) = \sqrt{\sum_{k=0}^{K=N_i} w_k^2} \tag{8.2}
\]

where \( N_i \) is number of events in the \( i \)-th bin. Table 8.1 displays, for every \( m_b \) bin, the bin content and the absolute and relative uncertainty inherited from ABCD-Region B.
8.4 Impact Assessment

A comprehensive summary of the relative impact of each category of systematic uncertainty on either the signal or the background yields is displayed in Table 8.2.

The leading contributions to the overall uncertainty on the simulated signal is consistently observed to be the resolution on the large-R jet mass, a crucial variable featuring heavily in the definition of the Signal Region. The overall uncertainty on the expected number of signal events ranges from 12-13% for a 1.2 TeV signal to around 19% for a highly massive 2.0 TeV resonance.

The overall uncertainty on the established background yield is approximately 9%, with leading factors being the uncertainties on the small-R jet $p_T$ reweighting and on the $R_{corr}$ scaling.

As can be gathered from Fig. 8.8, the leading contributors to the background model uncertainty shift between the various ranges of the examined $m_B$ spectrum: for $m_B$ 1.4 TeV and below, the uncertainty is driven mostly by the propagated B-region uncertainty and the $R_{corr}$ scaling contribution, whereas at the far end of the mass spectrum the B-region uncertainty and the small-R jet $p_T$ are by far the dominant

<table>
<thead>
<tr>
<th>Bin bounds</th>
<th>$n_i$</th>
<th>$\sigma(n_i)$</th>
<th>$\sigma(n_i)/n_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[900, 1050] GeV</td>
<td>45.6</td>
<td>3.7</td>
<td>8.1%</td>
</tr>
<tr>
<td>[1050, 1150] GeV</td>
<td>61.3</td>
<td>4.3</td>
<td>7.0%</td>
</tr>
<tr>
<td>[1150, 1250] GeV</td>
<td>47.8</td>
<td>3.7</td>
<td>7.8%</td>
</tr>
<tr>
<td>[1250, 1450] GeV</td>
<td>51.3</td>
<td>4.0</td>
<td>7.7%</td>
</tr>
<tr>
<td>[1400, 1600] GeV</td>
<td>28.1</td>
<td>2.9</td>
<td>10.3%</td>
</tr>
<tr>
<td>[1600, 1900] GeV</td>
<td>15.2</td>
<td>2.1</td>
<td>13.9%</td>
</tr>
<tr>
<td>[1900, 2300] GeV</td>
<td>7.5</td>
<td>1.8</td>
<td>23.3%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>256.8</td>
<td>8.8</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Table 8.1: Bin-wise impact of the statistical uncertainty on ABCD-Region B data propagated onto the background model.
8.4. Impact Assessment

contributions, with the remaining sources playing a minor role.

<table>
<thead>
<tr>
<th>Systematic</th>
<th>VLB M = 1.2 TeV</th>
<th>VLB M = 2.0 TeV</th>
<th>BKG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZBHB</td>
<td>WBHB</td>
<td>ZBHB</td>
</tr>
<tr>
<td>$b$-tagging</td>
<td>2.8%</td>
<td>2.8%</td>
<td>3.1%</td>
</tr>
<tr>
<td>JER</td>
<td>3.8%</td>
<td>1.7%</td>
<td>2.1%</td>
</tr>
<tr>
<td>JES</td>
<td>4.6%</td>
<td>2.0%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Large-R JES</td>
<td>2.1%</td>
<td>1.2%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Large-R JMR</td>
<td>10.0%</td>
<td>9.7%</td>
<td>16.9%</td>
</tr>
<tr>
<td>Large-R JMS</td>
<td>1.3%</td>
<td>7.7%</td>
<td>2.8%</td>
</tr>
<tr>
<td>(f)JVT</td>
<td>&lt; 0.1%</td>
<td>&lt; 0.1%</td>
<td>&lt; 0.1%</td>
</tr>
<tr>
<td>Luminosity</td>
<td>1.7%</td>
<td>1.7%</td>
<td>1.7%</td>
</tr>
<tr>
<td>MC Stats.</td>
<td>5.1%</td>
<td>5.6%</td>
<td>5.9%</td>
</tr>
<tr>
<td>VLB $p_T$ weights</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Small-R jet $p_T$ weights</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Model Stability</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>$R_{corr}$</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>ABCD-Reg. B Stats</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>12.6%</strong></td>
<td><strong>13.1%</strong></td>
<td><strong>19.0%</strong></td>
</tr>
</tbody>
</table>

Table 8.2: Relative effect of each group of systematics on the yields of the predicted background and two simulated signals with resonance mass respectively equal to 1.2 and 2.0 TeV, $\kappa = 0.5$. 
Chapter 9

Statistical Analysis

9.1 Statistical Fit Outline

The statistical interpretation of the results is carried out working under a frequentist approach by establishing a binned maximum likelihood (ML) fit on the observed distribution of the VLB system reconstructed invariant mass $m_B$.

The binned likelihood function constructed in this search can be seen as the joint probability of detecting, for each bin $i$ of the $m_B$ template, the number of observed data events $N_i$ given the expected contribution from the Standard Model background and the hypothetical signal, which can be written as $\mu s_i + b_i$.

The $\mu$ parameter, referred to as signal strength, regulates the magnitude of the contribution of the signal in units of the theoretical cross section, and within the fitting operation can be either left free and unconstrained as the main parameter of interest or be preemptively fixed to a specific value to test the statistical significance of the corresponding hypothesis. The background-only hypothesis, which is typically taken as the null hypothesis against which other hypotheses are tested, corresponds to $\mu = 0$.

In addition to the signal strength $\mu$, the prediction for the observed data in each bin depends on the values of a range of nuisance parameters $\vec{\theta}$, which reflect the potential alteration in the comprehensive data prediction owed to the systematic un-
certainties affecting either the signal or background models.

For each systematic uncertainty, the relative nuisance parameter is defined out of the provided "up" and "down" variated $m_B$ templates (see Chapter 8) as:

$$h_{\text{sysr}}(m_B, \theta) = h_0(m_B) + \theta \times \begin{cases} h_+(m_B) - h_0(m_B) & \text{if } \theta > 0 \\ h_-(m_B) - h_0(m_B) & \text{if } \theta < 0 \end{cases}$$

(9.1)

where $h_0$, $h_+$ and $h_-$ are respectively the nominal, "up" and "down" $m_B$ histograms for the sample and systematic under study.

While the value of the parameter of interest $\mu$ is typically left unconstrained, interpreting the "up" and "down" systematic variations as $\pm 1\sigma$ bounds on the effect of the relevant source on the nominal template suggests a constraint be added to the value of the corresponding nuisance parameter $\theta$. This is done adding a Gaussian term to the likelihood function for every nuisance parameter involved in the search, unless explicitly specified otherwise. The statistical uncertainty on the Monte Carlo simulations and the bin-by-bin uncertainty on the background model propagated from the statistical uncertainty in ABCD-Region B stand as notable exceptions of the Gaussian constraint rule: given the originally statistical nature of such uncertainties, the constraint is set to follow the Poisson distribution. To underline the difference between the Gaussian constrained and the bin-wise Poisson-constrained nuisance parameters, the latter are henceforth codified as "gamma" parameters $\vec{\gamma}$.

The two uncertainties on the background model associated to the kinematic reweighting procedure (see Sec. 8.3) are applied enforcing a full correlation scheme across the bins of the final observable $M_B$, and are therefore regulated by a single nuisance parameter each. An argument can be made that, since the kinematic weights ultimately arise from ratios of binned templates, the resulting uncertainties on each kinematic bin of the final observable should be treated as uncorrelated, much in the same way as the propagated statistical uncertainty on ABCD-Region B data. Nevertheless, a measure of cross-bin correlation is introduced by the kernel regression algorithm (Sec. A.1) employed to reduce the impact of statistical fluctu-
ation on the weights. An exact treatment of the cross-bin correlation for such uncertainty would require a full and cumbersome unfolding of the regression-induced correlations as well as the correlation between bins of the input reweighting variable templates and those of the final observable. A simple impact assessment of the effect of neglecting such effects is instead preferred. The two extreme schemes, which in turn assume $M_B$ bin uncertainties to be fully correlated or uncorrelated, are observed to produce exclusion limits on the resonance cross section differing by 3%, highlighting how a precise treatment of the cross-bin correlation in the reweighting uncertainties would have at most a minimal impact on the analysis results.

Given all these considerations, the full likelihood function is written as:

$$L(\mu, \bar{\theta}, \bar{\gamma}) = \prod_{i=0}^{N_{\text{bins}}} \left[ P(N_i \mid \mu s_i(\bar{\theta}_s) + b_i(\bar{\theta}_b)) \times P(\bar{\gamma}_i \mid \gamma_i) \right] \times \prod_{k=0}^{N_{\theta}} \left[ N(\bar{\theta}_k \mid \theta_k) \right]$$  (9.2)

in which:

- $P$ and $N$ represent respectively the Poisson and Normal probability density functions (PDFs);
- $N_i, s_i$ and $b_i$ are respectively the observed data events and the predicted signal and background contributions for the $i$-th $m_B$ bin;
- $\mu$ is the signal strength and parameter of interest (POI);
- $\bar{\theta}_s$ and $\bar{\theta}_b$ are the arrays of Gaussian-constrained nuisance parameters respectively regulating the impact of the signal and background systematic uncertainties in the fit;
- $\gamma_i$ are the Poisson constrained nuisance parameters modelling the uncertainties on each bin propagated from the statistical uncertainties on either the Monte Carlo simulation or the supporting ABCD-Region B dataset;
- $N(\bar{\theta}_k \mid \theta_k)$ is the Gaussian constraint applied on each nuisance parameter $\theta$;
- $P(\bar{\gamma}_i \mid \gamma_i)$ is the Poisson constraint applied to each "gamma" nuisance parameter.
The statistical fit operates as a multi-dimensional optimization problem returning the specific configuration of values for all model parameters that maximise the value of the likelihood of the observed data. The parameter values generated in this unconditional fit, which produce the maximal compatibility between the model and the data, are represented with the following notation:

\[ \hat{\mu}, \hat{\vec{\theta}}, \hat{\vec{\gamma}} \]  

Unconditional ML estimators

The likelihood maximization can also be executed setting one or more parameters to a specific, fixed value. In this scenario, the output values of the remaining parameters are referred to as conditional maximum likelihood estimates, typically referenced with the double-hat notation \(\hat{\hat{\theta}}\).

Running a conditional likelihood fit is an essential component of many procedures producing statistical inference on the data. An important example of such procedures is the ranking of the model systematics by their impact on the value of the parameter of interest \(\mu\), ultimately identifying the main limiting factors for the search in each region of the \(m_B\) phase space. This is accomplished running, for each nuisance parameter, a conditional fit in which the parameter under study is fixed to the ML value plus or minus its relative pre-fit uncertainty, and consequently evaluating the range \(\Delta \mu\) of values of the POI corresponding to the two conditional scenarios with respect to the full unconditional fit.

Conditional fits also play a dominant role in testing hypotheses on the data, where the behaviour of the likelihood is evaluated in two or more scenarios defined by specific values of the fit parameters, such as the null hypothesis \(\mu = 0\). The following subsections provide a more detailed account of how statistically supported information on the significance of an excess or exclusion limits on the production cross section can be drawn through dedicated instances of hypothesis testing.
9.1. Statistical Fit Outline

9.1.1 Significance of an excess

A claim of discovery of a new physics signal in the form of an excess in data must be substantiated by a statistically robust rejection of the background-only hypothesis, which follows from $\mu = 0$. In order to do so, it is necessary to define a test statistic with high resolving power with respect to such hypothesis. Let the profile likelihood $q_0$ [104] be defined as:

$$q_0 = \begin{cases} -2 \log \frac{L(\text{data} | 0, \hat{\theta})}{L(\text{data} | \hat{\mu}, \hat{\theta})} & \text{if } \hat{\mu} > 0 \\ 0 & \text{if } \hat{\mu} \leq 0 \end{cases} \quad (9.3)$$

For positive values of the unconditional POI estimate (corresponding to an excess, rather than a deficit), $q_0$ is defined as the logarithm of the ratio between the conditional likelihood in the background-only hypothesis and the unconditional likelihood of the observed data.

In the presence of a conspicuous excess, the background-only $\mu = 0$ hypothesis is highly penalized, and $q_0 \gg 1$, while for a small excess, where the unconditional POI estimate $\hat{\mu}$ is close to 0, the likelihood ratio approaches unity, with consequently $q_0 \approx 0$.

The statistical significance of the null hypothesis is quantified by computing the $p$-value $p_0$ corresponding to the measured value $q_0^{\text{obs}}$, which is defined as the probability of observing a value of $q_0$ equally or more unfavourable to the null hypothesis than $q_0^{\text{obs}}$.

$$p_0 = P(q_0 \geq q_0^{\text{obs}} | \text{null}) = \int_{q_0^{\text{obs}}}^{\infty} f(q_0 | \text{null}) dq_0 \quad (9.4)$$

where $f(q_0 | \text{null})$ is the probability density distribution of $q_0$ under the null hypothesis. For a fit with a single POI, the value of $p_0$ can be calculated explicitly under Wald’s approximation [135] for which $q_0$ is distributed as a chi-squared distribution with one degree of freedom.
The $p$-value of the background-only hypothesis is then converted into the *significance* $Z$ of the observed excess, defined out of the Gaussian distribution in such a way that:

$$p_0 = \int_{Z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2} dx \quad \rightarrow \quad Z = \Phi^{-1}(1-p) \quad (9.5)$$

The strength of the claim for the presence of a new signal is directly reflected by the measured value of the significance $Z$, with the conventional threshold for a discovery being set to $Z = 5$, corresponding to approximately a 1 in 3.5 million probability of $q_0 \geq q_0^{\text{obs}}$ under the null hypothesis.

### 9.1.2 Limit setting setup

Regardless of whether or not an excess is observed, the analyzer may want to assess the lowest value of the signal strength $\mu$ that can be excluded given the observed data and a specific confidence level. In a search for new or unobserved physics, this result can then be easily scaled into an *upper limit* for the production cross section of the signal which is not ruled out by the data.

At the basis of the limit-setting procedure lies the generalized form of the $q_0$ test statistic introduced in the excess significance computation, this time defined as the log-ratio between the conditional likelihood for a generic POI value $\mu$ and the unconditional likelihood:

$$q_\mu = \begin{cases} 
-2 \log \frac{L(\text{data} | \mu, \hat{\theta})}{L(\text{data} | \hat{\mu}, \hat{\theta})} & \text{if } \mu \geq \hat{\mu} \\
0 & \text{if } \mu < \hat{\mu} 
\end{cases} \quad (9.6)$$

Similarly to $q_0$, large values of $q_\mu$ suggest incompatibility between the examined hypothesis and the data, while for $\mu \sim \hat{\mu} \rightarrow q_\mu \sim 0$.

The *profile likelihood* estimator $q_\mu$ is engineered to be zero whenever the scanned value of $\mu$ is smaller than the unconditional ML estimator $\hat{\mu}$, so that in the presence of a greater excess in data than would be generated by the chosen value $\mu$, the
profile likelihood $q_\mu$ does not become large, incorrectly suggesting incompatibility between the data and the hypothesis.

The exclusion of a specific range of values of the POI $\mu$ given the observed $q^{\text{obs}}_\mu$ is carried out following the standard $CL_s$ frequentist approach [55], whereby a signal of strength $\mu$ is excluded at the $1 - \alpha$ confidence level (CL) if:

$$CL_s = \frac{p_{s+b}}{1-p_b} < \alpha \quad (9.7)$$

where $\alpha$ is the exclusion threshold, typically set at 5%, and $p_{s+b}$ and $p_b$ respectively the $p$-values corresponding to $q^{\text{obs}}_\mu$ respectively under the signal+background and background-only hypotheses:

$$p_{s+b} = \mathcal{P}( q_\mu < q^{\text{obs}}_\mu | s+b) \quad \text{and} \quad p_b = \mathcal{P}( q_\mu > q^{\text{obs}}_\mu | b-only) \quad (9.8)$$

### 9.1.3 Asimov Dataset and Expected Limits

The limit setting procedure outlined above allows to produce an estimate of the exclusion limit on $\mu$ well before the Signal Region data is available and unblinded. This is accomplished by writing out the likelihood and the $q_\mu$ test statistic for a set of fake data defined as the expected background-only contribution with all the nuisance parameters set to 0 (or the otherwise defined value corresponding to the Nominal template). This set of fake data is commonly referred to as an Asimov Dataset [104], and allows to estimate the expected exclusion threshold for $\mu$ given the signal and background models and their relative systematic uncertainties.

Running a profile likelihood test on an Asimov Dataset also produces information on the variance of the distribution of $\mu$ given the background only hypothesis, which is then converted to a $\pm 1\sigma$ confidence band about the expected exclusion limit for the POI accounting for the expected level of fluctuation of the real-data estimator $\hat{\mu}$ in the signal-less scenario.
9.2 Interpretation Strategy

As described in Chapter 2, and further discussed in Section 5.2.2, the production cross section and the experimental signature for the standalone VLB production depend on the mutual interplay of the phenomenological coupling constant regulating the interaction between the B and the W/Z and Higgs bosons:

\[ c_W = \kappa \times \sqrt{\frac{2\xi_W}{\rho_W}} \]  \hspace{1cm} (9.9)

\[ c_Z = \frac{M_Z}{M_W} \kappa \times \sqrt{\frac{2\xi_Z}{\rho_Z}} \]  \hspace{1cm} (9.10)

\[ c_H = \frac{1}{2} \frac{gM_B}{M_W} \kappa \times \sqrt{\frac{2\xi_H}{\rho_H}} \]  \hspace{1cm} (9.11)

where \( \xi_V \) is a dimensionless theoretical quantity determining the coupling hierarchy and \( \rho_V \) an equally dimensionless kinematic factor approximately equal to one for \( M_B \sim 1 \) TeV and above. It can be appreciated that all three couplings \( c_V \), within a given theoretical scenario for which \( \xi_V \) is fixed, scale with the same coupling strength \( \kappa \), so that ultimately both the production cross section and the decay width only depend on the critical parameter \( \kappa \).

The \( \kappa \)-rewighting mechanism implemented at the matrix element level of the sample generation process allows to predict the behaviour of the signal over a grid of \( (M_B, \kappa) \) configurations and likewise extract \( \kappa \)-specific 95\% CL\(_S\) exclusion limits on the signal strength \( \mu \) for a given resonance mass.

Given this framework, the results of the analysis can be presented in a number of forms, each highlighting a specific facet of the wider, multidimensional exclusion picture on the theoretical scenarios.
9.2. Interpretation Strategy

9.2.1 Mass-dependent limits on the production cross section

The most immediate level at which the search results can be presented follows the customary approach in ATLAS searches for new massive resonances, whereby the 95% CL exclusion limit on the production cross section times branching ratio is plotted against the mass of the resonance and compared with the theoretical value of $\sigma \times BR$.

Given that both the limit and $[\sigma \times BR]_{\text{theo}}$ depend on the theoretical scenario (e.g. multiplet state and value of $\kappa$), one dimensional limit vs. mass plots are generated for a range of notable values of $\kappa$ and a specific isospin multiplet state.

Under this picture, any resonance mass for the given value of $\kappa$ for which $\sigma_{\text{Lim}} \leq [\sigma \times BR]_{\text{theo}}$ is considered excluded at the 95% CL.

9.2.2 Mass-dependent limits on the coupling

Given that the analysis has access to a broad range of values of the coupling strength $\kappa$, it is natural to offer a comprehensive picture of the values of $\kappa$ that the search is sensitive to for each hypothetical resonance mass for a given multiplet state.

For each mass hypothesis, the expected/observed excluded value of $\kappa$ is the one that satisfies:

$$\sigma(M_B, \kappa)_{\text{Lim}} = \left[ \sigma(M_B, \kappa) \times BR[B \to bH]_{\text{theo}}(\kappa) \right]_{\text{theo}}$$ (9.12)

where the branching ratio is assumed to be largely independent from the resonance mass above the 1 TeV threshold [5]. Likewise, the expected $\pm 1(2)\sigma$ bands on the expected $\kappa$ limit are computed out of the Asimov dataset as the solution of:

$$\sigma^{\pm 1\sigma}(M_B, \kappa^{\pm 1\sigma})_{\text{Lim, Asimov}} = \left[ \sigma(M_B, \kappa^{\pm 1\sigma}) \times BR[B \to bH]_{\text{theo}}(\kappa^{\pm 1\sigma}) \right]_{\text{theo}}$$ (9.13)
9.2. Interpretation Strategy

9.2.3 Limits on theoretical model features

The interpretation protocols outlined in the previous sections, while powerful and intuitive, all require assuming a specific theoretical configuration under which the exclusion limits are then presented.

In this section, two further interpretation methods are outlined which allow a more comprehensive picture of the sensitivity of the analysis with respect of a broader range of theoretical scenarios, of which the previously discussed singlet and doublet isospin states are simply well motivated specific cases.

The first method proposes to display the smallest excluded resonance mass on a grid of values of two of the three $c_V$ couplings, given a specific relation defining the third (typically $c_H$) to one of the two grid couplings. This can be done by computing, for each point in the $(c_W, c_Z)$ plane, the resulting value of $\kappa$. Re-defining for ease of handling the coupling constants $c_V$ in the reduced $\bar{c}_V$ form as:

\begin{align}
\bar{c}_W &= c_W = \kappa \times \sqrt{\frac{2\xi_W}{\rho_W}} \\
\bar{c}_Z &= \frac{M_W}{M_Z} c_Z = \kappa \times \sqrt{\frac{2\xi_Z}{\rho_Z}} \\
\bar{c}_H &= \frac{2M_W}{gM_B} c_H = \kappa \times \sqrt{\frac{2\xi_H}{\rho_H}}
\end{align}

and given that $\xi_W + \xi_Z + \xi_H = 1$ by definition, allows to express $\kappa$ as a simple sum:

$$\kappa^2 = \frac{1}{2} \left[ \bar{c}_W^2 \cdot \rho_W + \bar{c}_Z^2 \cdot \rho_Z + [\bar{c}_H(\bar{c}_Z)]^2 \cdot \rho_H \right]$$

With $\rho_V$ approximately equal to one and $c_H$ being bound to the value of $c_Z$, $\kappa$ is uniquely determined throughout the entire $(c_W, c_Z)$ plane. Consequently, for each point of the plane, the $(M_B, \kappa)$ expected/observed limit grid is queried, and the smallest mass excluded for the relevant $\kappa$ is plotted on the z-axis, typically rendered by a chromatic scale in a 2D-plot.
The second interpretation method aims at producing a 2-Dimensional limit visualization showcasing the excluded resonance mass as a function of two key features of the VLB model phenomenology: the branching ratio of the $B$ into one of the three allowed decay modes and the total decay width of the resonance.

Given the reduced $\tilde{c}_V$ couplings, the branching ratio of the $B$ into the generic boson $A$ takes the form [51]:

$$\text{BR}[B \to qA] = \frac{\tilde{c}_A^2 \rho_A}{\tilde{c}_W^2 \rho_W + \tilde{c}_Z^2 \rho_Z + \tilde{c}_H^2 \rho_H}$$  \hspace{1cm} (9.18)$$

while the total decay width can be written as: [51]:

$$\Gamma(B) = \frac{g^2 M_B^3}{128\pi M_W^2} \times \left[ \tilde{c}_W^2 \rho_W^2 + \tilde{c}_Z^2 \rho_Z^2 + \tilde{c}_H^2 \rho_H^2 \right]$$  \hspace{1cm} (9.19)$$

Once again, Eqn.9.17 ensures that, for a given pair of $\Gamma \times \text{BR}$ values, $\kappa$ is uniquely determined for a given $c_Z \leftrightarrow c_H$ relation, so that the lowest excluded mass for each configuration can be extracted from the experimental limit grid in the $(M_B, \kappa)$ plane.
Chapter 10

Results

10.1 Unblinded Signal Region results

The distribution of the final observable $m_B$ in the Signal Region for 2015-2018 data, compared with the Standard Model background contribution as predicted according to the modelling procedure discussed in Chapter 7, is displayed in Figure 10.1. The expected $m_B$ distribution for a hypothetical 1.3 TeV, $\kappa = 0.3$ signal is also displayed for comparison.

A total of 262 events are observed in the Signal Region, in full compatibility with the Standard Model contribution to the dataset, predicted at 257 events with a total systematic uncertainty of 25 events.

The behaviour of several other kinematic variables of relevance in the Signal Region is displayed in Figure 10.2. No significant indication of mis-modelling is observed in the data/background comparison.

The largest deviation between the data and predicted background is detected in the kinematic bin defined by $1.25 < m_B < 1.4$ TeV, in which 67 events are observed out of the $51 \pm 6$ expected from Standard Model sources. The magnitude of the excess is roughly compatible with a possible contribution from a 1.3 TeV, $\kappa = 0.3$ signal with signal strength set to the nominal value of one. While the effective occurrence of a VLB signal under this $(M_B \times \kappa)$ configuration cannot be ruled out on the
10.1. Unblinded Signal Region results

Figure 10.1: Distribution of the VLB system invariant mass in the Signal Region for the predicted multi jet contribution (solid black line) and the observed full Run 2 data (dots). The shaded area about the multi jet prediction represents the quadrature sum of all systematic uncertainties affecting the background. The $m_B$ distribution for a 1.3 TeV $\kappa = 0.3$ VLB signal is overlaid to both the main distribution and the ratio plot as a red dashed line.

grounds of the observed data (as will be discussed in Sec. 10.3), the year-by-year breakdown of the data highlights that the excess with respect of the background is compatible with originating from an upward bin content fluctuation in the 2017 data alone (Fig. 10.3). The statistical significance of the excess in the aforementioned bin, computed as described in Sec. 9.1.1, is found to be $1.85\sigma$. Such value is smaller than the conventional $3\sigma$ threshold typically required to claim evidence of new phenomena, even without taking into account the further significance reduction brought about by correcting for the look elsewhere effect [136].
10.1. Unblinded Signal Region results

Comparison between data and Standard Model prediction for the Signal Region data for a number of key kinematic variables. The comprehensive systematic uncertainty on the background prediction is represented as the shaded band. The expected contribution from a 1.3 TeV, $\kappa = 0.3$ signal is overlaid for comparison.
10.1. Unblinded Signal Region results

Figure 10.3: Year by year breakdown of the observed data and the predicted Standard Model contribution. The $1.25 \text{ TeV} < m_B < 1.4 \text{ TeV}$ bin is highlighted by the red transparent bar.
10.2 Statistical Fit Results

10.2.1 Background-Only Fit

A background-only fit to the data is carried out in the form of a conditional maximum likelihood fit with the parameter of interest $\mu$ set to 0. This operation is set to assess the compatibility of the data with the background-only hypothesis by examining the output values of the nuisance parameters $\vec{\theta}$ and $\vec{\gamma}$, and identifying a level of tension between data and model in excessive or widespread pulls with respect to the nominal background configuration given by $\vec{\theta} = 0$ and $\vec{\gamma} = 1$.

Figure 10.4 displays the comparison between the observed data and the predicted Standard Model background before (left) and after (right) the fit. The output values of the nuisance parameters representing the impact of the signal and background systematics is displayed in Figure 10.5. None of the nuisance parameters regulating the background contribution ($r_{\text{Corr}_H2T2B}$, $c_{\text{H2T2B}}$, $V_{\text{ptModel}_H2T2B}$ and $J_{\text{ptModel}_H2T2B}$) is observed to be pulled away from the nominal zero-point in a significant manner, attesting the good global agreement between expected and observed data. The pull profile of the gamma parameters reflect the presence of a small localized excess in the fourth bin of the invariant mass distribution, with the corresponding gamma nuisance parameter pulled at approximately one sigma, for an approximate 5% increase in the expected bin content.

Table 10.1 provides a summary of the pre- and post-fit yields for signal and background in the background-only fit hypothesis.
### 10.2. Statistical Fit Results

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pre-Fit</th>
<th>Post-Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLB, 1.3 TeV, $\kappa = 0.3$</td>
<td>29.5 ± 4</td>
<td>/</td>
</tr>
<tr>
<td>Background estimation</td>
<td>257 ± 25</td>
<td>260 ± 17</td>
</tr>
<tr>
<td>Data</td>
<td></td>
<td>262</td>
</tr>
</tbody>
</table>

**Table 10.1:** Pre- and post-fit yields for signal and background in the background only hypothesis. The post-fit signal yield is accordingly zero by definition (conditional fit with $\mu = 0$).

**Figure 10.4:** Distribution of $m_B$ for the observed data and the background prediction before (left) and after (right) the fit to the data. The Z-Initiated (leading) and W-Initiated (sub-leading) contributions to a VLB signal with $M_B = 1.3$ TeV and $\kappa = 0.3$ are overlaid for reference.
10.2. Statistical Fit Results

Figure 10.5: Background-only pulls for the nuisance parameters regulating the expected shape and yields of the background contribution.
Table 10.2 displays the unconditional best fit results for the signal strength $\mu$ in a number of $M_\beta \times \kappa$ configurations. For each configuration, the uncertainty provided on $\mu$ represents the quadrature sum of the statistical and systematic contributions.

### 10.2.2 Unconditional Signal Strength Estimates

Table 10.2 displays the unconditional best fit results for the signal strength for a number of $(M_\beta \times \kappa)$ combinations. The values of $\hat{\mu}$ corresponding to the signal signatures most compatible with the localized excess in data are observed to be close to unity, and not compatible with zero within one sigma.

### 10.2.3 Systematics Ranking

As described in Section 9.1, the impact of each signal or background systematic on the unconditional Maximum-Likelihood estimate of the signal strength $\mu$ is evaluated running, for each resonance mass hypothesis, an array of conditional fits in
which the nuisance parameter under study is fixed to \( \theta_{\text{rank}} = \hat{\theta} \pm \Delta \theta \), while no condition is imposed on the remaining parameters. As the purpose of such fits is to determine the extent to which the correlation \( \Delta \theta \rightarrow \Delta \mu \) holds, the value of \( \mu \) is allowed to float freely.

The diagnostic plots showcasing the 10 most impactful systematics for each signal mass hypothesis are grouped in Figure 10.6. The impacts on \( \mu \) are represented by the blue and teal band, which are to be measured against the top axis scale. The dominant factor affecting the value of \( \mu \) is found to be the \( \gamma \) parameter regulating the first bin fluctuation for the 1.0 TeV hypothesis, the flat \( R_{\text{corr}} \) uncertainty for the 1.2 TeV hypothesis and the uncertainty on the small-R jet \( p_T \) shape correction for higher masses.

### 10.3 Observed Limits

Given the limit-setting statistical framework outlined in Section 9.1.2, as well as the result interpretation strategy discussed in Section 9.2, observed and expected limits were set on a number of features describing the wider theoretical picture for singly-produced VLB, as shall be presented and discussed in the following sections.

#### 10.3.1 Exclusion Plots: Cross Section

Figure 10.7 collects the one-dimensional exclusion plots on the production cross section times branching ratio of a Vector-Like B quark occurring as an isospin singlet with mass ranging from 1.0 to 2.0 TeV. Each plot refers to an individual value of the signal strength \( \kappa \). The values of the resonance mass that for a given \( \kappa \) are excluded at the 95\% CLs can be inferred as the range of the x-axis where the ex-

\(^{1}\)If the pre-fit uncertainty on the nuisance parameter is used, the systematic rankings are labelled "pre-fit", while the label "post-fit" corresponds to using the post-fit uncertainty on the relevant parameter.
10.3. Observed Limits

(a) $M_B = 1.0$ TeV

(b) $M_B = 1.2$ TeV

(c) $M_B = 1.4$ TeV

(d) $M_B = 1.6$ TeV

(e) $M_B = 1.8$ TeV

(f) $M_B = 2.0$ TeV

Figure 10.6: Ranking plots for the 10 leading systematics by impact on the signal strength $\mu$ in each resonance mass hypothesis. The blue and teal represent the impact (to be read against the top axis scale) on the value of $\mu$ of respectively a $+1\sigma$ and $-1\sigma$ pull of each nuisance parameter. The nuisance parameter pulls obtained within each conditional fit are represented by the position of the dots, to be evaluated against the bottom axis scale. The $\gamma$ nuisance parameter have a nominal value of 1.
10.3. Observed Limits

Expected/observed limit on the cross section is smaller than the theoretical value.

Table 10.3 provides a full breakdown of the expected/observed exclusion point for each \( \kappa \) scenario. The presence of an excess in the observed data over the predicted background in the \( 1.25 \text{ TeV} < m_B < 1.4 \text{ TeV} \) is reflected in the lower observed exclusion mass point for the values of \( \kappa \) most compatible with a narrow resonance.

<table>
<thead>
<tr>
<th>( \kappa )</th>
<th>Expected Exclusion</th>
<th>Observed Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>0.30</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>0.35</td>
<td>[1100, 1285] GeV</td>
<td>[1060, 1200] GeV</td>
</tr>
<tr>
<td>0.40</td>
<td>[1040, 1430] GeV</td>
<td>[1000, 1220] GeV</td>
</tr>
<tr>
<td>0.45</td>
<td>[1000, 1480] GeV</td>
<td>[1000, 1260] GeV</td>
</tr>
<tr>
<td>0.50</td>
<td>[1000, 1580] GeV</td>
<td>[1000, 1300] + [1580, 1800] GeV</td>
</tr>
<tr>
<td>0.60</td>
<td>[1000, 1700] GeV</td>
<td>[1000, 1860] GeV</td>
</tr>
<tr>
<td>0.70</td>
<td>[1000, 1780] GeV</td>
<td>[1000, 1870] GeV</td>
</tr>
<tr>
<td>0.80</td>
<td>[1000, 1820] GeV</td>
<td>[1000, 1880] GeV</td>
</tr>
<tr>
<td>0.90</td>
<td>[1000, 1860] GeV</td>
<td>[1000, 1940] GeV</td>
</tr>
</tbody>
</table>

Table 10.3: Summary table displaying the expected and observed mass intervals that are excluded at the 95\% CLs for each value of the signal strength \( \kappa \).
10.3. Observed Limits

Figure 10.7: Exclusion plots on the production cross section times branching ratio for a number of \( \kappa \) hypotheses. Expected (dashed line) and observed (solid line) 95% \( CL_s \) limits are compared with the theoretical value of \( \sigma(pp \rightarrow B \rightarrow bH) \), whose dependence on the resonance mass is conveyed as the red line.
10.3.2 Exclusion Plots: Phenomenology

Figure 10.8 displays the expected and observed exclusion limit on the phenomenological coupling $c_W$ defined as:

\[
  c_W = \kappa \sqrt{\frac{2\xi_W}{\rho_W}}
\]

(10.1)

For a Vector-Like B quark occurring as an isospin singlet with resonance mass above 1.0 TeV, $\xi_W = 0.5$ and $\rho_W \simeq 1$, which implies that $c_W \simeq \kappa$ to a very good approximation. For a given mass, the values of $c_W$ excluded at the 95% CL can be inferred as the range on the $y$-axis above the expected/observed exclusion line.

The broadening $\pm 1(2)\sigma$ confidence intervals about the expected exclusion limit in the higher range of the $M_B$ spectrum reflects the peculiar behaviour of the $pp \rightarrow B \rightarrow bH$ production chain for large-width resonances, whereby most of the events are produced off-shell with actual mass below the resonance pole mass (see Sec. 5.2.3 and Fig. 5.4). This, in turn, translates into most of the large-width events falling outside the acceptance of the analysis trigger, and results in an approximately constant experimental signature in the Signal Region for a broad range of values of $\kappa$, causing a very low expected resolving power on the coupling strength above $M_B \sim 1.5$ TeV, as evidenced in the exclusion plot.

Figure 10.9 shows on the color axis the expected (top) and observed (bottom) exclusion limit on the VLB mass for a given configuration of the two couplings $c_W$ and $c_Z$, with the third reduced coupling $\tilde{c}_H = \tilde{c}_Z$ set as the only theoretical assumption behind this exclusion limit presentation. The $\tilde{c}_Z = \tilde{c}_H$ assumption is well motivated theoretically, with the singlet, (T,B) and the (B,Y) doublets arising as specific cases obtained by setting respectively:

\[
  \tilde{c}_Z = \tilde{c}_H = \kappa / \sqrt{2} \rightarrow c_W = \tilde{c}_W = \kappa \quad \text{B singlet}
\]

(10.2)
Figure 10.8: Expected and observed exclusion limit on the $c_W$ coupling for a VLB singlet.

\[
\tilde{c}_Z = \tilde{c}_H = 0 \rightarrow c_W = c_W = \sqrt{2} \cdot \kappa \quad \text{(T,B) doublet} \quad (10.3)
\]

\[
\tilde{c}_Z = \tilde{c}_H = \kappa \rightarrow c_W = c_W = 0 \quad \text{(B,Y) doublet} \quad (10.4)
\]

\[
\tilde{c}_Z = \tilde{c}_H = \kappa / \sqrt{2}
\]

The solid black line overlaid to the exclusion results represents the position on the wider $(c_W, c_Z)$ phase space of all singlet configurations, which arise from the specific constraint: $c_Z = \frac{M_Z}{M_W} \frac{1}{\sqrt{2}} \times c_W$.

The apparent discontinuity in the mass exclusion axis in the observed limit is understood to originate from the behaviour of the observed data, which displays a moderate excess around 1.3 TeV and a small deficit from 1.6 TeV onwards. This causes the observed limit to fall short of the expected limits in the region where the latter was computed to be between 1.3 and 1.5 TeV, while for larger masses the search excludes larger resonance masses than was predicted with a fit on the Asimov dataset.
10.3. Observed Limits

Figure 10.9: 95% CL$_{s}$ exclusion limit on the resonance mass for the generic $c_W, c_Z$ combination. $\tilde{c}_Z = \tilde{c}_H$ is assumed throughout. The beige-coloured areas refer to combinations resulting in a value of $\kappa$ which either lies outside the available range or for which no mass exclusion is. All singlet configurations, which are obtained for $c_Z = \frac{M_Z}{M_W} \frac{1}{\sqrt{2}} \times c_W$, lie on the overlaid solid black line.

Finally, the expected and observed exclusion limits for the available values of $\kappa$ on the VLB mass are recast in terms of its total fractional decay width and its branching ratio into one of the three allowed decay channels. As explained in Sec. 9.2, both
quantities depend on the square sum of all the three reduced $\tilde{c}_V$ couplings. Once again, $\tilde{c}_Z = \tilde{c}_H$ serves as the sole underlying theoretical assumption for this visual interpretation of the results. Figure 10.10 shows the expected (left) and observed (right) mass limits as a function of the fractional B decay width and its branching ratio into either the tW channel. The $\Gamma/M_B$-dependent mass limit in the $B$ singlet scenario can be obtained drawing a vertical line in each plot corresponding to the value of the branching ratio into the relevant channel for a $B$ singlet: $BR[B \to tW] = 0.5$ and $BR[B \to bH] = 0.25$. 

**Figure 10.10:** Expected and observed limit on the VLB mass as a function of the branching ratio into the tW decay channels and the fractional decay width.
10.4 Comparison with the state of the art

The purpose of this section is compare the results of this search with the established knowledge on Vector-Like B quarks accumulated in previous published efforts at both ATLAS and CMS. A overview of the state of the art in VLB searches at the time of writing is provided in Section 2.3.

Given the comparatively complex theoretical picture of Vector Like Quarks, with critical observables such as the resonance width, production cross section and branching ratios depending on the theoretical scenarios, comparing results among different analyses or collaborations is often not trivial.

For the purpose of offering the fairest assessment of the outcome of the present analysis, the results were recast to match, whenever possible, the theoretical scenarios considered in each of the past searches that shall be discussed.

10.4.1 Single Production Searches

As of Autumn 2020, two searches for evidence of single-production of a Vector-Like B in LHC Run 2 data have been published:

- search for VLB decaying to the $B \rightarrow bH(\gamma\gamma)$ final state, an early-to-mid Run 2 (80 fb$^{-1}$, corresponding to the 2015, 2016 and 2017 datasets) ATLAS search [64];

- search for VLB decaying to the $B \rightarrow bH(b\bar{b})$ final state, a early-Run 2 (36 fb$^{-1}$, corresponding to the 2015 and 2016 datasets) CMS analysis [16] targeting the same final state as the search covered in this thesis.
10.4. Comparison with the state of the art

**ATLAS Search, \( B \to bH(\gamma\gamma) \)**

The ATLAS \( B \to bH(\gamma\gamma) \) search elected to set mass-dependent limits on the \( B \to bH \) cross section times branching ratio for \( \kappa = 0.5 \) and in the scenario where the \( B \) occurs as part of a weak iso-doublet along with the exotic Vector-Like \( Y \).

As outlined in Sec. 9.2, the doublet scenario is translated onto the simplified phenomenology by setting \( \xi_Z = \xi_H = 0.5 \) and consequently \( \xi_W = 0 \). For reference, the singlet scenario corresponds to \( \xi_w = 0.5 \) and \( \xi_Z = \xi_H = 0.25 \). The immediate implications of the doublet assumption are threefold:

- \( \xi_W = 0 \) implies \( c_W = 0 \), corresponding to a vanishing coupling between the \( B \) and the \( W \) boson. This in turn restricts the production channels to the sole \( bZ \to B \), and at the same time induces \( BR[B \to tW] = 0 \);

- having set \( \xi_H = \xi_Z \), the two channels will evenly share the total \( B \) width, which is left unaltered\(^2\), and have a 50% branching ratio each.

- setting \( \xi_Z^{\text{doub.}} = 0.5 = 2\xi_Z^{\text{sing.}} \) implies \( \tilde{c}_Z^{\text{doubl.}} = \sqrt{2}\tilde{c}_Z^{\text{sing.}} \). As the production cross section scales with the squared reduced couplings, \( \sigma(pp \to B)_{\text{doubl.}} \approx 2\sigma(pp \to B)_{\text{sing.}} \). Given that \( BR_H^{\text{doubl.}} = 2 \cdot BR_H^{\text{doubl.}} \), **four times** more \( B \to bH(b\bar{b}) \) events are expected in the doublet than in the singlet scenario, for a given mass and \( \kappa \).

The four times enhancement of the total \( \sigma(pp \to B \to bH) \) cross section for single VLB production into the relevant channel under the \( (B,Y) \) scenario translates in the expected and observed exclusion limits on the resonance mass being pushed towards higher values than in the singlet state.

However, as no significant differences are predicted to exist in signal event kinematics between the two scenarios, the excluded limit on the channel production\(^2\)

\(^2\)As stated in Eqn. 9.19, and given the definition of the \( \tilde{c}_V \), the total decay width is proportional to \( \kappa^2 \times (\xi_W + \xi_Z + \xi_H) \), which does not depend on the individual values of \( \xi_V \) defining the multiplet state.
10.4. Comparison with the state of the art

Figure 10.11: Expected and observed exclusion limits for a Vector-Like B occurring as part of a \((B,Y)\) weak isospin doublet extracted by the ATLAS Run 2 \(B \to bH(\gamma\gamma)\) search.

cross section, which is by construction independent of the signal normalization, is expected to be unaltered.

The \(B \to bH(\gamma\gamma)\) search calculated an expected exclusion limit on the resonance cross section (see Fig. 10.11) ranging upwards from 80 fb in the 1.0-1.6 TeV range, for \(\kappa = 0.5\). In the same range and for the same coupling strength, this search expects to exclude a production cross section greater than approximately 25 fb (Fig. 10.7b). Extrapolating the \(B \to bH(\gamma\gamma)\) limit to the full Run 2 data volume available to the present analysis, the projected \(\gamma\gamma\) limit is of the order of 60 fb, significantly inferior to the limit achieved in the present analysis.

CMS Search, \(B \to bH(\bar{b}b)\)

The CMS search for the same final state targeted by the present analysis chose to present their findings in terms of excluded \(\sigma(pp \to B \to bH)\) as a function of the resonance mass for a given value of the fractional resonance width \(\Gamma/M\).

In order for a more immediate comparison with the CMS results, the expected and observed exclusion limits of the present search on the VLB cross section
as a function of both the mass and $\Gamma/M$ are displayed in Figure 10.12. Figure 10.13 shows, on the other hand, the one-dimensional CMS exclusion limits for $\Gamma/M = 0.2$ and 0.3.

As can be appreciated comparing the two figures, the values of the cross sections excluded by the present search improve upon the CMS result by more than a factor of 5 throughout the overlapping section of the examined mass ranges (800 GeV to 1.8 TeV).
10.4. Comparison with the state of the art

Contrary to single production, the generation of a pair of Vector-Like Quarks in the final state is a purely QCD process whose cross section does not depend, for a narrow VLQ resonance, on the theoretical and phenomenological scenarios determined by the values of the $\xi_V$ and the coupling strength $\kappa$. The $\xi_V$ nevertheless affect the sensitivity of channel-specific searches by regulating the branching ratios of the VLQ in each of the three available decay modes.

Published pair-production analyses by ATLAS [54] and CMS [137] on respectively
10.4. Comparison with the state of the art

Figure 10.13: Expected and observed 95% \( CL_s \) exclusion limit on \( \sigma(pp \rightarrow B \rightarrow bH) \) extrapolated by the CMS \( B \rightarrow bH(b\bar{b}) \) search for two given values of the fractional resonance width [16].

early Run 2 (36 fb\(^{-1}\) integrated luminosity) and full Run 2 (135 fb\(^{-1}\)) have put forward interpretations of the search results in terms of 95\% \( CL_s \) limits on the resonance mass for a narrow \( B \) singlet and a \( (B,Y) \) doublet. The narrow-width approximation (NWA) ensures that the total cross section can be factorized as follows:

\[
\sigma(pp \rightarrow B\bar{B} \rightarrow \text{final state}) = \sigma(pp \rightarrow B\bar{B}) \times \text{BR}[B \rightarrow \text{final state}]
\]  

(10.5)

where \( \sigma(pp \rightarrow B\bar{B}) \) is completely model-independent and \( \text{BR}[B \rightarrow \text{final state}] \) does not depend on the coupling strength (see Eqn. 9.18).

The statistical combination of early Run 2 pair-production searches at ATLAS claims exclusion for \( B \) masses below 1.22 TeV in the singlet hypothesis and 1.15 TeV for a \( (B,Y) \) doublet. The recent CMS pair-production search, targeting both the \( B \rightarrow bH(b\bar{b}) \) and \( B \rightarrow bZ(b\bar{b}) \) final states, excluded, on the grounds of the full Run 2 dataset, resonance masses below 1.45 TeV in the \( (B,Y) \) hypothesis.

Figure 10.14 shows how the ATLAS and CMS limits on the \( B \) mass set by pair production searches compare with the singlet and doublet exclusion limits drawn in the present search, defined as the region of the \( M_B \times c_V \) space above the solid black line.

In both theoretical scenarios, the present search was able to exclude a larger-width resonance (corresponding to values of \( \kappa \geq 0.5 \) for singlet and \( \kappa \geq 0.35 \) for dou-
10.4. Comparison with the state of the art

![Figure 10.14: Mass dependent exclusion limit on the main coupling in the $B$ singlet (left) and $(B,Y)$ doublet scenarios. For a $(B,Y)$ doublet, $c_W = 0$, requiring the limits to be expressed in terms of $c_Z = M_Z/M_W \times \kappa$. In both plots, the exclusion limits on the $B$ resonance mass drawn by the most recent ATLAS and CMS pair-production results are represented by the vertical blue and red lines.](image)

bles) for masses well above the pair-production limits, owing to the quadratic scaling of the total single-production cross section with $\kappa$, as opposed to the $\kappa$-independent $\sigma(pp \rightarrow B\bar{B})$. Conversely, both pair production results outperform the present search for narrow resonances defined by $\kappa < 0.3$ (singlet) and $\kappa < 0.2$ (doublet), for which the single production cross section is too small to produce a statistically significant peak.
10.5 Summary and discussion

The previous sections provided a comprehensive picture of the results extracted out of the observed Signal Region data by means of a binned Maximum Likelihood fit and the related statistical tests outlined in Chapter 9.

The unblinded data was observed to be in good agreement with the data-driven prediction for the Standard Model background sources, as highlighted by targeted comparisons carried out on the spectra of the most relevant kinematic variables describing the event. A small localized excess in data is observed in the reconstructed VLB invariant mass bin ranging from 1.25 TeV to 1.4 TeV. The excess is found to be most compatible with a possible Vector-Like B singlet with resonance mass $M_B = 1.3$ TeV and coupling strength $\kappa = 0.25 \sim 0.3$, for which the best-fit value of the signal strength $\mu$ is found to be $1.2^{+0.7}_{-0.7}$. Expected and observed exclusion limits on a number of critical features of the single Vector-Like quark production phenomenology were drawn, highlighting a marked improvement in the sensitivity of the search with respect to the model-independent limits on the Vector-Like B mass extrapolated by pair-production searches at both ATLAS and CMS (Sec.2.3).

Beyond the improved limits on the excluded resonance mass, the search was able to cast mass-dependent statistical limits on the critical phenomenology couplings $c_V$ under the B singlet and $(B,Y)$ doublet theoretical scenario, and finally painted a more general picture evaluating the excluded resonance mass for a wider range of combinations of $(c_W,c_Z)$ and $(BR[B \rightarrow bH], \Gamma/M)$, going beyond the benchmark singlet and doublet hypotheses. Both these operations are in accordance with the common interpretation strategy for single-production Vector-Like quark searches at ATLAS defined in view of a future combination of all relevant VLQ searches.

A critical comparison of the analysis result with the published state of the art in single-production VLB searches at both ATLAS and CMS highlighted how the present search was able to significantly expand the region of the mass-coupling
theoretical phase space for which a Vector-Like B resonance is excluded according to the available LHC data.
Appendix A

Kernel-based extraction of reweighting functions in the background estimation

A.1 Smoothing Technique

Approximating $\rho_i(x_i)$ through the ratio between two data histograms is prone to bin-by-bin statistical fluctuations. This may result in sizably different weights applied to two events very close to each other in the $x_i$ space, but falling in neighboring bins that happened to fluctuate in opposite directions.

A weight function such as this is therefore understood to be sub-optimal for continuous variables, evidencing the need for a smoothing procedure to produce a more regular estimate.

As there is no a priori indication of a particular functional form describing the ratios, a Kernel Regression, which is a non parametric smoothing approach, is preferred over a parametric functional fit.

The founding principle of Kernel Regression is to produce a smooth estimate for the true functional form $\rho(x)$ underlying the ratio data points $(x_i, r_i)$ as a weighted average of the sample points $r_i$ themselves. The weight factors must depend on both...
A.2. Treatment of Extremes

the value of the $x$-variable where the output estimator $\hat{\rho}(x)$ is evaluated and the $x_i$ of each data point in the sum. The most common form for $\hat{\rho}(x)$ was suggested by Nadaraya and Watson [128] and employs a kernel function as the weight factor:

$$W(x) \equiv \hat{\rho}(x)_{NW} = \frac{\sum r_i \cdot K\left(\frac{x-x_i}{h}\right)}{\sum K\left(\frac{x-x_i}{h}\right)}, \quad \text{with } \int K(u)du = 1 \text{ and } K(u) = K(-u)$$

(A.1)

There are several popular choices for the kernel function, from a box-kernel which has non-zero value only where $|x-x_i| < h$, to a gaussian kernel. The choice of the $h$ parameter, called bandwidth, plays an important role in determining the features of $\hat{\rho}(x)$. To understand the importance of an appropriate choice of bandwidth it is useful to consider the two limits where $h$ is very small or very large with respect to the average separation between the $x_i$: for small $h$ the kernel will have a non negligible value only for $x \sim x_i$, and therefore $\hat{\rho}$ will be pulled strongly by each ratio point $r_i$, whereas a too large value of $h$ will result in all the terms in the average to have very similar importance, transforming $\hat{\rho}$ into the arithmetic average of the $r_i$.

The choice of kernel function fell on a Gaussian kernel with an added penalty term to reduce the value of $K_i$, and therefore the pulling power of the considered $r_i$, according to the uncertainty $\Delta r_i$ associated to the ratio point. This strategy is aimed at ensuring that the estimate for the ratio is as robust as possible with respect to the statistical fluctuations in the samples from which the ratio histogram originates.

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{\Delta r_i} \cdot e^{-\left(\frac{x-x_i}{\Delta r_i}\right)^2}, \quad \text{with } h = \frac{x_{\text{max}} - x_{\text{min}}}{k}$$

(A.2)

A.2 Treatment of Extremes

Kernel-based regression estimators are known to introduce some bias at the extremes of the $x$-range targeted. In order to mitigate this effect, independent linear fits of the tail of the ratios are set up and introduced in the estimation. To ensure
continuity and smoothness in the weight function, the ultimate \( \hat{\rho} \) is defined as follows:

- For \( x \in [x_{\min}, x_{L1}] \), linear fit of the left tail, fit extended between \( x_{\min} \) and \( x_{L2} \):

- For \( x \in [x_{L1}, x_{L2}] \), linear combination of the fit of the left tail and the Kernel estimate:

\[
\hat{\rho} = \rho \cdot \hat{\rho}_{fit} + (1 - \rho) \cdot \hat{\rho}_{NW}, \quad \text{with} \quad \rho(x) = \frac{x - x_{L1}}{x_{L2} - x_{L1}}
\]

- For \( x \in [x_{L2}, x_{H1}] \), the output of the kernel regression is used;

- For \( x \in [x_{H1}, x_{H2}] \), linear combination of the kernel regression output and the fit on the right tail, extended between \( x_{H1} \) and \( x_{\max} \). The smoothing factor \( \rho(x) \) is defined in a similar manner as the previous transition region;

- For \( x \in [x_{H2}, x_{\max}] \), linear fit of the right tail.

The choice of the transition zone boundaries is calibrated in order to best conform to the shape of the variable under exam.

### A.3 Choice of bandwidth parameter

As outlined in the previous section, the effectiveness of the non-parametric regression method used in the context of the background modelling relies on a sensible choice of the bandwidth parameter \( h \).

The Nadaraya-Watson estimator:

\[
\hat{\rho}(x)_{NW} = \frac{\sum r_i \cdot K\left(\frac{x - x_i}{h}\right)}{\sum K\left(\frac{x - x_i}{h}\right)}
\]

defines, for every sampling point \( x \), the value of the regression curve as a linear combination of the experimental ratio points \((x_i, r_i)\) whose coefficients depend on
the distance between sampling and experimental point $x - x_0$ relative to the arbitrary scale $h$.

Given that common choices of Kernel functions regulating the values of the coefficients are such that $K(0)$ is maximal, very large values of $h$ will result in all coefficients being approximately identical, whereas very small values of $h$ will cause only one (or at most two) coefficients to be non-vanishing, resulting in a regression curve excessively pulled by the experimental points.

This appendix details the definition and implementation of an algorithm to estimate and minimize the regression bias as a function of the value of the bandwidth $h$ in order for the background modelling procedure outlined in Chapter 7 to run optimally.

The first and most important step of the optimization procedure is the definition of a metric to estimate the regression bias.

In order to do so, the two data samples (events with at least 1 forward jet and events with no forward activity, corresponding to ABCD-Regions C and D respectively) used to extract the ratio histograms are randomly sorted into two sub-samples each, therefore generating two sets of independent experimental ratio points $R_A = (x_i, r^A_i)$ and $R_B = (x_i, r^B_i)$.

The regression procedure is then applied in parallel to both $R_A$ and $R_B$, producing in turn two independent regression curves $\hat{\rho}_A(x)$ and $\hat{\rho}_B(x)$. From the aforementioned objects, the two-fold Cross-Validation bias estimate $2-CV$ is then calculated as follows:

1. Define $R_A$ as the training data set and $R_B$ as its statistically independent testing sample;
2. Extract the regression $\hat{\rho}_A(x)$ from the training sample $R_A$;
3. Evaluate the $A-B$ bias comparing $\hat{\rho}_A(x)$ against the testing sample;
4. Repeat points 1-3 employing $R_B$ as training and $R_A$ as testing;
5. Average the two bias estimates.

And can be written as:

\[ 2-CV = \frac{1}{2} \left( e[A,B] + e[B,A] \right) \]  \hspace{1cm} (A.3)

The \( A - B \) bias \( e[A,B] \) was taken as the chi-squared of the data points of the testing sample with respect to the regression extracted from the training sample, as defined in Eqn. A.4

\[ e[A,B] = \chi^2[R_B,\hat{\rho}_A(x)] = \frac{1}{N.D.F.} \sum_{i}^{\text{bins}} \frac{(r^B_i - \hat{\rho}_A(x_i))^2}{\Delta(r^B_i)^2} \]  \hspace{1cm} (A.4)

with \((r^B_i)\) being the y-value of the experimental points \( R_B = (x_i, r^B_i) \), \((\Delta r^B_i)\) their associated statistical uncertainty and \( \hat{\rho}_A(x_i) \) the value of the training set regression curve in the experimental points \( x_i \).

As the value of \( 2-CV \) is primarily a function of the choice of bandwidth, the critical and only parameter of the Kernel function affecting the behaviour of \( \hat{\rho}_A(x_i) \), it is possible to iterate the calculation of \( 2-CV \) for a range of bandwidths in order to identify the optimal value of \( h \) at the minimum of the \( 2-CV \) shape.

In order to ensure a measure of statistical robustness against the fluctuations in the testing and training sets arising from the stochastic sampling employed in the splitting of the original data sample, the calculation of \( 2-CV(h) \), for each value of \( h \), is repeated 200 times, each with a different random seed employed in the initial random definition of sets A and B.

The variables on which a kernel-regression-based reweighting is applied are the transverse momenta of the small-\( R \) jet and the reconstructed VLB candidate.

In order to facilitate a simultaneous review of the bandwidth optimization procedure for each of the aforementioned variables, the results are presented not as a function of the absolute bandwidth \( h \), but of a reduced bandwidth defined as the ratio \( h/b \) between the bandwidth and the bin width.
A.3. Choice of bandwidth parameter

Figure A.1: Random Iteration N. 3. Regression of the small-\( R \) jet \( p_T \) ratio distribution. Testing samples \( R_A \) (left) and \( R_B \) (right) with the three regression curves extracted from the training samples \( R_B \) (left) and \( R_A \) (right). Different regression colours correspond to different values of the relative bandwidth.

Figures A.1, A.2, A.3, A.4 show some of the regression curves, for different values of \( h/b \) overlaid to the respective training data samples, showing how the value of bandwidth affects the features and behaviour of the regression curves.

Finally, figure A.5 shows the behaviour of the averaged 2-CV error as a function of the reduced bandwidth, showing clear and compatible minima of the regression bias for values of \( h/b \) between 0.9 and 1.0.
Figure A.2: Random Iteration N. 3. Regression of the VLB $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth.

Figure A.3: Random Iteration N. 128. Regression of the small-R jet $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth.
Figure A.4: Random Iteration N. 128. Regression of the VLB $p_T$ ratio distribution. Testing samples $R_A$ (left) and $R_B$ (right) with the three regression curves extracted from the training samples $R_B$ (left) and $R_A$ (right). Different regression colours correspond to different values of the relative bandwidth.

Figure A.5: Values of the two-fold cross-validation error with respect to the reduced bandwidth $h/b$. Values are averaged over 200 iterations with stochastic splitting of the original data sample in training and testing sets.
A.4 Uncertainty on reweighting functions

An estimate for the variance \[138\] on \(W(x)\) is given by the following expression:

\[
\text{Var} [\hat{\rho}(x)] \equiv \text{Var}[W(x)] = \frac{\sigma^2(x) \times \int K^2(u)du}{h \cdot n \cdot p(x)}, \tag{A.5}
\]

with \(n\) representing the number of data points and \(\sigma^2(x)\) the point-wise variance of the \(r_i\)s, describing, for a region about \(x\), the spread of the experimental points about the true value \(r(x)\). As only one \(r_i\) is available for each \(x_i\), the only handle on \(\sigma^2(x)\) is to assume \(\sigma^2(x) \sim \Delta r_i^2\) of the closest data point to \(x\). \(p(x)\) represents the probability density function regulating the distribution of the \(x_i\)s along the \(x\)-axis. As those are equidistant by construction, \(p(x) = \frac{1}{x_{\text{max}} - x_{\text{min}}}\). The denominator of Eqn. A.2 can be rewritten as:

\[
h \cdot n \cdot p(x) = \frac{x_{\text{max}} - x_{\text{min}}}{k} \times n \times \frac{1}{x_{\text{max}} - x_{\text{min}}} = \frac{n}{k}, \tag{A.6}
\]

where \(k\) is the scaling factor between the \(x\)-range and the bandwidth, set to 10 in the context of the present application. \(n/k\) therefore takes the meaning of the effective population of the sample with respect to the choice of partition of the range, and it is expected that increasing values of \(n/k\) would result in a smaller overall variance on \(W(x)\), as a higher number of data points in the interval about \(x\), where \(K\) is not negligible, will better constrain the shape of the estimator.
Appendix B

ABCD Modelling Limitations

The ABCD method is a popular and powerful tool to estimate the contribution from the sources of background in an inaccessible region of the phase space out of the existing understanding of the behaviour of data in other kinematically close, signal poor regions.

As with all data-driven estimation techniques, there exist, however, a number of potential pitfalls in the ABCD methodology that could potentially result in a biased prediction of the background in the blinded region. This appendix details how two such pitfalls were taken into consideration and ultimately found to produce either negligible or acceptable consequences, within the assigned systematics, onto the eventual model for the Signal Region background.

B.1 Background homogeneity

The simple scaling principle at the basis of the ABCD estimation, stating that given two uncorrelated axes defining four regions in the phase space, we can write:

\[ N_A = N_B \cdot \frac{N_C}{N_D} \]  

(B.1)

does not necessarily hold when applied to a composite background. This can be easily demonstrated assuming two event populations, X and Y, of similar magnitude,
B.1. Background homogeneity

for which it is assumed that the ABCD axes are completely uncorrelated, hence it is correct to write $X_A = X_B \cdot X_C / X_D$ and likewise for $Y$.

Applying the naive ABCD estimation to the full background ensemble in this scenario yields:

$$\tilde{N}_A = (X_B + Y_B) \times \frac{X_C + Y_C}{X_D + Y_D} \quad (B.2)$$

Which is in principle different from the real contribution in $A$, given by:

$$X_A + Y_A = X_B \times \frac{X_C}{X_D} + Y_B \times \frac{Y_C}{Y_D} \quad (B.3)$$

The naive estimation correctly approximates the true yields in $A$ only if one of the two sources is negligible (typically around the per-cent level) or when both sources scale in the same way through the $C \leftrightarrow D$ and $A \leftrightarrow B$ transitions. In the latter case, we can define the common scaling factor $k$ as:

$$k = \frac{X_C}{X_D} = \frac{Y_C}{Y_D} \quad \Rightarrow \quad \frac{Y_D}{X_D} = \frac{X_C}{Y_C} \quad (B.4)$$

and re-write the naive estimation in Eqn. B.2 as:

$$\tilde{N}_A = (X_B + Y_B) \times \frac{k(X_D + Y_D)}{X_D + Y_D} = k \cdot X_B + k \cdot Y_B = X_B \cdot \frac{X_C}{X_D} + Y_B \cdot \frac{Y_C}{Y_D} = X_A + Y_A \quad (B.5)$$

producing an unbiased estimation.

This section will take into exam two possible sources of background for the present search, assess their contribution in the ABCD Regions in the Higgs Peak window, and finally estimate the bias introduced in the background modelling procedure by ignoring their contribution and treating the data as entirely coming from a single background source in the custom ABCD-like procedure outlined in Chapter 7.
B.1. Background homogeneity

B.1.1 $Z+b$-jets

The production of a final state $Z$ boson in association with a $b$-jet has the potential to mimic the final state of the Signal Region when the boosted $Z \rightarrow b\bar{b}$ decays are reconstructed as a single large-$R$ jet.

We shall now produce an estimation of the number of SR $Z(b\bar{b})+b$-jet events applying the relevant approximated efficiency of the various selection steps ultimately defining the final Signal Region data sample.

The differential production cross section for highly boosted $Z+b$-jets events has been recently measured in the $Z \rightarrow e^+e^-/\mu^+\mu^-$ final state at ATLAS [139] to be $8 \pm 2 \cdot 10^{-6}$ pb/GeV in the kinematic bin defined by $p_T(Z) \in [500, 1000]$ GeV, roughly matching the full range of the Higgs Candidate $p_T$ observed in the data (see Fig.7.7a). The relevant integrated $Z(b\bar{b})+b$-jets cross section is therefore approximately:

$$\sigma(pp \rightarrow bZ(b\bar{b})|_{>500\text{GeV}} = 8 \cdot 10^{-6} \text{ pb/GeV} \times 500 \text{ GeV} \times 2.5 \approx 10\text{fb} \quad \text{(B.6)}$$

with the factor of 2.5 approximately rescaling the cross section by the $\text{BR}[Z \rightarrow b\bar{b}]/\text{BR}[Z \rightarrow \ell^+\ell^-]$ ratio.

The $b + Z(b\bar{b})$ yield in the signal region is estimated by multiplying the computed cross section by the following event selection efficiencies previously computed for signal and data events in Section 6.4. The full estimated event selection is detailed in Table B.1, resulting in an estimated 5 events in the Signal Region. The data yield in the Signal Region is of the order of 250 (Sec. 10.1), setting the $Z+b$-jet contribution at an approximate 2% level, well within the systematic uncertainties on the background prediction. The contribution from this source of background can therefore be safely ignored within the modelling procedure.
B.1. Background homogeneity

<table>
<thead>
<tr>
<th>Step</th>
<th>Efficiency</th>
<th>Eff. Type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large/Small R Jet $p_T$</td>
<td>$\sim 1$</td>
<td>/</td>
<td>Already in cross-section</td>
</tr>
<tr>
<td>$Z \rightarrow b\bar{b}$ acceptance</td>
<td>$\sim 1$</td>
<td>/</td>
<td>Already in cross-section</td>
</tr>
<tr>
<td>H2T2B</td>
<td>$\sim 0.5$</td>
<td>Signal</td>
<td>$\varepsilon_b = 70%$, conservative</td>
</tr>
<tr>
<td>log $\Delta R^*$ cut</td>
<td>$\sim 0.75$</td>
<td>Signal</td>
<td>Table 6.26</td>
</tr>
<tr>
<td>$p_T(Z)/M_{bZ}$ cut</td>
<td>$\sim 0.9$</td>
<td>Bkg, low-$M_{bZ}$</td>
<td>Table 6.25</td>
</tr>
<tr>
<td>J1B</td>
<td>$\sim 0.7$</td>
<td>Signal</td>
<td>$\varepsilon_b = 70%$</td>
</tr>
<tr>
<td>Forward jet cut</td>
<td>$\sim 0.2$</td>
<td>Background</td>
<td>Sec. 6.4.2</td>
</tr>
<tr>
<td>$M_{bb} \in [100, 135] \text{ GeV}$</td>
<td>0.07</td>
<td>/</td>
<td>Gaussian Estimate</td>
</tr>
</tbody>
</table>

Total Efficiency                | 0.003325 |

Run 2 Luminosity               | 139 fb$^{-1}$ |

Events                        | $10 \text{ fb} \cdot 139 \text{ fb}^{-1} \cdot 0.00475 \approx 5$ |

**Table B.1**: Estimated cutflow for the $Z+b$-jets background.

### B.1.2 All Hadronic $t\bar{t}$

Events where a pair of top-antitop quarks is produced and decays into a fully hadronic final state can potentially imitate the experimental signature targeted by the analysis, particularly as $W \rightarrow cs$ decays can produce additional spurious $b$-tags owing to the smaller charm rejecting power of common $b$-tagging algorithms with respect to light jets.

The contribution of all hadronic $t\bar{t}$ events to the four ABCD Regions in the Higgs mass peak window is estimated via a Pythia 8 Monte Carlo simulation, and is summarized in Table B.2, with the corresponding data yields also displayed for reference in table B.3. Comparing the $t\bar{t}$ and data yields, we observe that the contribution of $t\bar{t}$ events to the data is estimated to be approximately 2% in Regions C and D, and 10% in Regions B and A.

The top contamination in C and D is small enough to leave the value of the scaling
B.1. Background homogeneity

<table>
<thead>
<tr>
<th>ttbar, H2T2B, peak</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1 fwd jets</td>
<td>108 ± 10</td>
<td>19.8 ± 4.1</td>
</tr>
<tr>
<td>=0 fwd jets</td>
<td>468 ± 21</td>
<td>131 ± 11</td>
</tr>
</tbody>
</table>

\[ k_{FWD}^t = 0.230 ± 0.024 \quad k_{FWD}^{sub} = 0.150 ± 0.034 \]

Table B.2: Higgs Peak Region \((M_H \in [105,135]\,\text{GeV})\) yields for all hadronic \(t\bar{t}\) events in the H2T2B category. The displayed uncertainties on the \(t\bar{t}\) yields and the \(k_{FWD}\) scaling ratios only reflect the statistical uncertainty on the Monte Carlo simulation.

<table>
<thead>
<tr>
<th>data, H2T2B, Peak</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1 fwd jets</td>
<td>5315</td>
<td>262</td>
</tr>
<tr>
<td>=0 fwd jets</td>
<td>23831</td>
<td>972</td>
</tr>
</tbody>
</table>

\[ k_{FWD} = 0.222 ± 0.003 \]

Table B.3: Higgs Peak Region \((M_H \in [105,135]\,\text{GeV})\) data yields in the H2T2B category. The Region-A yield (red) represents the final estimate for the number of data events in the Signal Region, including the effects of the shape correction and \(R_{corr}\) re-scaling.

The shape correction procedure (Sec. 7.2.2) extracts event weights from the bin-wise ratios between the C and D Region data shapes of the VLB \(p_T\) and small-\(R\) jet \(p_T\). Given the available data and the simulated top sample, we can estimate the potential bias introduced in the C/D data ratios by not subtracting the top contribution, as it would be necessary to ensure that the ABCD-like correction is applied to a uniform event population.

Introducing \(c_i^{dat} (c_i^t, c_i^{QCD})\) and \(d_i^{dat} (d_i^t, d_i^{QCD})\) respectively as the Region C and D content of the \(i\)-th bin in the data \((t\bar{t}, \text{QCD})\) distribution of either the VLB \(p_T\) or...
the small-\(R\) jet \(p_T\), and defining the QCD contribution as the bin-wise difference between data and \(t\bar{t}\), we can write the bin ratio in data as:

\[
    r_{i}^{\text{dat}} = \frac{c_{i}^{\text{QCD}} + c_{i}^{t\bar{t}}}{d_{i}^{\text{QCD}} + d_{i}^{t\bar{t}}} = \frac{c_{i}^{\text{QCD}}}{d_{i}^{\text{QCD}}} \times \frac{1 + \frac{c_{i}^{t\bar{t}}}{c_{i}^{\text{QCD}}}}{1 + \frac{d_{i}^{t\bar{t}}}{d_{i}^{\text{QCD}}}} = r_{i}^{\text{QCD}} \times \lambda_{i}^{t\bar{t}}
\]

(B.8)

with \(\lambda_{i}^{t\bar{t}}\) quantifying the bias generated choosing not to subtract the top contribution before the weight extraction. The bin-by-bin value of \(\lambda_{i}^{t\bar{t}}\) was computed for both the variables on which event weights are calculated, and compared with the systematic uncertainty associated with the shape correction weights extracted from the variable under exam. The results are displayed in Figure B.1. As under no circumstance the value of \(\lambda_{i}^{t\bar{t}}\) was observed to be greater than the systematic uncertainty associated with the reweighting procedure, the top contribution in regions C and D can be altogether neglected in the derivation of the shape-correcting functions.
Figure B.2: Monte Carlo prediction for $t\bar{t}$ contribution to the Higgs Peak Region A (dots) and Region B (solid) data. The comparison is provided for $H2T2B$ (left) and $H2T1B$ Higgs Peak Region Data. $k_{FWD}$, the shape-correcting event weights as calculated on C and D data, and $R_{corr}$ are applied to B-Region events. The error bars and the shaded areas represent the statistical uncertainty on the Monte Carlo simulation.

Summarising, both $k_{FWD}$ and the shape-correcting event weights are unaffected by the top contribution. It is therefore acceptable to treat the data in region C and D as originating from a single source of background and proceed accordingly. The ratio between the top contribution in region A and B is found (Table B.2) to be statistically compatible with $k_{FWD}$ as computed on raw data (Table B.3).

If the top A and B shapes are also compatible (once B-Region data undergoes the full $k_{FWD}$ scaling plus shape correction based on C/D data), the top component in the Region B data will reproduce the top contribution in A once rescaled by $k_{FWD}$. Figure B.2 shows this to be the case for both the $J1B_H2T2B$ and $J1B_H2T1B$ categories in $t\bar{t}$ events. The good agreement in $H2T1B$ data (right plot) strongly suggests that an apparent downwards fluctuation in the second and third bin of the $H2T2$ histograms is purely of statistical nature, given the very poor statistics of the MC $t\bar{t}$ sample used in this category\(^1\).

Finally, Figure B.3 offers a comparison between the nominal prediction for the background and the background model obtained after subtracting the contribution

\(^1\)Out of the approximately $1.13 \cdot 10^{11}$ generated unweighted events in the inclusive sample, only 32 enter the $H2T2B$ Higgs Peak Region (a.k.a. Signal Region)
B.1. Background homogeneity

Figure B.3: Comparison between the Nominal background model derived neglecting the top contribution to the data (black, solid) and handling the subtracting the top before modelling, then re-adding the A-Region top prediction to the result (red dashed line).

of $t\bar{t}$ from Region B, applying the modelling procedure to the "top-purified" data and finally re-adding the Monte Carlo prediction for Region A $t\bar{t}$ to the result. To better understand the impact of neglecting the top contribution on the background model, the pre-fit statistical and systematic uncertainties on the model are provided summed in quadrature. For the H2T1B channel, Region-A data is also displayed. In both cases, the difference between the fully data-driven estimate and the one using the MC $t\bar{t}$ background subtraction is well within the established background systematics uncertainties.

B.1.3 Conclusion

Within this appendix, two additional sources of background were examined and found to yield no effects nor introduce any bias in the established background estimation.

The production of a $Z$ boson in association with a $b$-tagged jets was estimated to contribute as little as 1% of the data yields in the Signal Region, well below the final uncertainty on the predicted background.
The production of $t\bar{t}$ pairs, and their subsequent decay into a fully hadronic final state, was estimated through a Monte Carlo simulation to contribute around 7% of the Signal Region data, with a smaller contribution in the other data samples used within the modelling procedure. Despite the non-negligible contribution, $t\bar{t}$ events are predicted to scale across the $C \leftrightarrow D$ and $A \leftrightarrow B$ transitions in the same way as the dominant QCD background, meeting the condition for the viability of an ABCD-like model in the presence of more than one distinct sources of background.

As a result of this study, the data is treated as originating from a single, homogeneous source of background, requiring no subtraction of sub-leading contributions or additional systematic uncertainties.

**B.2 Signal Spillage in Regions B, C, D**

Another concern for the robustness and reliability of ABCD-driven background modelling techniques is the potential for bias to be introduced as a result of the presence of a non negligible amount of signal in region B, C and D, which are used to predict the yield and/or shape of the background in region A. The purpose of this appendix is to assess the bias on the background yield prediction, as well as a possible correction procedure to recover the original background shape in the presence of a signal.

For a quantitative assessment of the contamination, Table B.4 displays signal and data event yields in regions of the ABCD plane defined in the Higgs Peak Region. All events accounted for here belong to the most sensitive $H^{2T2B}$ channel.

The working assumption underlying the quantitative estimates in this appendix is that the established event count in region C, D and B are entirely determined by background processes. While this statement seems to contradict the spirit of this appendix, and it is obviously not true a priori, it is only used to create a starting point for the bias estimation, and is not featured in the proposed correction proce-
B.2. Signal Spillage in Regions B, C, D

The effect of the signal contamination on the background modelling can be twofold:

- The value of the yields scale factor $k_{FWD} = N_C/N_D$ is affected by the presence of signal in regions C and D;

- The change in the number of events and shape of the observables in region B due to the signal contamination is reflected on the background model.

The severity of the shift in $k_{FWD}$ due to the signal contamination is mitigated by the larger data yields in region C and D with respect to the less populated (and therefore more sensitive) region B. Assuming the presence of a signal compatible with the theoretical models for a vector-like $B$ quark of mass 1200 GeV, the scale factor would be thus affected:

$$k_{FWD} = \frac{N_{bg}^C}{N_{bg}^D} = 0.221 \pm 0.003 \quad \rightarrow \quad \tilde{k}_{FWD} = \frac{N_{bg}^C + N_{sig}^C}{N_{bg}^D + N_{sig}^D} = 0.223 \pm 0.003$$

Given that the shift is well within the statistical uncertainty on $k_{FWD}$, the presence of signal in regions C and D is considered negligible for the remainder of this appendix.

As stated earlier, the signal contamination in region B affects both the shape and the yields of the background model. The following identity can be intended to represent the behaviour of both the total number of events predicted for region A as well as

<table>
<thead>
<tr>
<th>Signal, $M = 1.2$ TeV</th>
<th>J0B</th>
<th>J1B</th>
<th>Data</th>
<th>J0B</th>
<th>J1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ fwd jets</td>
<td>62.3</td>
<td>80.0</td>
<td>$\geq 1$ fwd jets</td>
<td>5315</td>
<td>80.0</td>
</tr>
<tr>
<td>$=0$ fwd jets</td>
<td>38.9</td>
<td>51.1</td>
<td>$=0$ fwd jets</td>
<td>23831</td>
<td>972</td>
</tr>
</tbody>
</table>

Table B.4: Event yields in the four ABCD regions in the Higgs Peak region ($m_H \in [105, 135]$ GeV).
the content of each individual bin of the final observable:

\[
\tilde{N}_{A}^{pred} = k_{FWD} \times [N_{B}^{bg} + N_{B}^{sig}] = N_{A}^{bg} + k_{FWD} \cdot N_{A}^{sig} = N_{A}^{bg} + k_{FWD} \cdot \varepsilon_{B} N_{A}^{sig} \quad (B.9)
\]

Where the leakage factor \( \varepsilon_{B} \) is defined as the signal content in the generic ABCD region B relative to the signal yield in region A.

The bias introduced by the signal leakage in this scenario is then estimated taking the ratio of the data over background excess in the two scenarios where the signal contamination is either ignored (\( \Delta N_{cont} \)) or perfectly corrected (\( \Delta N_{corr} \)).

\[
1 - b = \frac{\Delta N_{cont}}{\Delta N_{corr}} = \frac{N_{A}^{obs} - \tilde{N}_{A}^{pred}}{N_{A}^{obs} - N_{A}^{bg}} = \frac{N_{A}^{sig} - k_{FWD} \cdot \varepsilon_{B} N_{A}^{sig}}{N_{A}^{sig}}
\]

\[
= 1 - k_{FWD} \cdot \varepsilon_{B} \simeq 0.85
\]

\[
\rightarrow b \simeq 0.15
\]

indicating an approximate 15% decrease in the sensitivity due to the contamination in region B.

### B.2.2 Correction Scheme

Let us consider the fundamental identity of the ABCD estimation (including the residual correlation correction derived in Sec. 7.2.3), which holds for the sources of background as observed in Fig. 7.3:

\[
N_{A}^{bg} = R_{corr} \cdot \frac{N_{B}^{bg} \times N_{C}^{bg}}{N_{D}^{bg}} \quad (B.10)
\]

In the simplified scheme where the contamination in D and C is assumed to be negligible, \( N_{C}^{bg} \) and \( N_{D}^{bg} \) can be replaced in Eqn. B.10 with the observed quantities \( N_{C}^{obs} \) and \( N_{D}^{obs} \). The unknown number of uncontaminated data events in region B, \( N_{B}^{bg} \), can be estimated as:
\[ N_{B}^{bg} = N_{B}^{obs} - \epsilon_{B} \times (N_{A}^{obs} - N_{A}^{bg}) \]  

(B.11)

where \( \epsilon_{B} \times (N_{A}^{obs} - N_{A}^{bg}) \) is a proxy for the signal contamination in region B, \( N_{B}^{bg} \).

Substituting Eqn. B.11 into Eqn. B.10, the only unknown quantity left is our target estimate for the uncontaminated region A background \( N_{A}^{bg} \), and the identity can therefore be treated as a first-degree equation in \( N_{A}^{bg} \), yielding:

\[ N_{A}^{bg} = \frac{k_{FWD}}{1 - \epsilon_{B} \cdot k_{FWD}} \times \left[ N_{B}^{obs} - \epsilon_{B} \times N_{A}^{obs} \right] \]  

(B.12)

B.2.3 Discussion

Equation B.12 can be applied to each bin in the final observable template, providing an alternative model for the A region background that accounts for the contamination in B. Several factors, however, limit the practical feasibility of this correction scheme:

- The bin-by-bin signal ratio between regions B and A, \( \epsilon_{B} \) is by default not constant throughout all the possible signals, making this correction model-dependent;

- The corrected model is partially correlated with the observed data in region A itself, which results in larger statistical fluctuations and a non-immediate statistical interpretation of the affecting uncertainties;

- As \( N_{A}^{obs} \) is blinded, the correction is only possible \emph{a posteriori}.

Given the conceptual and operational challenges of implementing the signal contamination correction, and the size of the contamination itself compared to the existent uncertainties on the model, no further action was undertaken to implement such a correction in the main analysis workflow.
Bibliography


https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.


[60] CMS Collaboration. Search for heavy resonances decaying into a vector boson and a Higgs boson in final states with charged leptons, neutrinos and b quarks at \( \sqrt{s} = 13 \) TeV, 2018.


[64] ATLAS Collaboration. Search for single production of a vector-like $B$ quark decaying into a bottom quark and a Higgs boson which decays into a pair of photons. 2018.


[67] CMS Collaboration. CMS Run 2 public luminosity summary plots.  
https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.


https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults.

[70] ATLAS Radiation Measurement and Simulation Public Results.  


[100] ATLAS Collaboration. ATLAS $b$-jet identification performance and efficiency measurement with $t\bar{t}$ events in $pp$ collisions at $\sqrt{s} = 13$ TeV. 2019.


[111] Luca Panizzi. Vector-Like B Quark NLO Cross-Sections in the Narrow Width Approximation under the 4 and 5 flavour scheme. Private Communication.


[114] Release 21 jet cleaning recommendations, atlas.  


[131] ATLAS Collaboration. Measurement prospects of the pair production and self-coupling of the Higgs boson with the ATLAS experiment at the HL-


[136]

